



UNIVERSIDADE FEDERAL DO AMAZONAS – UFAM

INSTITUTO DE COMPUTAÇÃO – ICOMP

PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA – PPGI

Uma Arquitetura Conceitual Integrando Diagnóstico Cognitivo e Modelos de Linguagem de Grande Escala para Avaliação de Habilidades

Arcanjo Miguel Mota Lopes

Manaus - AM

2026

Arcanjo Miguel Mota Lopes

Uma Arquitetura Conceitual Integrando Diagnóstico
Cognitivo e Modelos de Linguagem de Grande Escala para
Avaliação de Habilidades

Tese de Doutorado apresentada ao Programa de
Pós-Graduação em Informática da Universidade
Federal do Amazonas como parte dos requisitos
para a obtenção do título de Doutor em Ciência da
Computação.

Área de concentração: Inteligência Artificial.

Sub-área: Inteligência Artificial na Educação.

Orientador

Prof. Dr. José Francisco de Magalhães Netto

Universidade Federal do Amazonas – UFAM

Instituto de Computação – IComp

Manaus - AM

2026

Ficha Catalográfica

Elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

-
- L864a Lopes, Arcanjo Miguel Mota
 Uma Arquitetura Conceitual Integrando Diagnóstico Cognitivo e
 Modelos de Linguagem de Grande Escala para Avaliação de
 Habilidades / Arcanjo Miguel Mota Lopes. - 2026.
 194 f. : il., color. ; 31 cm.
- Orientador(a): José Francisco de Magalhães Netto.
 Tese (doutorado) - Universidade Federal do Amazonas,
 Programa de Pós-Graduação em Informática, Manaus, 2026.
1. Sistemas Tutores Inteligentes. 2. Avaliação de Habilidades. 3.
 Modelo de Diagnóstico Cognitivo. 4. Modelos de Linguagem de
 Grande Escala. I. Netto, José Francisco de Magalhães. II.
 Universidade Federal do Amazonas. Programa de Pós-Graduação
 em Informática. III. Título
-



Ministério da Educação
Universidade Federal do Amazonas
Coordenação do Programa de Pós-Graduação em Informática

FOLHA DE APROVAÇÃO

"UMA ARQUITETURA CONCEITUAL INTEGRANDO DIAGNÓSTICO COGNITIVO E MODELOS DE LINGUAGEM DE GRANDE ESCALA PARA AVALIAÇÃO DE HABILIDADES"

ARCANJO MIGUEL MOTA LOPES

Tese de Doutorado defendida e aprovada pela banca examinadora constituída pelos professores:

Prof. Dr. José Francisco de Magalhães Netto - **Presidente**

Prof. Dr. Raimundo da Silva Barreto - **Membro Interno**

Prof.^a Dr.^a Eliane Pozzebon - **Membro Externo**

Documento assinado digitalmente
gov.br ELIANE POZZEBON
Data: 05/08/2026 14:32:53-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Leônidas de Oliveira Brandão - **Membro Externo**

Documento assinado digitalmente
gov.br LEONIDAS DE OLIVEIRA BRANDAO
Data: 07/08/2026 10:39:21-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Vitor Bremgartner da Frota - **Membro Externo**

Documento assinado digitalmente
gov.br VITOR BREMGARTNER DA FROTA
Data: 07/08/2026 18:58:26-0300
Verifique em <https://validar.iti.gov.br>

Manaus, 30 de Julho de 2026.



Documento assinado eletronicamente por **José Francisco de Magalhaes Netto, Professor do Magistério Superior**, em 03/08/2026, às 10:57, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Raimundo da Silva Barreto, Professor do Magistério Superior**, em 03/08/2026, às 11:32, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufam.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3254890** e o código CRC **89AEAA28**.

Avenida General Rodrigo Octávio, 6200 - Bairro Coroadó I Campus Universitário Senador Arthur Virgílio Filho, Setor Norte - Telefone: (92) 3305-1181 / Ramal 1193
CEP 69080-900, Manaus/AM, coordenadorppgi@icomp.ufam.edu.br

Referência: Processo nº 23105.027847/2026-14

SEI nº 3254890

*Gostaria de homenagear as pessoas que de alguma maneira
contribuíram direta ou indiretamente.*

Agradecimentos

Agradeço primeiramente a DEUS por conceder, nos momentos difíceis, as forças e o discernimento necessários para seguir em frente e a oportunidade de crescer a cada obstáculo. Como bem escreveu Tomás de Kempis em A Imitação de Cristo: “Que não devemos nós a esse DEUS de bondade, e que lhe devemos por tantos benefícios? Ah! Em nossa miséria, outra coisa não temos a oferecer-Lhe senão nosso coração, e é tudo o que ele pede da sua pobre criatura”. É com esse sentimento de entrega e gratidão que concluo esta etapa.

À minha esposa, Márcia Carvalho, que, durante toda jornada em que estamos juntos, você esteve ao meu lado nas noites mal dormidas, nos momentos de insegurança e nas pequenas vitórias do caminho, sendo “forte”, com amor e ainda encontrando forças para me apoiar quando o cansaço me abatia. Minha fonte de motivação para levantar todos os dias, que qualquer coisa que eu fizesse por você, com você, seria sempre melhor.

À minha família, pelo amor e incentivo incondicional, que me impulsionaram a persistir, sempre presentes nas pequenas e grandes conquistas.

Agradeço, de maneira especial, ao Professor José Francisco de Magalhães Netto, pelas orientações, pelos conselhos nas minhas conquistas profissionais e por nossa amizade que tem se fortalecido ao longo dos anos. Suas contribuições acadêmicas e humanas foram essenciais para o desenvolvimento deste trabalho e para o meu crescimento intelectual.

Ao IComp e a todos os professores e servidores técnico-administrativos com quem tive a oportunidade de trabalhar durante minha atuação como professor substituto, pelos ensinamentos, pelo incentivo e pelas contribuições para minha formação profissional.

Aos professores e à coordenação do Projeto WebAcademy pelo suporte estrutural e computacional oferecido, que contribuiu diretamente para o desenvolvimento e para a obtenção dos resultados apresentados nesta tese.

À Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM) e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio concedido ao longo desta trajetória acadêmica. As bolsas de estudo e a infraestrutura de pesquisa contribuíram, de forma direta e indireta, para a realização desta tese.

Aos membros da banca examinadora pela disponibilidade, pela leitura atenta deste trabalho e pelas valiosas contribuições apresentadas durante o processo de avaliação. As

observações, sugestões e questionamentos foram fundamentais para o aprimoramento desta tese, contribuindo para o fortalecimento de seus aspectos teóricos, metodológicos e científicos.

Por fim, minha sincera gratidão a todos que, direta ou indiretamente, com cada gesto de incentivo, ensinamento ou apoio, foram essenciais para que este trabalho se tornasse realidade. Muito obrigado!

“Compra a verdade e não a vendas; também a sabedoria, a disciplina e o entendimento.”

(Provérbios 23:23)

Uma Arquitetura Conceitual Integrando Diagnóstico Cognitivo e Modelos de Linguagem de Grande Escala para Avaliação de Habilidades

Resumo

A avaliação de habilidades nos Sistema Tutor Inteligente (STI) depende da capacidade de representar o conhecimento a ser avaliado, inferir o estado cognitivo do estudante e utilizar essas informações para orientar o processo de aprendizagem. Entretanto, as abordagens convencionais frequentemente utilizam métricas agregadas de acertos e erros ou tratam de maneira desacoplada a representação das habilidades, o diagnóstico cognitivo e a adaptação pedagógica. Os Modelos de Diagnóstico Cognitivo (MDC) possibilitam superar parte dessas limitações ao fornecerem perfis multidimensionais e interpretáveis do domínio das habilidades. Porém, sua utilização requer a construção de relações entre itens de avaliação e os Componentes de Conhecimento (CC), geralmente representadas por uma Q-matriz. A elaboração dessa estrutura é predominantemente manual e dependente de especialistas, o que dificulta sua manutenção e escalabilidade. Nesse contexto, esta tese apresenta como principal contribuição a CAASA-ITS (*Cognitive-Adaptive Architecture for Skills Assessment in Intelligent Tutoring Systems*), uma arquitetura conceitual cognitivo-adaptativa para avaliação estruturada de habilidades em STI. A arquitetura integra, em um mesmo ciclo de avaliação e adaptação, a representação item-CC por meio da Q-matriz, o apoio de Modelo de Linguagem de Grande Escala (LLM) à geração e revisão dessas relações, a inferência do domínio de habilidades por MDC, a Modelagem do Estudante, a adaptação pedagógica e a visualização diagnóstica. A arquitetura Cognitive-Adaptive Architecture for Skills Assessment in Intelligent Tutoring Systems resulta de um processo incremental de pesquisa desenvolvido ao longo do doutorado. Inicialmente, um mapeamento da literatura permitiu identificar limitações na avaliação e na modelagem do estudante nos STI. Em seguida, uma arquitetura multiagente explorou mecanismos de recomendação e de intervenção pedagógica com base nas ações dos estudantes. A pesquisa avançou com a incorporação de um MDC em um STI, o que permitiu representar a relação entre itens e habilidades por meio de uma Q-matriz, além de produzir

perfis diagnósticos individuais e coletivos em um contexto educacional real. As limitações observadas nesse processo, especialmente a necessidade de construção manual da Q-matriz, motivaram um terceiro estudo, no qual os LLM foram investigados como mecanismo de apoio à geração das relações item-CC. As Q-matrizes produzidas foram avaliadas tanto estruturalmente, em comparação com aquelas definidas por especialistas, quanto funcionalmente, considerando seu impacto sobre os resultados do diagnóstico cognitivo. A consolidação desses resultados resultou na formalização da CAASA-ITS, que organiza os componentes desenvolvidos e avaliados durante a pesquisa de maneira modular e explana as responsabilidades associadas à autoria, à representação do domínio, à geração assistida da Q-matriz, ao diagnóstico cognitivo, à modelagem do estudante, à adaptação pedagógica e à visualização dos resultados. Os resultados demonstraram a viabilidade da utilização de MDC para produzir perfis interpretáveis de domínio de habilidades nos STI e indicaram que Q-matrizes geradas com o apoio da LLM podem alcançar um desempenho funcional comparável ao de matrizes elaboradas por especialistas em cenários controlados, apesar de apresentarem diferenças estruturais. Esses resultados fornecem evidências dos principais mecanismos que compõem a arquitetura proposta. A principal contribuição da tese é a formalização da CAASA-ITS como uma arquitetura que integra representação, inferência cognitiva, automação assistida e adaptação pedagógica em um processo integrado de avaliação de habilidades, oferecendo uma estrutura modular e escalável para o desenvolvimento de STI mais interpretáveis.

Palavras-chave: Sistemas Tutores Inteligentes; Avaliação de Habilidades; Modelo de Diagnóstico Cognitivo; Modelos de Linguagem de Grande Escala.

Uma Arquitetura Conceitual Integrando Diagnóstico Cognitivo e Modelos de Linguagem de Grande Escala para Avaliação de Habilidades

Abstract

The measurement of skills in Intelligent Tutoring Systems (ITS) is contingent upon the capacity to symbolise the knowledge under assessment, to deduce the student's cognitive condition, and to utilise this information to direct the learning process. However, conventional approaches frequently employ aggregated metrics of correct and incorrect answers or address skill representation, cognitive diagnosis, and pedagogical adaptation in a decoupled manner. Cognitive Diagnostic Models (CDM) have been shown to assist in overcoming some of these limitations by providing multidimensional, interpretable profiles of skill mastery. However, the utilisation of these tools necessitates the establishment of relationships between assessment items and Knowledge Components (KC), which are typically represented by a Q-matrix. The construction of this structure is predominantly a manual process that is dependent on expert knowledge, which hinders maintenance and scalability. In this context, the principal contribution of this thesis is CAASA (Cognitive-Adaptive Architecture for Structured Assessment), a conceptual architecture for structured skill assessment in ITS. The architecture integrates item-KC representation via a Q-matrix, Large Language Model (LLM) support for generating and reviewing these relationships, skill mastery inference using CDM, student modelling, pedagogical adaptation, and diagnostic visualisation, within a single assessment and adaptation cycle. The CAASA architecture is the result of an incremental research process developed throughout the doctoral study. A preliminary review of the literature revealed certain limitations with regard to the assessment and modelling of STIs. Subsequently, a multi-agent architecture was explored, with the aim of investigating recommendation and pedagogical intervention mechanisms based on student actions. The research progressed by incorporating a CDM into an ITS, enabling the representation of relationships between

items and KC via a Q-matrix, as well as the generation of individual and collective diagnostic profiles within a real-world educational context. The limitations observed during the process, specifically the necessity of manual construction of the Q-matrix, provided the motivation for a third study. This third study investigated the use of LLM as a mechanism to support the generation of item–KC relationships. The resulting Q-matrices were evaluated in terms of both structure by comparing them with expert-defined matrices and function by assessing their impact on cognitive diagnosis outcomes. The consolidation of these findings resulted in the formalisation of CAASA, which organises the components developed and evaluated during the research in a modular fashion and outlines the responsibilities associated with authoring, domain representation, assisted Q-matrix generation, cognitive diagnosis, student modelling, pedagogical adaptation, and result visualisation. The results demonstrated the feasibility of using machine-generated competence profiles to produce interpretable skill-mastery profiles in situated contexts. Furthermore, the results indicated that Q-matrices generated with language model support can achieve functional performance comparable to expert-crafted matrices in controlled scenarios, despite exhibiting structural differences. The findings presented herein provide evidence regarding the key mechanisms that constitute the proposed architecture. The thesis's principal contribution is the formalisation of CAASA as an architecture that integrates representation, cognitive inference, assisted automation, and pedagogical adaptation into an integrated skill assessment process, offering a modular and scalable framework for the development of more interpretable ITS.

Keywords: Intelligent Tutoring Systems; Skills Assessment; Cognitive Diagnosis; Large Language Models.

Lista de Ilustrações

Figura 1 – Ciclo investigativo da metodologia	10
Figura 2 – Histórico da pesquisa.	13
Figura 3 – Arquitetura tradicional de um STI.	19
Figura 4 – Exemplo padrão de um modelo BKT.	37
Figura 5 – Comparação entre BKT, DKT e DKVMN.	42
Figura 6 – Processos realizados para diagnóstico.	46
Figura 7 – Exemplo de uma Q-matriz com relação de itens e atributos.	46
Figura 8 – Fluxo simplificado do processamento de uma entrada textual em um LLM.	56
Figura 9 – Modelo de arquitetura <i>transformer</i>	57
Figura 10 – Exemplo simplificado do mecanismo de autoatenção.	58
Figura 11 – Exemplo de diferentes estágios de treinamento e inferência em LLMs. O <i>prompt</i> para geração de respostas é possível em diferentes estágios, como pré-treinamento, <i>instruction tuning</i> ou alinhamento.	60
Figura 12 – Modelo de gerenciamento de experiência de aprendizagem no sistema IWT.	76
Figura 13 – Geração e agregação de domínios de competências dentro do MM-Competences.	78
Figura 14 – Avaliação de competências na plataforma CAT.	80
Figura 15 – Competências alcançadas pelos alunos e validadas pelos professores, na Plataforma CAT.	81
Figura 16 – Ficha técnica de um aluno no módulo de avaliação do ANNETTE Project.	83
Figura 17 – Exemplo de passaporte de competências com avaliação concluída no ANNETTE Project.	83
Figura 18 – Arquitetura ArCARE.	85
Figura 19 – Atualização das habilidades do aluno no ArCARE.	86
Figura 20 – Arquitetura geral do CALEP System.	88
Figura 21 – Avaliação baseada em competência utilizando TRI no CALEP System.	89
Figura 22 – Arquitetura do Simplify ITS.	91
Figura 23 – <i>Dashboard</i> mostrando as competências e as atividades associadas no CaSS Project.	94

Figura 24 – Exemplo de relatório gerado para um aluno no sistema Stat-Knowlab.	96
Figura 25 – Ontologia <i>AlgosSkill</i> do Diag-Skill.	98
Figura 26 – Visão geral da arquitetura CAASA-ITS integrando diagnóstico cognitivo, geração de Q-matriz e mecanismos de devolutivas.	105
Figura 27 – Ciclo cognitivo-adaptativo da arquitetura CAASA-ITS.	118
Figura 28 – Arquitetura Multiagente proposta por Lopes e Netto (2021).	123
Figura 29 – Modelo de gerenciamento de atividades e diagnóstico cognitivo para uso no Sistema Tutor Inteligente.	127
Figura 30 – Interface de autoria do professor, voltada ao cadastro de atividades no STI e à associação dos CC a serem avaliadas pelo MDC.	129
Figura 31 – Pontuação das habilidades dominadas pelos alunos participantes medida pelo MDC G-DINA.	132
Figura 32 – Painel do perfil de habilidades do estudante <i>Estudante7</i> inferido pelo MDC.	133
Figura 33 – Visão geral do processo experimental, desde os itens de avaliação (<i>E</i>) e CC (<i>C</i>) até as Q-matrizes geradas pelo LLM e as subsequentes análises estruturais e estatísticas.	136
Figura 34 – Comparação estrutural das Q-matrizes geradas por LLM sob diferentes estratégias de <i>prompts</i>	138
Figura 35 – Melhoria progressiva no alinhamento estrutural das Q-matrizes geradas por LLM com diferentes estratégias de <i>prompts</i>	139
Figura 36 – Comparação ilustrativa entre uma Q-matriz definida por especialistas (à esquerda) e uma Q-matriz gerada por LLM sem nenhum exemplo, produzida por <i>qwen3 : 8b</i> (ao centro), juntamente com uma matriz de diferença binária (à direita), onde o valor 1 indica associações divergentes entre item e atributo.	139
Figura 37 – Comparação de RMSE e MAE entre diferentes estratégias de <i>promptings</i>	141

Lista de Quadros

Quadro 1 – Quadro comparativo dos trabalhos relacionados.	102
---	-----

Lista de Tabelas

Tabela 1 – Atributos de habilidades relacionados às operações de resolução de equação polinomial de primeiro grau.	129
Tabela 2 – Avaliação funcional de Q-matrizes definidas por especialistas e geradas por LLM usando o modelo DINA. Os valores representam a média $\pm$ desvio padrão entre as execuções.	140

Lista de Abreviaturas e Siglas

ADC Avaliação Dagnóstica Cognitiva

API Application Programming Interface

ArCARE Conceptual Framework of Educational Resources Adaptation in Virtual Learning Environments

BKT Bayesian Knowledge Tracing

CAASA-ITS Cognitive-Adaptive Architecture for Skills Assessment in Intelligent Tutoring Systems

CAI Computer-Assisted Instruction

CALEP Competency Based Learning Path Generation System

CaSS Competency and Skill System

CAT Competency Assessment Tool

CbKST Competence-based Knowledge Space Theory

CC Componentes de Conhecimento

CCI Curva Característica de Item

CoT Chain-of-Thought

DINA Deterministic Inputs Noisy “And” Gate

DKT Deep Knowledge Tracing

DKVMN Dynamic Key-Value Memory Network

ENEM Exame Nacional do Ensino Médio

G-DINA Generalized DINA Model

HL Hamming Loss

IA Inteligência Artificial

ICL Aprendizagem em Contexto

IWT Intelligent Web Teacher

JSI Jaccard Index

KSC Knowledge, Skill, Competence

KT Knowledge Tracing

LEIA *LearnIng Algebra*

LFA Análise de Fatores de Aprendizagem

LLM Modelo de Linguagem de Grande Escala

MAE Mean Absolute Error

MDC Modelos de Diagnóstico Cognitivo

MM-Competences Competences Memory Map

PFA Análise do Fator de Desempenho

PLN Processamento de Linguagem Natural

RMSE Root Mean Square Error

RNN Redes Neurais Recorrentes

RSM Rule Space Method

RST Rule Space Theory

Saeb Sistema de Avaliação da Educação Básica

STI Sistema Tutor Inteligente

TCT Teoria Clássica dos Testes

TRI Teoria de Resposta ao Item

xAPI *Experience API*

ZDP Zonas de Desenvolvimento Proximal

Sumário

Lista de Ilustrações	i
Lista de Quadros	iii
Lista de Tabelas	iv
Lista de Abreviaturas e Siglas	v
1 INTRODUÇÃO	1
1.1 Contexto	1
1.2 Problema	2
1.3 Motivação e Justificativa	4
1.4 Objetivos	5
1.4.1 Objetivo geral	6
1.4.2 Objetivos específicos	6
1.5 Questão de Pesquisa e Hipóteses	6
1.5.1 Questão de pesquisa principal	7
1.5.2 Questões norteadoras	7
1.5.3 Hipóteses	8
1.6 Metodologia da Pesquisa	8
1.7 Delimitação do Escopo da Tese	11
1.8 Evolução da Pesquisa e a Difusão dos Resultados	11
1.8.1 Publicações e colaborações	12
1.9 Aspectos Éticos e Integridade Acadêmica	15
1.10 Organização do Documento	15
2 FUNDAMENTAÇÃO TEÓRICA	17
2.1 Sistema Tutor Inteligente	17
2.1.1 Definição e evolução	17
2.1.2 Arquitetura clássica dos STI	18
2.1.2.1 Módulo do Domínio (Expert Model)	18
2.1.2.2 Módulo Tutor (Tutor Model)	19
2.1.2.3 Módulo do Estudante (Learner/Student Model)	20
2.1.2.4 Módulo de Interface (Interface Model)	23
2.1.2.5 Módulo de Controle	24

2.2	Avaliação da Aprendizagem nos Ambientes Educacionais	24
2.2.1	Funções e objetivos da avaliação	25
2.2.1.1	Avaliação formativa	25
2.2.1.2	Avaliação somativa	25
2.2.1.3	Avaliação diagnóstica	26
2.2.2	Elementos de uma avaliação baseada em habilidades	27
2.2.2.1	Componentes de conhecimento	27
2.2.2.2	Habilidade	28
2.2.2.3	Item	30
2.2.3	Avaliação Diagnóstica Cognitiva	31
2.2.4	Avaliação das Habilidades Cognitivas	32
2.2.4.1	Teoria Clássica dos Testes – TCT	33
2.2.4.2	Teoria de Resposta ao Item – TRI	34
2.2.4.3	Bayesian Knowledge Tracing – BKT	36
2.2.4.4	Análise do Fator de Desempenho – PFA	39
2.2.4.5	Deep Knowledge Tracing – DKT	41
2.2.4.6	Dynamic Key-Value Memory Network – DKVMN	41
2.2.5	Limitações	42
2.3	Diagnóstico Cognitivo e a Avaliação de Habilidades	44
2.3.1	Modelos de Diagnóstico Cognitivo	44
2.3.1.1	Configuração para a implementação da avaliação baseada em MDC	45
2.3.1.2	Definição dos atributos latentes	45
2.3.1.3	Q-matriz e a <i>Rule Space Theory</i>	47
2.3.2	Modelos clássicos para diagnóstico cognitivo	47
2.3.2.1	Modelo DINA	48
2.3.2.2	Modelo G-DINA	49
2.3.3	A avaliação funcional dos MDC	51
2.3.4	Limitações do MDC	53
2.4	IA no Apoio ao Processo de Avaliação	55
2.4.1	Modelo de Linguagem de Grande Escala	55
2.4.2	Famílias LLM	62
2.4.2.1	GPT – OpenAI	62

2.4.2.2	LLaMA – Meta AI	63
2.4.2.3	Gemini – Google	63
2.4.2.4	Phi – Microsoft	64
2.4.2.5	Mixtral – Mistral AI	65
2.4.2.6	Qwen – Alibaba Cloud	65
2.4.2.7	Granite – IBM Research	67
2.4.3	LLM no apoio à criação de avaliações da aprendizagem	68
2.4.4	Representação, inferência e automação usando LLM	70
2.4.5	Limitações	73
3	TRABALHOS RELACIONADOS	75
3.1	Intelligent Web Teacher (IWT)	76
3.2	Competences Memory Map – (MM-Competences)	77
3.3	Competency Assessment Tool – (CAT)	80
3.4	ANNETTE Project	82
3.5	Conceptual Framework of Educational Resources Adaptation in Virtual Learning Environments – (ArCARE)	84
3.6	Competency Based Learning Path Generation System – (CALEP)	87
3.7	Simplify ITS	90
3.8	Competency and Skill System – (CaSS)	92
3.9	Stat-Knowlab	95
3.10	Diag-Skills	97
3.11	Resumo do Capítulo	100
4	ARQUITETURA COGNITIVO-ADAPTATIVA	104
4.1	Evolução da Proposta	105
4.2	Problemas Abordados pela Arquitetura	106
4.3	Formalização Geral da Arquitetura	107
4.4	Módulo de Domínio	109
4.5	Módulo de Autoria e Gerenciamento da Q-Matriz	110
4.6	Módulo de Geração Assistida da Q-Matriz por LLM	111
4.7	Módulo do Estudante	113
4.8	Módulo de Diagnóstico	114
4.9	Módulo Pedagógico	115
4.10	Módulo Tutor e Módulo Especialista	116

4.11	Módulo de Dados e Registro de Interações	116
4.12	Módulo de Visualização e <i>Learning Analytics</i>	117
4.13	Integração entre os módulos	117
4.14	Resumo do Capítulo	119
5	RESULTADOS ALCANÇADOS	121
5.1	Estudo 1: Arquitetura Multiagente para Recomendação e Intervenção Pedagógica	121
5.1.1	Objetivo do estudo	122
5.1.2	Estrutura proposta	123
5.1.3	Resultados alcançados	124
5.1.4	Limitações identificadas	126
5.2	Estudo 2: Incorporação do MDC G-DINA em um STI	127
5.2.1	Objetivo do estudo	128
5.2.2	Definição dos CC e da Q-matriz	128
5.2.3	Aplicação com os estudantes	130
5.2.4	Resultados	131
5.2.5	Limitações identificadas	133
5.3	Estudo 3: Geração e avaliação de matriz-Q por LLMs	134
5.3.1	Objetivo do estudo	134
5.3.2	Configuração experimental	135
5.3.3	Avaliação estrutural da Q-matriz	137
5.3.4	Avaliação funcional	140
5.3.5	Limitações identificadas	142
5.4	Resumo do Capítulo	143
6	CONCLUSÕES FINAIS	145
6.1	Contribuições da Tese	147
6.2	Limitações	148
6.3	Trabalhos Futuros	150
	REFERÊNCIAS	152
	APÊNDICE A AVALIAÇÃO PRÉ- E PÓS-TESTE	175
	APÊNDICE B PIPELINE EXPERIMENTAL PARA GERAÇÃO E AVALIAÇÃO DE Q-MATRIZES	178

APÊNDICE C DECLARAÇÃO DE USO DE FERRAMENTAS DE INTELIGÊNCIA ARTIFICIAL	180
---	------------

Capítulo

1

INTRODUÇÃO

O Capítulo 1 apresenta os elementos que orientam o desenvolvimento desta tese. Inicialmente, é apresentado o contexto educacional e científico no qual a pesquisa está inserida, seguido da definição do problema investigado e da motivação e justificativa para sua realização. Em seguida, são apresentados os objetivos geral e específicos, bem como a questão de pesquisa e as hipóteses que orientaram a investigação. O capítulo também descreve a metodologia adotada, delimita o escopo da tese e apresenta a evolução da pesquisa e de sua difusão ao longo do período de desenvolvimento do trabalho.

1.1 Contexto

A capacidade de transmitir o conhecimento por meio do ensino constitui um dos pilares da inteligência humana. Nesse processo, o professor desempenha um papel de monitoramento do progresso dos aprendizes, permitindo a adaptação do ensino às necessidades cognitivas individuais. Com o advento dos Ambientes Virtuais de Aprendizagem, tornou-se decisivo rastrear o progresso dos estudantes, visando personalizar as experiências de aprendizado e uso ([Abdelrahman, Wang e Nunes, 2023](#); [Pinheiro e Rebouças, 2018](#)).

Nas instituições de ensino, os discentes, comumente, são avaliados por meio de instrumentos como provas e trabalhos, compostos por diferentes questões (itens) e recebem pontuações baseadas no seu desempenho. Diante disso, os educadores têm buscado metodologias e ferramentas que possibilitem o aprimoramento dos modelos avaliativos, com foco no desenvolvimento da proficiência em relação aos conceitos ou habilidades. Nesse contexto, as tecnologias têm contribuído de maneira significativa, especialmente no que diz respeito ao desenvolvimento cognitivo e à promoção da aprendizagem autônoma ([Laine et al., 2022](#); [Shute et al., 2016](#); [Shemshack e Spector, 2020](#); [Wang et al., 2023](#)). Um dos avanços tecnológicos de maior impacto na aprendizagem personalizada nos últimos anos é o Sistema Tutor Inteligente (STI) ([Akyuz, 2020](#); [Huang, Xu e Liu, 2025](#); [Liu, Zhang e Jia, 2024](#); [St-Hilaire et al., 2022](#)).

Os STI são ambientes de aprendizagem computadorizados que incorporam modelos computacionais multidisciplinares fundamentados nas áreas da Ciência Cognitiva, Aprendizagem de Máquina, Linguística Computacional e Inteligência Artificial (IA) (Graesser, Conley e Olney, 2012; Marouf et al., 2024; Wang et al., 2023). Nesses sistemas, a mensuração do desempenho dos discentes pode ser realizada por diferentes tipos de avaliação, tais como a somativa, a formativa e a diagnóstica. Os diferentes STI desenvolvidos promovem a transformação e a transferência do conhecimento, auxiliando na obtenção dos resultados de aprendizagem de maneira eficiente e eficaz. Entretanto, para alcançar essa eficiência, é necessário que o sistema armazene o estado de conhecimento latente do estudante. Por isso, há na sua arquitetura o Modelo do Estudante que é responsável por orientar os demais mecanismos que realizam a personalização do processo de ensino-aprendizagem individual (Bremgartner, 2017; Tenório, Isotani e Bittencourt, 2020; Silva et al., 2023).

À medida que o estudante interage com o sistema, suas informações são registradas ao longo do processo de instrução. O Modelo do Estudante é atualizado para definir as suas preferências e dificuldades. Esta atualização contínua é fundamental para garantir recomendações e o apoio pedagógico adequado. Por isso, um diagnóstico bem definido é primordial para o bom funcionamento do STI. Neste contexto, se destacam os métodos de avaliação que fornecem informações mais detalhadas sobre o conhecimento do estudante, especialmente os que realizam análises automáticas e que indicam o nível de proficiência e os aspectos que necessitam de melhoria (Tomi, 2025; Veenman e Cleef, 2019; Zaffalon et al., 2019).

Sendo assim, nos últimos anos, os Modelos de Diagnóstico Cognitivo (MDC) e a Avaliação Diagnóstica Cognitiva (ADC) têm-se destacado por fornecerem informações detalhadas sobre o desempenho dos estudantes, permitindo a identificação precisa das lacunas de conhecimento e das áreas que necessitam de intervenção (Ravand e Baghaei, 2020; Shi et al., 2021; Herliana et al., 2026). Por meio das ADC, os MDC podem fornecer um conjunto de inferências granulares em vez de uma estimativa numérica simples sobre o conhecimento de um único conceito, ao contrário da maioria das avaliações (Fernandes, 2024).

1.2 Problema

A avaliação do conhecimento dos estudantes nos STI constitui um dos principais desafios na área de Informática na Educação. Tradicionalmente, os resultados das avaliações são obtidos com base em métricas de desempenho observáveis, como o número de respostas

corretas e incorretas nas atividades, o que limita a identificação explícita das capacidades cognitivas subjacentes ao processo de aprendizagem. Por isso, as abordagens baseadas em ADC e MDC têm sido amplamente utilizadas, à medida que permitem inferir o domínio do conhecimento latente associado aos elementos da avaliação.

Conforme destacado por [Rambe \(2021\)](#), a implementação das ADC nos STI e em ambientes de ensino *online* apresenta desafios adicionais em relação ao modelo presencial, especialmente no que se refere à necessidade de representar explicitamente as relações entre os itens de avaliação (questões, exercícios ou problemas) e os Componentes de Conhecimento (CC) que se referem a uma unidade conceitual do domínio de conhecimento associada à qual se pretende medir ([Pelánek, 2025](#)). Essa relação é, intuitivamente, representada por meio de uma estrutura item–CC. Essa estrutura constitui um elemento central para o diagnóstico de habilidades e competências, bem como, para auxiliar na personalização da aprendizagem.

No entanto, apesar da sua relevância, o processo de construção da representação item–CC ainda é, atualmente, predominantemente manual, complexo e dependente de especialistas, exigindo coordenação entre educadores e revisões frequentes. Este cenário pode dar origem a inconsistências na modelagem do domínio de conhecimentos do STI, além de limitar a escalabilidade, conforme identificado por [Moore et al. \(2024\)](#) e [Pelánek \(2025\)](#).

Além disso, apesar dos avanços dos MDC, ainda persistem os desafios estruturais relacionados com a integração da representação do conhecimento dos itens, a inferência cognitiva e a utilização dos resultados no contexto dos sistemas adaptativos. Entre os principais obstáculos, destacam-se:

- **Complexidade e subjetividade na modelação do conhecimento:** os MDC exigem a definição da relação item–CC, que é frequentemente realizada manualmente e está sujeita aos vieses da interpretação dos especialistas ([Ding et al., 2025](#); [Pelánek, Effenberger e Čechák, 2022](#)).
- **Desacoplamento entre a representação, a inferência e a utilização pedagógica:** em muitos sistemas, a representação das habilidades, o processo de diagnóstico e a utilização dos resultados para a adaptação ou interpretação pedagógica são abordados de maneira independente, o que dificulta a criação de soluções integradas ([Matsuda et al., 2022](#); [Reddy, Labutov e Joachims, 2016](#)).
- **Escalabilidade limitada:** a inclusão de novos itens ou CC, exige-se uma revisão ou

reconstrução da relação item-CC, tornando o processo dispendioso nos ambientes educacionais dinâmicos.

Em paralelo, o surgimento e os avanços recentes dos Modelo de Linguagem de Grande Escala (LLM) ([Cruz-Castañeda e Amadeus, 2025](#); [Peikos e Stavrou, 2025](#); [Tang et al., 2025](#)) têm possibilitado a automação de tarefas relacionadas à representação de conhecimento, incluindo o apoio à identificação de relações entre itens-CC. No entanto, estas abordagens ainda são, frequentemente, exploradas de forma isolada, sem integração nos processos de avaliação diagnóstica e sem inserção em uma estrutura de *software* coerente. Embora os resultados no uso dos LLM confirmem o potencial para realizar as associações entre itens-CC, eles ainda não são suficientes para assegurar a confiabilidade dessas relações em contextos avaliativos.

A relação item-CC possui implicações diretas na interpretação do desempenho do estudante e, conseqüentemente, no diagnóstico cognitivo, exigindo metodologias que apoiem sua validação, rastreabilidade, consistência e alinhamento pedagógico. Nesse sentido, é necessário investigar formas de integrar o uso dos LLM nos processos sistemáticos de modelagem, verificação e validação, de modo que suas sugestões possam ser utilizadas não apenas como saídas automáticas, mas como elementos auditáveis e fundamentados em uma abordagem diagnóstica.

1.3 Motivação e Justificativa

A avaliação da aprendizagem é um componente central das principais tecnologias voltadas para o apoio ao ensino-aprendizagem, influenciando diretamente a modelagem do estudante e apoiando a tomada de decisões instrucionais e o planejamento de ambientes adaptativos. No entanto, as abordagens tradicionais de avaliação baseadas em testes padronizados e pontuações agregadas têm sido cada vez mais criticadas por sua capacidade limitada de capturar os processos cognitivos complexos ([Ahmad et al., 2024](#); [Ji et al., 2025](#)).

Por outro lado, nos últimos dez anos, os STI se consolidaram como uma tecnologia relevante na educação *online* por sua capacidade de proporcionar uma experiência de aprendizagem personalizada e adaptativa para inúmeros usuários. Essa evolução se manifesta tanto na adaptação de conteúdos quanto na aplicação de estratégias pedagógicas fundamentadas no perfil e no desempenho ([Pu et al., 2022](#); [Oliveira, Isotani e Bittencourt, 2023](#)).

Os trabalhos de [Lopes e Netto \(2020\)](#), [Han et al. \(2019\)](#), [VanLehn \(2006\)](#), [Huang, Xu e Liu](#)

(2025), [Pereira, Gomes e Primo \(2025\)](#), [Silva e Andrade \(2025\)](#) e [Storare, Coletto e Junior \(2025\)](#) demonstraram que os STI podem contribuir significativamente para a aquisição de conhecimento dos estudantes, favorecendo suporte para memorização, compreensão, aplicação, análise, síntese e avaliação.

A avaliação diagnóstica, nos STI, é frequentemente realizada através da aplicação de atividades ou testes avaliativos que estimam um indicador de conhecimento ou competência de um indivíduo com base nas suas respostas. Essas avaliações permitem identificar lacunas no conhecimento que requerem uma atenção especial ([Ravand e Baghaei, 2020](#)). Em geral, a modelagem do conhecimento do estudante envolve diversos aspectos e decisões, como o que será modelado (conhecimento, comportamento ou afeto), ou até os processos de aprendizagem, ou avaliação e os dados a serem incluídos nos modelos avaliativos ([Papamitsiou, Sharma e Giannakos, 2020](#)). Em paralelo, com o advento das novas teorias de mensuração da aprendizagem e das tecnologias, bem como o aumento das exigências educacionais, emergiu uma nova geração de teorias com a implementação das ADC nos sistemas adaptativos ([Sun, Wu e Xu, 2023](#); [Williamson, 2023](#)).

Associadas aos MDC, as ADC têm se destacado por sua capacidade de fornecer diagnósticos detalhados sobre o desempenho dos estudantes, permitindo a identificação granular de lacunas de conhecimento e de áreas que demandam intervenção pedagógica ([Shi et al., 2021](#); [Herliana et al., 2026](#)). Nesse contexto, os MDC ampliam o potencial das ADC ao possibilitar inferências mais sofisticadas sobre o domínio de habilidades específicas, superando a simples estimativa numérica agregada em um único score ou conceito, como ocorre na maioria dos modelos avaliativos tradicionais ([Fernandes, 2024](#)).

Diante dos elementos envolvidos, observa-se um interesse crescente na proposição de novos modelos e arquiteturas pedagógicas¹ que integrem tanto a modelagem do conhecimento do estudante quanto à capacidade dos instrumentos de avaliação. Nesse contexto, esforços estão sendo concentrados no desenvolvimento de abordagens inovadoras na área, conforme atestam os estudos recentes de [Alabdulhadi e Faisal \(2021\)](#), [Zhuang et al. \(2026\)](#) e [Zhou et al. \(2025\)](#).

1.4 Objetivos

Para atingir o propósito desta tese, seguem o objetivo geral e os objetivos específicos.

¹ De acordo com [Menezes, Junior e Aragón \(2020\)](#), as arquiteturas pedagógicas oferecem uma concepção integrada dos aspectos pedagógicos e tecnológicos na construção de proposta inovadora no campo das tecnologias educacionais.

1.4.1 Objetivo geral

Formalizar, instanciar e avaliar uma arquitetura conceitual cognitivo-adaptativa para a avaliação de habilidades em STI, integrando a representação das relações item-CC, o suporte dos LLM à geração e revisão dos itens, a inferência cognitiva por Modelos de Diagnóstico Cognitivo, a Modelagem do Estudante, adaptação pedagógica e a visualização diagnóstica.

1.4.2 Objetivos específicos

Para atingir o objetivo geral, a presente pesquisa enumera os seguintes objetivos específicos:

1. Formalizar os elementos conceituais e computacionais da arquitetura, definindo os conjuntos, relações e estruturas associados aos estudantes, itens de avaliação, CC, Q-matriz, respostas observadas e estados de domínio do conhecimento.
2. Instanciar a arquitetura em módulos funcionais, integrando representação de domínio, modelagem do estudante, tutoria, registro de interações, visualização diagnóstica e organização da avaliação de habilidades.
3. Avaliar a aplicação de MDC em STI para inferência do domínio de habilidades, utilizando a relação item-CC e as respostas dos estudantes para produzir perfis de habilidades individuais e coletivos.
4. Investigar e avaliar o uso dos LLM como mecanismo de apoio à geração e revisão da relação item-CC, considerando o alinhamento estrutural de referência e o impacto funcional no desempenho do MDC.

1.5 Questão de Pesquisa e Hipóteses

Diante dos objetivos apresentados, esta tese parte da premissa de que a avaliação de habilidades nos STI deve ser tratada como um processo estruturado composto por dimensões interdependentes.

A primeira fase do processo, refere-se à representação explícita das relações entre os itens de avaliação e os CC organizados por meio de uma estrutura chamada Q-matriz. A segunda corresponde à inferência do domínio cognitivo dos estudantes, realizada por meio do MDC. A terceira envolve o apoio computacional à construção e à validação da representação item-CC, incluindo o uso do LLM. A quarta diz respeito à integração desses elementos em uma

arquitetura modular capaz de apoiar a avaliação, a visualização diagnóstica e a adaptação pedagógica.

1.5.1 Questão de pesquisa principal

O problema central desta tese é expresso pela seguinte Questão de Pesquisa (QP):

QP: *Como formalizar, integrar e avaliar uma arquitetura para STI capaz de articular representação dos item-CC, inferência por MDC, modelagem do estudante e suporte à adaptação pedagógica, de modo a auxiliar a avaliação estruturada de habilidades de maneira consistente, interpretável e com potencial de escalabilidade?*

1.5.2 Questões norteadoras

Para orientar a investigação, foram definidas as seguintes Questões Norteadoras (QN):

- QN1: Como formalizar os elementos conceituais e computacionais necessários para representar estudantes, itens de avaliação, CC, Q-matriz, respostas observadas e estados de domínio do conhecimento em uma arquitetura para STI?
- QN2: Como os Modelos de Diagnóstico Cognitivo podem ser integrados ao STI para produzir inferências interpretáveis sobre o domínio de habilidades dos estudantes?
- QN3: Quais limitações emergem da construção manual da Q-matriz e como os mecanismos computacionais podem apoiar a vinculação dos itens aos CC, considerando o alinhamento dos estrutural com representações de referência e o desempenho funcional dos MDC?
- QN4: Como integrar a representação entre item e CC, inferência cognitiva, apoio à autoria, visualização diagnóstica e adaptação pedagógica em uma arquitetura modular para avaliação de habilidades?

As QNs organizam a investigação em torno dos principais componentes da tese. A QN1 orienta a formalização conceitual e computacional da arquitetura. A QN2 trata da integração dos Modelos de Diagnóstico Cognitivo ao processo de avaliação nos STI. A QN3 aborda as limitações da construção manual da Q-matriz e o uso dos LLM como suporte à geração e revisão da relação item-CC. A QN4 trata da integração desses elementos em uma arquitetura modular visando à avaliação estruturada de habilidades.

1.5.3 Hipóteses

Com base na questão de pesquisa e nas questões norteadoras, são formuladas as seguintes hipóteses:

- H1: A representação de habilidades cognitivas baseada na relação item-CC, organizada por meio da Q-matriz, fornece uma estrutura adequada para apoiar a implementação das ADC nos STI.
- H2: A integração do MDC no STI permite inferir perfis de habilidades interpretáveis a partir das respostas dos estudantes e da Q-matriz.
- H3: Q-matrizes geradas com apoio do LLM podem apresentar desempenho funcional comparável ao de matrizes construídas por especialistas em cenários controlados, mesmo quando há diferenças estruturais entre as representações.
- H4: A integração entre representação item-CC, inferência por MDC, apoio à geração assistida da Q-matriz por LLM e visualização diagnóstica permite estruturar uma arquitetura modular viável para avaliação de habilidades.

As hipóteses descritas acima estabelecem os pressupostos analisados ao longo da pesquisa. A H1 está relacionada à adequação da Q-matriz como representação da relação entre itens e CC. A H2 trata da capacidade dos Modelos de Diagnóstico Cognitivo de produzir perfis diagnósticos interpretáveis. A H3 considera a contribuição dos LLM para apoiar a construção da Q-matriz, distinguindo alinhamento estrutural e desempenho funcional. A H4 sintetiza a viabilidade da integração desses componentes na arquitetura CAASA-ITS.

A seção seguinte apresenta os procedimentos metodológicos adotados para investigar as questões de pesquisa e analisar as hipóteses formuladas.

1.6 Metodologia da Pesquisa

De acordo com [Prodanov e Freitas \(2013, pp. 43\)](#), a pesquisa científica consiste em um estudo planejado, fundamentado em bases teóricas e orientado por um problema, cujo objetivo é produzir conhecimento capaz de responder a questões ainda não resolvidas. Nesse sentido, a definição da metodologia envolve a escolha de procedimentos que garantam rigor, validade e coerência entre os objetivos propostos e os resultados obtidos ([Gerhardt e Silveira, 2009](#)).

Com base nas classificações propostas por [Gerhardt e Silveira \(2009\)](#), [Johnson e Christensen \(2014\)](#) e [Prodanov e Freitas \(2013\)](#), esta pesquisa caracteriza-se como de natureza aplicada, com abordagem quali-quantitativa e objetivos exploratórios e descritivos.

Sua natureza aplicada justifica-se pelo fato de que a pesquisa propõe uma solução voltada para um problema concreto da área de Informática na Educação, especificamente voltada para avaliação de habilidades nos STI. Assim, o objetivo não se restringe à compreensão do fenômeno, mas também à proposição de uma abordagem que possa ser utilizada em sistemas educacionais reais.

A abordagem quali-quantitativa decorre da necessidade de analisar tanto aspectos conceituais quanto experimentais. A dimensão qualitativa está presente na construção do modelo conceitual, na definição da arquitetura e na análise interpretativa dos resultados. Já a dimensão quantitativa está associada à realização de experimentos com o MDC e à avaliação de desempenho por meio de técnicas estatísticas.

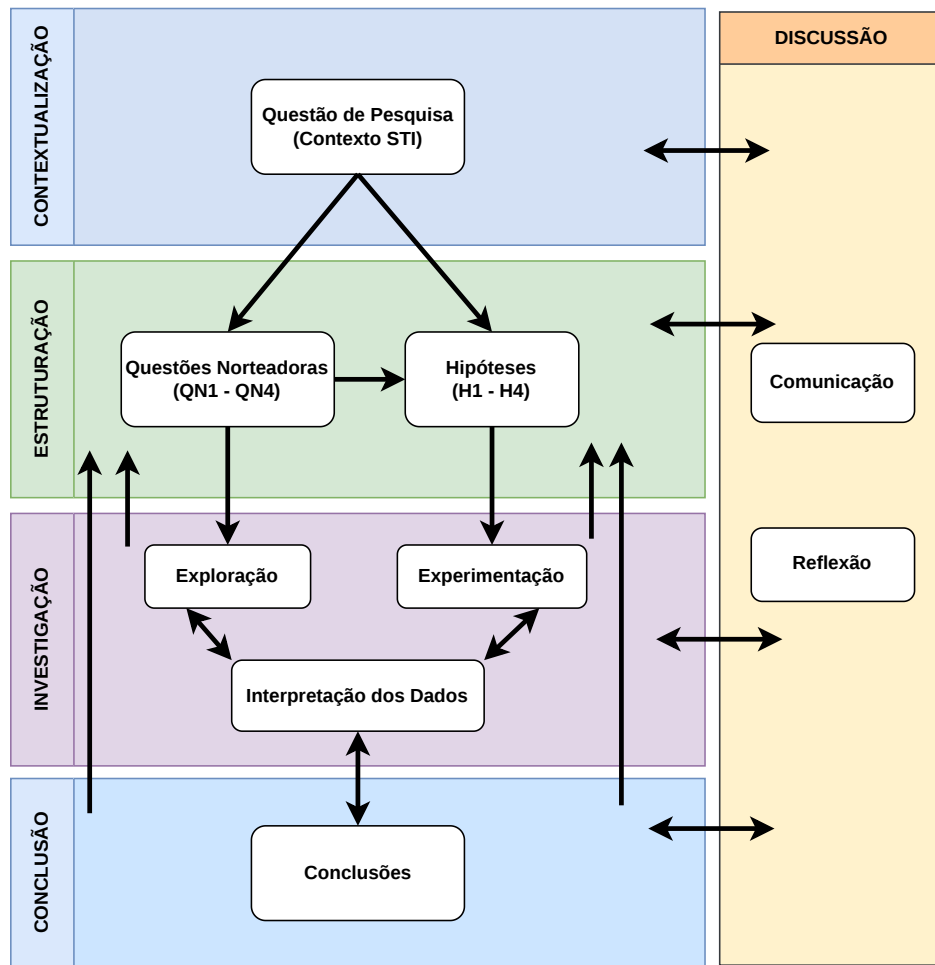
Quanto aos objetivos, a pesquisa tem caráter exploratório, pois investiga um problema ainda pouco estruturado na literatura que é a integração entre representação de habilidades, inferência cognitiva e automação. Além disso, o caráter descritivo deve-se ao fato de caracterizar os elementos envolvidos nesse processo e suas relações. A pesquisa tem natureza incremental, pois foi desenvolvida a partir de uma sequência de estudos.

A estratégia metodológica adotada nesta tese foi estruturada com base em um ciclo investigativo iterativo e adaptável, inspirado no arcabouço conceitual proposto por [Pedaste et al. \(2015\)](#). Cabe notar uma distinção de escopo nesta aplicação, que, embora o modelo original tenha sido concebido para mapear os processos cognitivos da aprendizagem baseada em investigação, nesta pesquisa ele foi transposto para o nível macro-metodológico, servindo como a estrutura operacional do desenvolvimento científico. Assim, o processo é organizado em etapas interdependentes: **Contextualização**, **Estruturação**, **Investigação** e **Conclusão**.

Essa abordagem permite compreender a pesquisa como um processo dinâmico e evolutivo, no qual os achados obtidos em cada etapa retroalimentam e refinam as fases subsequentes. A Figura 1 apresenta a organização da metodologia adotada, evidenciando a relação sistemática entre o problema de pesquisa, as questões norteadoras, as hipóteses e as etapas dos estudos conduzidos ao longo desta tese.

Nesta pesquisa, a etapa de **Contextualização** corresponde à definição do problema e à contextualização da avaliação de habilidades nos Sistema Tutor Inteligente. A etapa de **Estruturação** envolve a formulação das questões norteadoras (QN1–QN4) e das hipóteses

Figura 1 – Ciclo investigativo da metodologia



Fonte: adaptado de [Pedaste et al. \(2015\)](#).

(H1–H4), que estruturam o direcionamento científico.

A etapa de **Investigação** compreende duas dimensões: exploração e experimentação. A etapa de **Exploração** inclui a revisão da literatura e estudos iniciais, como o mapeamento dos STI ([Lopes e Netto, 2020](#)). Isso possibilitou identificar lacunas na área. A **Experimentação** abrange os estudos empíricos conduzidos ao longo da pesquisa, incluindo a aplicação do MDC em um STI ([Lopes e Netto, 2023](#)) e a investigação do uso de LLM para geração da vinculação de item-CC ([Lopes e Netto, 2026](#)). Essas etapas convergem para a interpretação dos dados, na qual são analisados os aspectos estruturais e funcionais das abordagens propostas.

Por último, a **Conclusão** é à consolidação dos resultados obtidos, culminando na proposição de uma abordagem arquitetural para avaliação de habilidades, integrando representação, inferência e automação. E, por fim, as atividades de **Discussão**, representadas por **Comunicação** e **Reflexão**, que ocorrem ao longo de todo o ciclo investigativo. A Comunicação está

associada às publicações científicas desenvolvidas durante o doutoramento, enquanto a reflexão corresponde ao processo contínuo de refinamento da pesquisa.

1.7 Delimitação do Escopo da Tese

Esta tese propõe uma abordagem arquitetural para a avaliação estruturada de habilidades voltada para os STI, integrando três dimensões principais: (i) a representação das relações item-CC, (ii) a inferência do domínio cognitivo por meio do MDC e (iii) a automação parcial dessa representação com apoio assistido de LLM. A proposta busca reduzir a dependência de processos manuais, aumentar a escalabilidade e preservar a interpretabilidade dos diagnósticos gerados.

Ressalta-se que a proposta desta tese não contempla a implementação completa e integrada de todos os componentes da arquitetura em um único sistema, tampouco a realização de experimentos em larga escala com estudantes reais. Essas limitações decorrem do escopo da pesquisa, cujo foco está na definição e validação dos elementos fundamentais do processo de avaliação e da sua organização arquitetural. Dessa forma, a integração completa da arquitetura em ambientes educacionais reais, bem como a condução de estudos longitudinais com estudantes, são consideradas como trabalhos futuros.

Sendo assim, o escopo desta pesquisa concentra-se na proposição, formalização e validação parcial de uma abordagem arquitetural para avaliação de habilidades nos STI, sustentada por evidências experimentais e pela integração dos estudos desenvolvidos ao longo desta linha de pesquisa.

1.8 Evolução da Pesquisa e a Difusão dos Resultados

A presente tese foi desenvolvida a partir de uma sequência de estudos realizados ao longo do período de doutoramento, refletindo uma evolução progressiva da pesquisa na área dos STI e avaliação de habilidades cognitiva dos estudantes.

Inicialmente, foi conduzido um estudo de mapeamento das produções brasileiras sobre os STI ([Lopes e Netto, 2020](#)), de natureza qualitativa, realizada em bases como SCOPUS, IEEE, CBIE e BDTD, buscando identificar as lacunas relacionadas à avaliação de habilidades nos STI e conhecimento, além disso, dos tipos de propostas para novas arquiteturas voltadas para avaliação de habilidades e competências. Essa primeira fase contribui diretamente para a validação da hipótese H1.

Posteriormente, a pesquisa avançou para a definição da arquitetura de um STI, estabele-

cendo bases estruturais para integração de componentes educacionais e apoio pedagógico. Como primeira proposta, foram exploradas abordagens baseadas em Agentes Pedagógicos, com foco na recomendação e intervenção a partir das ações dos estudantes (Lopes e Netto, 2021). Essa etapa permitiu investigar como as informações derivadas do comportamento dos discentes poderiam ser utilizadas para apoiar os processos adaptativos, constituindo o primeiro passo na direção da avaliação estruturada de habilidades.

Na sequência, foram incorporados mecanismos de análise e visualização de dados educacionais, permitindo melhor compreensão do comportamento dos estudantes no sistema (Lopes e Netto, 2023). A partir dessa evolução, foi implementada uma abordagem para avaliação de habilidades cognitivas baseada no MDC, permitindo a inferência do domínio de habilidades em um STI existente.

Por fim, a pesquisa avançou para a automação da representação das relações itens-CC, por meio do uso de LLM, investigando sua viabilidade estrutural e funcional (Lopes e Netto, 2026). A Figura 2 apresenta uma visão geral do histórico desta tese, evidenciando como os diferentes estudos contribuíram para a consolidação da abordagem proposta.

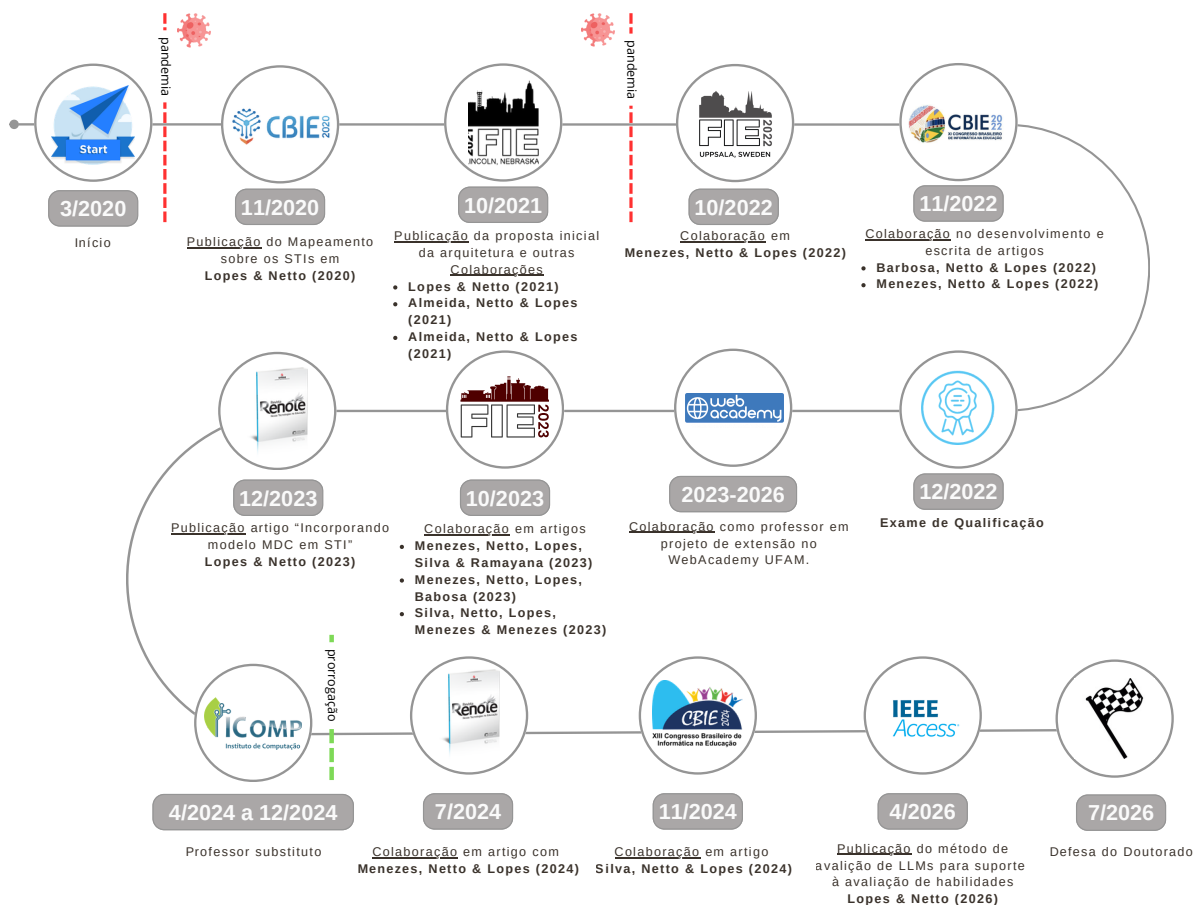
Além dos estudos teóricos e experimentais, esta pesquisa foi fortemente influenciada pela atuação do autor desta tese nas atividades de ensino e extensão, que proporcionaram um ambiente real para observação e refinamento das propostas desenvolvidas. A experiência como professor substituto e a participação nos projetos de extensão, contribuíram para a compreensão prática dos desafios associados à avaliação de habilidades nos ambientes educacionais reais. Essas atividades possibilitaram a observação do comportamento dos estudantes, o uso das ferramentas digitais no apoio ao ensino e a identificação das limitações nos processos tradicionais de avaliação.

Contudo, as experiências não configuram etapas formais de validação experimental, mas desempenham um papel relevante no amadurecimento da proposta, influenciando na definição das questões de pesquisa, na formulação das hipóteses e na evolução desta tese.

1.8.1 Publicações e colaborações

Esta seção apresenta as publicações científicas desenvolvidas no decorrer desta pesquisa, diretamente relacionadas ao tema da tese. Esses trabalhos refletem a evolução do problema investigado, bem como o desenvolvimento incremental das soluções propostas, abrangendo desde a identificação de lacunas na literatura até a proposição de abordagens voltadas à avaliação de habilidades nos STI.

Figura 2 – Histórico da pesquisa.



Fonte: Elaborado pelos autores (2026).

As publicações a seguir correspondem às contribuições centrais desta tese, evidenciando as diferentes etapas do processo investigativo:

- LOPES, A. M. M.; NETTO, J. F. de M. Sistemas Tutores Inteligentes: um Mapeamento das Produções Brasileiras. In: SBC. *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. Evento Virtual, Natal, RN, Brasil: SBC, 2020. p. 1313–1322. DOI: <https://doi.org/10.5753/cbie.sbie.2020.1313>. (Qualis A3)
- LOPES, A. M. M.; NETTO, J. F. D. M. Designing Pedagogical Agents Toward the Recommendation and Intervention Based on Students' Actions in an ITS. In: IEEE. *2021 IEEE Frontiers in Education Conference (FIE)*. Lincoln, NE, USA: IEEE, 2021. p. 1–8. DOI: <https://doi.org/10.1109/FIE49875.2021.9637406>. (Qualis A2)
- LOPES, A. M. M.; NETTO, J. F. d. M. Incorporando uma Abordagem para Avaliação de Habilidades Cognitivas em um Sistema Tutor Inteligente. *Revista Novas Tecnologias na*

Educação — RENOTE, SBC, v. 21, p. 178–187, 2023. DOI: <https://doi.org/10.22456/1679-1916.137738>. (Qualis A4)

- LOPES, A. M. M.; NETTO, J. F. D. M. Large Language Models for Learning Technologies: Structural and Functional Evaluation of Automated Q-Matrix Generation for Educational Assessment. *IEEE Access*, IEEE, v. 14, p. 58466–58483, 2026. DOI: <https://doi.org/10.1109/ACCESS> (Qualis A1)

Além das contribuições diretamente vinculadas a esta tese, o referencial teórico e as abordagens desenvolvidas também influenciaram na participação em outras pesquisas, realizadas em colaboração com diferentes autores, conforme descritas a seguir:

- ALMEIDA, T. O.; NETTO, J. F. de M.; LOPES, A. M. M. Recommendation Model of Personalized Teaching Materials in E-Learning Environments. In: IEEE. *2021 IEEE Frontiers in Education Conference (FIE)*. Lincoln, NE, USA: IEEE, 2021. p. 1–7. DOI: <https://doi.org/10.1109/FIE49875.2021.9637286>. (Qualis A2)
- ALMEIDA, T. O.; NETTO, J. F. D. M.; LOPES, A. M. M. Multi-agent System for Recommending Learning Objects in E-learning Environments. In: IEEE. *2021 IEEE Frontiers in Education Conference (FIE)*. Lincoln, NE, USA: IEEE, 2021. p. 1–5. DOI: <https://doi.org/10.1109/FIE49875> (Qualis A2)
- BARBOSA, F. M.; NETTO, J. F. de M.; LOPES, A. M. M. Uma Arquitetura de STI baseada nas Experiências de Aprendizagem Mediadas para o Ensino de Trigonometria. In: *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*. Manaus, Amazonas: SBC, 2022. p. 1066–1076. DOI: <https://doi.org/10.5753/sbie.2022.225210>. (Qualis A3)
- MENEZES, M. F. de; NETTO, J. F. de M.; LOPES, A. M. M. Visualização de dados educacionais: Desenvolvendo uma ferramenta para análise de dados de sistemas tutores inteligentes. *Revista Novas Tecnologias na Educação – RENOTE*, SBC, v. 22, n. 1, p. 207–216, 2024. DOI: <https://doi.org/10.22456/1679-1916.141548>. (Qualis A4)
- MENEZES, M. F. D.; NETTO, J. F. d. M.; LOPES, A. M. M.; SILVA, R. C.; MENEZES, R. A. The Proposal of a Dashboard for Analysis and Visualization of Educational Data of Intelligent Tutoring Systems. In: IEEE. *2023 IEEE Frontiers in Education Conference – (FIE)*. College Station, TX, USA: IEEE, 2023. p. 1–8. DOI: <http://doi.org/10.1109/FIE58773.2023.10343000>. (Qualis A2)

- MENEZES, M. F. D.; NETTO, J. F. d. M.; LOPES, A. M. M.; BARBOSA, F. M. D. Applying Educational Data Mining to Classify Students in an Intelligent Tutoring System for Algebra Instruction. In: IEEE. *2023 IEEE Frontiers in Education Conference – (FIE)*. College Station, TX, USA: IEEE, 2023. p. 1–8. DOI: <https://doi.org/10.1109/FIE58773.2023.10342993>. (Qualis A2)
- MENEZES, M. F. D.; NETTO, J. F. D. M.; LOPES, A. M. M. Educational Data Mining: Analysis Based On an Intelligent Tutoring System for Teaching Algebra. In: IEEE. *2022 IEEE Frontiers in Education Conference – (FIE)*. Uppsala, Sweden: IEEE, 2022. p. 1–7. DOI: <https://doi.org/10.1109/FIE56618.2022.9962484>. (Qualis A2)
- SILVA, R. C. da; NETTO, J. F. de M.; MENEZES, M. F. de; LOPES, A. M. M. Percepções dos Estudantes sobre o Laboratório Remoto de Ensino de Robótica e Programação ERPLAB. In: *Simpósio Brasileiro de Informática na Educação (SBIE)*. Rio de Janeiro, RJ: SBC, 2024. p. 951–960. DOI: <https://doi.org/10.5753/sbie.2024.242610>. (Qualis A3)
- SILVA, R. C. da; NETTO, J. F. de M.; LOPES, A. M. M.; MENEZES, M. F. D.; MENEZES, R. A. ERPLAB: Remote Laboratory for Teaching Robotics and Programming. In: *2023 IEEE Frontiers in Education Conference (FIE)*. College Station, TX, USA: IEEE, 2023. p. 1–7. DOI: <https://doi.org/10.1109/FIE58773.2023.10343379>. (Qualis A2)

1.9 Aspectos Éticos e Integridade Acadêmica

As ferramentas de IA generativa foram utilizadas como apoio à organização do documento de tese. A descrição desse uso é apresentada no Apêndice C.

1.10 Organização do Documento

Esta tese está organizada em seis capítulos. Este primeiro capítulo apresenta o contexto e a delimitação da pesquisa, o problema investigado, a questão de pesquisa e as questões norteadoras, as hipóteses, a motivação e a justificativa, os objetivos e os aspectos relacionados à ética e à integridade acadêmica.

O capítulo 2 apresenta a fundamentação teórica da pesquisa. Inicialmente, na Seção 2.1 são abordados os conceitos relacionados aos Sistema Tutor Inteligente, incluindo sua evolução e seus principais componentes: o Modelo do Estudante, o Modelo do Domínio, o Modelo Tutor e a Interface de interação. Na sequência, são discutidos os fundamentos da

avaliação da aprendizagem, do diagnóstico cognitivo e do uso de IA no apoio aos processos de avaliação.

Na Seção 2.2, são apresentadas as principais funções da avaliação da aprendizagem, com ênfase nas avaliações formativa, somativa e diagnóstica, além dos conceitos relacionados à avaliação de habilidades.

A seção 2.3 apresenta os fundamentos do diagnóstico cognitivo, incluindo os Modelos de Diagnóstico Cognitivo, a Q-matriz, a relação item-CC e os mecanismos utilizados para inferir os perfis de domínio dos estudantes.

A seção 2.4 aborda o uso de Modelo de Linguagem de Grande Escala como apoio ao processo de avaliação, com ênfase na representação das relações item-CC e na automação parcial e assistida da construção e revisão de Q-matrizes.

O capítulo 3 apresenta uma revisão narrativa dos trabalhos relacionados à pesquisa, considerando os estudos sobre os STI, modelagem do estudante, avaliação diagnóstica, MDC, representação de habilidades, construção e validação de relações item-CC, adaptação pedagógica e uso dos LLM em tarefas relacionadas à avaliação educacional.

O capítulo 4 apresenta a arquitetura CAASA-ITS, descrevendo seus princípios, módulos, componentes e fluxos de funcionamento. A arquitetura integra a representação item-CC, a autoria de avaliações, a geração assistida e a validação da Q-matriz, a inferência cognitiva por meio do MDC, a Modelagem do Estudante, a visualização diagnóstica e o apoio à adaptação pedagógica.

O capítulo 5 apresenta os resultados alcançados ao longo da pesquisa, organizados em três estudos que contribuíram progressivamente para a formalização da arquitetura CAASA-ITS. O primeiro estudo aborda uma arquitetura multiagente para recomendação e intervenção pedagógica para os STI; o segundo investiga a incorporação do modelo G-DINA ao STI LEIA para avaliação diagnóstica de habilidades; e o terceiro avalia o uso do LLM no apoio à geração assistida de Q-matrizes, considerando sua qualidade estrutural e seu desempenho funcional. O capítulo também descreve os aspectos de implementação, configuração experimental e análise computacional associados aos estudos realizados.

Por fim, o capítulo 6 apresenta as conclusões da tese, retomando os objetivos, a questão de pesquisa, as questões norteadoras e as hipóteses investigadas. O capítulo aponta para a arquitetura CAASA-ITS como contribuição central, além das específicas que foram fundamentais para conclusão desta tese. Por último, no capítulo é discutido as limitações identificadas e as possibilidades de trabalhos futuros.

Capítulo

2

FUNDAMENTAÇÃO TEÓRICA

2.1 Sistema Tutor Inteligente

A busca pela personalização do ensino é um dos ideais mais antigos e persistentes da pedagogia. A imagem do tutor humano, capaz de adaptar sua instrução às necessidades, ao ritmo e ao nível de conhecimento de cada aluno, é fundamental para a eficácia educacional. É nesse contexto que o Sistema Tutor Inteligente (STI) emergem como uma resposta tecnológica promissora.

Os STI são sistemas computacionais concebidos para fornecer instrução e respostas personalizadas sem a intervenção direta de um ser humano, visando replicar os aspectos mais cruciais da tutoração individual (Giraffa e Kohls-Santos, 2023).

2.1.1 Definição e evolução

O desenvolvimento da Instrução Assistida por Computador (do inglês *Computer-Assisted Instruction* – CAI) nas décadas de 1960 e 1970 marcou o início dos STI. A expressão “Sistema Tutor Inteligente” foi introduzida pela primeira vez em 1982. Seis anos depois, em 1988, a ideia foi amplamente consolidada na primeira conferência *Intelligent Tutoring Systems – ITS International Conferences*¹ (AlShaikh e Hewahi, 2021). Desde então, houve um aprimoramento nas pesquisas e estudos relacionados ao desenvolvimento, abrangendo uma gama de domínios de aprendizado, desde ciências exatas até ciências humanas (Oliveira, Isotani e Bittencourt, 2023).

O desenvolvimento dos CAI era influenciado pelo *behaviorismo skinneriano*, focando no condicionamento operante por meio de estímulos, respostas e reforços imediatos (Júnior e Silva, 2022). A arquitetura dos CAI era baseada em programas lineares, nos quais o conhecimento era fragmentado em pequenas unidades denominadas *frames*, apresentadas de maneira sequencial e inflexível. Nesse estágio, o sistema era passivo em relação ao raciocínio do aprendiz, limitando-se a registrar acertos e erros sem compreender a lógica subjacente à

¹ <https://iis-international.org/its/>

resposta.

Diferentemente, os STI são capazes de guiar o aprendiz por um currículo composto, geralmente, por uma série de tarefas, enquanto monitora o progresso do aprendiz (Silva, 2011). Nos STI, as técnicas de Inteligência Artificial (IA) são aplicadas para decidir como o conteúdo será apresentado, assim como os conteúdos que deverão ser apresentados e como modelar o conhecimento adquirido pelo estudante até o dado momento de uso do sistema. Esses sistemas são capazes de tomar decisões pedagógicas em tempo real (em execução), guiando o estudante no processo de aprendizagem e adaptando os conteúdos às características (Lopes e Netto, 2020; Zerkouk, Mihoubi e Chikhaoui, 2025).

2.1.2 Arquitetura clássica dos STI

Para compreender como os STI realizam a avaliação do estudante, é essencial destacar sua arquitetura funcional. Embora existam variações, ela é, geralmente, composta por quatro módulos principais que interagem de forma sinérgica para criar um ciclo de tutoração, avaliação e retroação imediata (*feedback*²).

Tradicionalmente modular, a arquitetura de um STI é projetada para separar as diferentes formas de conhecimento necessárias para uma instrução eficaz. Embora as implementações variem, a literatura apresenta uma arquitetura considerada “tradicional” ou também chamada de **meta-arquitetura** (Raabe, 2005), composta por quatro componentes que atuam para criar uma experiência de aprendizagem personalizada. A Figura 3 mostra a arquitetura considerada “clássica” dos STI e a interação entre os módulos que a compõem.

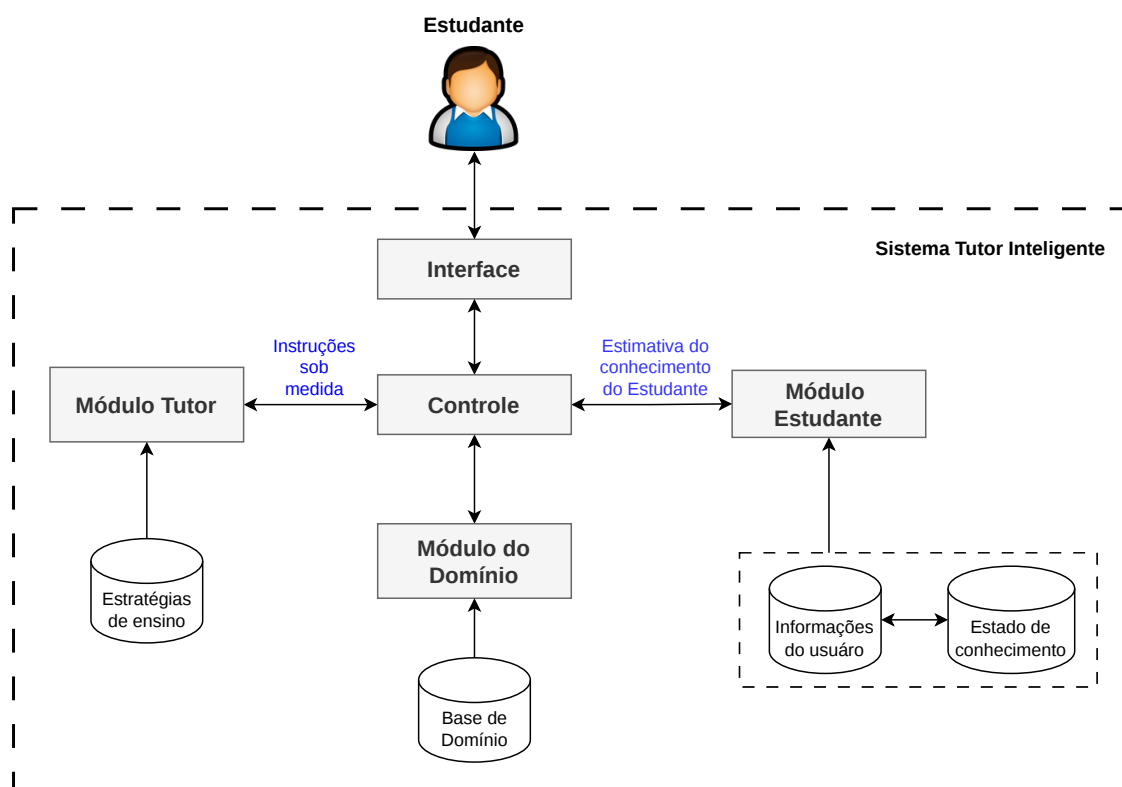
2.1.2.1 Módulo do Domínio (Expert Model)

O Módulo do Domínio representa a *expertise* do sistema. Ele contém os conceitos, regras e estratégias de resolução de problemas referentes ao conteúdo que será ensinado. Sua função é dupla: servir como fonte de conhecimento e padrão para avaliar o desempenho do estudante (Singh e Singh, 2022). Nos STI clássicos (Aleven, 2010; Jaques et al., 2013; Lopes et al., 2019), esse conhecimento é estruturado em redes semânticas ou regras de produção do tipo “IF-THEN”. Outros trabalhos, como os de Lefebvre et al. (2005) e Sklavakis e Refanidis (2014) organizam o conhecimento do domínio em ontologias baseadas na definição formal de termos e relações, permitindo ao sistema realizar inferências semânticas.

Com o avanço da tecnologia, o Módulo do Domínio evoluiu para dar suporte aos domí-

² Nesta tese, o termo *feedback* é entendido como devolutiva ou retorno pedagógico produzido a partir das evidências do processo de ensino e aprendizagem registradas pelo sistema. Esse retorno pode informar erros, acertos, lacunas de conhecimento, orientações de resolução ou recomendações de estudo.

Figura 3 – Arquitetura tradicional de um STI.



Fonte: adaptado de Raabe (2005).

nios menos estruturados. Nas arquiteturas recentes, ele pode ser representado por grafos de conhecimento, construídos automaticamente a partir de livros didáticos e de vídeos usando técnicas de Processamento de Linguagem Natural (PLN) (Castilhos e Jaques, 2024; Luise, Arevalillo-Herráez e Arnau, 2023). A construção manual da base de conhecimento, tradicionalmente custosa e dependente de especialistas, vem sendo substituída por mecanismos de fusão *cross-modal* que alinham explicações auditivas com conteúdo visual-textual para criar grafos instrucionais unificados (Deng e Yuan, 2026; Akyildiz e Eken, 2025; Wang et al., 2020).

2.1.2.2 Módulo Tutor (Tutor Model)

O Módulo Tutor é responsável pela estrutura didática e pedagógica, ou seja, reúne o conhecimento necessário para decidir sobre quais estratégias de ensino serão aplicadas ao conteúdo a ser trabalhado no STI. Em linhas gerais, ele abrange as estratégias de ensino, que se somam aos processos pedagógicos e à interação com o estudante durante o processo de aprendizagem.

Esse módulo encapsula as estratégias e táticas de ensino. Ele funciona como o “pedagogo” do sistema, decidindo “o Quê, Quando e Como?” ensinar (Abdelshiheed, Barnes e Chi, 2024;

[Zerkouk, Mihoubi e Chikhaoui, 2025](#)). Com base nas informações extraídas do Modelo do Estudante, este módulo define as **estratégias ensino** que corresponde ao planejamento de como ensinar e ao processo de adaptação. Essas estratégias incluem a seleção do nível de dificuldade, a definição da sequência dos conteúdos e a escolha da abordagem pedagógica mais apropriada como, por exemplo, o método socrático ou a aprendizagem baseada em casos. Essas estratégias podem ser aplicadas considerando o estado de conhecimento do aluno, ou afetividade e desempenho, bem como outras características individuais.

Além disso, o módulo contém as estratégias instrucionais, que são as ações concretas a serem realizadas durante a interação com o estudante. Segundo [Sottolare, Falco e Connor \(2014\)](#), elas são:

- **Feedback sensível ao erro:** fornecer mensagens específicas e oportunas quando o estudante comete erros, equilibrando frequência e nível de detalhes para evitar frustração ([Heeren e Jeuring, 2014](#));
- **Aprendizagem por domínio (*mastery learning*):** exige que o aprendiz alcance um nível satisfatório de compreensão antes de avançar para novos conteúdos ([Paiva, Ferreira e Frade, 2017](#));
- **Espaçamento e repetição adaptativa:** distribui revisões ao longo do tempo para fortalecer a retenção e promover uma aprendizagem duradoura;
- **Prompting metacognitivo:** incentiva o estudante a refletir, explicar e corrigir seu próprio raciocínio, favorecendo a construção de modelos mentais ([Kautzmann e Jaques, 2016](#));
- **Exemplos resolvidos (*worked examples*):** apresenta soluções passo a passo que gradualmente as partes são deliberadamente removidas ou atenuadas, estimulando a autonomia do aprendiz ([Lopes e Netto, 2021](#)).

Combinadas com as estratégias adaptativas baseadas nas informações do aprendiz, as técnicas instrucionais permitem que o STI personalize o ensino, aumentando sua eficácia da aprendizagem e promovendo um maior engajamento e persistência ([Holden e Sinatra, 2014](#)).

2.1.2.3 Módulo do Estudante (Learner/Student Model)

O Modelo do Estudante é descrito como componente vital do STI, pois a partir dele se permite personalizar a experiência da aprendizagem ([Farid et al., 2018](#)). Esse módulo

contém informações sobre o conhecimento, os padrões, os atributos e os comportamentos do estudante. Também são armazenadas informações coletadas sobre as concepções errôneas, lacunas de conhecimento e entre outras (Pelánek, 2017). Essas informações são fundamentais para que o sistema decida qual ação pedagógica deve ser tomada durante o ciclo de tutoramento. Com base no nível de conhecimento e no tipo de situação, o STI pode apresentar alguma informação ou ajudar na solução da atividade (Ferreira, 2018).

Segundo Sani, Bichi e Ayuba (2016), a modelagem do usuário tem origem no final da década de 1970, com a introdução dos conceitos de “modelo de usuário” e “modelagem do usuário”. Estes estudos iniciais contribuíram para o desenvolvimento de sistemas capazes de coletar informações e adaptar as respostas conforme aos seus perfis de usuários.

Inicialmente, a modelagem visava representar crenças, objetivos, intenções, preferências e comportamentos, sobretudo, em contextos de interação e comunicação com os sistemas computacionais. Neste processo, a utilização de estereótipos foi uma das primeiras estratégias utilizadas para diferenciar os usuários e inferir as suas preferências, tendo influenciando posteriormente as pesquisas sobre personalização e sistemas adaptativos (Sani, Bichi e Ayuba, 2016).

A forma como o conhecimento do aluno é representado e avaliado em relação ao conhecimento do especialista (Módulo de Domínio) é uma decisão de projeto. Segundo Vicari e Giraffa (2003) e Sani, Bichi e Ayuba (2016) as abordagens clássicas são:

- **Modelo de Sobreposição (*Overlay Model*):** esta é a abordagem mais simples, em que o conhecimento do aprendiz é considerado um subconjunto do conhecimento do especialista. O modelo consiste essencialmente em uma lista de tópicos ou competências da área em questão, sendo cada um marcado como “conhecido” ou “desconhecido”. Nesta perspectiva, os erros são interpretados como meras lacunas de conhecimento, ou seja, como ausência de uma informação correta. A avaliação, portanto, busca identificar quais partes do conhecimento do especialista o estudante ainda não adquiriu.
- **Modelo de Perturbação (*Buggy Model*):** esta abordagem é mais sofisticada porque reconhece que os erros do aprendiz não se deve à mera ausência de conhecimentos, mas sim à presença de conhecimentos incorretos ou mal formados, ou seja, às conhecidas “concepções errôneas” ou “bugs”. O modelo, portanto, não representa apenas o que o aluno sabe, mas também o que ele “sabe errado”. No modelo de perturbação, existe uma base de conhecimento e uma biblioteca de erros típicos. Este modelo permitem

um tratamento mais inteligente dos comportamentos incorretos do estudante; embora, ainda seja limitado, sobretudo devido à estrutura da biblioteca de erros (Vicari e Giraffa, 2003).

- **Modelo diferencial:** é considerado uma melhoria em relação ao Modelo de Sobreposição. Não pressupõe que todas as lacunas de conhecimento dos estudantes sejam indesejáveis. O conhecimento do aluno é dividido em duas categorias: o conhecimento que o aluno deveria ter e o conhecimento que não se poderia esperar que o aluno tivesse (Sani, Bichi e Ayuba, 2016). Em suma, essa modelagem, compara o desempenho do especialista com o do estudante, e não o seu conhecimento. O Modelo Diferencial continua a enfrentar a maioria das mesmas dificuldades que o Modelo *Overlay*, uma vez que continua a pressupor que o Modelo do Estudante é essencialmente um subconjunto do Modelo do Domínio.
- **Modelo Afetivo:** este modelo baseia-se em estudos que envolvem a área do conhecimento sobre da Computação Afetiva. Este modelo é utilizado para inferir os estados emocionais do aluno, como engajamento, confusão, frustração ou tédio, durante a interação (Maldaner, Pozzebon e Santos, 2022; Morais e Jaques, 2022). Como exemplo, se um erro cometido por um aluno em estado de alta frustração pode indicar um bloqueio afetivo em vez de uma lacuna de conhecimento, exigindo uma intervenção pedagógica diferente como, por exemplo, uma mensagem de encorajamento em vez de mais prática (Ascari, Gottardo e Pimentel, 2020).
- **Outros modelos:** abordagens adicionais incluem o **Modelo de Simulação** no qual o ambiente possui um modelo que representa a forma como o estudante pode ou deve se comporta-se em uma situação específica. Por conseguinte, este modelo permite prever o comportamento futuro do estudante (Vicari e Giraffa, 2003). Outro modelo, chamado de **Modelo de Crenças**, representa o grau de confiança do sistema de que o aluno compreende um determinado conceito (Bremgartner, 2017; Giraffa, 1999).

A escolha entre estes modelos conceituais não é apenas meramente técnica; reflete uma postura epistemológica e pedagógica. Por exemplo, um sistema que adota o Modelo de *Overlay* opera sob uma premissa mais instrutivista, segundo a qual a aprendizagem consiste em acumular fatos e procedimentos corretos. Em contraste, um sistema que utiliza o Modelo de Perturbação aproxima-se mais de uma visão construtivista, reconhecendo que os

estudantes constroem ativamente os seus próprios modelos mentais, que podem ser falíveis. Neste caso, a avaliação torna-se um ato de diagnóstico para compreender e corrigir essas construções mentais. Por outro lado, o Modelo Diferencial já utiliza uma abordagem baseada nas Zonas de Desenvolvimento Proximal (ZDP) de [Vygotsky \(1978\)](#).

Por fim, a arquitetura técnica do Modelo do Estudante não é pedagogicamente neutra. O modelo codifica uma teoria sobre a natureza do conhecimento e do erro que, por sua vez, define as estratégias disponíveis para o Módulo do Tutor. A eficácia de um STI depende criticamente da coerência entre a sua implementação técnica e a sua fundamentação pedagógica ([Zerkouk, Mihoubi e Chikhaoui, 2025](#)). Assim, o Modelo do Estudante funciona como uma representação computacional do aprendiz, permitindo ao tutor personalizar conteúdos, selecionar atividades e fornecer *feedbacks*.

2.1.2.4 Módulo de Interface (Interface Model)

A interface é o canal de comunicação entre o aluno e o sistema. Através da interface são apresentadas ao aluno as “perguntas do teste” (problemas, simulações e tarefas) e são recolhidas suas “respostas” (ações, entradas de texto e manipulação de objetos) ([Vicari e Giraffa, 2003](#)).

Uma interface bem projetada é crucial para o sucesso do STI, na medida que deve facilitar uma interação clara e envolvente, permitindo que o sistema acompanhe o progresso de forma não intrusiva ([Andrade, 2003](#)). Segundo [Jesus \(2003\)](#), é importante que a interface seja capaz de adaptar-se as diferentes necessidades de cada usuário, fazendo sugestões e fornecendo explicações. Os avanços em multimídia e das interfaces de *web* têm enriquecido significativamente as possibilidades de interação ([Júnior e Silva, 2022](#)) com a inserção de agentes de conversação ([Lopes, Netto e Lima, 2018](#); [Nye, Graesser e Hu, 2014](#); [Paladines e Ramirez, 2020](#)) e novos modelos de IA com uso de LLM ([Scarlatos et al., 2025](#)).

Nas arquiteturas contemporâneas dos STI, tem-se observado uma iniciativa cada vez mais frequente que é a inclusão do professor como utilizador ativo no processo de ensino e aprendizagem mediado pelo sistema. Como exemplo, os STI MAZK ([Valeriano, Corrêa e Pozzebon, 2019](#)), LEIA ([Lopes et al., 2019](#)), CURUMIM ([Barbosa, Netto e Lopes, 2022](#)) e o W-ITS ([Wong et al., 2010](#)).

Esta inclusão resulta do reconhecimento de que a adaptação pedagógica automatizada, embora relevante, não elimina a necessidade da mediação docente. O professor pode atuar na definição do domínio de conhecimento, na validação das habilidades modeladas, na seleção

ou revisão das estratégias de ensino, na interpretação dos diagnósticos produzidos a partir do Modelo do Estudante e na proposição de intervenções complementares (Raabe, 2005).

2.1.2.5 Módulo de Controle

O Módulo de Controle apresentado na Figura 3 pode ser compreendido como o componente responsável por coordenar a interação entre os demais módulos. A sua função é utilizar as informações sobre o estado de conhecimento do estudante, a base de domínio e as estratégias de ensino para orientar as decisões pedagógicas do sistema (Raabe, 2005). Nesta perspectiva, o controle pode ser relacionado às funções de *Outer Loop* e *Inner Loop* nos STI.

Segundo Essa (2016), o *Outer Loop* refere-se às decisões de sequenciamento pedagógico, como a seleção da próxima atividade, conteúdo ou habilidade a ser trabalhada. Já o *Inner Loop*, por sua vez, corresponde ao acompanhamento da resolução de uma tarefa específica, permitindo que o sistema interprete respostas, identifique erros, forneça *feedbacks*, dicas e instruções personalizadas (Pelánek, 2017). Dessa maneira, o Módulo de Controle atua como um mecanismo de coordenação pedagógica, transformando as informações do estudante e do domínio em decisões adaptativas durante o processo de aprendizagem (Raabe, 2005).

2.2 Avaliação da Aprendizagem nos Ambientes Educacionais

A avaliação constitui um dos principais processos no contexto educacional, pois possibilita a obtenção de evidências sobre a aprendizagem dos estudantes, além de subsidiar decisões relacionadas ao planejamento, ao acompanhamento e ao aperfeiçoamento do processo de ensino (Lara, Aljawarneh e Pamplona, 2020). Ao contrário de uma visão restrita à atribuição de notas, a avaliação é um processo contínuo de coleta, análise e interpretação de informações sobre o desempenho do estudante. Isso permite compreender seu estado de aprendizagem e orientar intervenções pedagógicas mais adequadas (Sun e Suzuki, 2013).

Nos STI, a avaliação exerce um papel ainda mais abrangente, pois as evidências produzidas durante a interação do estudante são utilizadas para construir representações computacionais de seu estado de conhecimento (Pelánek, 2017). Isso permite acompanhar sua evolução e apoiar mecanismos de adaptação pedagógica. Nessa perspectiva, a avaliação deixa de ser apenas um instrumento de verificação da aprendizagem para se tornar um componente responsável por alimentar o Modelo do Estudante e fornecer as informações necessárias para personalização do processo de ensino (Pelánek, 2025). Em resumo, as avaliações fornecem aos

professores informações sobre o que os estudantes aprenderam, quais dificuldades persistem e quais conteúdos ou habilidades devem ser retomados, contribuindo para a elaboração de estratégias de ensino mais eficazes.

Sendo assim, torna-se necessário compreender os fundamentos que sustentam a avaliação educacional. Por isso, nesta seção, são apresentadas as diferentes formas e objetivos da avaliação, evidenciando as características como formativa, somativa e diagnóstica. Em seguida, são discutidos os elementos que estruturam uma avaliação baseada em habilidades, incluindo os itens avaliativos e os Componentes de Conhecimento que representam o domínio do conhecimento. Por fim, são apresentados os fundamentos da Avaliação Diagnóstica Cognitiva, que constituem a base conceitual para os Modelos de Diagnóstico Cognitivo empregados nesta tese.

2.2.1 Funções e objetivos da avaliação

A avaliação da aprendizagem pode assumir diferentes funções de acordo com sua finalidade e com o momento em que seus resultados são utilizados no processo educacional ([Vaz e Nasser, 2021](#)). Entre essas funções, destacam-se a avaliação formativa, a avaliação somativa e a avaliação diagnóstica, que fornecem informações distintas sobre a aprendizagem dos estudantes e subsidiam diferentes decisões pedagógicas ([Bloom, Madaus e Hastings, 1971](#); [Freitas, Costa e Miranda, 2014](#); [Oliveira, 2002](#)).

2.2.1.1 Avaliação formativa

A avaliação formativa ocorre durante o processo de ensino e aprendizagem e tem como finalidade produzir informações que permitam acompanhar o progresso dos estudantes, identificar dificuldades e orientar ajustes nas ações pedagógicas. Dessa forma, as evidências obtidas por meio da avaliação são utilizadas para apoiar decisões que favoreçam a aprendizagem enquanto o processo educacional ainda está em desenvolvimento ([Sadler, 1989](#); [Black e Wiliam, 1998](#); [Oliveira, 2002](#)).

Nesse sentido, a avaliação formativa não se limita à verificação do desempenho, mas integra o próprio processo de ensino e aprendizagem, fornecendo informações que podem ser utilizadas pelo professor e pelo estudante para orientar as etapas seguintes da aprendizagem ([Barreira, Boavida e Araújo, 2006](#)).

2.2.1.2 Avaliação somativa

A avaliação somativa tem como finalidade sintetizar os resultados da aprendizagem alcançados pelo estudante ao término de uma unidade, período ou etapa do processo edu-

cacional (Vaz e Nasser, 2021; Freitas, Costa e Miranda, 2014). Seus resultados podem ser utilizados para expressar o desempenho alcançado e subsidiar decisões relacionadas, por exemplo, à atribuição de notas, certificação ou conclusão de uma etapa de aprendizagem (Bloom, Madaus e Hastings, 1971).

Enquanto a avaliação formativa utiliza as evidências produzidas para orientar ações durante o processo de aprendizagem, a avaliação somativa está principalmente relacionada à síntese dos resultados alcançados em determinado momento ou etapa (Harlen e James, 1997).

2.2.1.3 Avaliação diagnóstica

Por último, a avaliação diagnóstica tem como finalidade identificar os conhecimentos, habilidades, dificuldades e necessidades de aprendizagem dos estudantes, produzindo informações que possam subsidiar o planejamento e a adequação do processo de ensino (Bloom, Madaus e Hastings, 1971).

De acordo com Oliveira (2002), a avaliação diagnóstica (ou inicial) proporciona informações sobre a capacidade do estudante antes de iniciar um processo de ensino e aprendizagem. Freitas, Costa e Miranda (2014) caracterizam a avaliação diagnóstica como o objetivo de verificar a existência ou ausência de habilidades e conhecimentos pré-estabelecidos no início do processo avaliativo, bem como verificar se os alunos dominam os pré-requisitos necessários para novas aprendizagens. Sendo assim, permite que o professor realize uma sondagem, projeção ou uma retrospectiva sobre o desenvolvimento dos alunos, além de ajustar o programa da disciplina às condições identificadas (Ribeiro, 2014).

No último caso, a avaliação é vista como um instrumento que facilita a tomada de decisão e a mediação do conhecimento, tornando a avaliação diagnóstica uma ação norteadora para a aferição do conhecimento (Pinheiro e Rebouças, 2018). Briggs Derek C. Alonzo, Schwab e Wilson (2006) destacam que os professores começaram a utilizar ferramentas diagnósticas para compreender o estado de conhecimento dos alunos e, assim, orientar a instrução e o ensino.

De modo geral, as avaliações fornecem aos professores informações sobre o que os alunos aprenderam e o que ainda é necessário ensinar para eles (Lara, Aljawarneh e Pamplona, 2020). Por isso, os professores utilizam os resultados para aprimorar o ensino e, assim, elaborar um plano de instrução direcionado, visando facilitar o desenvolvimento educacional (Sun e Suzuki, 2013). Contudo, as ciências educacionais e as psicológicas têm se dedicado à

investigação de temas relacionados à inteligência e à cognição (Suehiro, Benfica e Cardim, 2015; Yamaguchi e Okada, 2020; Ravand e Baghaei, 2020), e com isso, têm surgido mecanismos direcionados para Avaliação Diagnóstica Cognitiva (ADC).

No contexto desta tese, a avaliação diagnóstica assume particular importância, pois está relacionada à identificação do estado de conhecimento do estudante. Essa perspectiva fornece a base para a discussão da avaliação cognitiva e dos Modelos de Diagnóstico Cognitivo apresentados na seção 2.3.

2.2.2 Elementos de uma avaliação baseada em habilidades

A modelagem do conhecimento é um dos principais elementos no desenvolvimento de STI e para aprendizagem adaptativa. Esses sistemas necessitam organizar grandes quantidades de conteúdo educacional, exercícios e atividades em estruturas que representem o conhecimento envolvido no processo de aprendizagem. Nesse contexto, surgem os conceitos de Componentes de Conhecimento, habilidades e itens, que desempenham um papel fundamental na representação cognitiva do domínio educacional.

2.2.2.1 Componentes de conhecimento

Os Componentes de Conhecimento (CC), do inglês *Knowledge Components*, podem ser caracterizados como habilidades, conceitos ou atributos cognitivos. De maneira geral, um CC representa uma unidade específica de conhecimento ou habilidade necessária para a resolução de um problema ou execução de uma tarefa (Pelánek, 2020). Por exemplo, exercícios envolvendo “soma de frações” podem pertencer ao mesmo CC relacionado à habilidade de “adição de frações”. Da mesma forma, exercícios sobre “localização de países europeus” podem estar associados ao CC relacionado ao “reconhecimento geográfico de países da Europa”. Essa estrutura possibilita que o sistema acompanhe o domínio de habilidades específicas ao em vez de considerar apenas o desempenho global do estudante.

Pelánek (2020) destaca que os CC podem possuir diferentes níveis de granularidade. Por exemplo, em um nível mais amplos, um CC pode representar uma habilidade geral, como “frações” ou em níveis mais específicos, esse mesmo domínio pode ser subdividido em habilidades menores, como “adição de frações com denominadores iguais” e “adição de frações com denominadores diferentes”. José (2022) destaca que a escolha do nível de granularidade influencia diretamente a precisão dos modelos de diagnóstico e a complexidade da modelagem do domínio.

2.2.2.2 Habilidade

O conceito de habilidade está diretamente relacionado à capacidade do estudante de aplicar conhecimentos para resolver problemas. Segundo Robson et al. (2020), habilidade (do inglês *skill*³) corresponde à capacidade de aplicar efetivamente conhecimentos e experiência para executar tarefas físicas ou mentais. Já Wang et al. (2015) descrevem as habilidades como capacidades cognitivas relacionadas ao uso de conceitos e regras para a resolução de problemas, enquanto Patel, Kinshuk e Russell (2000) associam as habilidades cognitivas aos processos de análise, interpretação e tomada de decisão.

Segundo Schunk (2012), as habilidades podem ainda ser classificadas de acordo com seu nível de especificidade. **Habilidades gerais** estão associadas a competências amplas, como pensamento crítico e resolução de problemas (não questões), enquanto **habilidades específicas** são vinculadas a domínios particulares, como resolução de equações ou manipulação de estruturas algébricas.

Neste trabalho, adota-se a distinção entre CC e habilidade. Conforme Pelánek (2020), os CC podem ser compreendidos como unidades estruturais utilizadas para organizar conteúdos, itens e conceitos dentro do domínio educacional. Sendo assim, ao longo desta tese, o termo CC é utilizado para se referir aos conceitos ou elementos do domínio associados aos itens de avaliação. Já o termo habilidade é utilizado para representar o nível de domínio demonstrado pelo estudante em relação a esses CC (Pelánek, 2020).

Essa distinção é particularmente importante para os Modelos de Diagnóstico Cognitivo (MDC), pois permite separar a estrutura de representação do conhecimento do processo de inferência do domínio (Desmarais, Beheshti e Naceur, 2012). Sendo assim, a relação item-CC é utilizada como base estrutural, enquanto o domínio das habilidades corresponde ao estado cognitivo inferido a partir dos padrões de resposta observados por (Emond, 2020; Lopes e Netto, 2026).

A relação entre itens e os CC também é discutida na Teoria dos Espaços de Conhecimento de Falmagne e Doble (2013). De acordo com Düntsch e Gediga (1995), para estabelecer uma conexão entre as estruturas de conhecimento e modelos baseados na avaliação de habilidades é necessário formalizar essa relação entre os itens do domínio e os atributos de habilidades

³ **Skill ou Ability:** é comum que na tradução do inglês para o português as palavras *Ability* ou *Skill* tenham o mesmo significado de habilidade. No entanto, “*Ability*” costuma ser algo natural, uma capacidade, enquanto “*Skill*” pode ser algo adquirido por meio do treinamento e da prática. Disponível em: <<https://www.englishexperts.com.br/forum/skill-x-ability-habilidade-qual-a-diferenca-t17876.html>>. Acessado em 08 de maio de 2026.

cognitivas.

Nesse contexto, Heller, Ünlü e Albert (2013) propõem uma formalização na qual um domínio de conhecimento é representado por um conjunto não vazio de itens (Q), enquanto (S) denota um conjunto não vazio de habilidades, ou CC, consideradas relevantes para a resolução dos itens em (Q). A relação entre itens e habilidades é definida por meio de uma função chamada de *skill multimap*, representada pela tripla $((Q, S, \mu))$, em que:

$$\mu: Q \rightarrow 2^{2^S} \quad (2.1)$$

Na notação utilizada, 2^S representa o conjunto das partes de S , isto é, o conjunto formado por todos os subconjuntos possíveis de habilidades contidas em S . Por sua vez, 2^{2^S} representa o conjunto das partes de 2^S , ou seja, o conjunto de todas as possíveis coleções desses subconjuntos. Dessa forma, a função μ associa cada item $q \in Q$ não necessariamente a uma única habilidade ou a um único conjunto de habilidades, mas a uma coleção de conjuntos de habilidades que podem ser suficientes para sua resolução.

Assim, para cada item $q \in Q$, $\mu(q)$ corresponde a uma coleção não vazia de subconjuntos não vazios de S :

$$\mu(q) \subseteq 2^S, \quad \mu(q) \neq \emptyset \quad (2.2)$$

$$\forall C \in \mu(q), \quad C \subseteq S \quad \text{e} \quad C \neq \emptyset \quad (2.3)$$

Cada elemento ($C \in \mu(q)$) é denominado competência e representa um conjunto de habilidades suficiente para a resolução do item (q). Desse modo, a formalização permite modelar a existência de diferentes caminhos de solução para um mesmo item, uma vez que mais de uma competência pode ser associada a (q). Assim, uma resposta correta a um item não implica, necessariamente, a identificação de um único conjunto de habilidades dominadas pelo estudante, mas indica que ao menos uma das habilidades associadas ao item pode ter sido mobilizada.

Quando as habilidades associadas a cada item são minimamente suficientes, isto é, quando nenhum conjunto distinto de habilidades contém outro, a *skill multimap* satisfaz a condição de incomparabilidade aos pares:

$$\forall C_1, C_2 \in \mu(q), \quad C_1 \neq C_2 \Rightarrow (C_1 \not\subseteq C_2 \wedge C_2 \not\subseteq C_1) \quad (2.4)$$

Nesse caso, (Q, S, μ) é denominada uma *skill function*. Essa condição expressa a ideia de que os conjuntos de habilidades associados a um item devem representar combinações mínimas de habilidades suficientes para sua resolução, evitando redundâncias entre essas combinações.

A partir de uma *skill multimap* $((Q, S, \mu))$, [Heller, Ünlü e Albert \(2013\)](#) definem ainda uma função-problema (p), responsável por determinar quais itens podem ser resolvidos a partir de um conjunto disponível de habilidades ($T \subseteq S$). Essa função é dada por:

$$p: 2^S \rightarrow 2^Q \quad (2.5)$$

$$p(T) = \{q \in Q \mid \exists C \in \mu(q) \text{ tal que } C \subseteq T\} \quad (2.6)$$

Assim, $(p(T))$ representa o conjunto de itens que podem ser resolvidos quando o estudante dispõe do conjunto de habilidades (T). Essa formulação estabelece uma ponte entre habilidades, itens e estados de conhecimento, pois cada subconjunto de habilidades pode determinar um subconjunto de itens solucionáveis. Além disso, a função problema é monotônica em relação à inclusão de conjuntos, isto é:

$$T_1 \subseteq T_2 \Rightarrow p(T_1) \subseteq p(T_2) \quad (2.7)$$

com as seguintes condições de fronteira:

$$p(\emptyset) = \emptyset \quad (2.8)$$

$$p(S) = Q \quad (2.9)$$

Essas propriedades indicam que, sem habilidades disponíveis, nenhum item é solucionável, enquanto o domínio completo das habilidades em (S) permite resolver todos os itens em (Q).

2.2.2.3 Item

Os itens correspondem às questões, exercícios ou problemas apresentados aos estudantes durante o processo de avaliação da aprendizagem. Eles podem assumir diferentes formatos, como questões de múltipla escolha, respostas abertas, exercícios interativos ou problemas passo a passo. Conforme descrito por [Pelánek \(2020\)](#), os itens representam objetos de prática utilizados pelos sistemas adaptativos para coletar evidências sobre o desempenho

do estudante e subsidiar mecanismos de personalização e adaptação (Pelánek, Effenberger e Čechák, 2022).

Cada item pode exigir o domínio de um ou mais CC, estabelecendo uma relação entre o conteúdo avaliado e as habilidades cognitivas necessárias para sua resolução. Assim, a modelagem da relação entre itens e os CC constitui uma etapa fundamental nos sistemas adaptativos e MDC.

2.2.3 Avaliação Diagnóstica Cognitiva

A Avaliação Diagnóstica Cognitiva (ADC) foi introduzida como um novo método de medição educacional capaz de fornecer informações diagnósticas detalhadas sobre o nível de proficiência dos examinandos no domínio do conhecimento (Javidanmehr e Sarab, 2017; Williamson, 2023).

A demanda por ADC aumentou, devido às avaliações tradicionais não terem conseguido satisfazer a expectativa de resposta e transparência nos resultados de diagnóstico (Javidanmehr e Sarab, 2017). Elas não servem apenas para realizar um diagnóstico cognitivo, mas também para determinar até que ponto os alunos possuem a proficiência ou as características de interesse Briggs Derek C. Alonzo, Schwab e Wilson (2006).

A ADC pode ser considerada um avanço em relação aos resultados obtidos pelos esforços que as avaliações educacionais têm promovido (Javidanmehr e Sarab, 2017). Para que sua popularização e operacionalização se concretizem, é necessário que as ADC sejam aplicadas tanto em contextos de sala de aula quanto em testes de larga escala (Torre e Minchen, 2014; Ravand e Baghaei, 2020; Williamson, 2023).

Bollela, Borges e Troncon (2018) destacam que as questões objetivas ou, também chamadas de testes de múltipla escolha, são um dos métodos mais utilizados em todo o mundo para avaliar habilidades cognitivas. No contexto brasileiro, avaliações em larga escala como Sistema de Avaliação da Educação Básica (Saeb)⁴ e o Exame Nacional do Ensino Médio (ENEM)⁵ também se estruturam a partir da mensuração de competências e habilidades cognitivas, embora apresentem finalidades, matrizes e modelos avaliativos distintos. Primi et al. (2001) descrevem que o ENEM, desde sua formulação inicial, procurou deslocar o foco da simples reprodução de conteúdos para a avaliação de processos cognitivos mais amplos, como interpretação, resolução de problemas, estabelecimento de relações e aplicação de conhecimento em situações contextualizadas.

⁴ <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/saeb>

⁵ <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem>

As diferenças entre Saeb e ENEM podem ser compreendidas a partir do modo como cada uma concebe e operacionaliza as habilidades cognitivas. Saeb tem caráter sistêmico e diagnóstico, sendo voltado à avaliação de educação básica em diferentes etapas escolares, especialmente por meio das matrizes de referências organizadas em descritores associados aos conteúdos e habilidades curriculares (Silva et al., 2025; INEP, 2026b).

Desse modo, o Saeb se aproxima de um modelo de avaliação de proficiência escolar, voltado para monitorar a qualidade da educação básica comparando os resultados entre anos, etapas e sistemas de ensino (Gusmão e Amorim, 2023). Já o ENEM se aproxima de um modelo de avaliação de desempenho individual ao final da educação básica, com forte ênfase na aplicação de conhecimentos, na interdisciplinaridade e na resolução de problemas (Pereira et al., 2023; INEP, 2026a).

Segundo Soares e Nascimento (2011), essas avaliações se diferenciam principalmente por três aspectos: i) o uso e a forma de aplicação do modelo de avaliação Teoria de Resposta ao Item (TRI); ii) a abordagem filosófica adotada na construção dos itens; e iii) a definição do universo avaliado. Embora ambas sejam voltadas à mensuração de habilidades cognitivas, o Saeb privilegia a produção de indicadores educacionais em larga escala, enquanto o ENEM articula a avaliação cognitiva à certificação, à seleção e ao acesso ao ensino superior (Scherer, 2025).

2.2.4 Avaliação das Habilidades Cognitivas

As habilidades podem ser classificadas conforme o grau de especificidade. Por exemplo, há as que são chamadas de “Habilidades Gerais”, que se aplicam a uma ampla variedade de disciplinas, como pensamento crítico e resolução de problemas. Há também as “Habilidades Específicas”, que são úteis apenas em certos domínios, como fatoração de polinômios e resolução de problemas de raiz quadrada (Schunk, 2012).

Zompero, Laburú e Vilaça (2019) classificam as habilidades cognitivas como de baixa ordem e de alta ordem. As habilidades de baixa ordem são definidas como: conhecer, recordar ou lembrar informação, aplicar o conhecimento de maneira simples, ou utilizar algoritmos memorizados em situações familiares para resolver exercícios. Já as habilidades de alta ordem são definidas como processos cognitivos relacionados à investigação, à resolução de problemas (não exercícios), à tomada de decisão, ao pensamento crítico e avaliativo. Bollela, Borges e Troncon (2018) destacam que o princípio básico da avaliação de habilidades pode ser definido em modo restrito ou amplo. De forma restrita, é possível verificar se o

estudante atingiu os objetivos instrucionais estabelecidos, ou seja, se aprendeu o que deveria. No sentido amplo, significa compreender o processo de como são obtidas as informações sobre os estudantes.

Marinho e Marti (2016), Ravand e Baghaei (2020), Torre e Minchen (2014), apresentam um conceito mais realista aos objetivos desta tese, no qual a habilidade cognitiva é uma propriedade latente do indivíduo, não passível de mensuração direta por sua própria natureza. Algumas abordagens teóricas foram desenvolvidas para quantificar essa variável, mas diferem na interpretação dos resultados de erros e acertos.

Nas avaliações educacionais tradicionais, algumas abordagens são, geralmente, baseadas na TRI ou na Teoria Clássica dos Testes (TCT), sendo a pontuação de um aluno determinada pela identificação de sua localização ao longo de um único (*continuum*) de proficiência que está baseado em um traço latente (Javidanmehr e Sarab, 2017; Torre e Minchen, 2014).

2.2.4.1 Teoria Clássica dos Testes – TCT

A Teoria Clássica dos Testes (TCT) constitui uma das abordagens tradicionais da psicometria para análise dos resultados obtidos por meio de instrumentos de avaliação. Na TCT, a análise concentra-se principalmente no escore obtido pelo indivíduo no conjunto de itens que compõem o teste (Pasquali, 2009; Sartes e Souza-Formigoni, 2013).

No contexto de ambientes educacionais, Marinho e Marti (2016) destacam que o desempenho dos estudantes pode ser representado a partir da quantidade de respostas corretas obtidas no instrumento de avaliação. Em sua formulação clássica, a pontuação observada é compreendida como a combinação entre um escore verdadeiro e um componente de erro de medida. Pasquali (2009) representa essa relação pela seguinte equação:

$$T = V + E \quad (2.10)$$

em que T representa o escore observado do indivíduo, obtido a partir dos pontos alcançados no teste; V representa o escore verdadeiro, correspondente à magnitude do atributo que se pretende mensurar na ausência do erro de medida; e E representa o erro associado a medida (Pasquali, 2009).

Na TCT, as propriedades dos itens são analisadas a partir das respostas fornecidas pelos indivíduos. Para itens dicotômicos, por exemplo, o índice de dificuldade pode ser expresso pela proporção de indivíduos que respondem corretamente ao item, enquanto para itens politômicos, outras estatísticas podem ser empregadas de acordo com a forma de

pontuação adotada (Pasquali, 2009). Sartes e Souza-Formigoni (2013) destacam que uma das limitações da TCT está relacionada à dependência existente entre as características dos itens, o instrumento utilizado e a amostra de indivíduos avaliados.

Dessa forma, estatísticas dos itens e escores obtidos em diferentes instrumentos ou populações devem ser interpretadas considerando as características da aplicação. Essas limitações contribuíram para o desenvolvimento e a adoção de abordagens psicométricas que estabelecem uma relação probabilística entre as respostas aos itens e os traços latentes dos indivíduos, entre elas a Teoria de Resposta ao Item (TRI) (Pasquali, 2009; Sartes e Souza-Formigoni, 2013).

2.2.4.2 Teoria de Resposta ao Item – TRI

A Teoria de Resposta ao Item (TRI) compreende uma família de modelos probabilísticos que estabelecem relações entre o desempenho observado nos itens de um instrumento, as características desses itens e um traço latente do indivíduo, como uma habilidade ou proficiência (Hambleton e Swaminathan, 1985; Andrade e Valle, 1998; Andrade, Tavares e Valle, 2000). O traço latente, geralmente representado por θ , corresponde a uma característica que não pode ser observada diretamente, mas que pode ser estimada a partir das respostas fornecidas aos itens.

Diferentemente da TCT, cuja análise está predominantemente centrada no escore obtido no teste, a TRI modela a probabilidade de uma determinada resposta em função do nível do traço latente do indivíduo e das características dos itens (Hambleton e Swaminathan, 1985; Andrade e Valle, 1998; Andrade, Tavares e Valle, 2000; Pasquali, 2009). Dessa forma, a probabilidade de resposta correta não representa, por si só, a proficiência do indivíduo, mas resulta da relação entre sua proficiência e os parâmetros do item.

Marinho e Marti (2016) discutem a aplicação da TRI na modelagem da proficiência de estudantes em contextos educacionais. Soares e Nascimento (2011) por sua vez, destacam sua utilização em avaliações educacionais de larga escala, nas quais escalas de proficiência permitem analisar o desempenho dos estudantes e realizar comparações entre diferentes aplicações.

Soares e Nascimento (2011) explicam que a aplicação da TRI envolve a estimação dos parâmetros dos itens por meio de modelos estatísticos. Para itens dicotômicos, a relação entre o traço latente e a probabilidade de uma resposta pode ser representada pela Curva Característica do Item (CCI). Para itens politômicos, modelos específicos permitem representar

as probabilidades associadas às diferentes categorias de resposta. A análise das propriedades dos itens constitui uma das aplicações da TRI, permitindo examinar seu comportamento ao longo da escala do traço latente (Reeve e Fayers, 2005).

De acordo com Andrade, Tavares e Valle (2000), Hambleton e Swaminathan (1985), Soares e Nascimento (2011) e Monteiro, Leitão e Barreto (2021), diferentes modelos da TRI podem ser especificados de acordo com os parâmetros considerados. Entre os modelos logísticos para itens dicotômicos estão o Modelo Logístico de Um Parâmetro (ML1P), que considera o parâmetro de dificuldade; o Modelo Logístico de Dois Parâmetros (ML2P), que considera os parâmetros de dificuldade e discriminação e o Modelo Logístico de Três Parâmetros (ML3P), que acrescenta o parâmetro relacionado à probabilidade de acerto ao acaso.

O ML3P é expresso pela seguinte equação Andrade, Tavares e Valle (2000) e Soares e Nascimento (2011):

$$P(U_{ij} = 1 | \theta_j, a_i, b_i, c_i) = c_i + (1 - c_i) \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}} \quad (2.11)$$

em que:

$P(U_{ij} = 1 | \theta_j, a_i, b_i, c_i)$: probabilidade do indivíduo j com habilidade θ_j acertar o item i ;

a_i : parâmetro de discriminação;

b_i : parâmetro de dificuldade;

c_i : é a probabilidade de acerto ao acaso do item i .

De acordo com Soares, Soares e Santos (2021), o ML2P pode ser obtido fazendo $c_i = 0$ e o ML1P fazendo $c_i = 0$ e $a_i = 1$.

Sartes e Souza-Formigoni (2013), com base em Hambleton e Swaminathan (1985), destacam algumas propriedades da TRI que representam avanços em relação à TCT. Uma delas é a invariância dos parâmetros, segundo a qual, quando o modelo apresenta ajuste adequado aos dados e seus pressupostos são atendidos, as estimativas do traço latente tendem a ser menos dependentes do conjunto específico de itens utilizado, desde que os itens mensurem o mesmo traço. De forma correspondente, os parâmetros dos itens tendem a ser menos dependentes do grupo de indivíduos utilizado em sua estimação (Hambleton e Swaminathan, 1985; Andrade, Tavares e Valle, 2000).

Essa propriedade não deve ser entendida como uma independência absoluta entre os parâmetros, os itens e as amostras, uma vez que sua obtenção depende da adequação do modelo aos dados e da qualidade da estimação. Também não deve ser confundida com o

pressuposto de independência local da TRI, segundo o qual, dado um determinado nível do traço latente, as respostas de um indivíduo aos diferentes itens são estatisticamente independentes (Hambleton e Swaminathan, 1985; Sartes e Souza-Formigoni, 2013; Reeve e Fayers, 2005).

Outra característica da TRI é a possibilidade de posicionar itens e indivíduos em uma mesma escala do traço latente. Nessa escala, a localização dos itens pode ser relacionada aos seus parâmetros de dificuldade, enquanto a localização dos indivíduos representa seus níveis estimados de proficiência. Isso permite analisar quais itens fornecem maior informação em diferentes regiões da escala (Andrade, Tavares e Valle, 2000; Sartes e Souza-Formigoni, 2013).

Além disso, diferentemente de determinadas formulações da TCT, a TRI não pressupõe que o erro de medida seja constante para todos os indivíduos. A precisão da medida pode variar ao longo da escala do traço latente, dependendo das características dos itens utilizados. A TRI também possibilita avaliar a precisão da medida sem exigir instrumentos estritamente paralelos (Sartes e Souza-Formigoni, 2013; Hambleton e Swaminathan, 1985).

A TRI constitui uma abordagem consolidada na psicometria e tem sido aplicada em diferentes contextos de avaliação educacional (Andrade, Tavares e Valle, 2000; Sartes e Souza-Formigoni, 2013). Embora permita estimar a proficiência dos indivíduos e analisar propriedades dos itens, a TRI, em suas formulações tradicionais, não foi concebida para fornecer diretamente perfis diagnósticos detalhados de domínio e não-domínio de múltiplas habilidades específicas. Para produzir informações mais detalhadas sobre os pontos fortes e as dificuldades dos estudantes, diferentes Modelos de Diagnóstico Cognitivo (MDC) têm sido propostos (Javidanmehr e Sarab, 2017; Ravand e Baghaei, 2020).

2.2.4.3 Bayesian Knowledge Tracing – BKT

O *Bayesian Knowledge Tracing* (BKT), também conhecido como Rastreamento Bayesiano do Conhecimento, é uma das primeiras abordagens voltadas à modelagem da evolução do estado de conhecimento do aprendiz. Esse modelo pode ser considerado uma das primeiras propostas contrárias à hipótese de que os estados de conhecimento são estáticos. Este método foi introduzido para realizar o Rastreio de Conhecimento (do inglês *Knowledge Tracing* – KT) nos ambientes de aprendizagem adaptativos (Abdelrahman, Wang e Nunes, 2023; Corbett e Anderson, 1994).

Na sua formulação original, o BKT supõe que cada item avalia um único conceito. Dessa forma, as questões são associadas a apenas um CC. Com base em uma sequência de

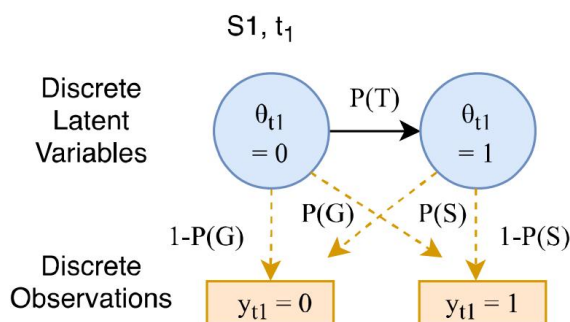
respostas do estudante, o modelo infere, a cada instante de tempo, o estado de domínio da habilidade. Essa abordagem é particularmente relevante para os STI que utilizam exercícios e suportes instrucionais (*scaffolding*) como principais mecanismos de aprendizagem, além de monitorar o domínio de habilidades ao nível granular para decidir quais os próximos passos instrucionais.

O BKT utiliza uma estrutura temporal baseada em uma Cadeia de *Markov*. Em uma Cadeia de *Markov*, um sistema é representado por um conjunto de estados e pelas possíveis transições entre esses estados ao longo do tempo. No BKT, para cada habilidade modelada, esses estados representam, de forma simplificada, se o estudante ainda não domina ou já domina determinada habilidade. As transições representam a possibilidade de mudança entre esses estados durante as interações do estudante com as atividades.

Considerando t como o instante atual da interação e $t-1$ como o instante imediatamente anterior, a propriedade de Markov estabelece que a probabilidade do estado no instante t depende diretamente do estado no instante $t-1$, sem depender explicitamente de toda a sequência de estados anteriores. Dessa forma, o estado de conhecimento estimado pelo BKT é atualizado de forma sequencial à medida que novas respostas do estudante são observadas. Além disso, o BKT pode ser representado como uma Rede *Bayesiana* Dinâmica, uma classe de modelos voltada à representação de informações que evoluem ao longo do tempo (Jaques et al., 2013).

A Figura 4 mostra uma estrutura padrão de BKT para uma sequência com duas observações de desempenho, sendo uma incorreta e a outra correta, bem como os estados de domínio inferidos, que variam de “não aprendido” (valor 0), para “aprendido” (valor 1).

Figura 4 – Exemplo padrão de um modelo BKT.



Fonte: Sapountzi et al. (2018).

Conforme a Figura 4, o BKT é composto por quatro parâmetros, os quais são estimados a

partir de dados coletados durante a construção de um modelo específico para cada habilidade. A probabilidade inferida depende principalmente desses parâmetros, utilizados para prever o domínio de um estudante em uma determinada habilidade com base em sua sequência cronológica de tentativas corretas e incorretas em questões associadas.

As quatro probabilidades fundamentais do BKT são:

- $P(L_0)$: probabilidade do estudante já conhecer a habilidade antes de tentar resolver o primeiro problema;
- $P(T)$: probabilidade do estudante aprender a habilidade após uma oportunidade de prática, mesmo sem conhecê-la anteriormente;
- $P(G)$: probabilidade do estudante responder corretamente a uma questão por meio de adivinhação (tentativa e erro), mesmo sem conhecer a habilidade, também chamada de *guess*;
- $P(S)$: probabilidade do estudante responder incorretamente a uma questão apesar de conhecer a habilidade, também chamada de *slip*.

Em um ambiente de aprendizagem que utiliza BKT, cada item deve estar associado à habilidade cuja probabilidade de domínio será estimada. Assim, quando o estudante responde a um item relacionado a determinada habilidade, a resposta observada é utilizada como evidência para atualizar a estimativa de domínio dessa habilidade. A probabilidade de o estudante dominar a habilidade no instante t , caso sua resposta ao item associado seja correta, é definida por:

$$P(L_t | \text{Correto}) = \frac{P(L_{t-1})(1 - P(S))}{P(L_{t-1})(1 - P(S)) + (1 - P(L_{t-1}))P(G)} \quad (2.12)$$

Se a resposta for incorreta, a probabilidade será atualizada conforme a Equação 2.13:

$$P(L_t | \text{Incorreto}) = \frac{P(L_{t-1})P(S)}{P(L_{t-1})P(S) + (1 - P(L_{t-1}))(1 - P(G))} \quad (2.13)$$

Para representar de forma geral a resposta observada no instante t , define-se $R_t \in \{\text{Correto}, \text{Incorreto}\}$. Assim, após a atualização da probabilidade de domínio com base na resposta do estudante, estima-se a possibilidade de aprendizagem da habilidade durante a interação com o problema por meio da Equação 2.14:

$$P(L_t) = P(L_t | R_t) + (1 - P(L_t | R_t)) P(T) \quad (2.14)$$

Inicialmente, a probabilidade de domínio da habilidade é atualizada a partir da resposta observada, utilizando-se a Equação 2.12 ou a Equação 2.13, conforme a resposta seja correta ou incorreta, respectivamente. Em seguida, considera-se a probabilidade de transição $P(T)$ para estimar a possibilidade de o estudante ter aprendido a habilidade durante a interação com o problema, conforme a Equação 2.14.

Para os STI que utilizam o algoritmo BKT, o estudante pode ser instruído a praticar resolução de itens semelhantes relacionados aos CC específico sempre que o sistema identificar que ele ainda não possui domínio suficiente. Um exemplo comum é considerar que o domínio ainda não foi alcançado quando a probabilidade estimada de conhecimento for inferior a 95% (Sapountzi et al., 2018).

Na literatura, diversas extensões do BKT foram propostas, incluindo a contextualização das estimativas dos parâmetros de adivinhação (*guess*) e deslize (*slip*) (Baker, Corbett e Alevan, 2008), a estimação da probabilidade de transição a partir do uso de recursos de ajuda e a estimação da probabilidade inicial de conhecimento do estudante (Pardos e Heffernan, 2010). Além disso, a dificuldade dos itens também já foi integrada ao modelo (Pardos e Heffernan, 2012).

2.2.4.4 Análise do Fator de Desempenho – PFA

A Análise do Fator de Desempenho (PFA) foi adaptada a partir da Análise de Fatores de Aprendizagem (LFA) (Pavlik, Cen e Koedinger, 2009), com o objetivo de permitir a criação de um “Modelo de *Overlay*” individual capaz de rastrear previsões para os estudantes em relação a habilidades específicas. Dessa maneira, o modelo possibilita a adaptação da instrução e a realização de intervenções de remediação com base no desempenho atual (Gong, Beck e Heffernan, 2011).

A LFA acumula informações sobre a aprendizagem do estudante i considerando uma ou mais habilidades k . Sua formulação pode ser representada por:

$$\log P_{ij} = \theta_i + \sum_{k \in KCs} (\beta_k + \gamma_k n_{ik}) \quad (2.15)$$

em que:

- KCs (Componentes de Conhecimento – CC): $[C_1, C_2, \dots, C_n]$;

- θ_i : habilidade ou proficiência do estudante i ;
- β_k : viés associado à habilidade k ;
- γ_k : viés associado às tentativas anteriores na habilidade k ;
- n_{ik} : número de tentativas do estudante i na habilidade k .

Esse modelo pode ser compreendido como uma extensão do modelo *Rasch*, pertencente à TRI unidimensional. O modelo *Rasch* possui uma forma equivalente à Equação 2.15 quando γ_k é definido como zero e há apenas um único valor de β_k .

Posteriormente, Pavlik, Cen e Koedinger (2009) adaptaram o modelo LFA para torná-lo sensível aos indicadores de desempenho de aprendizagem. O modelo PFA permite representar a conjunção de habilidades ao somar as contribuições de todas as habilidades necessárias para um determinado desempenho. Esse tipo de modelo “compensatório” de aprendizagem multihabilidades possibilita que a ausência de domínio em um CC seja compensada pela presença de outro, e permite a representação de efeitos conjuntivos.

Assim como o BKT, a PFA também refuta a suposição de conhecimento estático, pois modela a aprendizagem ao longo do tempo e considera múltiplas habilidades simultaneamente (Pavlik, Cen e Koedinger, 2009). Sua estrutura básica é dada por:

$$\log P_{ij} = \sum_{k \in KCs} (\beta_k + \gamma_k s_{ik} + \rho_k f_{ik}) \quad (2.16)$$

em que:

- KCs (Componentes de Conhecimento – CC): $[C_1, C_2, \dots, C_n]$;
- β_k : viés associado à habilidade k ;
- γ_k : viés associado às tentativas bem-sucedidas na habilidade k ;
- ρ_k : viés associado às tentativas malsucedidas na habilidade k ;
- s_{ik} : número de tentativas bem-sucedidas do estudante i na habilidade k ;
- f_{ik} : número de tentativas malsucedidas do estudante i na habilidade k .

Ao contrário da LFA, a PFA não considera explicitamente o parâmetro de proficiência do estudante (θ_i), pois supõe que essa proficiência não pode ser estimada previamente em situações adaptativas (Pavlik, Cen e Koedinger, 2009).

2.2.4.5 Deep Knowledge Tracing – DKT

O Aprendizado Profundo, do inglês *Deep Learning*), nos anos recentes, passou a ser utilizado nas tarefas de KT. Entretanto, de forma semelhante ao que ocorreu em diversos outros domínios, seu desempenho passou a superar abordagens consideradas o estado da arte ([Abdelrahman, Wang e Nunes, 2023](#)).

[Piech et al. \(2015\)](#) apresentaram o *Deep Knowledge Tracing* (DKT). De modo análogo ao BKT, o DKT usa dados compostos por tentativas associadas a habilidades e seus respectivos resultados de desempenho para prever tentativas futuras em uma sequência de interações. Para cada instante de tempo t , o modelo codifica a habilidade praticada e a resposta do estudante em um vetor de entrada do tipo *one-hot* ([Minn et al., 2018](#)). A camada de saída fornece a probabilidade prevista de que o estudante responda corretamente ao próximo problema no instante $t + 1$.

O DKT utiliza Redes Neurais Recorrentes, do inglês *Recurrent Neural Networks* (RNN), para representar dinamicamente o espaço latente de conhecimento dos estudantes ao longo das interações com as atividades. Dessa maneira, é possível inferir a evolução do conhecimento do estudante durante uma atividade a partir do seu histórico de desempenho anterior.

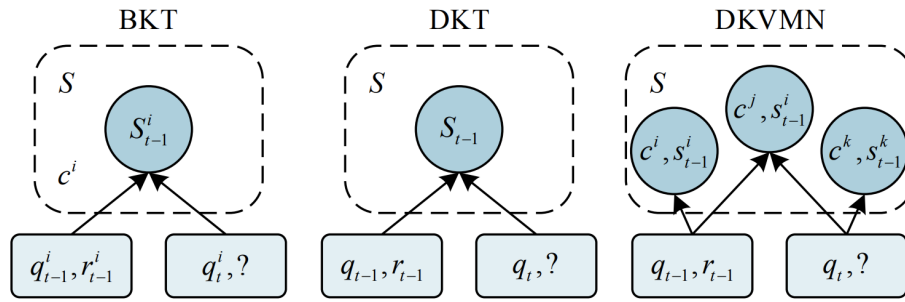
2.2.4.6 Dynamic Key-Value Memory Network – DKVMN

Por fim, destaca-se outra arquitetura de Redes Neurais Profundas denominada *Memory Networks*, que integra um componente de memória de longo prazo com operações de leitura e escrita. Essa arquitetura busca superar uma limitação das RNN que é a dificuldade em memorizar grandes quantidades de informações relevantes provenientes de interações passadas.

Originalmente, as *Memory Networks* foram aplicadas a tarefas de perguntas e respostas em linguagem natural. Entretanto, essa arquitetura passou a ser utilizada para avaliar o domínio de conceitos por estudantes, por meio de uma estrutura específica denominada *Dynamic Key-Value Memory Network* (DKVMN) ([Zhang et al., 2017](#)).

Segundo a teoria da DKVMN, as estimativas do nível de domínio do estudante em habilidades específicas são armazenadas em espaços de memória, sendo controladas por operações de leitura e escrita, que são mediadas por mecanismos adicionais de atenção. O modelo incorpora uma matriz: uma chamada de *key memory*, responsável por armazenar as representações dos conceitos; e a outra de *value memory*, responsável por armazenar e atualizar informações sobre a compreensão do estudante, ou seja, o estado de conhecimento

Figura 5 – Comparação entre BKT, DKT e DKVMN.



Fonte: [Man \(2024\)](#).

associado a cada conceito.

[Man \(2024\)](#) mostra na Figura 5 que o DKVMN combina as vantagens dos modelos BKT e DKT. Assim como o BKT, ele permite acompanhar estados de domínio associados a conceitos específicos; e, assim como o DKT, utiliza uma arquitetura neural capaz de aprender relações mais complexas a partir dos dados. Na imagem, observa-se que o DKVMN representa vários conceitos de conhecimento, como c^i , c^j e c^k , cada um associado ao seu respectivo estado de domínio anterior, como S_{t-1}^i , S_{t-1}^j e S_{t-1}^k . Quando o estudante responde a um exercício anterior, q_{t-1}, r_{t-1} , o modelo identifica quais conceitos estão relacionados a essa interação e atualiza apenas os estados de conhecimento correspondentes.

Na Figura 5, a nova questão q_t não depende de um único estado geral ou de apenas um conceito isolado, mas pode estar associada a diferentes conhecimentos armazenados na memória do modelo. O algoritmo DKVMN, lê os estados de domínio dos conceitos relevantes e utiliza essas informações para prever se o estudante responderá corretamente à nova questão. Após a resposta, apenas os conceitos relacionados ao exercício são atualizados, mantendo os demais estados de conhecimento preservados ([Man, 2024](#)).

2.2.5 Limitações

A aplicação de avaliações de habilidades em ambientes educacionais proporcionam uma compreensão mais aprofundada do processo de aprendizagem, sobretudo porque as habilidades cognitivas são características latentes do indivíduo, não podendo ser observadas diretamente. Nesse sentido, modelos e instrumentos avaliativos inferem o nível de domínio do estudante a partir de evidências observáveis, tais como respostas corretas e incorretas em itens de avaliação. Essa perspectiva possibilita que a avaliação deixe de ser apenas um mecanismo de classificação, passando a funcionar como um recurso para identificar dificuldades, acompanhar o progresso e subsidiar a tomada de decisões pedagógicas.

Entre os principais benefícios dessa implementação, destaca-se a possibilidade de obter informações mais detalhadas sobre o desempenho dos estudantes. Avaliações apoiadas em modelos como a Teoria Clássica dos Testes (TCT), a Teoria de Resposta ao Item (TRI), o Bayesian Knowledge Tracing (BKT), o Análise do Fator de Desempenho (PFA), o Deep Knowledge Tracing (DKT) e o *Dynamic Key-Value Memory Network* (DKVMN) permitem diferentes formas de representar o conhecimento do aprendiz. Enquanto abordagens mais tradicionais, como a TCT, possibilitam uma análise simples baseada na quantidade de acertos, modelos mais sofisticados, como a TRI, permitem estimar a proficiência considerando características dos itens, como dificuldade, discriminação e probabilidade de acerto ao acaso.

Um ponto positivo é que essas avaliações podem contribuir para a personalização do ensino. Ao identificar quais habilidades foram dominadas e quais ainda necessitam de reforço, o sistema pode oferecer atividades mais adequadas ao nível de cada estudante. Essa abordagem se aproxima da função diagnóstica da avaliação, pois fornece informações para que professores e os sistemas educacionais compreendam não apenas o resultado final da aprendizagem, mas também os processos e as dificuldades enfrentada durante o percurso formativo.

No entanto, apesar das vantagens, a implementação de avaliações de habilidades também apresenta limitações. A TCT, por exemplo, depende fortemente do conjunto de itens aplicados e da amostra de estudantes avaliados, o que pode dificultar comparações entre diferentes instrumentos ou populações. A TRI supera parte dessas limitações ao posicionar itens e estudantes em uma mesma escala de proficiência, mas exige maior rigor estatístico, amostras adequadas e itens bem calibrados para que seus resultados sejam confiáveis.

Por outro lado, o BKT possui boa interpretabilidade, pois acompanha habilidades específicas. No entanto, geralmente supõe que cada item está associado a um único conceito, o que pode simplificar excessivamente situações reais de aprendizagem, nas quais uma mesma questão pode exigir múltiplas habilidades. O PFA avança ao considerar o histórico de sucessos e fracassos em múltiplos componentes de conhecimento, mas ainda depende de uma definição adequada das habilidades envolvidas em cada item. Já os modelos baseados em aprendizagem profunda, como o DKT e o DKVMN, podem alcançar maior capacidade preditiva, porém, demandam maior volume de dados e poder computacional, além de apresentarem menor transparência em relação à interpretação dos resultados.

Sendo assim, a escolha do modelo avaliativo deve equilibrar simplicidade, interpretabilidade, precisão e viabilidade de implementação. Os modelos mais simples são mais fáceis de

aplicar e interpretar, mas oferecem diagnósticos menos detalhados. Por outro lado, modelos mais complexos conseguem capturar relações mais ricas entre estudantes, itens e habilidades, porém exigem bases de dados consistentes, infraestrutura apropriada e uma atenção metodológicas para evitar interpretações equivocadas.

Portanto, a implementação de avaliações de habilidades representa um avanço significativo para a análise da aprendizagem, pois possibilita a transformação de dados de desempenho em informações úteis para o diagnóstico educacional. No entanto, para que essas avaliações sejam efetivas, é necessário estarem apoiadas em modelos capazes de representar adequadamente a relação entre itens, habilidades e os estados de conhecimento dos estudantes. Essa necessidade conduz à discussão dos Modelos de Diagnóstico Cognitivo (MDC), apresentados no próximo capítulo, os quais buscam oferecer uma representação mais refinada dos atributos cognitivos dominados ou não dominados pelos aprendizes, ampliando a capacidade diagnóstica das avaliações educacionais.

2.3 Diagnóstico Cognitivo e a Avaliação de Habilidades

2.3.1 Modelos de Diagnóstico Cognitivo

Contrastando com a natureza descritiva dos modelos TRI e TCT, os MDC foram desenvolvidos como uma estrutura psicométrica alternativa para fornecer informações diagnósticas por meio da classificação dos examinados em relação a um conjunto de habilidades. Conhecidos como modelos de classe latente restritos, os MDC modelam as respostas em função de variáveis latentes discretas em cada item ([Torre e Minchen, 2014](#)).

Os MDC são uma família de modelos psicométricos de variáveis latentes discretas que têm sido amplamente utilizados nas ciências educacionais (avaliações), psicológicas (personalidades), sociais e biológicas para classificar indivíduos com perfis únicos de atributos ([Ravand e Baghaei, 2020](#)). Esses modelos visam realizar observações ruidosas (“noisy”) discretas multivariadas que assumem a existência de atributos latentes discretos capazes de explicar ou governar as variáveis observadas ([Ma, Ouyang e Xu, 2022](#); [Shi et al., 2021](#)).

[Ma, Minchen e Torre \(2020\)](#) e [Torre e Minchen \(2014\)](#) chamam de traços latentes todas as características não observadas diretamente, como, por exemplo, a capacidade de leitura dos alunos, a inteligência emocional ou a qualidade de vida dos examinados. Essas características são medidas por meio de respostas de testes projetados sob medida. Dessa forma, a partir das respostas, é possível inferir os atributos latentes subjacentes.

Os MDC podem fornecer inferência aprimorada sobre os perfis de atributos latentes dos sujeitos com base nas respostas observadas. Além disso, é possível detectar os subgrupos latentes correspondentes de uma população também podem ser detectados com base nos perfis de atributos latentes inferidos (Ma, Ouyang e Xu, 2022). Esses atributos podem representar habilidades, *misconceptions*⁶ ou estratégias de resolução de problemas (Shi et al., 2021). Segundo Ravand e Robitzsch (2015), os MDC podem ser empregados de duas maneiras: (a) como análise *post-hoc* (*retrofitting*) de testes não diagnósticos existentes, a fim de extrair informações detalhadas, ou (b) para projetar um conjunto de itens ou tarefas que visam diagnosticar atributos de habilidades latentes.

Os MDC têm se tornado cada vez mais relevantes quando aplicados à temática educacional, pois demonstraram o potencial de identificar os pontos fortes e fracos dos alunos auxiliando significativamente a instrução e a aprendizagem em sala de aula (Shi et al., 2021). Alguns deles já foram aplicados em diferentes áreas da educação como, por exemplo, em teste de língua estrangeira (Aryadoust, 2021; George e Robitzsch, 2015; Liu, 2019; Min e He, 2022), operação com Álgebra e Fração (Liu et al., 2020; Kuo, Chen e Torre, 2018; Medrano, 2016; Wafa, Hussani e Pazhman, 2020; Yuan et al., 2022) e na análise do conhecimento de professores em formação curricular na teoria da Análise de Dados (Wu, Xu e Zhang, 2021).

2.3.1.1 Configuração para a implementação da avaliação baseada em MDC

Segundo Liu (2019), para implementar um MDC é necessário seguir algumas etapas como (i) especificar as habilidades necessárias para solucionar uma questão específica e (ii) especificar quais itens (questões) medem quais habilidades. A Figura 6, mostra a dinâmica para realizar o diagnóstico cognitivo dos estudantes.

2.3.1.2 Definição dos atributos latentes

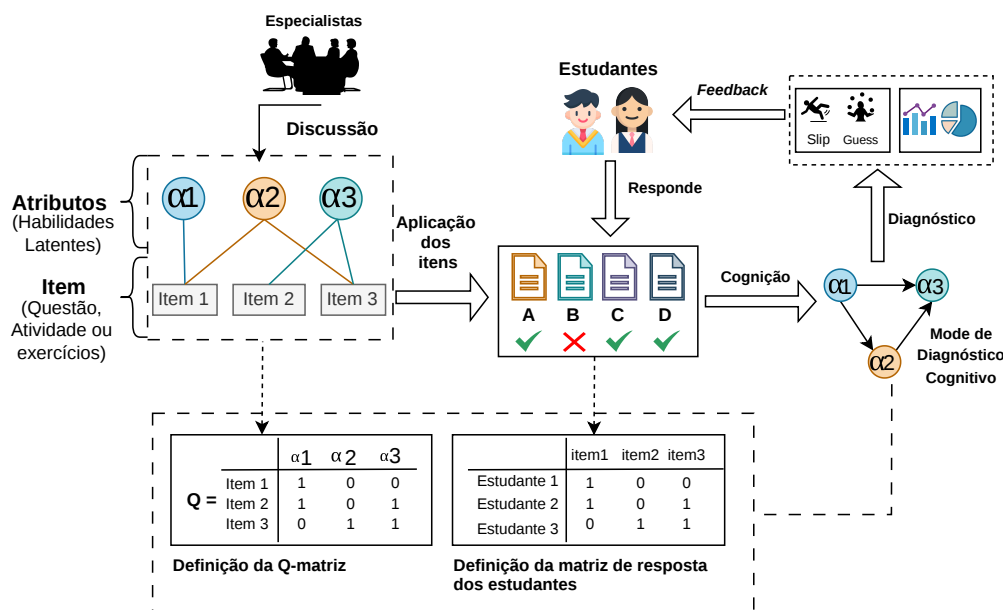
Primeiramente, para definir os atributos⁷ envolvidos em um teste, podem ser consultados de diversas fontes primárias da literatura (Ravand e Robitzsch, 2015). Geralmente, os atributos (conceitos ou habilidades latentes) necessários para cada item (problema, questão ou exercício) devem ser previamente definidos por um especialista e a sua interpretação é feita em uma estrutura chamada Q-matriz. A Figura 7, mostra como a relação entre itens e atributos é feita e como uma Q-matriz é criada.

A relação item-atributos são especificadas por meio de uma Q-matriz binária $J \times K$.

⁶ Uso de algum conceito errôneo, um mal-entendido baseado vieses pessoais, suposições falsas ou conclusões incorretas (Carle, 1993).

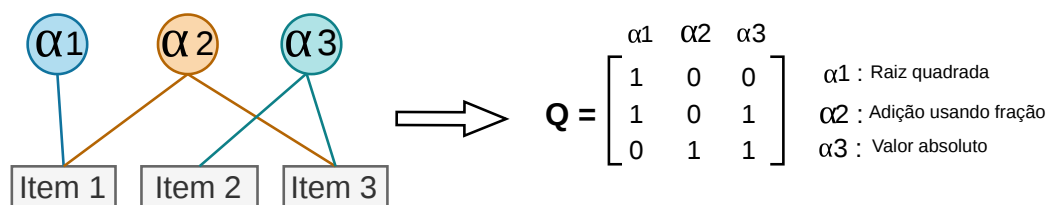
⁷ O termo mais genérico para denominar as habilidades latentes usadas pelos MDC é conhecido como “atributos”, os quais são tipicamente assumidos como variáveis binárias (1 ou 0)

Figura 6 – Processos realizados para diagnóstico.



Fonte: adaptado de [Lopes e Netto \(2023\)](#).

Figura 7 – Exemplo de uma Q-matriz com relação de itens e atributos.



Fonte: Elaborado pelos autores (2026).

A Q-matriz (Q) de entrada é o conjunto $q_{jk} \in \{0,1\}$, que indica a ausência ou presença, respectivamente, de uma ligação entre o j -ésimo item e o k -ésimo atributo latente. A Q-matriz também pode ser expressa por meio de um gráfico, como mostra a Figura 7.

Em uma segunda matriz, são armazenadas as respostas dos examinados. A Q-matriz X , de dimensões $I \times J$ é composta por respostas binárias para os itens de um teste. O elemento x_{ij} indica que i está correto o item j ($x_{ij} = 1$) ou ($x_{ij} = 0$) para não correto.

Para avaliar os examinados de um teste usando um MDC, são necessárias duas Q-matrizes. Primeiramente cria-se a Q-matriz, conforme mostra a Figura 7, que contém para cada item j (linhas) e o atributo k (colunas), os elementos q_{jk} que indicam se o item j requer o atributo k para solução. Nesse caso, $q_{jk} = 1$ significa que a habilidade é requerida ou $q_{jk} = 0$

para não requerida.

2.3.1.3 Q-matriz e a *Rule Space Theory*

[Tatsuoka \(1984\)](#) introduziu originalmente a Q-matriz no âmbito da *Rule Space Theory* (RST), que se baseia na premissa de que os alunos aplicam regras cognitivas específicas ao resolver problemas. O *Rule Space Method* (RSM) foi proposta como uma abordagem para realizar diagnósticos detalhados das estratégias de resolução de problemas ou da aplicação de regras pelos alunos, combinando especificações cognitivas com análises estatísticas fundamentadas na TRI ([Durand, Belacel e Goutte, 2015](#)). Uma premissa central da RST é que os estudantes devem dominar conceitos ou habilidades específicas, denominados **atributos de conhecimento**, para responder corretamente aos itens de avaliação.

O RSM foi conduzido em duas fases. Primeiramente, foram definidos os atributos relevantes necessários para a construção de padrões de resposta ideais. Em seguida, foi realizada a classificação estatística por meio de análises baseadas na TRI para mapear as respostas observadas dos alunos aos padrões ideais predefinidos ([Oeda e Iizumi, 2022](#)). Conforme essa estrutura, todos os padrões de resposta possíveis, derivados de regras corretas e incorretas identificadas por especialistas, são representados em um espaço multidimensional. As respostas dos alunos são então comparadas a esses pontos de referência, e o ponto mais próximo indica a regra que o aluno provavelmente está aplicando ([Barnes et al., 2005](#)).

Essa abordagem, transcende os modelos puramente procedimentais, pois constrói explicitamente um espaço de conhecimento no qual todas as respostas podem ser mapeadas sistematicamente e contrastadas com as previsões baseadas em especialistas ([Desmarais, Beheshti e Naceur, 2012](#); [Ravand e Baghaei, 2020](#)). Dessa forma, a Q-matriz se torna um instrumento essencial para representar a relação entre itens de avaliação e atributos de conhecimento.

2.3.2 Modelos clássicos para diagnóstico cognitivo

Os Modelos de Diagnóstico Cognitivo (MDC) (do inglês *Cognitive Diagnostic Models*) constituem uma abordagem psicométrica voltada à inferência do domínio de habilidades específicas por parte dos estudantes, representadas como atributos latentes ([Rupp, Templin e Henson, 2010](#); [Torre, 2009a](#)). Ao contrário das abordagens tradicionais, como a TCT e a TRI, que fornecem pontuações globais, os MDC permitem a identificação de perfis multidimensionais de domínio, oferecendo informações mais detalhadas sobre o processo de aprendizagem ([Rupp, Templin e Henson, 2010](#)).

A relação entre itens e atributos é formalizada por meio da Q-matriz, proposta originalmente por Tatsuoaka (1984), na qual cada elemento q_{jk} indica se o atributo k é requerido para a resolução do item j . Destacam-se dois modelos fundamentais na literatura: o modelo DINA (*Deterministic Inputs Noisy “And” Gate*) e o modelo G-DINA (*Generalized DINA Model*) (Torre, 2009b; Torre, 2011).

2.3.2.1 Modelo DINA

O modelo DINA (*Deterministic Inputs Noisy “And” Gate*) é um modelo não compensatório, ou seja, o acerto de um item depende do domínio completo de todos os atributos requeridos. Nesse caso, a ausência de qualquer atributo não pode ser compensada pelo domínio de outros atributos (Torre e Minchen, 2014).

Seja X_{ij} a resposta do estudante i ao item j , em que $X_{ij} = 1$ indica resposta correta e $X_{ij} = 0$ indica resposta incorreta. Seja também α_{ik} o domínio do estudante i em relação ao atributo k , de modo que $\alpha_{ik} = 1$ indica domínio e $\alpha_{ik} = 0$ indica não domínio. A Q-matriz é representada por q_{jk} , em que $q_{jk} = 1$ indica que o item j exige o atributo k , e $q_{jk} = 0$ indica que esse atributo não é exigido.

A primeira etapa do modelo DINA consiste em calcular a resposta latente ideal η_{ij} :

$$\eta_{ij} = \prod_{k=1}^K \alpha_{ik}^{q_{jk}}, \quad (2.17)$$

O símbolo $\prod$ representa um produtório, isto é, a multiplicação sucessiva dos termos considerados para todos os atributos k , de 1 até K . Assim, a expressão equivale a multiplicar os valores $\alpha_{ik}^{q_{jk}}$ correspondentes a cada atributo considerado no modelo.

Na expressão 2.17, $\eta_{ij} = 1$ quando o estudante domina todos os atributos exigidos pelo item. Se faltar pelo menos um atributo necessário, tem-se $\eta_{ij} = 0$. Assim, essa equação representa a lógica “AND” do modelo, segundo a qual todos os atributos requeridos precisam estar presentes para que o estudante seja considerado apto a responder corretamente ao item.

Entretanto, o desempenho observado do estudante não é completamente determinístico. Um estudante que domina todos os atributos podem errar por descuido, enquanto um estudante que não domina todos os atributos podem acertar por palpite ou por meio de uma estratégia alternativa. Por isso, o modelo DINA incorpora dois parâmetros:

- s_j : parâmetro de *slip*, que representa a probabilidade de erro mesmo quando o estudante domina todos os atributos exigidos pelo item;

- g_j : parâmetro de *guessing*, que representa a probabilidade de acerto mesmo quando o estudante não domina todos os atributos exigidos pelo item.

A probabilidade de resposta correta no modelo DINA é dada por:

$$P(Y_{ij} = 1 | \alpha_i) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}}, \quad (2.18)$$

A equação 2.18 pode ser interpretada de forma simples. Quando $\eta_{ij} = 1$, isto é, quando o estudante domina todos os atributos exigidos, a probabilidade de resposta correta é $1 - s_j$. Quando $\eta_{ij} = 0$, ou seja, quando falta ao menos um atributo necessário, a probabilidade de resposta correta é g_j .

De forma equivalente, a probabilidade pode ser escrita como:

$$P(X_{ij} = 1 | \alpha_i) = g_j + (1 - g_j - s_j)\eta_{ij} \quad (2.19)$$

A equação 2.19 deixa mais clara a transição entre os dois grupos considerados pelo modelo: estudantes plenamente qualificados para o item e estudantes não plenamente qualificados. Por essa razão, o DINA é considerado parcimonioso e interpretável, pois utiliza apenas dois parâmetros por item, independentemente da quantidade de atributos envolvidos (Torre, 2009b; Liu et al., 2020).

2.3.2.2 Modelo G-DINA

Como extensão do DINA, o modelo G-DINA⁸ foi proposto por Torre (2011) com o objetivo de permitir maior flexibilidade na modelagem das relações entre atributos e desempenho.

O modelo G-DINA (*Generalized DINA Model*) é considerado um modelo saturado, pois não impõe restrições prévias sobre a forma como os atributos interagem na determinação da probabilidade de acerto de um item. Isso significa que, para cada item, todas as combinações possíveis de domínio dos atributos requeridos são consideradas explicitamente. Dessa forma, cada padrão de habilidades, como dominar apenas um atributo, dois atributos ou todos eles, pode estar associado a uma probabilidade distinta de resposta correta (Shi et al., 2021).

Em termos práticos, isso significa que o modelo não pressupõe, a priori, se a relação entre os atributos é compensatória ou não. Em vez disso, ele estima parâmetros específicos para capturar tanto os efeitos individuais de cada atributo quanto as interações entre eles. Dessa maneira, o G-DINA consegue representar diferentes processos cognitivos subjacentes à

⁸ Disponível em: <<https://cran.r-project.org/web/packages/GDINA/index.html>>. Acessado em 04 de maio de 2026.

resolução dos itens, oferecendo maior flexibilidade em comparação a modelos mais restritivos, como o DINA (Torre e Minchen, 2019).

Seja K_j^* o número de atributos exigidos pelo item j . Nesse caso, o modelo G-DINA pode considerar até $2^{K_j^*}$ perfis latentes diferentes para esse item. Isso significa que cada combinação de atributos dominados e não dominados pode ter uma probabilidade própria de resposta correta.

Sua forma geral é expressa pela Equação 2.20 :

$$P(Y_{ij} = 1 | \alpha_i) = \delta_{j0} + \sum_k \delta_{jk} \alpha_{ik} + \sum_{k < k'} \delta_{jkk'} \alpha_{ik} \alpha_{ik'} + \dots + \delta_{j12 \dots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{ik}, \quad (2.20)$$

Na equação 2.20, α_{ij}^* representa o vetor reduzido de atributos exigidos pelo item j . O parâmetro δ_{j0} representa a probabilidade base de acerto; δ_{jk} representa o efeito principal do atributo k ; $\delta_{jkk'}$ representa o efeito de interação entre dois atributos; e o último termo representa a interação conjunta entre todos os atributos requeridos pelo item.

De maneira simplificada, essa formulação possibilita responder à seguinte questão: *como cada atributo, individualmente ou em combinação com outros, contribui para a probabilidade de acerto do item?* Assim, o G-DINA não pressupõe que todos os atributos precisam ser dominados simultaneamente, como ocorre no DINA. Em vez disso, ele permite estimar como diferentes padrões de domínio influenciam a resposta do estudante.

Por essa razão, o G-DINA é considerado um modelo saturado, isto é, ele não impõe restrições fortes sobre a forma como os atributos interagem. Cada combinação possível de domínio pode apresentar uma probabilidade distinta de acerto. Essa flexibilidade possibilita a representação de relações tanto compensatórias quanto não compensatórias entre os atributos. Em uma relação compensatória, o domínio de um atributo pode compensar parcialmente a ausência de outro. Essa um relação não compensatória, todos os atributos são necessários simultaneamente para que haja uma alta probabilidade de acerto (Fernandes, 2024).

Uma característica relevante do G-DINA é que ele funciona como um modelo geral que engloba diversos modelos reduzidos como casos particulares, mediante restrições nos parâmetros (Torre, 2011). Entre os principais modelos derivados, destacam-se, além do MDC DINA, o *Deterministic Input, Noisy "Or" gate* (DINO) que é um modelo disjuntivo - não compensatório (Templin e Henson, 2006); *Additive Cognitive Diagnostic Model* (A-CDM) modelo aditivo compensatório (Torre, 2011); *Linear Logistic Model* (LLM) modelo log-linear

(Henson, Templin e Willse, 2009); *Reduced Reparameterized Unified Model* (RRUM) (Zhang et al., 2015). Essa variedade permite selecionar o modelo mais adequado de acordo com a natureza dos dados e das habilidades avaliadas.

Contudo, a maior flexibilidade do G-DINA também implica maior complexidade. Como esse modelo estima efeitos principais e interações entre atributos, o número de parâmetros tende a aumentar rapidamente à medida que cresce o número de atributos exigidos por um item. Por isso, sua aplicação requer uma quantidade maior de dados e mais cuidado na especificação da Q-matriz (Ma e Torre, 2020).

2.3.3 A avaliação funcional dos MDC

Além de estimar os parâmetros e classificar os estudantes em perfis de domínio, os MDC também podem ser avaliados quanto à sua capacidade preditiva. Nesse caso, busca-se verificar quão próximas estão as probabilidades estimadas pelo modelo das respostas observadas dos estudantes.

A *Root Mean Square Error* (RMSE) e o *Mean Absolute Error* (MAE) são métricas amplamente utilizadas para avaliar modelos (Hodson, 2022). Ambas medem a diferença entre a resposta observada X_{ij} e a probabilidade prevista pelo modelo $\hat{P}_{ij}$.

O RMSE é definido como:

$$RMSE = \sqrt{\frac{1}{NJ} \sum_{i=1}^N \sum_{j=1}^J (X_{ij} - \hat{P}_{ij})^2} \quad (2.21)$$

De forma intuitiva, o RMSE representa a magnitude dos erros entre os valores observados e os valores previstos pelo modelo. Como as diferenças são elevadas ao quadrado antes do cálculo da média, erros maiores exercem maior influência sobre o resultado. Assim, valores próximos de zero indicam maior proximidade entre as previsões do modelo e as respostas observadas.

Já o MAE é definido como:

$$MAE = \frac{1}{NJ} \sum_{i=1}^N \sum_{j=1}^J |X_{ij} - \hat{P}_{ij}| \quad (2.22)$$

Nas equações 2.21 e 2.22, N representa o número de estudantes, J representa o número de itens, X_{ij} é a resposta observada e $\hat{P}_{ij}$ é a probabilidade de resposta correta estimada pelo modelo. Quanto menores os valores de RMSE e MAE, maior é a proximidade entre as respostas previstas e as respostas observadas, o que indica melhor desempenho preditivo do modelo

(Yamaguchi e Zhang, 2023). Em outras palavras, essas métricas não indicam diretamente quais habilidades cognitivas foram dominadas ou não, mas ajudam a avaliar se o modelo está produzindo previsões compatíveis com as respostas reais.

A principal diferença entre as duas métricas está na influência dos erros de maior magnitude. O RMSE atribui maior peso a essas discrepâncias devido à elevação dos erros ao quadrado (Kang, Yang e Zeng, 2019). Já o MAE utiliza o valor absoluto dos erros, fornecendo uma medida direta da diferença média entre a predição e a observação.

Para ilustrar o uso dessas métricas, considera-se um cenário simplificado com dois estudantes ($N = 2$) e três itens ($J = 3$). A matriz de resposta X e a matriz de probabilidades previstas pelo modelo $\hat{P}$ são dadas por:

$$X = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \quad (2.23)$$

$$\hat{P} = \begin{bmatrix} 0,90 & 0,30 & 0,80 \\ 0,20 & 0,70 & 0,60 \end{bmatrix} \quad (2.24)$$

Nesse exemplo, cada elemento de X representa uma resposta observada, sendo que 1 indica acerto e 0, erro. Já cada elemento de $\hat{P}$ representa a probabilidade estimada de acerto pelo modelo para o estudante e o item correspondente.

Para calcular o MAE, considera-se a média dos erros absolutos:

$$MAE = \frac{|1 - 0,90| + |0 - 0,30| + |1 - 0,80| + |0 - 0,20| + |1 - 0,70| + |1 - 0,60|}{6} \quad (2.25)$$

$$MAE = \frac{0,10 + 0,30 + 0,20 + 0,20 + 0,30 + 0,40}{6} = \frac{1,50}{6} = 0,25 \quad (2.26)$$

Para o cálculo do RMSE, os erros são elevados ao quadrado antes da média:

$$RMSE = \sqrt{\frac{(1 - 0,90)^2 + (0 - 0,30)^2 + (1 - 0,80)^2 + (0 - 0,20)^2 + (1 - 0,70)^2 + (1 - 0,60)^2}{6}} \quad (2.27)$$

$$RMSE = \sqrt{\frac{0,01 + 0,09 + 0,04 + 0,04 + 0,09 + 0,16}{6}} = \sqrt{0,0717} \approx 0,268 \quad (2.28)$$

Nesse caso, o MAE igual a 0,25 indica que, em média, as probabilidades previstas diferem das respostas observadas em 0,25 na escala de 0 a 1, o que corresponde a 25 pontos

percentuais. Já o RMSE igual a aproximadamente 0,268 representa uma magnitude de erro de aproximadamente 0,268 nessa mesma escala. O valor ligeiramente superior ao MAE decorre da maior influência exercida pelos erros de maior magnitude no cálculo do RMSE.

Esse exemplo mostra que ambas as métricas avaliam a distância entre as respostas observadas e as probabilidades estimadas, mas com sensibilidades diferentes. Em MDC, essa comparação é útil para verificar se o modelo está produzindo probabilidades coerentes com o desempenho real dos estudantes.

No contexto desta tese, essas métricas são relevantes, pois permitem avaliar não apenas a estrutura formal da Q-matriz, como também seu impacto funcional no desempenho do modelo. Assim, uma Q-matriz pode ser avaliada tanto em termos estruturais, por sua proximidade com uma matriz definida por especialistas, quanto em termos funcionais, pelo desempenho preditivo obtido quando utilizada em modelos como o DINA ou o G-DINA.

2.3.4 Limitações do MDC

Apesar dos avanços proporcionados pelos MDC, sua aplicação em contextos educacionais reais ainda enfrenta limitações importantes, que não estão necessariamente relacionadas aos modelos, mas à forma como o conhecimento é representado (Fu et al., 2024; Rupp, Templin e Henson, 2010). A Q-matriz, por exemplo, ocupa um papel central, pois estabelece explicitamente a relação entre itens de avaliação e atributos cognitivos, servindo de base para a inferência do estado de conhecimento dos estudantes (Tatsuoka, 1984; Torre, 2009a).

Essa dependência estrutural implica que a qualidade das inferências produzidas por modelos como DINA e o G-DINA está diretamente condicionada à especificação correta da Q-matriz. Em outras palavras, a Q-matriz define quais habilidades (conceitos) são necessárias(os) para resolver cada item (questão) e, portanto, qualquer inconsistência nessa representação pode comprometer todo o processo de diagnóstico (Rupp, Templin e Henson, 2010). Estudos recentes mostram que erros na especificação da Q-matriz podem levar a estimativas incoerentes de parâmetros, aumento do erro preditivo e interpretações cognitivas equivocadas, afetando diretamente a validade dos resultados obtidos pelos MDC (Torre, 2009a; Lopes e Netto, 2026).

A construção da Q-matriz é realizada por especialistas, que analisam os itens e identificam os atributos envolvidos em sua resolução (Ravand e Baghaei, 2020). Embora esse processo seja teoricamente fundamentado, ele apresenta limitações relevantes. A primeira delas é a subjetividade. Diferentes especialistas podem associar habilidades distintas ao

mesmo item, especialmente em domínios complexos ou multidisciplinares (Pelánek, 2020; Yamaguchi, 2025). Além disso, a ausência de critérios padronizados pode gerar inconsistências na definição dos atributos e na forma como eles se relacionam aos itens (Matsuda et al., 2022). A isso se soma o alto custo do processo, que demanda tempo, experiência e múltiplas rodadas de validação, dificultando sua aplicação em larga escala (Ding et al., 2025).

Como alternativa, métodos orientados a dados (do inglês *Data-Driven*) têm sido propostos para validar ou estimar a Q-matriz a partir das respostas dos estudantes (Barnes, 2005; Desmarais, Beheshti e Naceur, 2012; Li e Chen, 2025). No entanto, essas abordagens também apresentam restrições importantes. Em geral, elas dependem de grandes volumes de dados, envolvem alto custo computacional e, frequentemente, assumem a existência de uma matriz inicial próxima da estrutura real (Xiong et al., 2024). Além disso, muitos desses métodos são direcionados ao refinamento da Q-matriz, e não à sua construção inicial, o que mantém a necessidade de especialistas como parte essencial do processo (Köhn et al., 2025).

Outro desafio importante é a escalabilidade. Nos ambientes educacionais digitais, como os STI, a quantidade de itens e CC aumenta rapidamente, tornando inviável a construção manual da relação item-atributo (Matsuda et al., 2022). Além disso, a necessidade de atualização contínua da matriz, para refletir mudanças curriculares ou novos conteúdos, torna ainda mais evidente a limitação das abordagens tradicionais em cenários dinâmicos (Pelánek, 2020).

Diante das limitações apresentadas, surge a necessidade de abordagens mais escaláveis e consistentes para apoiar a construção da Q-matriz. Nesse contexto, os LLM têm sido explorados como uma alternativa promissora, especialmente por sua capacidade de operar diretamente sobre o conteúdo textual dos itens (Li et al., 2024b; Tang et al., 2025). Esses modelos podem ser empregados em tarefas de “Rotulação de Conhecimento” (do inglês *Knowledge Tagging*), sugerindo possíveis associações entre itens e atributos (CC) sem depender diretamente dos dados de resposta dos estudantes.

Nesse sentido, uma contribuição importante desta tese é a adoção de uma abordagem de avaliação em dois níveis: o primeiro é estrutural, que analisa o alinhamento da Q-matriz gerada com apoio de LLM em relação aos padrões definidos por especialistas; o segundo é funcional, que avalia seu impacto no desempenho dos MDC. É importante ressaltar que essa automação não corresponde à automatização integral do MDC ou do processo avaliativo. O MDC continua dependendo de uma Q-matriz previamente definida e validada, que explicita quais CC são requeridos por cada item, além das respostas observadas dos estudantes. Assim,

nesta tese, o LLM atua especificamente como mecanismo de apoio à construção e à revisão da relação item-CC, enquanto a inferência do domínio das habilidades permanece sendo realizada pelo MDC a partir da Q-matriz validada.

2.4 IA no Apoio ao Processo de Avaliação

A integração da IA nos processos educacionais evoluiu de ferramentas de apoio para sistemas capazes de auxiliar diferentes etapas da aprendizagem e da avaliação. Entre essas aplicações, destacam-se a geração e a revisão de itens, a anotação de informações pedagógicas e o apoio à representação das relações entre itens e habilidades. Nesta tese, o interesse concentra-se particularmente na automação parcial e assistida da construção da relação item-CC utilizada na Q-matriz, etapa que tradicionalmente demanda esforço de especialistas. Esta seção discute como os LLM podem apoiar esse processo, sem substituir a validação pedagógica realizada pelo professor ou especialista.

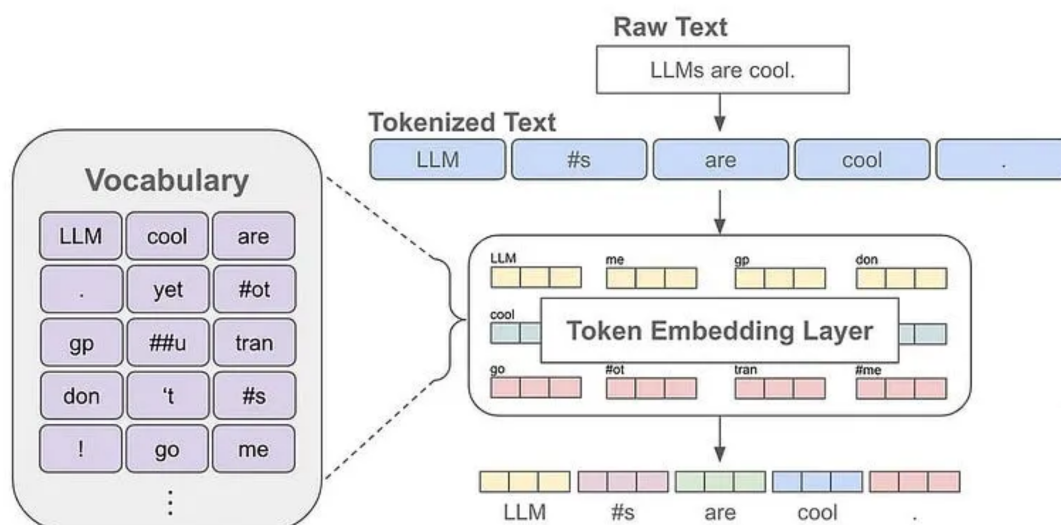
2.4.1 Modelo de Linguagem de Grande Escala

Os Modelo de Linguagem de Grande Escala, ou *Large Language Model* (LLM), são modelos computacionais de IA baseados em redes neurais profundas, treinados com grandes volumes de dados (*corpus*) textuais para processar e gerar linguagem natural (Naveed et al., 2025). Em aplicações generativas, o modelo recebe uma entrada que pode conter uma instrução, uma pergunta, um contexto ou exemplos destinados a orientar a geração da resposta. Esse conjunto de informações fornecido ao modelo é comumente denominado *prompt*. A partir dessa entrada, o LLM produz uma saída por meio da previsão sequencial de unidades textuais chamadas *tokens*. Um *token* pode corresponder a uma palavra, parte de uma palavra, número, símbolo ou sinal de pontuação. Assim, a resposta produzida pelo modelo é construída de forma progressiva, a partir da estimativa dos próximos *tokens* mais prováveis em um contexto específico (Brown et al., 2020).

Antes de ser processado pelo modelo, o conteúdo textual de entrada passa por uma etapa de *tokenização*. Nessa etapa, a sequência textual é dividida em *tokens*, que são posteriormente convertidos em representações numéricas chamadas de *embeddings*. Essas representações permitem que os *tokens* sejam projetados em um espaço vetorial, no qual relações sintáticas e semânticas podem ser processadas pelas camadas da rede neural. Dessa maneira, palavras, fragmentos de palavras e símbolos deixam de ser tratados apenas como elementos puramente textuais, passando a ser manipulados matematicamente pelo modelo. A Figura 8 exemplifica

o fluxo de processamento de uma entrada em um LLM.

Figura 8 – Fluxo simplificado do processamento de uma entrada textual em um LLM.



Fonte: Souza (2024).

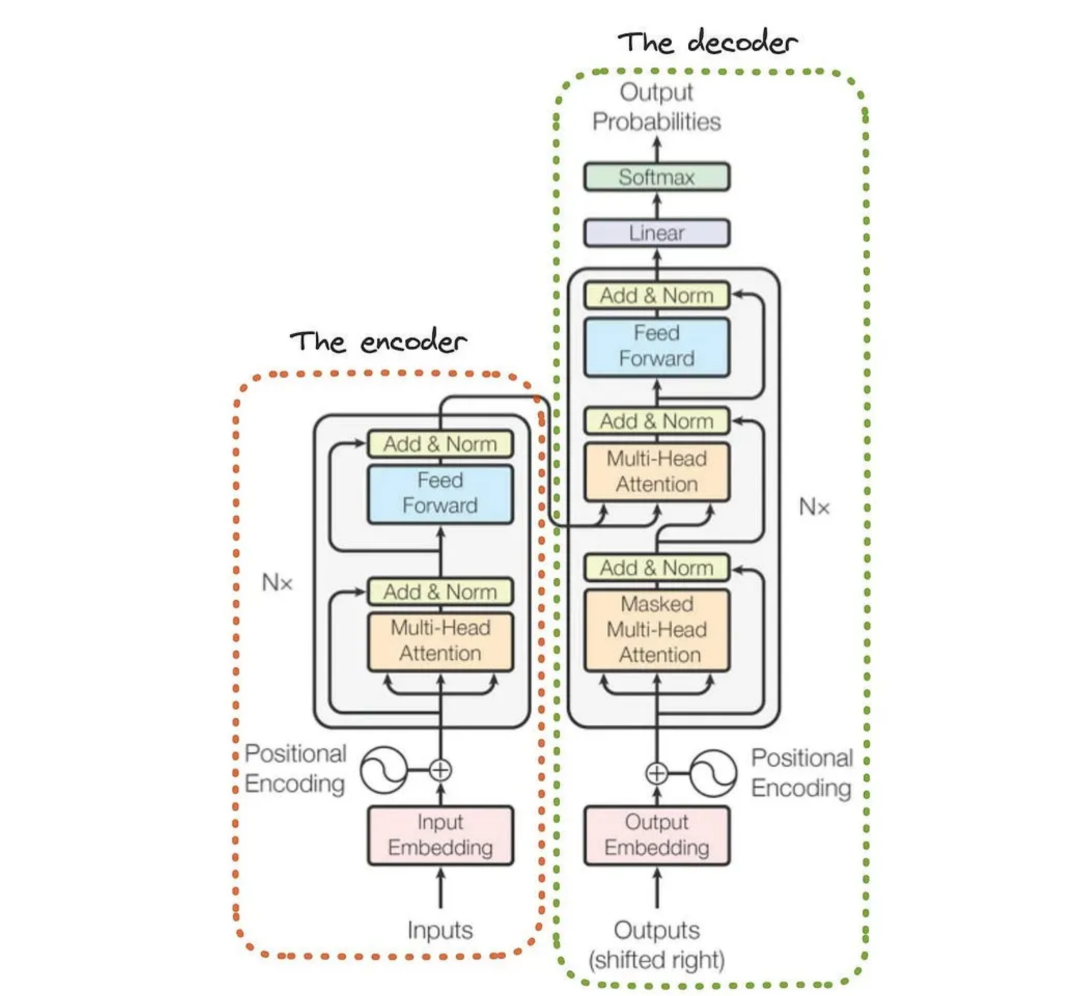
Conforme ilustra a Figura 8, a entrada textual é inicialmente convertida em *tokens*. Em seguida, esses *tokens* são transformados em *embeddings*. Essas representações passam pelas camadas do modelo de Aprendizagem Profunda (*deep learning*). Por fim, o modelo calcula uma distribuição de probabilidade sobre o vocabulário e seleciona o próximo *token*. Esse processo é repetido sucessivamente até a produção da resposta final.

Outro conceito importante é a chamada janela de contexto. Ela define a quantidade máxima de *tokens* que o modelo consegue “observar” em uma interação. Essa característica é relevante, pois determina a quantidade da entrada e da resposta anterior que permanece disponível para a geração dos próximos *tokens*. Em tarefas educacionais, a janela de contexto influencia diretamente a possibilidade de incluir, no mesmo *prompt*, o enunciado da questão, as alternativas, a lista de habilidades, exemplos de anotação, critérios de qualidade e o formato esperado da resposta. Modelos com janelas de contexto maiores podem analisar documentos mais longos ou conjuntos maiores de itens avaliativos, enquanto modelos com janelas menores exigem *prompts* mais sintéticos ou divisão do conteúdo em partes.

A base arquitetural da maioria dos LLM modernos é conhecida como *Transformer*, proposta por Vaswani et al. (2017). A Figura 9 ilustra parte dessa arquitetura. Originalmente, sua função era atender a tarefas de tradução automática, mas passou a ser amplamente utilizada no processamento de linguagem natural devido à sua capacidade de representar dependências entre diferentes partes de uma sequência textual. Diferentemente das arqui-

teturas recorrentes, o *Transformer* permite processar os *tokens* de maneira mais eficiente, considerando relações entre diferentes posições da sequência textual.

Figura 9 – Modelo de arquitetura *transformer*.



Fonte: Vaswani et al. (2017).

As LLMs podem ser classificadas em três tipos principais de arquitetura, que diferem na forma como processam a **atenção** e geram o texto. Segundo Naveed et al. (2025) são eles:

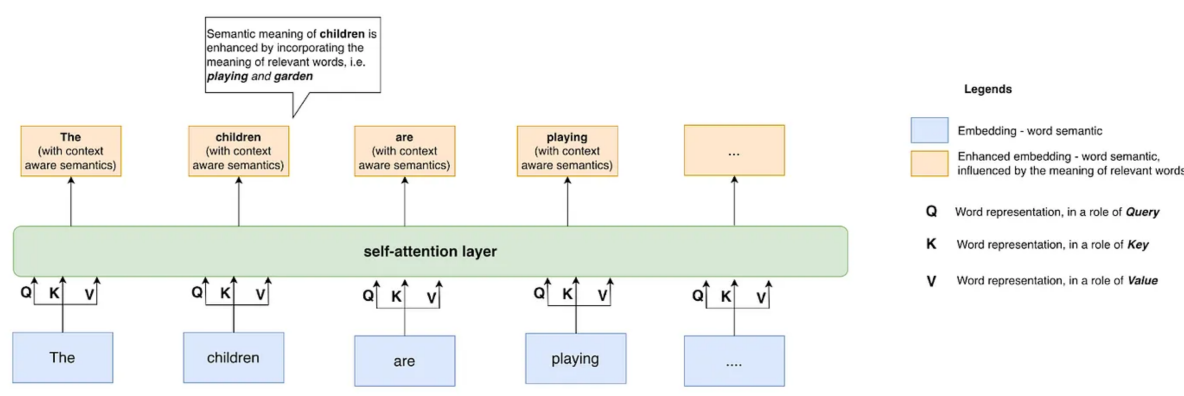
- **Encoder-Decoder:** Utilizado por modelos como o T5 e BART. Enquanto o *encoder* processa a sequência de entrada inteira de forma bidirecional, enquanto o *decoder* gera a saída sequencialmente. Esse tipo de arquitetura é ideal para tarefas de tradução e sumarização.
- **Causal Decoder (Decoder-only):** É a arquitetura mais popular atualmente. Nela, o modelo recebe uma sequência de *tokens* e gera sua continuação de forma autorregressiva, ou seja, produz um *token* por vez, considerando os *tokens* anteriores. Essa configuração

é apropriada para tarefas de geração de textos, resposta a perguntas, elaboração de explicações, criação de itens avaliativos e anotação de informações pedagógicas. Modelos como GPT, LLaMA, Qwen e Granite seguem essa lógica, embora apresentem variações arquiteturais e diferentes estratégias de treinamento (Naveed et al., 2025).

- **Prefix Decoder** (*Non-Causal Decoder*): Utilizado por modelos como o GLM-130B. Este modelo combina características das duas arquiteturas, permitindo atenção bidirecional para um “prefixo” de entrada e atenção causal para o restante a ser gerado.

O mecanismo de atenção (*Self-Attention*) é um dos elementos centrais do *Transformer*. De modo simplificado, a atenção permite que o modelo atribua pesos diferentes aos *tokens* da janela de contexto, indicando quais partes do texto são mais relevantes para representar um *token* ou para produzir a próxima saída. Por exemplo, em uma frase longa, o modelo pode relacionar um verbo a um complemento distante, ou associar um pronome ao termo ao qual ele se refere. Essa capacidade é importante para lidar com dependências, comuns em textos naturais.

Figura 10 – Exemplo simplificado do mecanismo de autoatenção.



Fonte: Parakhiya (2023).

A Figura 10 mostra um exemplo de autoatenção semelhante ao apresentado por Vaswani et al. (2017). Nesse tipo, as conexões entre os *tokens* indicam quais as palavras ou partes da sequência receberam maior peso durante o processamento. Cada cabeça de atenção pode capturar diferentes tipos de relação, como dependências sintáticas, relações semânticas, referências entre termos ou padrões de organização textual. Em tarefas educacionais, esse mecanismo pode fazer com que o modelo estabeleça relação entre partes do enunciado, alternativas, conceitos envolvidos e habilidades necessárias para resolução de uma questão.

A arquitetura original do *Transformer* utiliza o mecanismo de atenção em múltiplas cabeças (*Multi-Head Attention*). Isso significa que o modelo não calcula apenas uma única relação de atenção, mas várias relações em paralelo. Cada cabeça pode aprender a observar aspectos diferentes da sequência. Por exemplo, uma cabeça pode capturar relações gramaticais, enquanto outra pode destacar as dependências semânticas ou padrões relacionados à posição dos *tokens*. Em seguida, essas diferentes representações são combinadas para formar uma representação contextual mais completa. É importante ressaltar que a maioria dos LLM atuais utiliza uma configuração *decoder-only* (Naveed et al., 2025).

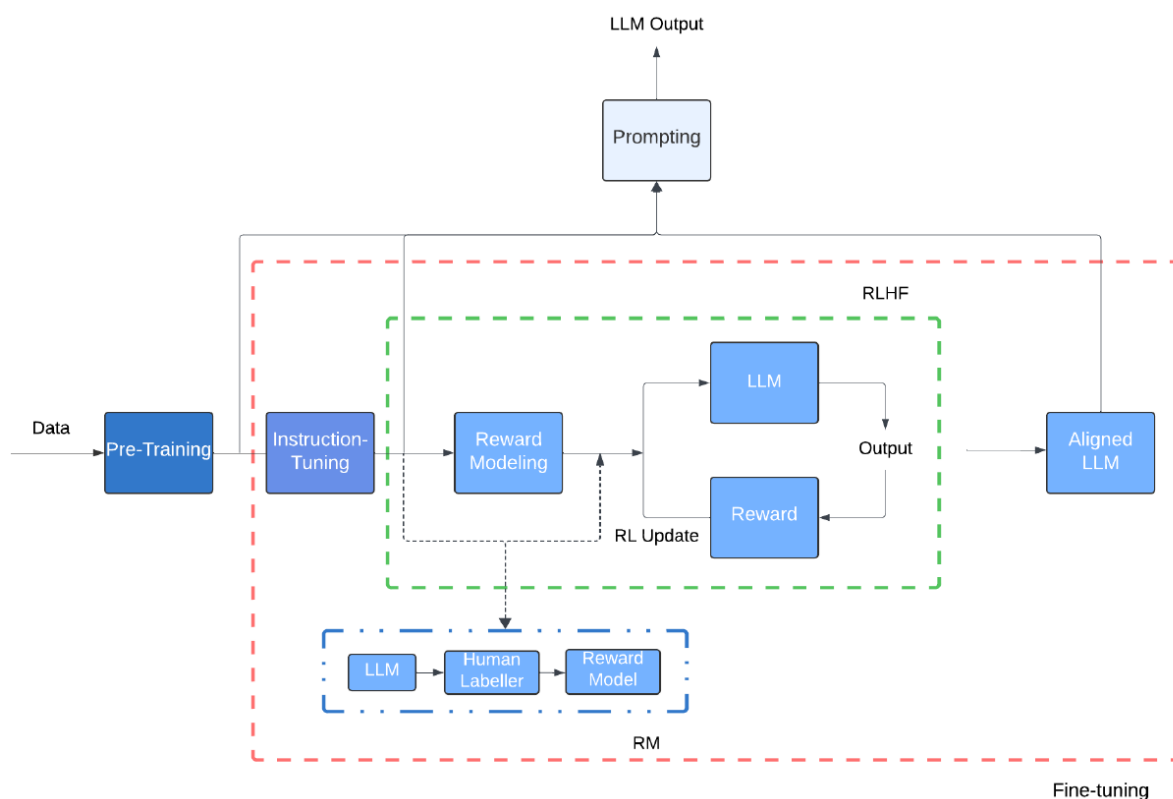
A tarefa principal de um LLM é prever o próximo *token* de uma sequência. O modelo recebe uma sequência textual, calcula uma distribuição de probabilidade sobre o vocabulário e estima qual *token* deve aparecer em seguida. A diferença entre a previsão do modelo e o *token* correto presente no conjunto de treinamento é medida por uma função de perda, frequentemente baseada em entropia cruzada. Em seguida, essa perda é propagada de volta pelas camadas da rede por meio de retropropagação, ou *backpropagation*, permitindo a atualização dos parâmetros do modelo por métodos de otimização baseados em gradiente (Goodfellow, Bengio e Courville, 2016).

A Figura 11 pode representar o processo de treinamento de uma LLM. O modelo recebe uma sequência de *tokens*, gera uma previsão para o próximo *token*, compara essa previsão com o *token* real, calcula o erro e atualiza seus parâmetros. Esse ciclo é repetido em larga escala com grandes volumes de textos. Dessa forma, o modelo ajusta seus parâmetros internos para aprimorar progressivamente sua capacidade de prever sequências linguísticas.

Nesse contexto, é importante diferenciar dois usos da palavra “peso”. Os pesos de atenção são valores calculados dinamicamente durante o processamento de uma entrada específica. Eles indicam quanto cada *token* deve considerar os demais *tokens* da sequência. Já os parâmetros do modelo são os pesos aprendidos durante o treinamento e armazenados na rede neural. São esses parâmetros que são atualizados durante a etapa de otimização. Enquanto os pesos de atenção variam conforme o texto de entrada, os parâmetros do modelo representam o conhecimento estatístico acumulado durante o treinamento (Rahman e Borera, 2024).

Após o treinamento, o modelo é utilizado na fase de inferência. Nessa etapa, ele recebe um *prompt* e gera uma resposta *token a token*. A cada passo, o modelo considera os *tokens* da entrada e os *tokens* já gerados para calcular a probabilidade do próximo *token*. O processo continua até que um critério de parada seja atingido, como o limite máximo ou a geração de um *token* de finalização. Essa dinâmica explica por que a formulação do *prompt* influencia

Figura 11 – Exemplo de diferentes estágios de treinamento e inferência em LLMs. O *prompt* para geração de respostas é possível em diferentes estágios, como pré-treinamento, *instruction tuning* ou alinhamento.



Fonte: Naveed et al. (2025).

diretamente a qualidade da resposta produzida.

A Aprendizagem em Contexto (*In-Context Learning* - ICL) (Brown et al., 2020; Naveed et al., 2025) é uma das capacidades mais relevantes e úteis das Modelo de Linguagem de Grande Escala, especialmente para a rotulação de conhecimento. A ICL é a habilidade do modelo de realizar uma nova tarefa apenas com exemplos fornecidos no *prompt*, sem que seus parâmetros internos sejam atualizados, ou seja, sem *fine-tuning*. Segundo (Brown et al., 2020) esta capacidade se manifesta em dois regimes principais:

- **Zero-Shot:** nesse caso, o modelo recebe apenas uma instrução sem exemplos e deve realizar a tarefa. Exemplo: “Classifique o texto a seguir como positivo ou negativo.”
- **Few-Shot:** O modelo recebe alguns exemplos da tarefa (ou demonstrações) junto com a instrução. Exemplo: “Texto: ‘Amei o filme’ -> Positivo. Texto: ‘Um tédio’ -> negativo. Texto: ‘Muito bom’ -> ?” . Essa técnica é particularmente valiosa para o *Knowledge Tagging*, pois permite guiar o modelo com exemplos corretos de associação entre

questão e CC (Moore et al., 2024; Li et al., 2024a).

- **Raciocínio em Cadeia de Pensamento** (*Chain-of-Thought*) (CoT): É uma técnica de *prompt* que incentiva o modelo a gerar passos intermediários de raciocínio antes de fornecer a resposta final. Isso é feito por meio de instruções como “Vamos pensar passo a passo” ou pelo fornecimento de exemplos que demonstram o raciocínio explícito, conforme (Wei, Carvalho e Stamper, 2025). O CoT melhora drasticamente o desempenho do modelo em tarefas que exigem lógica, múltiplas etapas de inferência e, crucialmente para esta pesquisa, a determinação de se uma questão requer um determinado conceito matemático (Li et al., 2024a).

Além das estratégias citadas acima, existem estratégias de revisão, como *self-review* ou *self-reflection*, nas quais o modelo é solicitado a revisar, justificar ou corrigir sua própria resposta (Maity e Deroy, 2024; Quincozes et al., 2025).

Além do uso por *prompt*, algumas LLM podem ser adaptadas por meio de técnicas conhecidas como *fine-tuning* e *prompt-tuning*. O *fine-tuning* consiste em dar continuidade ao treinamento do modelo, ou de parte dele, com dados específicos de um domínio ou tarefa. Essa estratégia pode aprimorar o desempenho em contextos especializados, mas exige dados de qualidade, maior custo computacional e atenção à generalização. O *prompt-tuning*, por sua vez, busca ajustar representações associadas ao *prompt*, preservando a maior parte dos parâmetros originais do modelo. Em contextos educacionais, essas técnicas podem ser úteis quando se deseja adaptar o modelo a um currículo, específico, um idioma, uma área de conhecimento ou um tipo de avaliação.

Os LLM também podem ser classificados conforme sua forma de acesso. Modelos fechados (*closed-source*), como GPT e Gemini, são, geralmente, acessados por meio de API (*Application Programming Interface*) ou *interfaces* comerciais. Esses modelos geralmente apresentam alto desempenho e facilidade de uso, mas oferecem menor transparência quanto aos dados de treinamento, detalhes arquiteturais e reprodutibilidade experimental (Ding et al., 2025). Modelos abertos (*open-sources*) ou parcialmente abertos, como LLaMA, Mixtral, Qwen, Phi e Granite, permitem maior controle experimental, execução local em alguns cenários e adaptação a pesquisas acadêmicas. Conforme Manchanda et al. (2024) essa distinção é relevante, pois influencia custos, privacidade dos dados, reprodutibilidade dos experimentos e possibilidade de integração com sistemas educacionais.

2.4.2 Famílias LLM

A literatura recente fornece um panorama diversificado de LLM que têm sido aplicados com sucesso em diferentes contextos educacionais. A Tabela 5 do *Survey* de [Naveed et al. \(2025\)](#) constitui uma referência para comparar os principais modelos, apresentando detalhes arquiteturais, como o número de camadas, o tipo de atenção e a função de ativação. No contexto específico desta tese, alguns modelos se destacam pelos resultados alcançados em uma variedade de pesquisas educacionais.

2.4.2.1 GPT – OpenAI

A LLM “Generative Pre-trained Transformer” – GPT⁹ (Transformador Pré-treinado Generativo) é um modelo *decoder-only* e *closed-source* desenvolvido pela OpenAI¹⁰, reconhecido por seu alto desempenho em benchmarks de raciocínio e compreensão de linguagem. De acordo com [Naveed et al. \(2025\)](#), o modelo possui aproximadamente 1,8 trilhão de parâmetros, distribuídos em 120 a 150 camadas (dependendo da configuração), com tamanho de estado oculto estimado entre 20.000 e 32.000 dimensões. A arquitetura do GPT-4 incorpora, especificamente, mecanismos atenção de múltiplas cabeças com mecanismo de consulta agrupada (*grouped-query attention*), que reduz os custos computacionais durante a decodificação, além de utilizar a função de ativação GeLU ([Cruz-Castañeda e Amadeus, 2025](#); [Naveed et al., 2025](#)).

O treinamento do GPT-4 foi realizado com taxa de aprendizado de 6×10^{-5} , tamanho *batch* de 32 mil a 200 mil *tokens*, e um *dropout* de 0,1. A janela de contexto máxima durante o treinamento variou entre 20.000 e 32.768 *tokens* conforme ([Bengesi et al., 2024](#)). Por ser um modelo proprietário, seu acesso se dá exclusivamente por meio de API, o que limita a reprodutibilidade e a transparência das pesquisas, embora garanta maior controle sobre a segurança e o alinhamento ([Manchanda et al., 2024](#)). O GPT-4 demonstrou desempenho superior em tarefas de raciocínio matemático e compreensão de linguagem, alcançando 85,5% de acurácia em três disparos (3-shot) no *benchmark* MMLU e 94,2% no GLUE ([Raiaan et al., 2024](#)).

⁹ <https://chatgpt.com>

¹⁰ <https://openai.com>

2.4.2.2 LLaMA – Meta AI

A família LLaMA¹¹ (acrônimo para *Large Language Model Meta AI*) representa um marco na democratização do acesso ao LLM de alto desempenho¹². Lançada originalmente em 2023, a primeira versão LLaMA-1 foi disponibilizada em tamanhos de 7, 13, 33 e 65 bilhões de parâmetros (Touvron et al., 2023a). Sua arquitetura *decoder-only* incorpora normalização RMS pré-camada, função de ativação SwiGLU que é uma variante da GLU que combina Swish com *gating linear* e *embeddings* posicionais rotacionais RoPE (Naveed et al., 2025). O modelo foi treinado com 1,4 trilhão de *tokens* utilizando taxa de aprendizado de $1,5 \times 10^{-4}$, *batch* de 4 milhões de *tokens* e pesos (*weight decay*) de 0,1. O treinamento utilizou 2.048 GPUs A100 de 80 gigabytes ao longo de 21 dias (Naveed et al., 2025; Raiaan et al., 2024).

A versão LLaMA-2, lançada em julho de 2023, expandiu significativamente as capacidades do modelo. Foram disponibilizadas versões de 7, 13, 34 e 70 bilhões de parâmetros, treinadas com 2 trilhões de *tokens* (Touvron et al., 2023b). As principais inovações incluem: (i) duplicação do tamanho do contexto de 2.048 para 4.096 *tokens*; (ii) introdução da atenção de consulta agrupada (*grouped-query attention*) para maior eficiência inferencial; (iii) licenciamento comercialmente para uso empresarial; e (iv) disponibilização de uma versão “Chat” com alinhamento via RLHF (*Reinforcement Learning from Human Feedback*), tornando o modelo mais seguro e útil para diálogo (Naveed et al., 2025; Touvron et al., 2023b). O treinamento do LLaMA-2 utilizou 2.000 GPUs A100 de 80 gigabytes ao longo de 25 dias (Raiaan et al., 2024).

O LLaMA-3, mais recente, representa um salto adicional em escala e capacidade. De acordo com Naveed et al. (2025), ele foi treinado com um conjunto de dados sete vezes maior que do LLaMA-2, com aproximadamente 128 mil *tokens* de contexto e 405 bilhões de parâmetros em sua maior versão. Sua arquitetura mantém o mecanismo de atenção de consulta agrupada e a função de ativação SwiGLU, com 126 camadas, 128 cabeças de atenção e um tamanho de estado oculto de 16.384 dimensões.

2.4.2.3 Gemini – Google

O Gemini¹³ é o LLM do Google¹⁴. De forma mais ampla, é uma família de modelos multimodal projetados para processar várias modalidades ou tipos de dados, incluindo áudio, imagens, código de software, texto e vídeo. Anunciada em 2023 e 2024, representa a evolução

¹¹ <https://www.llama.com/>

¹² <https://www.llama.com/llama-downloads/>

¹³ <https://gemini.google/br/about/?hl=pt-BR>

¹⁴ <https://about.google/>

da Google para competir diretamente com o GPT-4¹⁵.

O Gemini 1.0 Ultra foi o primeiro modelo autorregressivo a atingir um desempenho comparável ao humano no *benchmark* MMLU, com 83,7% de acurácia em cinco disparos (Naveed et al., 2025). Sua arquitetura é baseada em *decoder-only* com mecanismo de atenção de múltiplas cabeças e incorporação posicional RoPE. O modelo foi treinado utilizando a infraestrutura de TPUs v4 e v5e da Google, embora detalhes específicos dos parâmetros não tenham sido divulgados publicamente (Naveed et al., 2025).

Posteriormente, foi lançado o Gemini 1.5 Pro, que introduziu uma arquitetura MoE (*Mixture-of-Experts*) com uma janela de contexto de até 2 milhões de *tokens*, permitindo o processamento de documentos extremamente longos, como livros inteiros ou grandes bases de código (Naveed et al., 2025). O modelo obteve 87,8% no *benchmark HellaSwag* e 74,4% no *HumanEval* para tarefas de programação (Raiaan et al., 2024). Embora seja predominantemente um modelo de código fechado, a Google tem disponibilizado versões reduzidas como o Gemini Nano para dispositivos móveis e para a comunidade de pesquisa. Sua arquitetura multimodal sugere potencial para aplicações educacionais que envolvam não apenas texto, mas também imagens, gráficos e vídeos instrucionais (Naveed et al., 2025).

2.4.2.4 Phi – Microsoft

A família Phi¹⁶ representa uma abordagem distinta no desenvolvimento de LLMs, focada na eficiência e na qualidade dos dados de treinamento em detrimento da escala bruta. O Phi-2, com apenas 2,7 bilhões de parâmetros, foi treinado em um *corpus* de dados sintéticos e de alta qualidade, incluindo livros didáticos e materiais educacionais cuidadosamente selecionados (Javaheripi et al., 2023).

Sua arquitetura *decoder-only* utiliza mecanismo de atenção de consulta agrupada, normalização RMS, função de ativação SwiGLU e incorporação posicional RoPE (Lu et al., 2024). Apesar de seu tamanho modesto, o Phi-2 superou modelos muito maiores, como o LLaMA-2-7B e o Mistral-7B, em tarefas de raciocínio matemático e codificação, com acurácia de 90,2% no MATH-500 (Naveed et al., 2025). Seu treinamento foi realizado em 14 dias utilizando infraestrutura da Microsoft (Raiaan et al., 2024).

O Phi-3, lançado posteriormente, expandiu a família para 3,8 bilhões de parâmetros de sua versão mini, 7 e 14 bilhões, mantendo o foco em dados educacionais de qualidade e adicionando suporte a uma janela de contexto de até 128 mil *tokens* (Naveed et al., 2025).

¹⁵ <https://www.ibm.com/br-pt/think/topics/google-gemini>

¹⁶ <https://azure.microsoft.com/pt-br/products/phi>

A importância do Phi-2 para a pesquisa educacional é particularmente relevante, pois [Wei, Carvalho e Stamper \(2025\)](#) o utilizaram como base para o algoritmo *KCluster*, uma abordagem inovadora para descoberta de CC que combina a capacidade inferencial da LLM com técnicas de clusterização.

2.4.2.5 Mixtral – Mistral AI

O Mixtral 8x7B é um modelo de código aberto que popularizou a arquitetura MoE (Mixture-of-Experts) entre a comunidade científica. Ao contrário de modelos densos, que ativam todos os parâmetros para cada *token* processado, o Mixtral utiliza oito MoE, ativando apenas dois deles por *token*. Isso resulta em um modelo com 47 bilhões de parâmetros totais, mas que executa apenas 13 bilhões por *token* durante a inferência, combinando alta capacidade com eficiência computacional. O modelo foi treinado com 32 mil *tokens* de contexto, utilizando uma taxa de aprendizado de 3×10^{-4} e otimizador *AdamW*. A arquitetura incorpora atenção de consulta agrupada, normalização RMS, incorporação posicional RoPE e função de ativação SwiGLU ([Jiang et al., 2024](#)).

O Mixtral demonstrou desempenho competitivo ou superior ao LLaMA-2-70B em diversos benchmarks, com a vantagem de ser significativamente mais rápido e econômico. Na área de pesquisa educacional, ([Li et al., 2024a](#)) reportaram que o Mixtral-8x7B-Instruct apresentou um dos melhores desempenhos entre os modelos de código aberto avaliados na tarefa de *Knowledge Tagging* de conceitos matemáticos, com 88,94% de acurácia em configurações de *few-shot*, superando vários concorrentes de porte similar.

2.4.2.6 Qwen – Alibaba Cloud

O Qwen¹⁷ (acrônimo para *Quantum Wen*), também conhecido como Tongyi Qianwen, é uma família de LLMs desenvolvida pela Alibaba Cloud¹⁸, representando um dos mais significativos esforços da indústria chinesa no desenvolvimento de LLMs para o idioma chinês e para aplicações multilíngues ([Bai et al., 2023](#)). De acordo com [Naveed et al. \(2025\)](#), o Qwen se destaca por seu foco em desempenho multilíngue, eficiência computacional e capacidades de raciocínio matemático e científico, características que o tornam particularmente relevante para aplicações educacionais em contextos não anglófonos¹⁹, incluindo o português brasileiro ([Cruz-Castañeda e Amadeus, 2025](#)).

O Qwen adota uma arquitetura *decoder-only* com otimizações específicas para balancear

¹⁷ <https://qwen.ai/home>

¹⁸ <https://www.alibabacloud.com/>

¹⁹ Refere-se a países, comunidades ou indivíduos cuja língua materna ou oficial não é o inglês.

capacidade e eficiência. A família Qwen é composta por modelos que variam de 1,8 bilhão a 72 bilhões de parâmetros. A versão Qwen1.5-72B-Chat sendo a mais robusta para tarefas de raciocínio complexo (Bai et al., 2023). A arquitetura do Qwen incorpora atenção de múltiplas cabeças com otimizações para redução de memória, incorporação posicional RoPE, normalização RMS e função de ativação SwiGLU Naveed et al. (2025). Uma característica distintiva do Qwen é seu vocabulário de aproximadamente 150.000 *tokens*, significativamente maior que o de modelos ocidentais como o GPT-4, que possui cerca de 100.000 *tokens*, para acomodar a riqueza morfológica do chinês e de outros idiomas asiáticos (Bai et al., 2023).

De acordo com Bai et al. (2023) o treinamento do Qwen foi realizado na infraestrutura da Alibaba Cloud, utilizando GPUs NVIDIA A100 em configuração distribuída. O modelo foi pré-treinado em um *corpus* multilíngue com mais de 3 trilhões de *tokens*, abrangendo chinês, inglês e outros idiomas asiáticos, além de código-fonte e dados científicos. Essa escala de treinamento é comparável à de outros modelos da mesma geração, como o LLaMA-2 (2 trilhões de *tokens*) e o GPT-3 (300 bilhões de *tokens*) (Naveed et al., 2025; Raiaan et al., 2024).

Embora detalhes específicos sobre o tempo de treinamento e o número exato de GPUs não sejam amplamente divulgados, a Tabela 3 no trabalho de Naveed et al. (2025) fornece informações comparativas sobre a infraestrutura de treinamento de diversas LLM, posicionando o Qwen na categoria de modelos que exigem recursos computacionais significativos, porém acessíveis para instituições de pesquisa com infraestrutura adequada.

O Qwen está disponível em duas modalidades principais. A primeira é a API comercial da Alibaba Cloud, acessível mediante pagamento, que oferece as versões mais recentes e otimizadas do modelo. A segunda, e mais relevante para a pesquisa acadêmica, é a disponibilização de versões *open-source* no Hugging Face, sob licenças que permitem uso acadêmico e, em alguns casos, comercial (Bai et al., 2023). Diferentemente de modelos como o GPT-4 (completamente *closed-source*) ou o LLaMA-2 (totalmente *open-source*), o Qwen adota uma estratégia híbrida: versões de médio porte até 7B parâmetros são completamente abertas, enquanto versões maiores acima de 14B requerem aprovação para uso comercial. Esta política de disponibilização tem contribuído para a adoção do Qwen pela comunidade internacional, especialmente em estudos que envolvem idiomas de baixo recurso e aplicações multilíngues (Manchanda et al., 2024).

O Qwen1.5-72B-Chat²⁰ tem demonstrado desempenho competitivo com modelos de porte similar em *benchmarks* internacionais. De acordo com Bai et al. (2023), o modelo

²⁰ <https://qwen.ai/blog?id=qwen1.5>

alcança resultados notáveis em tarefas de raciocínio matemático em *benchmark* GSM8K, compreensão de leitura SQuAD e MMLU de conhecimento geral. Estudos independentes citados por [Cruz-Castañeda e Amadeus \(2025\)](#) indicam que o Qwen se posiciona entre os melhores modelos *open-source* disponíveis para tarefas multilíngues, com desempenho superior ao LLaMA-2 em tarefas envolvendo idiomas asiáticos. Embora dados específicos de *benchmarks* padronizados não sejam amplamente divulgados na literatura analisada, o Qwen é reconhecido por seu equilíbrio entre desempenho e eficiência computacional ([Li et al., 2024a](#)).

2.4.2.7 Granite – IBM Research

A família de modelos de LLM Granite²¹, desenvolvida pela IBM Research²², representa uma iniciativa no desenvolvimento de LLM de código aberto com foco em aplicações empresariais, incluindo código, chamada de funções e compreensão multimodal de documentos ([Granite Team, 2024](#)). Todos os modelos são disponibilizados sob licença Apache 2.0, que permite uso comercial e acadêmico irrestrito.

O Granite 3.0 é composto por duas arquiteturas distintas. Os modelos densos incluem o Granite-3.0-2B com 2,5 bilhões de parâmetros e o Granite-3.0-8B com aproximadamente 8,1 bilhões de parâmetros, ambos com 40 camadas, atenção GQA (*Grouped-Query Attention*) com 8 cabeças KV, normalização RMS, ativação SwiGLU e *embeddings* posicionais RoPE ([Granite Team, 2024](#)). Os modelos MoE incluem o Granite-3.0-1B-A400M com 400M parâmetros ativados, 1,3B totais e o Granite-3.0-3B-A800M com 800M ativados, 3,3B totais. Ambos têm ativação de 8 *experts* por *token* em um total de 32 e 40 *experts*, respectivamente, e utilizam a implementação ScatterMoE sem descarte de *tokens* ([Granite Team, 2024](#)).

Os modelos densos foram treinados em 12 trilhões de *tokens*, enquanto os modelos MoE, em dois estágios, foram treinados com 10 trilhões de *tokens*, conforme citado por ([Granite Team, 2024](#)). O treinamento empregou μ P (*Maximal Update Parameterization*) para transferência de hiperparâmetros de modelos pequenos para grandes, e o *Power Scheduler* para transferência da taxa de aprendizado entre diferentes tamanhos de *batch* (lotes) ([Granite Team, 2024](#)). O pós-treinamento incluiu SFT (*Supervised Fine-Tuning*) com dados de inglês (40%), código (25%), matemática (5%) e segurança (30%), seguido de alinhamento com ensemble de PPO, BRAIn e Best-of-N Sampling ([Granite Team, 2024](#)).

²¹ <https://www.ibm.com/br-pt/granite>

²² <https://research.ibm.com/>

O modelo Granite-20B-FUNCTIONCALLING²³, especializado em chamadas de funções, foi treinado com QLoRA sobre o Granite-20B-CODE-INSTRUCT, utilizando 142 mil exemplos distribuídos em sete tarefas: detecção de nomes de funções, preenchimento de parâmetros, chamada aninhada, encadeamento de funções, chamadas paralelas e geração de respostas (Abdelaziz et al., 2024). No *Berkeley Function Calling Leaderboard*, o modelo alcançou a quarta posição geral (84,71%), sendo o melhor entre os modelos de licença aberta, com a menor taxa de alucinação ($< 0,1$) em avaliações fora de domínio (Abdelaziz et al., 2024).

O Granite Vision é um modelo multimodal para compreensão de documentos visuais com aproximadamente 3 bilhões de parâmetros totais. Este modelo combina *encoder* visual SigLIP com resolução de 384×384 pixels, utilizando a técnica “AnyRes” para processamento de imagens de alta resolução através de grade de até 10 *patches*. Ele também possui um projetor MLP de duas camadas com ativação GELU; e o Granite-3.1-2B-Instruct como LLM base (Granite Vision Team et al., 2025). O treinamento em três estágios utilizou aproximadamente 20 milhões de pares imagem-texto, incluindo datasets sintéticos proprietários (DocFM-VQA com 19,9M QAs, DocFM-VQA-Charts com 5,5M QAs) e datasets públicos (DocVQA, ChartQA, TextVQA, AI2D, InfoVQA). O modelo introduziu vetores de atenção esparsos para classificação de segurança, alcançando 96,2% de acurácia no VLGard (Granite Vision Team et al., 2025).

2.4.3 LLM no apoio à criação de avaliações da aprendizagem

Os LLM vêm sendo estudadas como ferramentas para auxiliar na elaboração de avaliações de aprendizagem, principalmente por sua capacidade de produzir e organizar textos educacionais a partir de instruções em linguagem natural. No contexto avaliativo, esses modelos podem auxiliar na geração de questões, alternativas, gabaritos, justificativas, comentários e diferentes tipos de itens. Essa aplicação é relevante, pois a construção manual de avaliações demanda tempo, domínio de conteúdo, conhecimento pedagógico e atenção aos critérios técnicos de qualidade. Maity e Deroy (2024) discutem esse potencial ao tratar da geração automática de questões e da avaliação de respostas com LLMs, considerando diferentes estratégias de uso, como *zero-shot*, *few-shot chain-of-thought prompting*, *fine-tuning* e *prompt-tuning*.

A geração automática de questões é uma das principais aplicações dos LLM na construção de avaliações educacionais. Por exemplo, Varma et al. (2025) propõem um sistema para criar e validar as questões escolares, destacando que os LLM podem produzir itens (questões) a partir de tópicos ou materiais de referência, desde que a saída seja revisada em relação

²³ <https://huggingface.co/ibm-granite/granite-20b-functioncalling>

à adequação pedagógica e ao nível de conhecimento dos estudantes. Nesse sentido, [Shirkar, Langhi e Agrawal \(2025\)](#) propõem a geração de questionários em ambientes de Sistema de Gestão de Aprendizagem, indicando que os LLM podem apoiar a criação de avaliações em diferentes formatos e reduzir parte do esforço manual dos professores. Tal abordagem aproxima os LLM de sistemas educacionais personalizados, nos quais a avaliação pode ser ajustada ao conteúdo, ao desempenho e às dificuldades dos estudantes.

A literatura também destaca que a qualidade das avaliações geradas depende da forma como os LLM são instruídas. [Quincozes et al. \(2025\)](#) analisam as técnicas de Engenharia de *Prompt* para geração de questões de múltipla escolha na área de Engenharia de Requisitos, comparando estratégias como *zero-shot*, *few-shot*, *chain-of-thought*, *exemplar-guided* e *template-based*. Os autores destacam que a qualidade das questões produzidas está relacionada aos critérios de fluidez, diversidade, relevância e complexidade, além da consistência entre julgamentos automatizados e avaliações humanas. Assim, a criação de avaliações com o apoio de LLM não depende apenas do modelo utilizado, mas também da clareza das instruções, dos critérios de qualidade e do alinhamento entre o *prompt*, o conteúdo e o objetivo avaliativo.

Apesar desse potencial, a qualidade pedagógica dos itens gerados continua sendo um desafio. [Nogueira et al. \(2025\)](#) analisam o uso de LLM para a geração automática de itens avaliativos em Língua Portuguesa, voltados para avaliações brasileiras utilizadas em larga escala. Os autores destacam que a elaboração de itens envolve a definição das habilidades avaliadas, das formas de aferição, dos níveis de dificuldade e da adequação pedagógica, o que torna a produção manual uma tarefa especializada e custosa. Além disso, o estudo indica que as tecnologias atuais ainda não resolvem completamente o problema de se gerar itens de alta qualidade em português, sendo necessárias estratégias de revisão, triagem e validação por especialistas.

Outro ponto relevante está relacionado à geração de alternativas e de distratores em questões de múltipla escolha. [Chico et al. \(2025\)](#) investigam a geração automática de questões com distratores significativos em língua portuguesa, utilizando modelos como Sabia-3 e GPT-4o-mini. O estudo mostra que a incorporação de princípios educacionais nos *prompts* melhora a relevância e a diversidade dos distratores gerados, indicando que os LLM podem contribuir para a elaboração de alternativas mais plausíveis e mais alinhadas aos objetivos da avaliação. Esse aspecto é importante nas avaliações diagnósticas porque os distratores não devem funcionar apenas como respostas incorretas, mas como elementos capazes de indicar

dificuldades, concepções equivocadas (*misconception*) ou lacunas de aprendizagem.

O mapeamento dos itens de avaliação consiste em associar cada item a informações pedagógicas e cognitivas, tais como a habilidade avaliada, o nível de dificuldade, o tipo de conhecimento mobilizado, o objetivo de aprendizagem, o gabarito, a justificativa e os possíveis erros esperados (Briggs Derek C. Alonzo, Schwab e Wilson, 2006; Desmarais, Beheshti e Naceur, 2012). Nesse sentido, as LLM podem auxiliar no mapeamento inicial dos itens, sugerindo quais habilidades estão envolvidas em uma questão, quais conceitos são necessários para resolvê-la, qual é a justificativa do gabarito e quais dificuldades podem estar relacionadas aos distratores. Tal possibilidade é relevante para avaliações diagnósticas e para sistemas educacionais inteligentes, pois facilita a interpretação das respostas dos estudantes não apenas como acertos ou erros, mas também como evidências de domínio das habilidades.

Entretanto, a anotação automática ou semiautomática de itens deve ser tratada como uma etapa de apoio à autoria, e não como uma substituição da validação humana. Santos et al. (2025) investigaram os LLM leves para gerar questões matemáticas no português brasileiro e indicaram que a avaliação pedagógica manual ainda demanda tempo, mas que alternativas semiautomatizadas podem contribuir para tornar esse processo mais escalável, desde que preservem a relevância pedagógica da análise. Dessa forma, no contexto desta pesquisa, as LLM podem ser entendidas como ferramentas de auxílio à criação, revisão e anotação inicial de avaliações, mantendo o professor e os especialistas como responsáveis pela validação final dos itens e de suas associações com habilidades.

2.4.4 Representação, inferência e automação usando LLM

A utilização de LLM para apoiar a avaliação da aprendizagem pode ser articulada em três dimensões complementares: representação, inferência e automação assistida. A representação refere-se à forma como os elementos da avaliação são modelados computacionalmente, incluindo a relação entre itens e CC. A inferência diz respeito aos mecanismos utilizados para estimar o estado de conhecimento do estudante a partir das respostas observadas. Já a automação assistida corresponde ao uso de técnicas computacionais para reduzir o esforço envolvido em etapas de autoria, anotação e revisão dessas representações.

No escopo desta tese, essas dimensões possuem responsabilidades distintas. O LLM atua no apoio à construção e à revisão da relação item-CC, enquanto o MDC realiza a inferência cognitiva a partir da Q-matriz validada e das respostas dos estudantes. Portanto, a automação proposta não elimina a necessidade de definição da relação item-habilidade nem automatiza

integralmente o processo de diagnóstico, mas auxilia uma etapa específica de construção da representação necessária ao funcionamento do MDC. Um ponto central na representação é a explicitação da relação entre itens avaliativos e habilidades. Essa relação é operacionalizada por meio da Q-matriz, uma estrutura binária que indica se um determinado item requer um conceito ou uma habilidade latente, também sendo considerado como CC. Descrita como uma matriz, a Q-matriz tem linhas que representam os itens e as colunas os atributos de habilidades. O valor 1 é usado para indicar que a habilidade é requerida pelo item, e o valor 0, para indicar que não é requerida.

Essa forma de representação é importante, pois permite que a avaliação não seja analisada apenas pela pontuação total do estudante. Ao associar cada item a um CC, é possível interpretar o desempenho como evidência de domínio ou não domínio de componentes específicos do conhecimento. Essa perspectiva está alinhada aos MDC, nos quais as respostas dos estudantes são utilizadas para estimar perfis de domínio de habilidades. Nas propostas de [Lopes e Netto \(2023\)](#) e [Villanueva et al. \(2018\)](#), os MDC são tratados como modelos estatísticos voltados à classificação diagnóstica, utilizados para identificar habilidades latentes do aprendiz e apoiar a modelagem do conhecimento em sistemas inteligentes.

Nesse contexto, a inferência ocorre quando o sistema utiliza as respostas dos estudantes e a relação item-CC para estimar o estado de conhecimento do aprendiz. Esta tese descreve que os MDC lidam com a presença ou ausência de conceitos ou habilidades latentes e consideram parâmetros de incerteza, como *slip* e *guessing*. O *slip* ocorre quando o estudante domina uma habilidade, mas erra o item. Já o *guessing* ocorre quando o estudante não domina a habilidade, mas acerta o item. Assim, a inferência diagnóstica não se limita a contabilizar acertos e erros, mas busca estimar quais habilidades foram mobilizadas corretamente, quais estão ausentes e quais respostas podem ter sido influenciadas por fatores ruidosos.

A automação surge como uma necessidade, pois a construção dessas estruturas é trabalhosa e depende de especialistas. A especificação das habilidades requeridas para resolver cada item, a criação da Q-matriz, a elaboração de distratores e a validação pedagógica dos itens demandam tempo e conhecimento técnico. Essa limitação também é apontada por [Nogueira et al. \(2025\)](#), que afirma que a elaboração de itens de avaliação em larga escala envolve etapas complexas, como a definição de habilidades, aferição, construção de itens, a diversificação de componentes e a adequação pedagógica. Dessa forma, a automação pode atuar como apoio à produção inicial dessas estruturas, sem eliminar a necessidade de validação por especialistas.

Nesse cenário, os LLM podem apoiar a automação da representação ao sugerir associações entre itens e CC, gerar justificativas para essas associações e propor anotações pedagógicas. [Lopes e Netto \(2026\)](#) investigam essa possibilidade ao avaliar o uso de LLM na geração automática de Q-matrizes, considerando a associação entre itens e os CC como uma representação interpretável da relação item-atributo. O estudo propõe avaliar essa geração em dois níveis: o alinhamento estrutural com Q-matrizes definidas por especialistas e o impacto funcional no desempenho preditivo de um MDC DINA, usando métricas como RMSE e MAE. Essa abordagem é relevante, pois não apenas avalia se as LLM marcam as habilidades de maneira semelhante à do especialista, como também se essa marcação mantém a utilidade diagnóstica do modelo.

Os resultados descritos por [Lopes e Netto \(2026\)](#) indicam que estratégias de *few-shot* podem melhorar o alinhamento estrutural em relação a configurações *zero-shot*, enquanto o uso de CoT produz refinamentos mais limitados. Apesar das variações estruturais entre as Q-matrizes geradas por diferentes estratégias de *prompt*, o desempenho preditivo do MDC permaneceu estável e comparável ao obtido com Q-matrizes definidas por especialistas. Esse resultado sugere que a automação da anotação item-CC pode ser usada como apoio à modelagem de avaliações diagnósticas, desde que seja acompanhada por critérios de avaliação estrutural e funcional.

A automação também pode contribuir para a revisão e validação preliminares dos itens gerados. [Quincozes et al. \(2025\)](#) mostram que a avaliação cruzada por LLM pode apresentar consistência razoável em relação à avaliação docente, embora os autores também indiquem que as técnicas de *prompt* são influenciadas pela arquitetura de cada modelo. [Santos et al. \(2025\)](#), por sua vez, reforçam que a análise pedagógica dos itens gerados por LLM ainda demanda revisão humana, mas apontam a necessidade de alternativas semiautomatizadas para tornar essa avaliação mais escalável. Dessa forma, a automação pode atuar em diferentes etapas do processo, como a geração do item, a anotação das habilidades, a produção de justificativas, a triagem de qualidade e o apoio à validação.

Assim, a integração entre representação, inferência e automação permite que os LLM sejam posicionados não apenas como ferramentas de geração textual, mas também como componentes de apoio à modelagem diagnóstica da aprendizagem. A representação organiza a relação entre itens, habilidades e evidências, enquanto a inferência utiliza essas relações para estimar o domínio dos estudantes. A automação, por sua vez, apoia a criação, a anotação e a revisão dessas estruturas. No contexto desta tese, essa articulação é relevante, pois

aproxima a geração automática de avaliações da lógica do MDC, mantendo a preocupação com a interpretabilidade, a validade pedagógica e a função do especialista na revisão das associações produzidas.

2.4.5 Limitações

Esta seção apresentou os fundamentos relacionados ao uso dos LLM no apoio à criação, anotação e automação de avaliações de aprendizagem. Inicialmente, foram apresentados os conceitos, como tokenização, janela de contexto, autoatenção, treinamento e inferência, que são considerados necessários para a adequada interpretação do potencial e das limitações.

Em seguida, diferentes famílias de LLM, como GPT, LLaMA, Gemini, Phi, Mixtral, Qwen e Granite, foram apresentadas, com destaque para suas características arquiteturais, formas de acesso, capacidades de processamento e possíveis aplicações educacionais. Essa apresentação permitiu situar os modelos como tecnologias com diferentes níveis de acesso, custo computacional, capacidade inferencial e adequação a tarefas educacionais baseadas em linguagem natural.

Na seção 2.4.3, são discutidas as aplicações dos LLM na geração automática de avaliações. A literatura analisada indica que esses modelos podem auxiliar na produção de questões, alternativas, gabaritos, justificativas, *feedbacks* e variações de itens avaliativos (Maity e Deroy, 2024; Varma et al., 2025; Shirkar, Langhi e Agrawal, 2025). Também ficou evidente que a qualidade dos itens gerados depende não apenas do modelo utilizado, mas também da forma como as instruções são elaboradas. Neste sentido, a Engenharia de *Prompt* é um fator relevante para orientar a geração de questões mais coerentes com os objetivos pedagógicos (Quincozes et al., 2025). Apesar desse potencial, os estudos revisados também mostram que a geração automática de itens ainda apresenta limitações, sobretudo no que se refere à validade pedagógica, clareza, adequação ao nível de ensino e à qualidade dos distratores (Nogueira et al., 2025; Chico et al., 2025).

Esta seção, também, destacou que a anotação das avaliações de aprendizagem é uma etapa importante para os sistemas educacionais inteligentes e para as avaliações diagnósticas. Ao se associarem a cada item as informações, como a habilidade avaliada, o objetivo de aprendizagem, o nível de dificuldade, gabarito, justificativa e os possíveis erros esperados, é possível interpretar as respostas dos estudantes como evidências de domínio ou não-domínio de habilidades. Nesse contexto, as LLM podem auxiliar a anotação inicial dos itens, sugerindo relações item-CC, mas não substituem a validação humana. Essa posição é reforçada por

estudos que indicam a necessidade de uma revisão pedagógica especializada para garantir a qualidade dos materiais gerados ([Santos et al., 2025](#); [Nogueira et al., 2025](#)).

Por fim, foi discutida a relação entre representação, inferência e automação. A representação organiza a relação entre itens, habilidades e conceitos de aprendizagem. A inferência usa essas relações para estimar o estado de conhecimento dos estudantes, enquanto a automação auxilia na criação, anotação e revisão dessas estruturas. Essa articulação é compatível com a lógica dos MDC, nos quais a Q-matriz representa a relação entre itens e atributos latentes, possibilitando a estimativa de perfis de domínio dos estudantes com base nas suas respostas.

Neste sentido, pesquisas recentes indicam que os LLM podem auxiliar a geração automática de Q-matrizes e a associação de item com os CC, desde que os resultados sejam avaliados tanto em termos estruturais quanto funcionais ([Lopes e Netto, 2026](#); [Moore et al., 2024](#); [Li et al., 2024a](#)). Com base nessas informações, o próximo capítulo apresenta os trabalhos relacionados que fundamentam e delimitam a proposta desta tese.

Capítulo

3

TRABALHOS RELACIONADOS

Este capítulo apresenta os trabalhos relacionados à avaliação e à modelagem do conhecimento do estudante, bem como ao uso para realizar o acompanhamento da aprendizagem. Os trabalhos selecionados buscaram representar, de diferentes formas, o conhecimento, as habilidades e as competências dos estudantes, além de apoiar os processos de avaliação e fornecer subsídios para personalização, recomendação e a tomada de decisão pedagógica.

A identificação e a seleção dos trabalhos foram conduzidas por meio de uma revisão narrativa da literatura, desenvolvida ao longo da evolução desta pesquisa. As revisões narrativas possibilitam a análise e a síntese de estudos relevantes sobre uma temática, favorecendo a discussão de questões conceituais e a integração de diferentes perspectivas teóricas e metodológicas (Baumeister e Leary, 1997; Rother, 2007). Diferentemente das revisões sistemáticas, esse tipo de revisão não pressupõe necessariamente uma busca exaustiva ou um protocolo rígido de seleção dos estudos, sendo sua abrangência orientada pelo objetivo e pelo contexto da análise (Grant e Booth, 2009; Snyder, 2019).

Inicialmente, foram delimitados os principais eixos relacionados ao problema investigado, abrangendo representação do conhecimento, habilidades e competências, modelagem do estudante, avaliação, diagnóstico e adaptação pedagógica. A identificação dos trabalhos partiu da literatura levantada nas etapas exploratórias da pesquisa, incluindo o mapeamento anterior sobre os STI (Lopes e Netto, 2020), e foi complementada por buscas realizadas conforme novos componentes foram sendo incorporados à investigação.

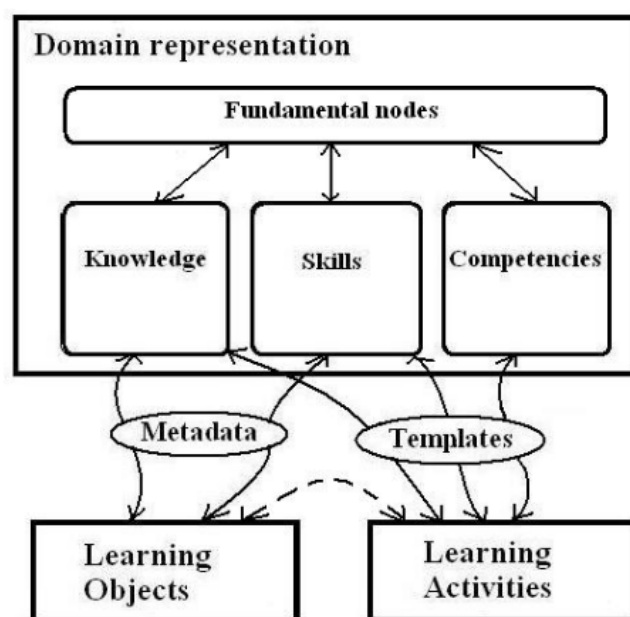
Os estudos apresentados foram selecionados por sua relevância conceitual e funcional para o problema da tese e analisados considerando três aspectos: (i) a forma de representação do domínio do conhecimento, (ii) o mecanismo utilizado para avaliar ou inferir o estado de conhecimento do estudante e (iii) o suporte oferecido ao professor ou ao sistema educacional. Por fim, os trabalhos foram tabulados e comparados de modo a evidenciar aproximações, diferenças e contribuições para o delineamento da arquitetura proposta.

3.1 Intelligent Web Teacher (IWT)

[Albano \(2011\)](#) apresenta um modelo para geração de experiências de aprendizagem personalizadas em ambientes de *e-learning*, com foco no ensino de Matemática. O trabalho parte da distinção entre conhecimento, habilidades e competências, considerando que a competência matemática envolve não apenas aspectos cognitivos, mas também elementos metacognitivos e não cognitivos. Para isso, o modelo utiliza a plataforma *Intelligent Web Teacher* (IWT), que integra características de LCMS (*Learning Content Management System*), sistemas adaptativos e a representação explícita do conhecimento por meio de ontologias.

A plataforma IWT é estruturada a partir de três modelos principais: o Modelo de Conhecimento, o Modelo do Aprendiz e o Modelo Didático. O primeiro representa o domínio por meio de ontologias, objetos de aprendizagem e metadados. O Modelo do Estudante registra informações sobre o estado cognitivo do estudante, suas preferências de aprendizagem e regras de evolução. Já o Modelo Didático, por sua vez, define as formas mais adequadas de transferência de conhecimento, considerando tanto o domínio representado quanto as características do estudante. A partir desses elementos, o sistema é capaz de gerar trilhas de aprendizagem personalizadas, considerando os objetivos estabelecidos pelo professor, o conhecimento prévio do estudante e os objetos de aprendizagem disponíveis. A Figura 12 mostra o modelo proposto por [Albano \(2011\)](#).

Figura 12 – Modelo de gerenciamento de experiência de aprendizagem no sistema IWT.



Fonte: [Albano \(2011\)](#).

A Figura 12 apresenta um modelo no qual a representação do domínio está organizada em três níveis: conhecimento, habilidades e competências. Conforme descrito em Albano (2011), no nível de conhecimento, são considerados elementos como definições, teoremas, exemplos e propriedades. No nível de habilidades, são representados métodos de cálculo, procedimentos e capacidades de resolução de problemas. Já no nível de competências, são abordadas as habilidades matemáticas mais amplas, como raciocínio, modelagem, comunicação, representação, uso de símbolos e ferramentas. Com base nessa organização, o sistema usa os objetivos definidos pelo professor e as informações do Modelo do Estudante para oferecer uma experiência de aprendizagem personalizada.

Essa organização é relevante para esta tese, pois evidencia a necessidade de se estruturar previamente o domínio educacional antes da personalização ou da avaliação da aprendizagem. No modelo mostrado na Figura 12, a estrutura serve para auxiliar na seleção dos objetos de aprendizagem e organizar os recursos visando à personalização. Além disso, as competências se estendem pelo domínio matemático, não estando vinculadas a um conhecimento ou habilidade isolada. Ainda assim, o desempenho em atividades relacionadas a competências depende do domínio prévio dos conhecimentos e habilidades envolvidos. Esse ponto é importante, pois ajuda a diferenciar a avaliação de uma competência ampla da avaliação de habilidades específicas.

Portanto, o trabalho de Albano (2011) se aproxima desta tese ao tratar da relação entre conhecimento, habilidades e competências em ambientes de *e-learning*. Sua contribuição está mais relacionada à organização conceitual do domínio e à geração de experiências personalizadas de aprendizagem. A proposta desta tese avança em outra direção, ao utilizar essa estruturação do domínio como base para a criação de avaliações diagnósticas, para o relacionamento entre itens e CC e visando a estimativa automática do nível de conhecimento do estudante.

3.2 Competences Memory Map – (MM-Competences)

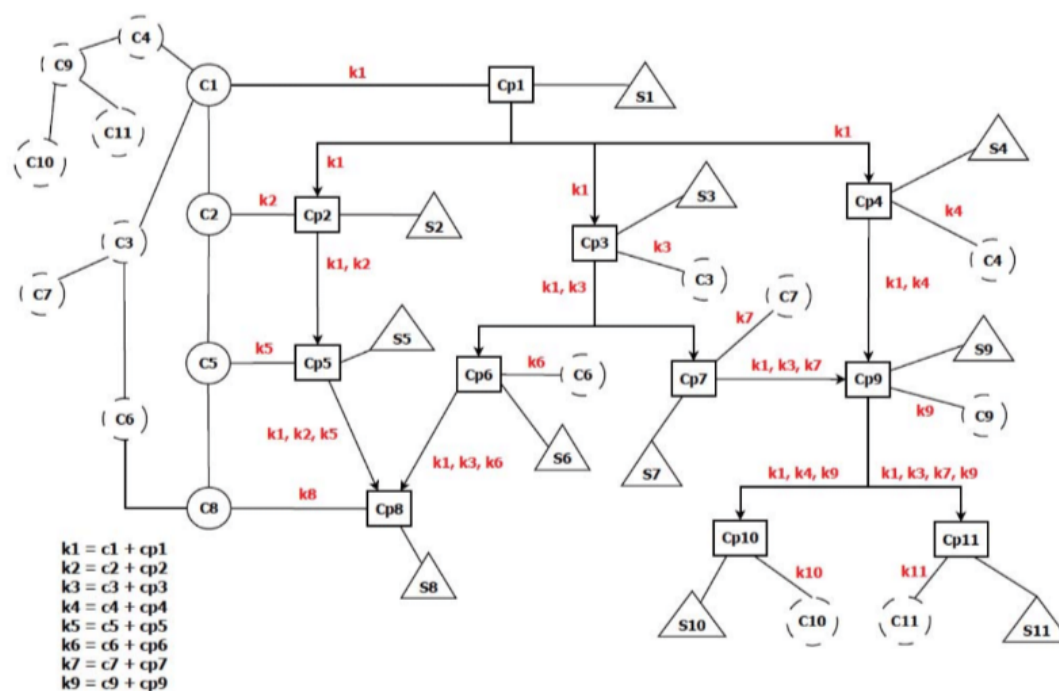
O trabalho de Ramirez e Sanchez (2012) é um modelo formal chamado de Mapa de Memória de Competências (do inglês, *Competences Memory Map*) (MM-Competences), que pode ser implementado computacionalmente. Segundo os autores, trata-se de um modelo que propõe a representação de competências capaz de caracterizar seus componentes de maneira integrada e simples.

O principal objetivo, de acordo com Ramirez e Sanchez (2012), o MM-Competences foi

projetado para modelar o estado mental de uma pessoa. Para isso, o modelo é composto por três componentes denominados, Competência (*Competences*), Habilidades (*Skills*) e Conceitos (*Concepts*). [Ramirez e Sanchez \(2012\)](#) descrevem que um *Concept* é modelado por uma unidade chamada de *Concepts-RU* e uma *Skill* é representada por uma unidade chamada *Skill-RU*. O MM-Competences engloba as duas unidades em uma única estrutura chamada de *Competences-RU*, capaz de descrever a relação comportamento e interação entre elas.

[Ramirez e Sanchez \(2012\)](#) descrevem o *Competences-RU* como um grafo direcionado, no qual os nós representam as competências e os arcos representam as associações entre elas. A direção dessas associações indicam o nível de generalização de uma competência específica na estrutura, ou seja, as competências de nível superior apontam para as de nível inferior; as competências de nível inferior são, na verdade, usadas como componentes de uma de nível superior. A competência de nível mais alto de uma estrutura específica determina o objetivo principal de toda a estrutura, e as competências mais baixas ou “atômicas” são semelhantes às habilidades primitivas genéricas. A Figura 13 ilustra um exemplo da aplicação do MM-Competences.

Figura 13 – Geração e agregação de domínios de competências dentro do MM-Competences.



Fonte: [Ramirez e Sanchez \(2012\)](#).

Na Figura 13 os *Concepts-RU* são representados pelos círculos, as *Skill-RU* são representadas pelos retângulos e *Competences-RU* são os triângulos. Os domínios, destacados

em vermelho, são denominados por k_i . [Ramirez e Sanchez \(2012\)](#) detalham que o esquema mostrado na Figura 13, é uma proposta de domínio de competências que se origina da relação entre as unidades *Competences-RU* e *Concepts-RU*, visando representar o contexto de um conjunto de competências interligadas, denominado **Estrutura de Competências**. O conceito de Estrutura de Competências é utilizado para associar um conjunto de *Competences-RU* para formar uma estrutura de grafo direcionado com uma competência raiz e uma sequência de execução.

[Ramirez e Sanchez \(2012\)](#) destacam que o modelo MM-Competences pode ser utilizado em aplicações educativas para a modelagem do conhecimento, das habilidades e das competências dos estudantes. A avaliação do modelo é finalizada quando não há mais associações de competências de níveis inferiores rotuladas com ele. Segundo os autores, essa propriedade de domínio possibilita definir, de maneira totalmente flexível, quantas subcompetências forem necessárias para a aprendizagem.

Quanto à tese, o trabalho de [Ramirez e Sanchez \(2012\)](#) é relevante por apresentar uma formalização computacional para conhecimento, habilidades e competências. Essa organização se aproxima da proposta desta pesquisa, uma vez que também considera a necessidade de estruturar habilidades e competências de modo que possam ser manipuladas por sistemas computacionais. Além disso, o modelo contribui para a compreensão de como habilidades podem ser associadas a conceitos e organizadas em estruturas hierárquicas ou agregativas. Esse aspecto é importante para a construção de avaliações orientadas para o domínio de habilidades.

No entanto, o MM-Competences possui um foco diferente da proposta desta tese. Embora se concentre na representação formal das competências e nas relações entre seus componentes, o modelo não apresenta um mecanismo para inferir o domínio de habilidades dos estudantes a partir de suas respostas nas atividades avaliativas. Nesse caso, a avaliação está mais relacionada à organização estrutural das competências do que à avaliação diagnóstica do conhecimento do aprendiz. Assim, o MM-Competences contribui como uma base conceitual e estrutural para a modelagem de competências em sistemas educacionais.

A presente tese avança ao propor um arcabouço que não apenas representa habilidades associadas a itens avaliativos, mas também utiliza o MDC para estimar as habilidades dominadas. Enquanto o MDC enfatiza a representação computacional de conceitos, habilidades e competências, esta pesquisa concentra-se na criação, aplicação e análise diagnóstica de avaliações orientadas por habilidades. O objetivo é apoiar os professores na identificação dos

pontos fortes e fracos dos estudantes e na tomada de decisões pedagógicas.

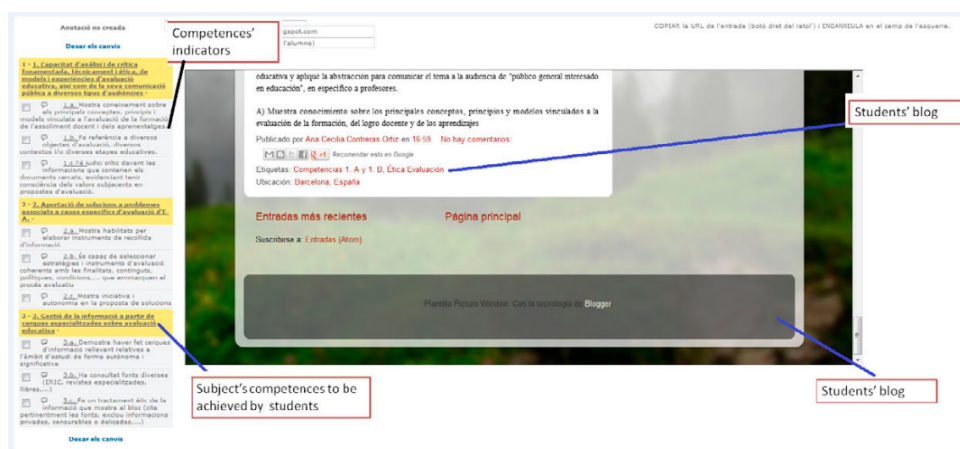
3.3 Competency Assessment Tool – (CAT)

Ion, Cano e Cabrera (2016) desenvolveram a plataforma *Competency Assessment Tool* (CAT), para apoiar experiências de avaliação baseada em competências no ensino superior. O trabalho parte do princípio de que a avaliação por competências deve considerar não apenas o conhecimento teórico, mas também habilidades, atitudes e a forma como esses recursos recursos são mobilizados em situações contextualizadas. Nesse sentido, a competência é compreendida como a capacidade de articular diferentes habilidades cognitivas e sociais para solucionar problemas ou tomar decisões em um contexto específico.

A proposta do CAT está associada ao uso de *blogs* como instrumentos de acompanhamento da aprendizagem. Conforme demonstrado na Figura 14, a plataforma possibilita que os professores especifiquem quais competências serão avaliadas, as organizem em indicadores e acompanhem as postagens dos estudantes. Cada estudante ou grupo cria um *blog* e associa suas postagens aos indicadores definidos.

Por sua vez, o professor utiliza a plataforma para verificar se os indicadores foram alcançados ou não, além de fornecer comentários qualitativos sobre o desempenho dos estudantes. Dessa maneira, o CAT funciona como uma ferramenta de apoio aos retornos pedagógicos formativos, à autorregulação da aprendizagem e à reflexão dos estudantes sobre o seu próprio desenvolvimento.

Figura 14 – Avaliação de competências na plataforma CAT.

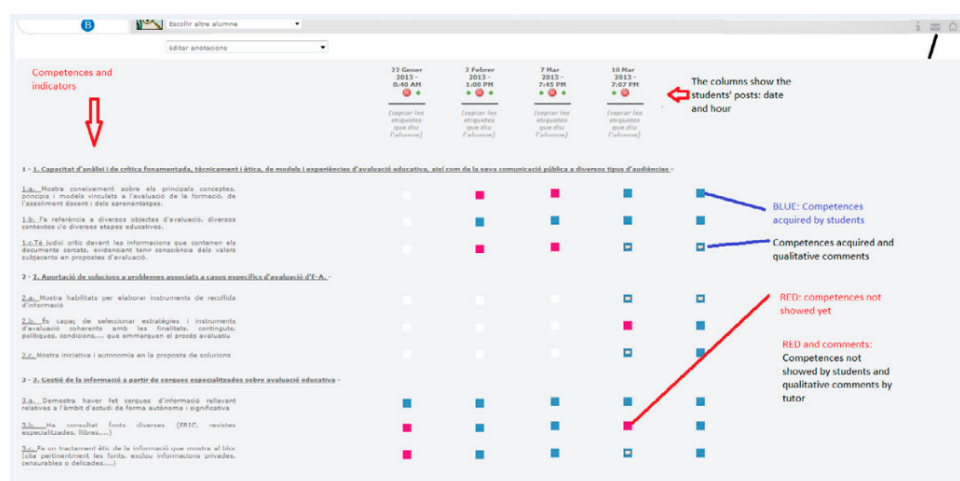


Fonte: Ion, Cano e Cabrera (2016)

Segundo Ion, Cano e Cabrera (2016), para iniciar a avaliação da aquisição de competências, o professor deve seguir quatro etapas: 1) identificar, por meio da plataforma CAT, as

competências a serem avaliadas; 2) dividir cada competência em uma série de indicadores. Em seguida são definidas as atividades específicas a serem desenvolvidas pelos alunos em formato de *blog* com acesso livre e aberto. Então, na etapa três, os alunos ingressam na plataforma e criam seus *blogs*, marcando cada entrada conforme os indicadores atribuídos. Por fim, na quarta etapa (4), cada professor apresenta suas considerações sobre o trabalho realizado. Isso possibilita acessar os indicadores de competência com base no que os alunos postaram e avaliar se foram alcançados (em azul) ou não (em vermelho). Além disso, é possível adicionar comentários, como mostra a Figura 15.

Figura 15 – Competências alcançadas pelos alunos e validadas pelos professores, na Plataforma CAT.



Fonte: Ion, Cano e Cabrera (2016)

Em relação à tese, o trabalho de Ion, Cano e Cabrera (2016) é relevante por apresentar uma ferramenta computacional voltada à avaliação de competências e ao acompanhamento do desenvolvimento dos estudantes. Assim como nessa pesquisa, o CAT reconhece a importância de explicitar competências e indicadores para orientar o processo avaliativo. Além disso, a proposta valoriza as devolutivas como elemento central para apoiar a aprendizagem, principalmente quando se considera a necessidade de fornecer informações diagnósticas para professores e estudantes.

Contudo, o CAT difere da proposta desta tese em alguns aspectos. A plataforma concentra-se na avaliação de competências por meio de *blogs*, indicadores e validação docente, sendo fortemente dependente da análise qualitativa e do acompanhamento realizado pelo professor. Além disso, a identificação das competências alcançadas pelos estudantes ocorre por meio da marcação manual de indicadores e da interpretação docente.

Dessa forma, o trabalho de [Ion, Cano e Cabrera \(2016\)](#) contribui para esta tese ao destacar a importância de ferramentas que apoiem a avaliação baseada em competências, as devolutivas formativas e a autorregulação da aprendizagem. Entretanto, esta tese avança em outra direção ao incorporar o MDC para estimar o domínio de habilidades a partir da relação entre estudante, item e habilidade. Enquanto o CAT enfatiza a avaliação reflexiva e qualitativa de competências em atividades mediadas por *blogs*, esta tese busca estruturar um arcabouço capaz de apoiar a criação de avaliações orientadas por habilidades e de produzir diagnósticos automatizados sobre o conhecimento dos estudantes.

3.4 ANNETTE Project

A plataforma ANNETTE descrita em [Anzieu e Buhai \(2017\)](#) está vinculada à necessidade de organizar perfis profissionais, resultados de aprendizagem, qualificações, certificações e processos de reconhecimento de competências entre diferentes instituições e empregadores. A proposta é baseada no uso de um modelo de competências focado na mobilidade profissional e na padronização de perfis no contexto europeu, considerando referências como o Sistema Europeu de Créditos do Ensino e Formação Profissionais (ECVET) e o Quadro Europeu de Qualificações (EQF)¹.

De acordo com [Anzieu e Buhai \(2017\)](#) a ferramenta especificada no projeto busca atender diferentes usuários, como empregados, empregadores, avaliadores, gestores e organizações de treinamento. Para os empregados, o sistema permite o registro de competências, qualificações, evidências e trajetórias profissionais. Para os empregadores, possibilita associar competências a perfis de trabalho, verificar lacunas entre o perfil do trabalhador e o perfil esperado para uma função, além de apoiar processos de avaliação e certificação. Para as instituições de treinamento, a ferramenta auxilia na identificação dos blocos de competências e resultados de aprendizagem necessários para a composição de programas formativos alinhados às demandas profissionais.

A plataforma ANNETTE, tem como dados de entrada, 1) características básicas do funcionário, 2) competências avaliadas por um colaborador; 3) certificações dos funcionários e; 4) perfis de trabalho definidos conforme a taxonomia de trabalho com áreas nucleares. Cabe destacar, que este último, considera o perfil de competências, que são: *Knowledge* (Competência Cognitiva); *Skills* (Competência técnica) e *Competence* (Atitude, Comportamento e Competência Pessoal) (KSC). A Figura 16, mostra um exemplo de dados de entrada para

¹ <https://europass.europa.eu/en/european-qualifications-framework-eqf>

avaliação de habilidades a partir da plataforma ANNETTE.

Figura 16 – Ficha técnica de um aluno no módulo de avaliação do ANNETTE Project.

DATASHEET							
PERSONAL INFORMATION	CV	JOB PROFILE/KSCs	TRAINING RECORDS	CAREER PATH	EVALUATIONS	EVIDENCES	
JOB REQUIREMENTS							
KNOWLEDGE (COGNITIVE COMPETENCE)	Theoretical EQF level (1-8)	Current EQF level (1-8)	Host EQF level	External Certifications	Evidences	Comments	
Electrical installations	4	6			✓ Attachments		
Electrical instrumentation	4	5			✓		
Electrical equipment	4	4			✓		

■ Mandatory
■ Optional

Fonte: Anzieu e Buhai (2017)

De modo geral, segundo Anzieu e Buhai (2017), o *software* permite que os usuários listem e verifiquem quais blocos oficiais de competências eles possuem e quais precisam adicionar para atender ao cargo anunciado. Além disso, o candidato entenderá qual tipo de aprendizado ou treinamento, formal ou não, é necessário para adquirir as competências adicionais identificadas por meio da análise de lacunas. Em resumo, a comparação dos KSC do candidato com os previstos para a vaga deve ser facilitada, conforme mostra a Figura 22.

Figura 17 – Exemplo de passaporte de competências com avaliação concluída no ANNETTE Project.

Skills Passport

Competence Records Training Records Organisation History Evidence Book Training Status [Add a new Record](#)

All Order by date completed Descending Refresh

	Electrical & Electronic Engineering...	24/01/2017	Source: n/a	View Edit Archive	Manage supporting files
	Company Induction to include Health...	08/01/2013	Source: n/a	View Edit Archive	Manage supporting files
	Triple Bar Existing Sites	04/01/2012	Source: n/a	View Edit Archive	Manage supporting files

Tom Black
helen.greaves@nsan.co.uk

Employer: Areva
Job role: Annette Electrical Supervisor
Last verified assessment: 24/01/2017 10:57:43

Skills Passport

Competence Records Training Records Organisation History Evidence Book Training Status

Job Role Courses

Course requirements that have been created from assigned Job Roles.

Job Role	Course	Completed	Date
Annette Electrical Supervisor	Company Induction to include Health and Safety Awareness	No	
Annette Electrical Supervisor	Electrical & Electronic Engineering HND	No	
Annette Electrical Supervisor	Human Performance Practitioner - 8, Human Factors	No	

Fonte: Anzieu e Buhai (2017)

O ANNETTE Project, em relação à tese, é relevante por apresentar uma abordagem sistematizada para gestão de competências e habilidades em um ambiente computacional. Assim como nesta pesquisa, o projeto reconhece a importância de explicitar competências, habilidades e evidências de aprendizagem para apoiar processos de avaliação e de tomada de decisão. A proposta, também, mostra a necessidade de organizar perfis individuais, registrar evidências e permitir que diferentes usuários acompanhem o desenvolvimento ou a validação de competências ao longo do tempo.

No entanto, o ANNETTE Project difere em relação ao domínio de aplicação e ao tipo de avaliação realizada. O projeto está direcionado à gestão de competências profissionais no setor nuclear, com foco em certificação, mobilidade profissional, reconhecimento externo e adequação aos perfis de trabalho. A avaliação ocorre principalmente por meio de registros, evidências, autoavaliação, validação por gestores e análise de aderência aos perfis profissionais. Dessa maneira, o sistema não tem o objetivo de modelar atividades educacionais ou inferir automaticamente o domínio de habilidades cognitivas a partir das respostas dos estudantes nos itens avaliativos. Sendo assim, enquanto o ANNETTE Project enfatiza a gestão de competências para certificação e mobilidade profissional, esta tese busca apoiar professores na criação de avaliações orientadas por habilidades e na identificação diagnóstica dos pontos fortes e fracos dos estudantes.

O ANNETTE Project contribui para essa pesquisa ao demonstrar a importância de sistemas computacionais direcionados à gestão estruturada de competências, perfis individuais e evidências de desempenho. A presente proposta, entretanto, avança no contexto educacional ao incorporar modelos psicométricos capazes de estimar o domínio de habilidades latentes. Isso permite uma análise mais detalhada da aprendizagem dos estudantes e oferece subsídios para intervenções pedagógicas baseadas em dados.

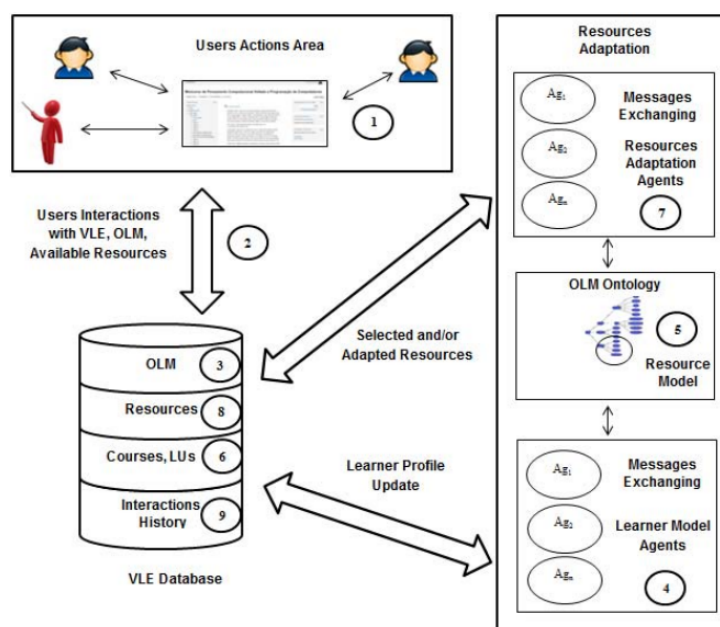
3.5 Conceptual Framework of Educational Resources Adaptation in Virtual Learning Environments – (ArCARE)

Bremgartner, Netto e Menezes (2017) apresentam o ArCARE (*Conceptual Framework of Educational Resources Adaptation in Virtual Learning Environments*) um *framework* conceitual voltado à adaptação e recomendação de recursos educacionais em Ambientes Virtuais de Aprendizagem. A proposta busca superar a apresentação uniforme de recursos aos estudantes, considerando características individuais para apoiar a personalização do processo de

aprendizagem. Para isso, o ArCARE utiliza uma arquitetura baseada em agentes de *software* e uma ontologia para representação de um *Open Learner Model*.

Conforme mostra a Figura 18, a arquitetura é organizada em três componentes principais: o módulo de adaptação de recursos, a base de dados do ambiente educacional e o espaço de interação dos usuários. O módulo de adaptação concentra a camada de inteligência do *framework* e é composto por agentes responsáveis pelo gerenciamento do Modelo do Estudante e por agentes responsáveis pela adaptação dos recursos educacionais. A base de dados armazena informações sobre cursos, recursos, interações e características dos estudantes, enquanto o espaço de interação representa as ações realizadas pelos usuários no ambiente.

Figura 18 – Arquitetura ArCARE.

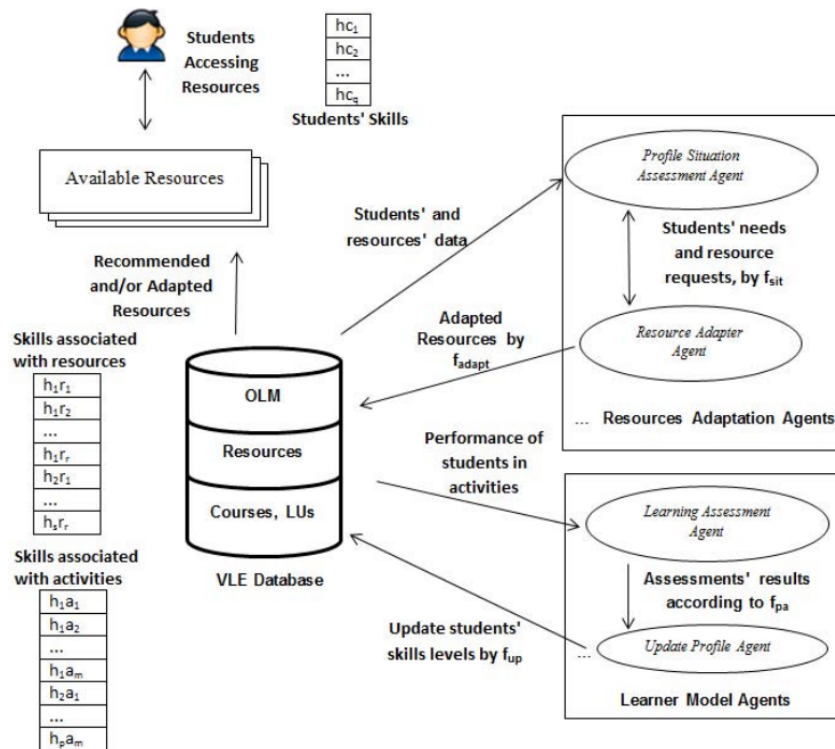


Fonte: [Bremgartner, Netto e Menezes \(2017\)](#).

Na arquitetura ArCARE, o Modelo do Estudante é atualizado continuamente a partir das interações realizadas no ambiente. A Figura 19 apresenta o ciclo de atualização das habilidades dos estudantes.

Entre as informações consideradas estão as habilidades, interesses, estilos de aprendizagem, frequência e histórico de desempenho nas atividades. O arcabouço possui um *Learning Assessment Agent*, responsável por avaliar as atividades realizadas pelos estudantes e identificar alterações nos níveis de habilidades, e um *Update Profile Agent*, que utiliza essas informações para atualizar o Modelo do Estudante. Posteriormente, o *Profile Situation*

Figura 19 – Atualização das habilidades do aluno no ArCARE.



Fonte: Bremgartner, Netto e Menezes (2017).

Assessment Agent analisa a situação do estudante e identifica os recursos mais adequados, enquanto o *Resource Adapter Agent* realiza sua recomendação ou adaptação no ambiente.

Conforme ilustrado na Figura 19, um dos aspectos relacionados à esta tese está na associação entre atividades e habilidades. No ArCARE, uma atividade pode envolver diferentes habilidades (Componentes de Conhecimento), e o processo de avaliação busca identificar o desempenho do estudante em cada uma delas. Dessa maneira, as informações obtidas durante a realização das atividades são utilizadas para atualizar os níveis de habilidades registrados no Modelo do Estudante e, posteriormente, orientar a seleção de recursos educacionais. O trabalho também associa habilidades aos próprios recursos, permitindo que a recomendação considere a proximidade entre as necessidades identificadas no estudante e as características dos recursos disponíveis.

Além disso, o ArCARE apresenta aproximações importantes por considerar habilidades como elementos do Modelo do Estudante, relacioná-las às atividades educacionais e utilizar os resultados da avaliação para orientar a adaptação pedagógica. Assim como na arquitetura proposta nesta pesquisa, existe um ciclo no qual as interações do estudante produzem evidências sobre seu estado de aprendizagem; essas informações atualizam seu modelo e, posteriormente, são utilizadas para apoiar decisões relacionadas às atividades ou aos recursos

educacionais.

Entretanto, as propostas diferem quanto à forma de representação e inferência das habilidades. No ArCARE, os níveis de habilidades são atualizados a partir do desempenho nas atividades e de um conjunto de informações do estudante, sendo utilizados juntamente com outras características, como estilos de aprendizagem e interesses, para realizar a adaptação dos recursos. Na arquitetura proposta nesta tese, a relação entre os itens de avaliação e os CC é explicitamente organizada por meio da Q-matriz, e as respostas dos estudantes são utilizadas por MDC para estimar os perfis de domínio das habilidades. Além disso, a arquitetura desta tese contempla o apoio de LLM à geração e revisão das relações item-CC, processo que não faz parte do escopo do ArCARE.

Sendo assim, o ArCARE contribui para esta tese ao demonstrar a importância da integração entre Modelo do Estudante, avaliação de habilidades e adaptação de recursos em ambientes educacionais. O trabalho reforça a necessidade de utilizar as informações produzidas durante as atividades dos estudantes para atualizar seus modelos e orientar intervenções pedagógicas. A presente tese avança em uma direção complementar ao estruturar essa relação a partir da representação item-CC, da Q-matriz e da inferência por MDC, de maneira integrada para apoiar a avaliação estruturada de habilidades.

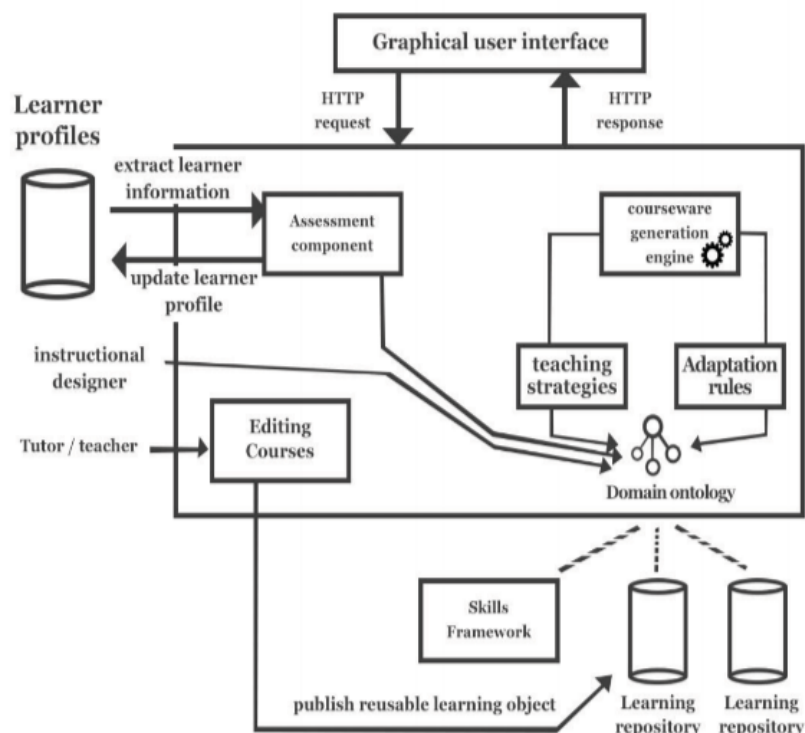
3.6 Competency Based Learning Path Generation System – (CALEP)

Segundo [Hnida, Idrissi e Bennani \(2018\)](#) *Competency based Adaptive Learning Path generation system* (CALEP) é um sistema inteligente para *e-learning* voltado à geração automática de trilhas de aprendizagem personalizadas com base no nível atual de competência do estudante. A proposta parte do problema de que os ambientes educacionais *online* precisam considerar as diferenças individuais dos estudantes, como conhecimentos prévios, ritmo de aprendizagem e nível de competência, para reduzir dificuldades como sobrecarga cognitiva, baixa motivação e evasão dos cursos.

A arquitetura do CALEP é organizada como uma aplicação Web, baseada no modelo cliente-servidor. Seus principais componentes são: a ontologia de domínio, o componente de avaliação, o mecanismo de geração de material didático e o componente de *feedback*. A Figura 20 resume a arquitetura do CALEP System.

A ontologia de domínio funciona como a base de conhecimento do sistema, armaze-

Figura 20 – Arquitetura geral do CALEP System.



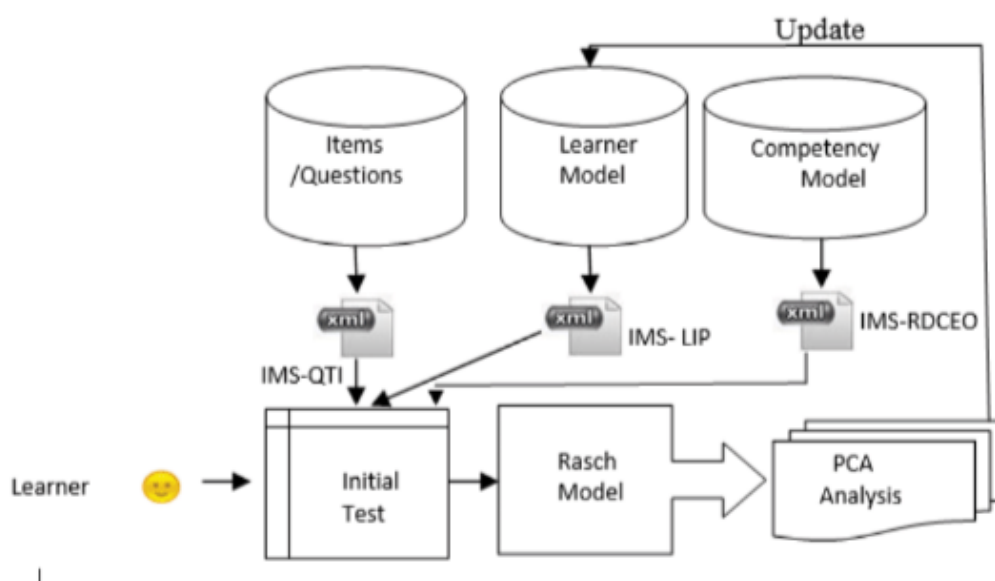
Fonte: [Hnida, Idrissi e Bennani \(2018\)](#).

nando descrições de competências, conceitos do curso, objetos de aprendizagem, estratégias de ensino e regras de adaptação. Essa estrutura possibilita a organização dos conteúdos e a seleção de recursos de aprendizagem de acordo com os objetivos educacionais e com as características do estudante. O professor usa a ferramenta para publicar recursos reutilizáveis de aprendizagem em relacionados à abordagem baseada em competências. O mecanismo de geração de material usa ontologia de domínio para extrair os conceitos-chave e suas propriedades. Ela, também, extrai estratégias de ensino e regras de adaptação de modo a organizar esses conceitos em uma sequência de ensino-aprendizagem apropriada.

O diagnóstico do nível do aluno no início da sessão de aprendizagem é realizado pelo componente de avaliação. Seu objetivo é avaliar o desempenho do aluno em relação a um padrão de desempenho, identificar seus pontos fortes e fracos, e comparar seu nível atual com o esperado. A Figura 21 apresenta o esquema de funcionamento do componente de avaliação do CALEP System.

O componente de avaliação estima o nível de competência do estudante no início da sessão de aprendizagem. Para tanto, utiliza-se o modelo de *Rasch* que é uma aplicação do TRI. Com base nas respostas do estudante a um conjunto de questões, avalia-se seu desempenho. Com base nesse diagnóstico inicial, o perfil do aluno é atualizado e utilizado

Figura 21 – Avaliação baseada em competência utilizando TRI no CALEP System.



Fonte: [Hnida, Idrissi e Bennani \(2018\)](#)

pelo mecanismo de geração de trilhas de aprendizagem. Esse mecanismo seleciona e organiza os conceitos em uma sequência personalizada, por meio de uma variação do algoritmo *Harmony Search*, considerando restrições como a ordem crescente de complexidade, os conceitos já dominados, as relações entre os conceitos, o tempo total disponível e o caminho crítico de aprendizagem.

Em relação à tese, o trabalho de [Hnida, Idrissi e Bennani \(2018\)](#) é relevante por propor uma arquitetura modular para personalização da aprendizagem com base em competências. Assim como nesta pesquisa, o CALEP reconhece que a competência é uma característica do estudante que deve ser considerada nos sistemas educacionais. Além disso, o sistema utiliza um componente avaliativo para identificar o nível de cada aluno e orientar decisões posteriores. Essas características se aproximam da proposta desta tese, que é utilizar informações diagnósticas para apoiar a tomada de decisões pedagógicas.

No entanto, o CALEP difere desta tese em termos de foco e modelo avaliativo utilizado. Enquanto a principal finalidade do CALEP é gerar trilhas de aprendizagem personalizadas, enquanto esta tese tem como foco a construção de um *framework* para criação, aplicação e análise de avaliações orientadas por habilidades. Ademais, o CALEP utiliza a TRI para estimar o nível de competência do estudante, ao passo que a presente tese utiliza o MDC, que permite identificar o domínio ou não domínio de habilidades específicas associadas aos itens avaliativos.

Essa diferença é relevante, pois a TRI tende a posicionar o estudante em uma escala de proficiência, ao passo que os MDC fornecem um diagnóstico mais detalhado dos atributos das habilidades latentes envolvidos no desempenho. Embora o CALEP possibilite adaptar a sequência de aprendizagem a partir do nível estimado do estudante, seu objetivo central não é identificar detalhadamente os pontos fortes e fracos em relação a cada habilidade associada aos itens.

Dessa maneira, o CALEP contribui para esta tese ao demonstrar a importância de arquiteturas adaptativas, ontologias de domínio e componentes de avaliação em sistemas educacionais inteligentes. Entretanto, a presente proposta avança ao concentrar-se na avaliação diagnóstica baseada na relação entre estudante, item e habilidade, utilizando MDC para estimar o domínio de habilidades cognitivas. Enquanto o CALEP direciona o diagnóstico para a geração automática de trilhas de aprendizagem, esta tese busca estruturar um modelo conceitual que auxilie professores na criação de avaliações orientadas por habilidades e na interpretação diagnóstica dos resultados dos estudantes.

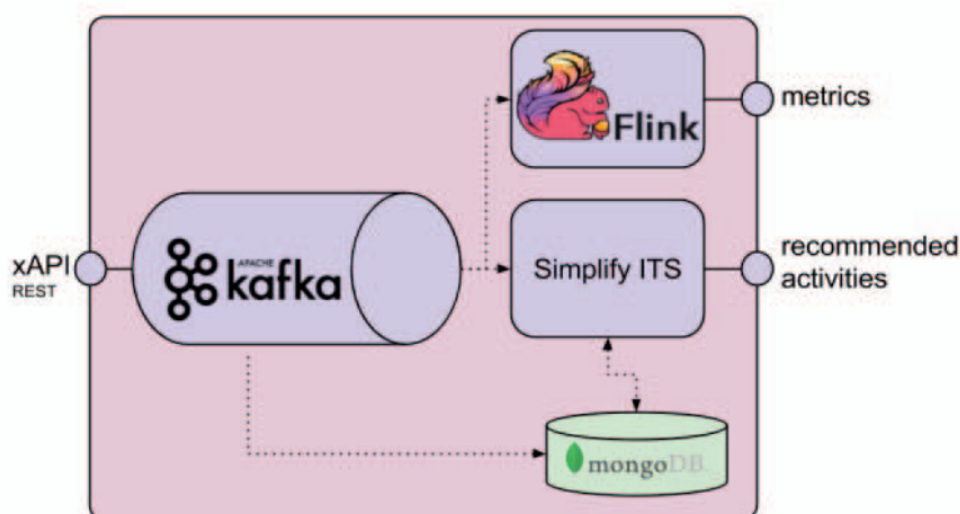
3.7 Simplify ITS

[Villanueva et al. \(2018\)](#) apresentam o Simplify ITS, um STI baseado em MDC e estratégias de aprendizagem espaçada. A proposta tem o objetivo de auxiliar os professores na adaptação das experiências de aprendizagem às necessidades dos estudantes. Para tanto, o sistema estima automaticamente o perfil de habilidades do aluno com base em suas interações com atividades educacionais, utilizando essas informações para recomendar novas atividades ou revisões de conteúdos já trabalhados.

O Simplify ITS combina modelagem psicométrica e recomendação adaptativa. Ele utiliza o MDC para identificar quais habilidades os estudantes dominam e quais não dominam, permitindo gerar um perfil de habilidades com base em suas respostas às atividades. No sistema, o MDC DINA é adotado por sua estrutura simples e interpretação direta. Isso porque que o estudante precisa dominar todas as habilidades requeridas por uma atividade para respondê-la corretamente (*compensatory*). Com base nessa estimativa, o sistema recomenda atividades relacionadas às habilidades ainda não dominadas e propõe revisões para reforçar as habilidades previamente já adquiridas.

A plataforma possui vários fluxos de trabalho, onde o professor gerencia as atividades. Inicialmente, o professor escolhe manualmente as primeiras atividades, que atendem às suas necessidades. Em seguida, a plataforma Simplify ITS decide qual deve ser a próxima atividade.

Figura 22 – Arquitetura do Simplify ITS.



Fonte: Villanueva et al. (2018).

Por fim, o aluno é apresentado a uma atividade que tentará solucionar. A pontuação será submetida pelo sistema ou pelo professor (Villanueva et al., 2018).

Outro aspecto importante da proposta é o uso das estratégias de *interleaving* e *spaced learning*. Além de recomendar atividades direcionadas às habilidades não dominadas, o Simplify ITS considera que habilidades já aprendidas podem ser esquecidas ao longo do tempo. Por isso, o sistema organiza revisões em intervalos progressivos, inspiradas em outros sistemas educacionais, buscando manter o desempenho do estudante nessas habilidades. Dessa maneira, a recomendação não se limita à aquisição de novas habilidades, mas também considera a retenção do conhecimento.

Do ponto de vista técnico, o Simplify ITS foi projetado considerando a interoperabilidade. O sistema utiliza a especificação *Experience API* (xAPI) para coletar e organizar registros das interações dos estudantes, como, por exemplo, a criação de atividades, a associação de metadados, a resolução de atividades e a pontuação. Esses registros permitem a construção do modelo de inferência utilizado pelo sistema. Além disso, sua arquitetura é baseada em serviços *RESTful* e utiliza tecnologias como Apache Kafka, Apache Flink e MongoDB, para garantir escalabilidade, disponibilidade e integração com diferentes plataformas educacionais.

O trabalho descrito por Villanueva et al. (2018) é um dos mais próximos da presente proposta, pois também utiliza MDC para estimar o perfil de habilidades dos estudantes. Ambos os trabalhos consideram a relação entre atividades, habilidades e o desempenho do aluno como base para produzir informações úteis ao processo de ensino e aprendizagem.

Ademais, o Simplify ITS destaca a relevância de oferecer dados aplicáveis ao professor, um aspecto igualmente presente nesta tese, que tem como objetivo apoiar a tomada de decisões pedagógicas com base em diagnósticos de domínio de habilidades.

No entanto, a proposta desta tese difere em alguns pontos. Enquanto o Simplify ITS está centrado na recomendação adaptativa de atividades e na revisão espaçada de habilidades, esta tese tem como foco principal a criação, aplicação e análise de avaliações orientadas por habilidades. Dessa forma, mesmo que a recomendação de atividades possa ser uma consequência futura, o cerne da presente proposta está na estruturação da avaliação diagnóstica, na associação entre itens e habilidades, bem como na interpretação dos resultados para professores e estudantes.

Outra diferença está no modelo de diagnóstico utilizado. O Simplify ITS utiliza o MDC DINA, que pressupõe uma relação não compensatória entre as habilidades. Ou seja, o estudante precisa dominar todas as habilidades exigidas por um item para ter maior probabilidade de sucesso. A presente tese, por sua vez, propõe utilizar e comparar diferentes MDCs, com destaque para o MDC G-DINA. Essa escolha amplia a análise da relação entre alunos, itens e habilidades, visando uma estrutura mais flexível para a avaliação diagnóstica em diferentes domínios.

Nesse contexto, o Simplify ITS contribui diretamente para esta pesquisa demonstrando a viabilidade de integrar MDC aos sistemas educacionais inteligentes e de utilizar o perfil de habilidades dos estudantes para apoiar decisões pedagógicas. NO entanto, vai além ao propor uma estrutura própria para autoria e a aplicação de avaliações diagnósticas baseadas em habilidades, com foco na adaptação a diferentes domínios do conhecimento e na análise do domínio de habilidades latentes. Enquanto, o Simplify ITS enfatiza a recomendação adaptativa e a retenção por meio da aprendizagem espaçada, esta tese concentra-se na construção de um ambiente para avaliação diagnóstica que auxilie o professor, desde a modelagem dos itens até a interpretação dos resultados.

3.8 Competency and Skill System – (CaSS)

[Robson \(2019\)](#) e colaboradores ([Robson et al., 2020](#)) apresentam o *Competency and Skill System* (CaSS) como uma infraestrutura aberta para apoiar a educação, o treinamento e o desenvolvimento profissional baseados em competências. A proposta surge da limitação das avaliações centradas apenas nos resultados da aprendizagem, pois esses resultados, geralmente, são definidos com base no que pode ser medido dentro de uma atividade educacional

específica. Segundo os autores, as competências desejadas muitas vezes extrapolam o contexto de uma única atividade, sendo necessário analisá-las a partir de diferentes evidências de desempenho, como experiências educacionais, atividades profissionais, registros de sistemas e observações externas.

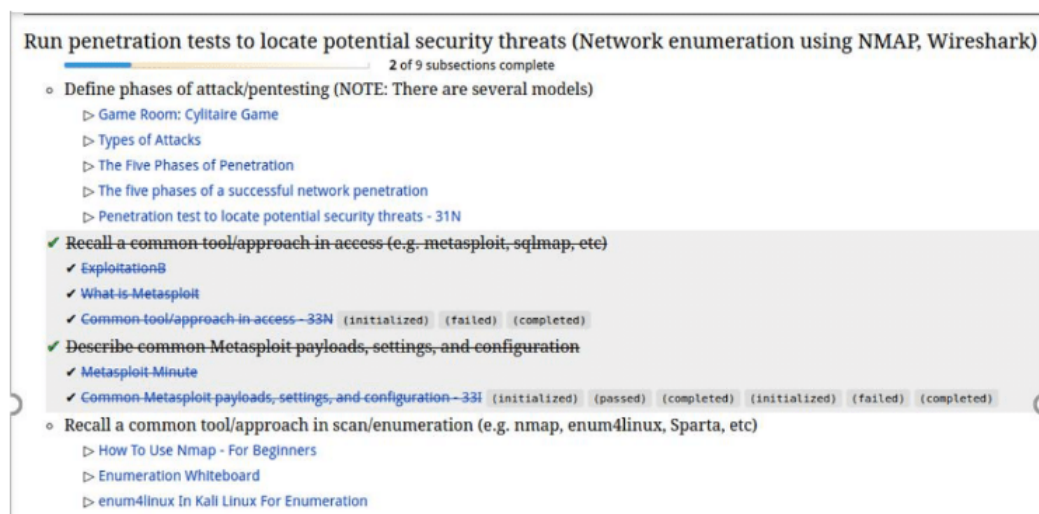
O CaSS foi desenvolvido para armazenar, organizar, compartilhar e relacionar competências em estruturas digitais denominadas *Competency Frameworks*. Essas estruturas podem representar conhecimentos, habilidades, atitudes, objetivos de aprendizagem, resultados de aprendizagem ou outros elementos relacionados ao que uma pessoa deve saber ou ser capaz de fazer. Cada competência pode ter um identificador único, descrição, nível, relações com outras competências, um contexto, rubricas, recursos associados e metadados. Essa organização permite que as competências sejam usadas por diferentes sistemas e compartilhadas em formatos legíveis por humanos e por máquinas.

O CaSS é implementado como um aplicativo *web* que expõe sua funcionalidade e seus dados a outros sistemas por meio de microserviços *RESTful* e dados vinculados. O CaSS é composto basicamente por um Kit de Desenvolvimento de *Software* (SDK) para Aprendizagem baseada em Competências, que visa economizar tempo e dinheiro dos usuários no desenvolvimento e na gestão de competências e *Frameworks* de competências. O repositório da estrutura do CaSS permite que as competências sejam definidas, organizadas em estruturas estruturadas associadas a um trabalho, tarefa, assunto ou um domínio de conhecimento, além de serem compartilhadas de acordo com os padrões do setor. O CaSS permite que todas as partes interessadas visualizem, compreendam e interpretem as competências (Robson et al., 2020).

Conforme mostrado na Figura 23, as ferramentas de edição do CaSS permitem que os usuários pesquisem estruturas de competência em seu repositório. Para compartilhar os dados, o CaSS pode coletá-los por meio de suas próprias API. No entanto, isso exige a adição de terminais da API proprietários que devem ser mantidos ao longo de seus ciclos de vida das fontes de dados e do sistema que os coleta. De acordo com Robson (2019), o CaSS faz isso com o *Alignment Object* do Schema.org², que especifica as competências ensinadas, avaliadas ou requeridas por uma atividade. Essa prática é suficiente quando a conclusão de uma atividade, ação ou avaliação é relatada, fornecendo evidências sobre a manutenção ou não de uma competência. No entanto, não é suficiente quando os dados relatados devem ser interpretados para a criação de uma medida significativa.

² <https://schema.org/>

Figura 23 – *Dashboard* mostrando as competências e as atividades associadas no CaSS Project.



Fonte: [Robson e Poltrack \(2017\)](#)

Um aspecto central do CaSS é o uso de *assertions*, isto é, declarações sobre o domínio ou não domínio de uma competência de um indivíduo, equipe ou organização. Elas podem ser produzidas por professores, sistemas tutores, dispositivos, avaliações, ambientes de simulação ou pelo próprio usuário. Cada *assertion* pode conter evidências, nível de confiança, data, agente responsável e informações sobre sua validade ao longo do tempo. Assim, o CaSS não considera uma evidência isolada como um verdade absoluta, mas como uma indicação que deve ser interpretada em conjunto com outras evidências disponíveis.

No entanto, há uma limitação, onde, as *assertion* podem entrar em conflito, se geradas em momentos diferentes, ou vir de uma variedade de fontes com confiabilidade variável e ser feitas com diferentes níveis de confiança. Geralmente, a confiabilidade e a validade de uma avaliação não são conhecidas e, mesmo no caso de avaliações realizadas por psicometristas profissionais, é ingênuo considerar os resultados da avaliação como prova de que um indivíduo possui uma competência específica.

Diante disso, as reivindicações de competência coletadas pelo CaSS podem não se basear em apenas uma avaliação, podendo ser derivadas de autodeclarações ou feitas sob a suposição de que um aluno que concluiu uma atividade que está relacionada as habilidades e os conhecimentos abordados. Assim, a veracidade dessas *assertions* pode estar sujeita à interpretações, não devendo ser totalmente confiável. [Robson \(2019\)](#), destaca a necessidade de criação de novos métodos para compreender isso e formar opiniões razoáveis e defensáveis sobre quais competências uma pessoa, equipe ou organização possui.

No contexto desta tese, o CaSS é relevante por oferecer uma base conceitual e tecnológica para a organização de competências, habilidades, evidências e perfis de aprendizagem em estruturas computacionais reutilizáveis. Assim como nesta pesquisa, o CaSS reconhece que a avaliação deve considerar a relação entre o desempenho do estudante e as competências ou habilidades associadas. Há também a aproximação com a proposta desta tese no que se refere ao uso de estruturas digitais para representar habilidades, conectar atividades avaliativas para essas habilidades e apoiar decisões educacionais baseadas em evidências.

Assim, os trabalhos sobre o CaSS contribuem para esta pesquisa ao demonstrarem a importância de representar competências de uma forma estruturada, interoperável e fundamentada em evidências. Eles também reforçam a necessidade de diferenciar competências, objetivos de aprendizagem e resultados observáveis. A presente tese avança ao aplicar essa lógica ao contexto da avaliação diagnóstica com MDC, buscando produzir informações mais detalhadas sobre o domínio de habilidades.

3.9 Stat-Knowlab

[Chiusole et al. \(2020\)](#) apresentam o Stat-Knowlab, um STI desenvolvido para apoiar a aprendizagem de estatística básica no ensino superior. O sistema é fundamentado na Competence-based Knowledge Space Theory (CbKST). A CbKST é uma extensão da *Knowledge Space Theory* voltada à avaliação formativa de conhecimentos e competências. A proposta parte da ideia de que a aprendizagem pode ser representada por estados de competência, isto é, um conjunto de habilidades dominadas por um estudante em determinado momento.

A arquitetura do Stat-Knowlab é composta por dois módulos principais: o módulo de avaliação e o módulo de aprendizagem. Esses módulos interagem continuamente, compartilhando informações sobre o estado atual de conhecimento do estudante. O Módulo de Avaliação realiza uma avaliação adaptativa do estado de competência do estudante, buscando identificar quais habilidades ele domina e quais não domina. Ao término da avaliação, o sistema gera um relatório formativo contendo informações sobre o que o estudante já sabe, o que não sabe e o que ele está pronto para aprender. A Figura 24 mostra um relatório individual gerado para o aluno.

O Módulo de Aprendizagem usa o resultado da última avaliação para personalizar a trajetória do estudante. Com base no estado de competência identificado, o sistema oferece objetos de aprendizagem relacionados às habilidades já dominadas e àquelas que o estudante está pronto para aprender. Essas últimas correspondem ao conceito de *outer fringe*, que

Figura 24 – Exemplo de relatório gerado para um aluno no sistema Stat-Knowlab.

TOPIC: Chi-square test

In overall, your current knowledge about this topic cover the 14%. So far, your knowledge is very poor.

 **WHAT I ALREADY KNOW**

- Computing the degrees of freedom in a contingency table

 **WHAT I DO NOT KNOW YET**

- Computing the Chi-square for a contingency table $r \times c$
- Taking the statistical decision by using the the Chi-square hypothesis testing
- Estimating the expected frequency in a contingency table
- Reading the Chi-square distribution tables
- Computing the margional frequencies in a contingency table
- Conducting the hypothesis testing by using the Chi-square test

 **WHAT I AM READY TO LEARN**

- Taking the statistical decision by using the the Chi-square hypothesis testing
- Reading the Chi-square distribution tables
- Computing the margional frequencies in a contingency table

Fonte: [Chiusole et al. \(2020\)](#).

representa o conjunto de habilidades que podem ser aprendidas, pois seus pré-requisitos já foram dominados. Dessa forma, o sistema evita apresentar conteúdos que ainda são difíceis, organizando a aprendizagem de acordo com a condição atual do estudante.

O trabalho também descreve a relação entre itens e habilidades por meio de um *skill map*, que é uma estrutura que associa cada item avaliativo ao conjunto de habilidades necessárias para resolvê-lo. Essa ideia se aproxima da Q-matriz utilizada nos MDC, pois ambas estabelecem uma relação entre itens e atributos. No Stat-Knowlab, essa relação é utilizada para a construção de estruturas de conhecimento e competência, possibilitando que o sistema estime o estado do estudante e oriente seu percurso de aprendizagem.

O sistema foi aplicado ao domínio da estatística básica, abrangendo conteúdos como estatística descritiva, combinatória e probabilidade, *t*-teste Student, *Z*-teste e teste qui-quadrado. No total, o Stat-Knowlab avalia 55 habilidades por meio de 86 itens. Os autores também apresentam estudos empíricos realizados com estudantes universitários, nos quais são analisados os dados de navegação, frequência de uso, proporção de habilidades dominadas e transições entre estados de competência. Os resultados obtidos sugerem que o sistema pode ser útil para o monitoramento de processos de aprendizagem e o auxílio à

personalização do ensino.

Em relação à tese, o Stat-Knowlab se destaca por propor uma arquitetura educacional que integra avaliação de habilidades, diagnóstico do desempenho do estudante e personalização da aprendizagem. Assim como nesta pesquisa, o trabalho considera que os instrumentos avaliativos devem estar associados às habilidades e o desempenho dos estudantes pode ser utilizado para inferir quais habilidades foram dominadas. Além disso, a proposta reforça a importância de oferecer informações formativas sobre o estado da aprendizagem, aspecto também presente nesta tese.

No entanto, o sistema se diferencia da presente proposta em relação ao modelo teórico e psicométrico utilizado. Enquanto o sistema é baseado na CbKST, esta tese propõe uma estrutura baseada em MDC. Na CbKST, o foco está na organização dos estados de competência e nas relações de pré-requisito entre habilidades, o que permite identificar o que o estudante domina e o que ele está pronto para aprender. Nos MDC, por sua vez, a inferência ocorre a partir dos padrões de resposta dos estudantes aos itens, considerando a matriz de habilidades e estimando perfis latentes de domínio.

Outra diferença está no escopo da proposta. O Stat-Knowlab foi desenvolvido para o domínio de estatística básica, com estruturas de conhecimento e objetos de aprendizagem específicos desse conteúdo. Esta tese, por sua vez, busca propor um *framework* mais geral para a criação, aplicação e análise de avaliações orientadas por habilidades, com possibilidade de adaptação a diferentes domínios de conhecimento.

Por fim, o Stat-Knowlab contribui para esta tese ao demonstrar uma arquitetura educacional que utiliza habilidades, itens e estados de competência para apoiar a avaliação formativa e a personalização da aprendizagem. O trabalho também destaca a importância de representar explicitamente as relações entre habilidades e itens e de usar os resultados da avaliação para orientar decisões pedagógicas. Entretanto, a presente proposta avança em uma direção distinta ao adotar o MDC como base para estimar o domínio de habilidades latentes e estruturar um *framework* voltado à autoria, à aplicação e à análise diagnóstica de avaliações orientadas por habilidades.

3.10 Diag-Skills

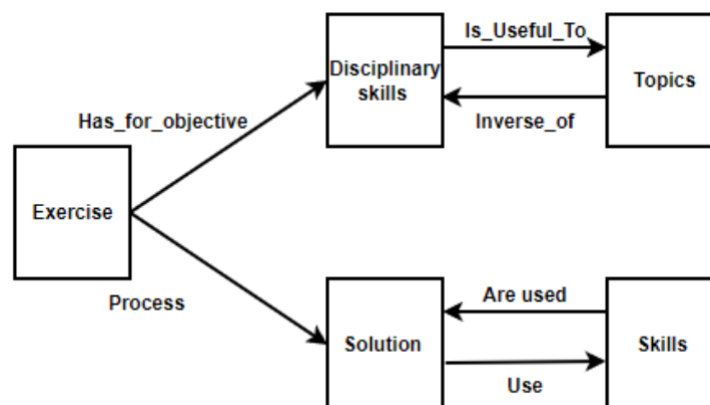
[Rahmouni et al. \(2021\)](#) apresentam o Diag-Skills, um sistema de diagnóstico de habilidades desenvolvido para STI. O trabalho parte do problema de que diagnosticar o estado de conhecimento do estudante é uma tarefa marcada por incerteza, imprecisão, contradição e

ausência de informação. De acordo com os autores, um estudante pode acertar uma questão por acaso ou errar mesmo que domine a habilidade avaliada. Por isso, o sistema busca lidar com as imperfeições dos dados de avaliação antes de definir se uma habilidade foi adquirida.

A proposta combina um modelo semântico de domínio, representado por uma ontologia, com a teoria das funções de crença, também conhecida como teoria de *Dempster-Shafer*. O modelo de domínio organiza tópicos, habilidades e recursos de aprendizagem e representa relações de pré-requisito entre habilidades.

Conforme ilustrado na Figura 25, a ontologia *Algoskill* estrutura os exercícios, as soluções, os tópicos e as habilidades, indicando como determinados tópicos são úteis para o desenvolvimento de habilidades disciplinares e como as soluções mobilizam habilidades específicas. Essa organização permite que o sistema diagnostique não apenas as habilidades diretamente avaliadas, como também estime o estado das habilidades não avaliadas com base nas conexões existentes entre elas.

Figura 25 – Ontologia *AlgosSkill* do Diag-Skill.



Fonte: [Rahmouni et al. \(2021\)](#)

O processo de diagnóstico do Diag-Skills é composto por cinco etapas. Inicialmente, as notas obtidas nas avaliações são transformadas em massas de crença, que estão associadas aos estados de aquisição, não aquisição, ignorância e conflito. Em seguida, essas crenças são revisadas à medida que novas evidências são recebidas. Em seguida, as informações são propagadas entre habilidades relacionadas, considerando tanto os pré-requisitos quanto as habilidades dependentes. As crenças propagadas são, então, combinadas com as crenças locais, o que permite identificar possíveis inconsistências. Por fim, as distribuições de crença são traduzidas em estados linguísticos, tais como: adquirido, não adquirido, provavelmente adquirido, provavelmente não adquirido, indeterminado ou conflitante.

O sistema foi implementado em Python e avaliado em dois estudos. O primeiro estudo foi realizado em uma disciplina de programação e algoritmos, com foco em habilidades relacionadas a estruturas iterativas, tabelas e busca de elementos. O segundo estudo foi realizado em uma disciplina de usabilidade e acessibilidade de interfaces e envolveu habilidades de prototipagem com o Adobe XD³. Os resultados indicaram boa capacidade preditiva do modelo, com acurácia de 84% no primeiro estudo e 92% no segundo. Além disso, quando não havia avaliação efetiva disponível para algumas habilidades, a comparação com o julgamento do professor indicou alto nível de concordância.

O trabalho de [Rahmouni et al. \(2021\)](#) é relevante por abordar diretamente o diagnóstico de habilidades em STI. Assim como nesta pesquisa, o Diag-Skills parte do princípio de que a aprendizagem pode ser representada por habilidades associadas a atividades avaliativas, e de que o diagnóstico dessas habilidades pode contribuir para a personalização do ensino. Há também aproximação com a proposta desta tese no que se refere ao uso de uma estrutura de domínio para relacionar habilidades e apoiar a interpretação dos resultados dos estudantes. Contudo, o Diag-Skills difere desta tese quanto ao método de diagnóstico utilizado.

De acordo com [Rahmouni et al. \(2021\)](#), o Diag-Skills utiliza funções de crença para representar e combinar evidências incertas, possibilitando o enfrentamento de conflitos, ignorância e contradições entre informações. A presente tese, por sua vez, propõe o uso do MDC para estimar o domínio de habilidades latentes a partir da relação entre estudante, item e habilidade. Nesse caso, o foco não está apenas em propagar crenças entre habilidades relacionadas, mas, também, em modelar estatisticamente os padrões de resposta dos estudantes, a fim de identificar quais habilidades foram ou não dominadas.

Outra diferença está no papel da avaliação. No Diag-Skills, as notas das avaliações são transformadas em crenças e utilizadas para inferir o estado das habilidades. Já nesta tese, os itens avaliativos são estruturados a partir de uma matriz, de modo que cada item possa ser associado às habilidades necessárias para sua resolução. Essa estrutura possibilita o uso de modelos diagnósticos para estimar perfis de domínio e analisar o desempenho dos estudantes de maneira mais detalhada.

Sendo assim, o Diag-Skills contribui para esta tese ao demonstrar a importância de sistemas que possam diagnosticar habilidades e lidar com as incertezas presentes nos dados educacionais. O trabalho também reforça a necessidade de representar as relações entre item-atributo, principalmente quando o objetivo é apoiar decisões pedagógicas e personalizar

³ <https://helpx.adobe.com/br/xd/get-started.html>

o percurso de aprendizagem. Entretanto, a presente proposta avança em outra direção ao organizar uma estrutura baseado em MDC, com foco na criação, aplicação e análise de avaliações orientadas por habilidades. Enquanto o Diag-Skills enfatiza a combinação de evidências e a propagação de crenças em um modelo semântico, esta tese concentra-se na estimativa diagnóstica do domínio de habilidades latentes com base nas respostas dos estudantes.

3.11 Resumo do Capítulo

Neste capítulo, são apresentados trabalhos que abordam a individualização da aprendizagem, a representação de habilidades e competências, a avaliação formativa e o diagnóstico do estudante em sistemas educacionais. Em geral, observa-se que a análise de habilidades e competências passou a ter um papel importante em propostas de ensino adaptativo e avaliação educacional.

Essas abordagens visam organizar informações sobre o conhecimento do estudante, seu perfil cognitivo e seu percurso de aprendizagem, enquanto procuram oferecer ao professor melhores condições para acompanhar o processo educacional. As avaliações computadorizadas também se tornaram uma alternativa recorrente para a avaliação de estudantes em diferentes contextos. Além de permitirem maior flexibilidade na aplicação dos itens, essas avaliações devem ser organizadas de modo a produzir estimativas precisas em menos tempo (Vie et al., 2017).

Sendo assim, é necessário identificar e gerenciar os pontos fortes e fracos do conhecimento dos estudantes (Ravand e Robitzsch, 2015; Villanueva et al., 2018). Essa necessidade justifica o uso de modelos que forneçam informações detalhadas do que uma pontuação geral de desempenho. Os trabalhos analisados apontam diferentes abordagens para essa finalidade. Algumas propostas, como IWT, MM-Competences, CAT, CALEP, ArCARE e ANNETTE, concentram-se na representação de competências, indicadores, perfis e objetos de aprendizagem. Essas abordagens destacam a importância de explicitar os elementos avaliados e de relacioná-los aos processos de ensino, avaliação e certificação.

Outras propostas se aproximam do objetivo desta tese, pois tratam do diagnóstico de habilidades em sistemas educacionais. O Simplify ITS utiliza MDC DINA, para apoiar a recomendação adaptativa de atividades (Villanueva et al., 2018). O Diag-Skills (Rahmouni et al., 2021) emprega funções de crença para lidar com incerteza, conflito e ignorância no diagnóstico de habilidades. Já o Stat-Knowlab (Chiusole et al., 2020) utiliza a CbKST para

realizar avaliação adaptativa e organizar o percurso de aprendizagem com base no estado de competência do estudante. Esses trabalhos demonstram que a associação entre avaliação, habilidades e personalização da aprendizagem tem sido explorada por diferentes modelos teóricos e computacionais.

A adoção de modelos psicométricos, como o DINA em [Villanueva et al. \(2018\)](#) e a TRI em [Hnida, Idrissi e Bennani \(2018\)](#), pode ser compreendida pela limitação das formas tradicionais de avaliação em fornecer devolutivas diagnósticas. As avaliações tradicionais, muitas vezes, indicam apenas o desempenho geral do estudante, sem explicitar quais habilidades foram dominadas ou não dominadas. Nesse sentido, os MDC apresentam uma vantagem em relação aos modelos centrados apenas em proficiência, pois permitem associar o desempenho observado a estruturas de conhecimento subjacentes ([Javidanmehr e Sarab, 2017](#); [Yeung, 2019](#); [Ma, Minchen e Torre, 2020](#)). Dessa forma, as especificações de habilidades utilizadas pelos MDC podem apoiar intervenções mais pontuais, voltadas às necessidades de um estudante ou de um grupo.

Apesar das contribuições identificadas, observa-se que os trabalhos analisados não integram, em uma mesma proposta, a elaboração autoral de avaliações, a associação explícita entre itens e habilidades, a aplicação dos itens e a análise diagnóstica por MDC. Esse ponto define o principal diferencial desta tese, que propõe uma estrutura educacional voltada à criação, aplicação e análise de avaliações orientadas por habilidades. O Quadro 1 apresenta uma síntese comparativa dos trabalhos relacionados e destaca o posicionamento da tese.

Quadro 1 – Quadro comparativo dos trabalhos relacionados.

Trabalho	Objetivo	Representação	Diagnóstico/Avaliação	Saída	Limitação em relação à tese
IWT	Personalização da aprendizagem	Ontologias; habilidades; competências	Modelo do Estudante; regras de adaptação	Percursos e objetos personalizados	Não estrutura um processo de avaliação diagnóstica item-habilidade.
MM-Competences	Modelagem formal de competências	Grafos; conceitos; habilidades; competências	Estrutura conceitual	Modelo computacional de competências	Formaliza competências, mas não contempla autoria de itens.
CAT	Avaliação formativa	Competências; indicadores; blogs	Validação docente; análise qualitativa	Feedback formativo	A avaliação depende da interpretação docente e não produz perfil diagnóstico automatizado.
CALEP	Trilhas de aprendizagem	Ontologia; competências; objetos de aprendizagem	TRI; modelo de Rasch	Trilhas personalizadas	Usa a avaliação para selecionar trilhas, mas não detalha o domínio de habilidades.
ANNETTE Project	Gestão profissional	KSC; perfis; evidências	Registros; validação institucional	Lacunas de competência	Não há avaliação diagnóstica de aprendizagem.
ArCARE	Adaptação e recomendação de recursos em AVA	Ontologia OLM; habilidades; competências; perfil do estudante	Avaliação de desempenho e atualização dos níveis de habilidades	Perfil aberto do estudante e recursos adaptados/recomendados	Depende da definição prévia das habilidades associadas às atividades e aos recursos educacionais. Não contempla automação da relação item-CC.
Simplify ITS	Tutoria adaptativa	Atividades; habilidades	MDC DINA	Recomendação; revisão espaçada	Não há autoria e análise pedagógica.
CaSS	Interoperabilidade	Frameworks; evidências; assertions	Declarações de competência	Perfis compartilhados	Não possui ciclo avaliativo de criação, aplicação e diagnóstico de habilidades.
Diag-Skills	Diagnóstico em STI	Ontologia; habilidades; pré-requisitos	Funções de crença	Estado da habilidade	Não organiza a avaliação explícita de itens e habilidades.
Stat-Knowlab	Avaliação adaptativa	CbKST; skill map	KST; BLIM	Estado de competência; habilidades prontas para aprender	Não há análise diagnóstica de avaliações construídas pelo professor.
CAASA-ITS	Avaliação diagnóstica	Matriz de habilidades; estudante-item-habilidade	Modelos de Diagnóstico Cognitivo	Perfil de domínio; pontos fortes e fracos; suporte para LLM	Integra autoria, aplicação e análise diagnóstica de avaliações orientadas por habilidades.

O Quadro 1 indica que a literatura já dispõe de propostas voltadas para a representação de habilidades e competências, além da personalização da aprendizagem, a avaliação formativa e ao diagnóstico de habilidades. No entanto, as contribuições tratam esses elementos de maneira parcial ou com diferentes enfoques. Enquanto algumas propostas enfatizam a organização conceitual das competências; outras priorizam a recomendação de atividades, a gestão de perfis, a avaliação adaptativa ou a interoperabilidade entre sistemas.

A proposta desta tese diferencia-se por concentrar-se no processo avaliativo orientado por habilidades. Nesse sentido, a matriz de relação item-CC tem um papel central, pois deixa explícita a relação entre os itens avaliativos e as habilidades necessárias para resolvê-los. A partir dessa estrutura, os MDC permitem estimar perfis de domínio dos estudantes, oferecendo informações mais detalhadas sobre habilidades dominadas e não dominadas. Assim, a estrutura proposta dessa tese busca apoiar o professor desde a elaboração da avaliação até a interpretação diagnóstica dos resultados, favorecendo uma análise mais específica do desempenho dos estudantes.

Capítulo

4

ARQUITETURA COGNITIVO-ADAPTATIVA

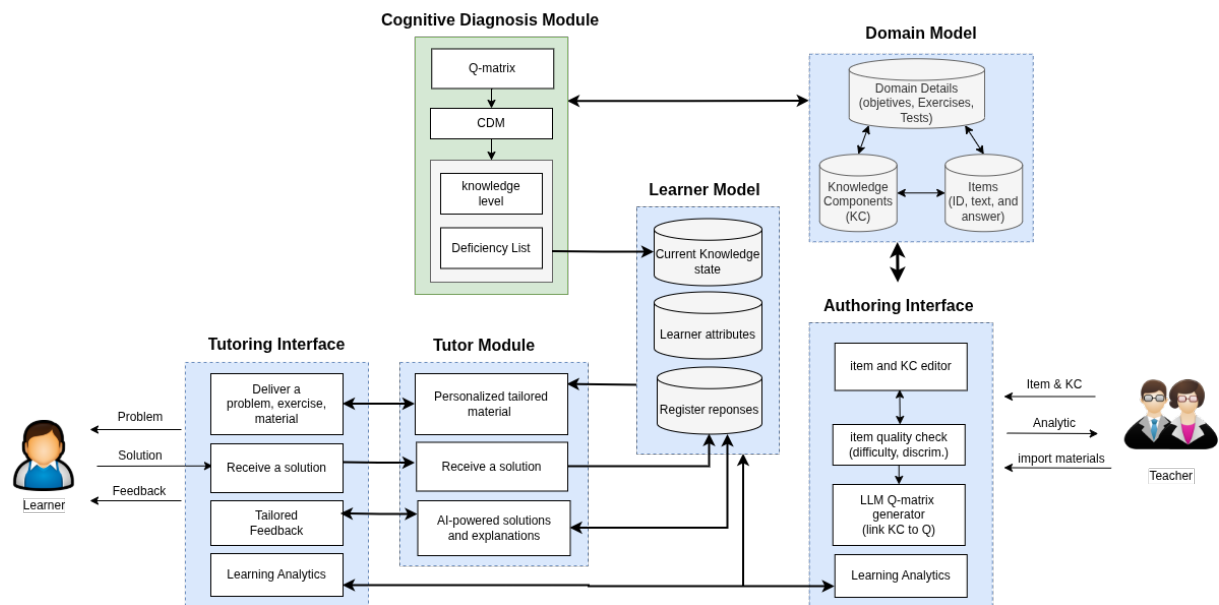
Este capítulo apresenta a proposta da tese, denominada *Cognitive-Adaptive Architecture for Skills Assessment in Intelligent Tutoring Systems* (CAASA-ITS). A proposta consiste em uma arquitetura para Sistema Tutor Inteligente (STI) que integra Modelos de Diagnóstico Cognitivo (MDC), gerenciamento de Q-matriz, modelagem do estudante, mecanismos de adaptação pedagógica e suporte de Modelo de Linguagem de Grande Escala (LLM) para auxiliar na vinculação entre os itens de avaliação e os CC.

A proposta parte das limitações identificadas nos trabalhos relacionados, nos quais foram observados três problemas recorrentes. O primeiro, está relacionado à dificuldade de representar explicitamente quais habilidades ou Componentes de Conhecimento (CC) são avaliados por cada item.

O segundo problema é a dependência de processos manuais na construção da relação item-CC em uma Q-matriz, o que exige conhecimento especializado e dificulta a ampliação dos bancos de questões. Por fim, o terceiro problema refere-se à separação entre o diagnóstico produzido pelos modelos avaliativos e as ações pedagógicas executadas pelo sistema. Isso faz com que os resultados da avaliação nem sempre se convertam em intervenções, recomendações ou devolutivas compreensíveis para professores e estudantes.

A CAASA-ITS é uma arquitetura cognitivo-adaptativa que organiza esses elementos em um ciclo contínuo de autoria, avaliação, diagnóstico e adaptação. Nesse ciclo, o professor define ou valida itens e habilidades, enquanto o sistema auxilia na construção da Q-matriz. Os estudantes, por sua vez, resolvem as atividades, e o módulo de diagnóstico estima o domínio das habilidades. Por fim, o módulo pedagógico utiliza esse diagnóstico para fornecer apoio gradual por meio de mensagens informativas, recomendações de objetos de aprendizagem e novas atividades. Dessa maneira, a avaliação deixa de ser tratada como um registro de acertos e erros, passando a ser utilizada como um mecanismo para inferir o estado de conhecimento do estudante em relação a habilidades específicas. A Figura 26 apresenta uma visão geral da arquitetura CAASA-ITS.

Figura 26 – Visão geral da arquitetura CAASA-ITS integrando diagnóstico cognitivo, geração de Q-matriz e mecanismos de devolutivas.



Fonte: Elaborado pelos autores (2026).

A Figura 26 destaca a integração entre os módulos de domínio, estudante, diagnóstico cognitivo, tutoria, autoria e visualização. A estrutura também evidencia o papel da matriz-Q como elemento de ligação entre os itens de avaliação e os componentes de conhecimento, bem como a inserção de um módulo baseado em LLM para apoiar a vinculação de itens e CC. Essa organização possibilita a representação do ciclo contínuo entre autoria, resolução de atividades, registro de respostas, diagnóstico de habilidades e adaptação pedagógica.

4.1 Evolução da Proposta

A arquitetura CAASA-ITS é resultado de um percurso de pesquisa desenvolvido em etapas. Inicialmente, investigaram-se arquiteturas de STIs e o uso de agentes pedagógicos para apoiar a recomendação e a intervenção em atividades de álgebra. Essa etapa permitiu compreender a necessidade de organizar o domínio do conhecimento, acompanhar as ações dos estudantes e oferecer apoio pedagógico com base no desempenho observado (Lopes e Netto, 2021).

Com base nesse estudo, foi desenvolvido um modelo multiagente para o STI LEIA, voltado ao ensino de equações polinomiais de primeiro grau. A arquitetura projetada por Lopes e Netto (2021) inclui agentes de *software* responsáveis pelo domínio, estudante e tutoria. Esses agentes têm a função de seleção de atividades, acompanhamento das interações e definição de estratégias para fornecer apoio pedagógico ao estudante. Essa etapa evidenciou que a

adaptação pedagógica depende de uma representação clara do domínio e das habilidades envolvidas em cada atividade.

Em seguida, a pesquisa avançou para a incorporação de um MDC. O estudo com o MDC G-DINA em [Lopes e Netto \(2023\)](#) demonstrou a possibilidade de se estimar perfis de domínio de habilidades a partir das respostas dos estudantes, por meio de uma Q-matriz construída com o apoio de especialistas. O trabalho em questão estudou a aplicação do MDC no domínio das equações de primeiro grau. Nessa etapa, foram definidos atributos de habilidades associados aos itens de avaliação. Os resultados foram apresentados ao professor por meio de um painel de visualização. Essa etapa mostrou que o diagnóstico cognitivo pode oferecer uma visão mais detalhada do conhecimento do estudante do que uma simples pontuação total em uma avaliação.

Apesar dos avanços, uma limitação central permaneceu: a construção da Q-matriz continuava dependente da análise manual dos especialistas. Por isso, a etapa seguinte da pesquisa investigou o uso de LLMs para automatizar ou apoiar a vinculação entre itens e componentes de conhecimento. Nessa investigação, a geração de Q-matriz foi formulada como uma tarefa de *Knowledge Tagging* (ver seção 2.4.4) multirrótulo, na qual o modelo recebe o enunciado do item e uma lista de CC e retorna as associações previstas. A avaliação foi conduzida em dois níveis: estrutural, comparando a Q-matriz gerada com uma matriz definida por especialistas; e funcional, avaliando o impacto da nova Q-matriz no desempenho preditivo e, para isso, ela foi testada em um MDC.

Com base nesses estudos, a tese consolida CAASA-ITS como uma arquitetura que integra três dimensões: a dimensão diagnóstica, baseada em MDC; a dimensão autoral, baseada no gerenciamento de itens, habilidades e Q-matriz; e a dimensão adaptativa, responsável por utilizar os resultados que utiliza o diagnóstico para orientar a ação pedagógica do STI.

4.2 Problemas Abordados pela Arquitetura

Os estudos relacionados (ver capítulo 3) mostram que muitos dos sistemas educacionais adaptativos utilizam os Modelos de Estudante e mecanismos de recomendação, mas nem sempre possuem uma estrutura formal para associar itens de avaliação às habilidades mensuradas. Em alguns casos, os sistemas apresentam recomendações baseadas em desempenho geral, em trilhas de aprendizagem ou em regras previamente definidas. Outros empregam ontologias, grafos de competência ou modelos de inferência, mas mantêm a avaliação dependente de anotações manuais ou de interpretações subjetivas dos especialistas.

Nos sistemas baseados em MDC, a Q-matriz possui um papel importa, pois define quais habilidades são necessárias para resolução de cada item. Quando essa matriz é inconsistente, incompleta ou difícil de manter, o diagnóstico realizado pode ser comprometido. Além disso, nos ambientes reais, novos itens podem ser adicionados com frequência, exigindo uma revisão constante da estrutura item-atributo. Tal processo torna a manutenção do sistema onerosa para professores e desenvolvedores.

A arquitetura proposta atua nesse ponto. Seu objetivo é oferecer uma estrutura que permita: criar e gerenciar CC; associar itens a habilidades por meio de uma Q-matriz; apoiar a geração dessa matriz com LLM; validar as associações geradas; estimar o domínio de habilidades dos estudantes com MDC; e utilizar os resultados para adaptação ou tomada de decisões pedagógicas.

É importante destacar que a arquitetura não elimina o papel do professor. Ele permanece sendo o responsável pela validação pedagógica das associações item-CC e pela interpretação dos resultados no contexto da turma. O LLM atua como apoio à autoria, reduzindo o esforço inicial de classificação dos itens e oferecendo sugestões que podem ser aceitas, ajustadas ou rejeitadas.

4.3 Formalização Geral da Arquitetura

Seja $(U = u_1, u_2, \dots, u_N)$ o conjunto de estudantes cadastrados no sistema, $(E = e_1, e_2, \dots, e_J)$ o conjunto de itens de avaliação, e $(C = c_1, c_2, \dots, c_K)$ o conjunto de CC ou habilidades cognitivas avaliadas. Cada item (e_j) pode estar associado a um ou mais CC (c_k) , pois uma mesma atividade pode exigir diferentes operações cognitivas ou procedimentos de resolução.

A relação entre itens e CC é representada por uma Q-matriz binária:

$$Q = [q_{jk}]J \times K, \quad q_{jk} \in \{0, 1\} \quad (4.1)$$

em que:

$$q_{jk} = \begin{cases} 1, & \text{se o item } e_j \text{ requer o Componentes de Conhecimento } c_k; \\ 0, & \text{caso contrário.} \end{cases} \quad (4.2)$$

A matriz de respostas dos estudantes é definida por:

$$X = [x_{ij}]N \times J, \quad x_{ij} \in \{0, 1\} \quad (4.3)$$

em que $(x_{ij} = 1)$ indica que o estudante (u_i) respondeu corretamente ao item (e_j) , enquanto $(x_{ij} = 0)$ indica resposta incorreta. Essa matriz registra o desempenho observável do estudante, enquanto a Q-matriz representa a estrutura cognitiva esperada para a resolução dos itens.

O estado de conhecimento do estudante é representado por um vetor de domínio:

$$\alpha_i = (\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{iK}) \quad (4.4)$$

em que (α_{ik}) representa o estado de domínio do estudante (u_i) em relação ao CC (c_k) . Em uma representação binária, $(\alpha_{ik} = 1)$ indica domínio e $(\alpha_{ik} = 0)$ indica não domínio. Em uma representação probabilística, adotada em modelos como G-DINA, o estado de domínio pode ser estimado como:

$$\hat{\pi}_{ik} = P(\alpha_{ik} = 1 \mid X_i, Q) \quad (4.5)$$

em que $(\hat{\pi}_{ik})$ representa a probabilidade estimada de domínio do estudante (u_i) sobre o componente (c_k) , considerando suas respostas (X_i) e a estrutura da Q-matriz.

A partir de um limiar de domínio (τ) , o conjunto de habilidades dominadas pelo estudante pode ser definido como:

$$K_i^+ = c_k \in C \mid \hat{\pi}_{ik} \geq \tau \quad (4.6)$$

e o conjunto de habilidades não dominadas pode ser definido como:

$$K_i^- = c_k \in C \mid \hat{\pi}_{ik} < \tau \quad (4.7)$$

O conjunto (K_i^-) constitui a lista de deficiências do estudante e é utilizado pelo módulo pedagógico para selecionar as atividades, devolutivas ou intervenções adequadas ao seu estado atual de conhecimento.

Em geral, o processo de diagnóstico da arquitetura CAASA-ITS pode ser representado por uma função:

$$M_{CDM} : (X, Q) \rightarrow \Pi \quad (4.8)$$

em que (M_{CDM}) representa o MDC utilizado e $(\Pi = [\hat{\pi}_{ik}]N \times K)$ representa a matriz de probabilidades de domínio dos estudantes em relação aos componentes de conhecimento.

A geração assistida da Q-matriz por LLM pode ser representada pela função:

$$f_{LLM} : (E, C, P) \rightarrow \hat{Q} \quad (4.9)$$

em que (P) representa a estratégia de *prompting* utilizada e ($\hat{Q}$) representa a Q-matriz sugerida pelo modelo. Essa matriz não é incorporada diretamente ao sistema sem validação. A arquitetura prevê uma função de validação pedagógica:

$$\nu_T(\hat{Q}) \rightarrow Q^{(r)} \quad (4.10)$$

em que (ν_T) representa a validação realizada pelo professor ou especialista, e ($Q^{(r)}$) representa a versão revisada da Q-matriz. Assim, a Q-matriz é tratada como uma estrutura versionada, podendo ser atualizada conforme novos itens, novos CC ou novas evidências diagnósticas sejam incorporados ao sistema.

4.4 Módulo de Domínio

O Módulo de Domínio representa o conhecimento a ser ensinado, avaliado e monitorado. Esse modelo deriva do *Domain Agent* descrito em [Lopes e Netto \(2021\)](#), e sua função é auxiliar o professor na classificação de equações de acordo com conceitos matemáticos e níveis de dificuldade. Na arquitetura CAASA-ITS, esse módulo é ampliado para outros domínios, organizando não apenas questões específicas, mas também objetivos de aprendizagem, CC, relações de pré-requisito, níveis de dificuldade e versões da Q-matriz.

Formalmente, o domínio pode ser definido como:

$$D = \langle O, E, C, R, Q \rangle \quad (4.11)$$

onde (O) representa o conjunto de objetivos de aprendizagem, (E) o conjunto de itens de avaliação, (C) o conjunto de CC, (R) o conjunto de relações pedagógicas entre componentes e (Q) a matriz de associação item e CC.

O conjunto (R) permite representar relações entre CC, como pré-requisito, dependência, equivalência ou progressão. Uma relação de pré-requisito pode ser expressa por:

$$R \subseteq C \times C \quad (4.12)$$

em que $((c_a, c_b) \in R)$ indica que o componente (c_a) deve ser dominado antes de (c_b). Essa estrutura permite que o sistema selecione atividades não apenas com base nas habilidades

não dominadas, mas também considerando a ordem pedagógica de desenvolvimento dessas habilidades.

Cada item (e_j) (questão, exercício, problema) é representado por um conjunto de metadados:

$$e_j = \langle id_j, enunciado_j, resposta_j, dificuldade_j, tipo_j, q_j \rangle \quad (4.13)$$

em que (q_j) corresponde à linha da Q-matriz associada ao item (e_j). Essa representação permite que o sistema identifique quais habilidades estão envolvidas na resolução de cada item e utilize essa informação tanto no diagnóstico quanto na adaptação pedagógica.

[Lopes e Netto \(2023\)](#) utilizam o exemplo no domínio de equações de primeiro grau: os CC são *propriedades de igualdade, simplificação de frações, resolução de equações com constantes em ambos os lados, resolução de equações com variáveis e constantes em ambos os lados e aplicação da propriedade distributiva*.

Em outros domínios, a mesma estrutura pode ser reutilizada, desde que os CC sejam redefinidos de acordo com os objetivos da área ou do item.

4.5 Módulo de Autoria e Gerenciamento da Q-Matriz

O Módulo de Autoria permite que o professor crie, edite e organize itens de avaliação e os CC. Este módulo é derivado da interface do professor proposta em trabalhos anteriores ([Lopes e Netto, 2021](#); [Lopes et al., 2019](#)), na qual o docente insere atividades e relaciona cada uma delas às habilidades ou conceitos avaliados. Na nova arquitetura, essa função é ampliada para incluir a edição da Q-matriz. O controle de versão das associações entre item e CC, e a validação de sugestões geradas por LLM.

O módulo possui três funções principais. A primeira é a criação de itens, em que o professor registra o enunciado, a resposta esperada, o tipo de questão e o nível de dificuldade. A segunda função, é a criação ou seleção dos CC. A terceira é a validação da associação entre itens e CC, representada pela Q-matriz.

A Q-matriz pode ser construída de três formas: manualmente, por meio de definições diretas do professor, de maneira assistida, a partir de sugestões produzidas por LLM, ou de maneira híbrida, combinando uma matriz inicial definida por especialistas com sugestões computacionais e revisões posteriores. A função de autoria pode ser representada por:

$$A_T : (E, C) \rightarrow Q_T \quad (4.14)$$

em que (A_T) representa o processo de autoria conduzido pelo professor e (Q_T) representa a Q-matriz definida manualmente. Quando há suporte por LLM, a matriz sugerida é dada por:

$$\hat{Q}_{LLM} = f_{LLM}(E, C, P) \quad (4.15)$$

A matriz validada pode então ser representada por:

$$Q^{(r)} = V(Q_T, \hat{Q}_{LLM}) \quad (4.16)$$

onde (V) representa o processo de validação humana e computacional. Essa formulação permite compreender que o LLM não substitui o professor, mas atua como mecanismo de apoio para reduzir o esforço de classificação inicial dos itens.

4.6 Módulo de Geração Assistida da Q-Matriz por LLM

Este módulo é responsável por sugerir associações entre itens de avaliação e CC. Essa função é incorporada à arquitetura CAASA-ITS a partir dos resultados de (Lopes e Netto, 2026), nos quais a geração assistida da Q-matriz foi tratada como uma tarefa de *Knowledge Tagging* multirrotulo.

Na prática, o mecanismo que fundamenta este módulo foi operacionalizado no Estudo 3 (ver Seção 5.3) por meio de um pipeline computacional no qual o texto de cada item de avaliação e um conjunto previamente definido de CC são fornecidos como entrada ao LLM. A partir dessas informações, um *prompt* orienta o modelo a identificar quais componentes estão relacionados à resolução do item. A resposta produzida pelo modelo é então processada e convertida em um vetor binário, no qual o valor 1 representa a associação entre o item e o CC, e o valor 0 representa sua ausência. A concatenação desses vetores produz a Q-matriz sugerida pelo modelo.

Na implementação experimental, a comunicação com os LLM foi realizada por meio da API do Ollama, utilizando modelos executados localmente. Foram investigadas estratégias de *prompting* do tipo *zero-shot*, *few-shot* e *few-shot* com CoT. As Q-matrizes resultantes foram avaliadas estruturalmente por meio das métricas HL e JSI e, em seguida, funcionalmente por meio do modelo DINA, considerando as métricas RMSE e MAE. Os detalhes da configuração experimental e dos resultados são apresentados na Seção 5.3.

É importante distinguir essa implementação experimental da instanciação completa do módulo na arquitetura CAASA-ITS. O Estudo 3 forneceu a implementação e as evidências

empíricas do mecanismo de geração e avaliação das associações item–CC. Na arquitetura proposta, esse mecanismo é incorporado como um serviço de apoio ao Módulo de Autoria: as associações sugeridas pelo LLM não são utilizadas diretamente pelo MDC, mas são encaminhadas ao professor ou especialista para validação, ajuste ou rejeição antes de comporem a versão validada da Q-matriz.

Com base nessa operacionalização, o funcionamento do Módulo de Geração Assistida da Q-Matriz pode ser representado, formalmente, pela associação entre cada item de avaliação e o conjunto de CC considerado. Dado um item (e_j) e o conjunto (C) de CC, o LLM deve produzir um vetor binário:

$$\hat{q}_j = (\hat{q}_{j1}, \hat{q}_{j2}, \dots, \hat{q}_{jK}) \quad (4.17)$$

em que ($\hat{q}_{jk} = 1$) indica que o modelo associou o item (e_j) ao componente (c_k), e ($\hat{q}_{jk} = 0$) indica ausência de associação. A matriz completa sugerida pelo LLM é formada pela concatenação dos vetores de todos os itens:

$$\hat{Q}_{LLM} = [\hat{q}_1 \ \hat{q}_2 \ : \ \hat{q}_J] \quad (4.18)$$

A geração pode ser realizada com diferentes estratégias de *prompting*, como *zero-shot*, *few-shot* e *few-shot* com CoT. A estratégia utilizada é representada por (P), de modo que:

$$\hat{Q}_{LLM}^{(p)} = f_{LLM}(E, C, p) \quad (4.19)$$

em que ($p \in P$) representa uma estratégia específica de instrução. Tal formulação possibilita a comparação entre diferentes estratégias de geração, verificando-se, assim, quais produzem como saída, matrizes mais próximas das matrizes de referência ou mais adequadas ao funcionamento do modelo diagnóstico.

A avaliação da matriz sugerida ocorre em dois níveis. No primeiro nível estrutural ($\hat{Q}_{LLM}$), a matriz é comparada com uma matriz de referência (Q_{ref}), que é normalmente definida por especialistas e representa a estrutura considerada correta ou esperada. A comparação entre as duas matrizes permite verificar o quanto a saída do modelo se aproxima desse padrão. Para isso, são utilizadas métricas como *Hamming Loss* (HL) e *Jaccard Index* (JSI), que quantificam, respectivamente, o grau de divergência e o grau de similaridade entre as matrizes. No nível funcional, a matriz sugerida é utilizada como entrada de um MDC, e seu impacto é analisado por métricas como RMSE e MAE. Essa avaliação permite verificar se a

Q-matriz gerada mantém a coerência estrutural e produz desempenho preditivo adequado de diagnóstico.

4.7 Módulo do Estudante

O Módulo do Estudante deriva do *Learner Model* proposto em [Lopes e Netto \(2021\)](#). Na arquitetura, o modelo é baseado no *Overlay Model*, segundo o qual o conhecimento do estudante é entendido como um subconjunto do conhecimento representado pelo sistema. Na CAASA-ITS, essa ideia é mantida, mas é combinada com as estimativas probabilísticas produzidas pelo MDC.

O perfil do estudante pode ser formalizado como:

$$P_i = \langle id_i, dados_i, H_i, X_i, L_i, \alpha_i, K_i^+, K_i^- \rangle \quad (4.20)$$

em que (id_i) é o identificador do estudante, $(dados_i)$ representa informações cadastrais ou pedagógicas, (H_i) representa o histórico de atividades, (X_i) é o vetor de respostas do estudante, (L_i) representa os registros de interação, (α_i) é o vetor de domínio estimado, (K_i^+) representa as habilidades dominadas e (K_i^-) representa as habilidades não dominadas.

Os registros de interação são representados por:

$$L_i = l_{i1}, l_{i2}, \dots, l_{it} \quad (4.21)$$

em que cada registro (l_{it}) pode conter informações como o item acessado, o tempo de resolução, o número de tentativas, os erros cometidos, as dicas solicitadas, as mensagens de devolutivas recebidas ou os passos de resolução. Esses registros ampliam a modelagem do estudante, pois permitem observar não apenas o resultado da resposta, mas também o comportamento durante a atividade.

A atualização do estado de conhecimento ocorre sempre que novas respostas são registradas. Essa atualização pode ser representada por:

$$P_i^{(t+1)} = Update(P_i^{(t)}, x_{ij}, l_{it}, Q, M_{CDM}) \quad (4.22)$$

em que $(P_i^{(t)})$ é o perfil do estudante antes da nova interação, (x_{ij}) a resposta ao item (e_j) , (l_{it}) o registro da interação, (Q) a Q-matriz e (M_{CDM}) o MDC utilizado.

4.8 Módulo de Diagnóstico

O MDC estima o domínio dos CC a partir das respostas dos estudantes e da Q-matriz. Esse módulo foi construído e testado na arquitetura de [Lopes e Netto \(2023\)](#), na qual o MDC G-DINA foi incorporado em um STI LEIA para produzir perfis de habilidades dos estudantes.

Dado o conjunto de respostas (X) e a Q-matriz (Q), o Módulo de Diagnóstico estima a matriz de domínio:

$$\Pi = M_{CDM}(X, Q) \quad (4.23)$$

em que ($\Pi = [\hat{\pi}_{ik}] N \times K$) contém a probabilidade de domínio de cada estudante (u_i) em relação a cada componente (c_k).

Este módulo facilita a incorporação de diferentes MDCs. O G-DINA foi adotado como referência por sua flexibilidade para modelar diferentes efeitos dos atributos na probabilidade de acerto (ver explicação na seção 2.3.2.2). O DINA também pode ser utilizado em experimentos controlados, principalmente quando o objetivo é avaliar o impacto de diferentes Q-matrizes em um modelo não compensatório.

No modelo DINA, a resposta latente ideal do estudante (u_i) ao item (e_j) é definida por:

$$\eta_{ij} = \prod_{k=1}^K \alpha_{ik}^{q_{jk}} \quad (4.24)$$

em que ($\eta_{ij} = 1$) indica que o estudante domina todos os componentes requeridos pelo item, e ($\eta_{ij} = 0$) indica ausência de pelo menos um componente necessário. A probabilidade de acerto pode ser expressa por:

$$P(X_{ij} = 1 \mid \alpha_i) = g_j + (1 - s_j - g_j)\eta_{ij} \quad (4.25)$$

em que (g_j) representa o parâmetro de acerto ao acaso e (s_j) representa o parâmetro de erro mesmo com domínio dos atributos. Essa estrutura permite analisar a relação entre o domínio latente do estudante e o desempenho observado nas atividades.

O resultado do Módulo de Diagnóstico é composto por três saídas principais: o vetor de domínio do estudante, a lista de habilidades não dominadas e os indicadores agregados para acompanhamento da turma. Essas informações são enviadas ao Módulo Pedagógico e ao Módulo de Visualização (*Interface Model*).

4.9 Módulo Pedagógico

O Módulo Pedagógico é derivado do *Tutor Agent* de [Lopes e Netto \(2021\)](#). Ele utiliza o estado de conhecimento do estudante para selecionar atividades, organizar a sequência de aprendizagem, definir estratégias de apoio e fornecer respostas personalizadas. Na CAASA-ITS, esse módulo é orientado diretamente pelo Módulo de Diagnóstico.

A seleção de atividades é realizada com base na lista de habilidades não dominadas (K_i^-). O conjunto de itens candidatos à recomendação para o estudante (u_i), denotado por E_i^{cand} , é formado pelos itens que avaliam pelo menos uma habilidade ainda não dominada por esse estudante:

$$E_i^{cand} = \{e_j \in E \mid \exists c_k \in K_i^-, q_{jk} = 1\} \quad (4.26)$$

Esse conjunto contém os itens que avaliam pelo menos uma habilidade ainda não dominada pelo estudante. No entanto, a seleção final também deve considerar o nível de dificuldade, as relações de pré-requisito e o histórico de interações. Assim, a recomendação pedagógica pode ser representada por:

$$Rec_i = Select(E_i^{cand}, P_i, R, dificuldade) \quad (4.27)$$

em que (*Select*) representa a função de seleção pedagógica, (P_i) é o perfil do estudante, (R) as relações entre os CC e (*dificuldade*) representa os níveis de complexidade dos itens.

A devolutiva também é produzida com base no diagnóstico. Para um item (e_j), o sistema pode identificar quais componentes de conhecimento foram exigidos e quais deles não estão dominados pelo estudante. Assim, a devolutiva pode ser formalizada como:

$$F_{ij} = Feedback(e_j, q_j, K_i^-, L_i) \quad (4.28)$$

em que (F_{ij}) representa a devolutiva fornecida ao estudante (u_i) após a interação com o item (e_j), (q_j) representa a linha correspondente da matriz-Q, (K_i^-) representa as habilidades não dominadas e (L_i) representa o histórico de interação.

Esse módulo também pode acionar estratégias no formato de um suporte estruturado (*scaffolding*). Quando o estudante apresenta dificuldade em um componente específico, o sistema pode oferecer dicas graduais, explicações passo a passo, atividades mais simples ou revisão de pré-requisitos. Quando o estudante demonstra domínio, o sistema pode reduzir intervenções e avançar para atividades de maior complexidade.

4.10 Módulo Tutor e Módulo Especialista

Na arquitetura descrita em [Lopes e Netto \(2021\)](#), o *Expert Module* é responsável por analisar as soluções dos estudantes em cada passo da resolução das equações, verificando operações como divisão, soma de termos semelhantes e transformação algébrica. Na CAASA-ITS, essa função é preservada como parte do Módulo Tutor e do Módulo Especialista.

O Módulo Especialista contém as regras, procedimentos ou modelos capazes de avaliar a solução esperada para cada item. Em domínios procedimentais, como álgebra, ele pode comparar cada passo da resolução do estudante com uma sequência esperada ou com um conjunto de transformações válidas. Em domínios conceituais, pode verificar a resposta final, as justificativas ou as evidências de uso de determinados CC. Com isso, a avaliação realizada pelo Módulo Especialista pode ser representada por:

$$r_{ij} = Eval(e_j, a_{ij}, S_j) \quad (4.29)$$

em que (a_{ij}) representa a resposta ou sequência de ações do estudante (u_i) no item (e_j) , (S_j) representa a solução esperada ou o conjunto de soluções válidas, e (r_{ij}) representa o resultado da avaliação. Esse resultado pode assumir forma binária, quando apenas acerto ou erro é considerado, ou forma detalhada, quando são registrados erros específicos e passos intermediários.

O Módulo Tutor utiliza essa avaliação para decidir a próxima ação instrucional. Se o estudante erra um item que exige determinado CC, o sistema verifica se esse componente também aparece em outros erros anteriores. Quando há recorrência, o diagnóstico é reforçado e uma intervenção mais direcionada pode ser recomendada.

4.11 Módulo de Dados e Registro de Interações

O Módulo de Dados deriva do *Data Agent* de [Lopes e Netto \(2021\)](#). Sua função é registrar as ações dos estudantes, organizar os dados de resposta, manter o histórico de interação e repassar essas informações para os Módulos de Diagnóstico, Tutoria e *Analytics*.

Cada interação do estudante pode ser representada por uma tupla, onde:

$$l_{it} = \langle u_i, e_j, a_{it}, r_{it}, tempo_{it}, dica_{it}, feedback_{it} \rangle \quad (4.30)$$

em que (u_i) identifica o estudante, (e_j) identifica o item, (a_{it}) representa a ação realizada, (r_{it}) representa o resultado da ação, $(tempo_{it})$ o tempo de interação, $(dica_{it})$ indica se houve

solicitação de dica e ($feedback_{it}$) registra a mensagem ou dica fornecida.

Esses dados cumprem duas funções. A primeira é alimentar a matriz de respostas (X), utilizada pelo MDC. A segunda é enriquecer o perfil do estudante com informações comportamentais, tais como o número de tentativas, a dependência de dicas e padrões de erro. Embora o diagnóstico cognitivo seja calculado com base na Q-matriz e nas respostas, os registros de interação ajudam o sistema e o professor a compreender melhor o processo de aprendizagem do estudante.

4.12 Módulo de Visualização e *Learning Analytics*

O Módulo de Visualização é derivado do *Analytics Agent* da arquitetura de [Lopes e Netto \(2021\)](#) e do *Dashboard* implementado na proposta de [Lopes e Netto \(2023\)](#). Sua função é apresentar os resultados do diagnóstico de maneira compreensível para professores e estudantes.

Para os professores, o sistema fornece informações individuais e agregadas da turma, tais como o domínio de habilidades de cada estudante; as habilidades com maior índice de não domínio; a evolução temporal; desempenho por item e; os padrões de erro. Para os estudantes, o sistema apresenta seu próprio perfil de domínio, destacando habilidades já dominadas, habilidades em desenvolvimento e atividades recomendadas.

A visualização do estado do conhecimento pode ser representada por:

$$Dash_T = View(\Pi, X, Q, L) \quad (4.31)$$

para o professor é:

$$Dash_i = View(P_i, K_i^+, K_i^-, Rec_i) \quad (4.32)$$

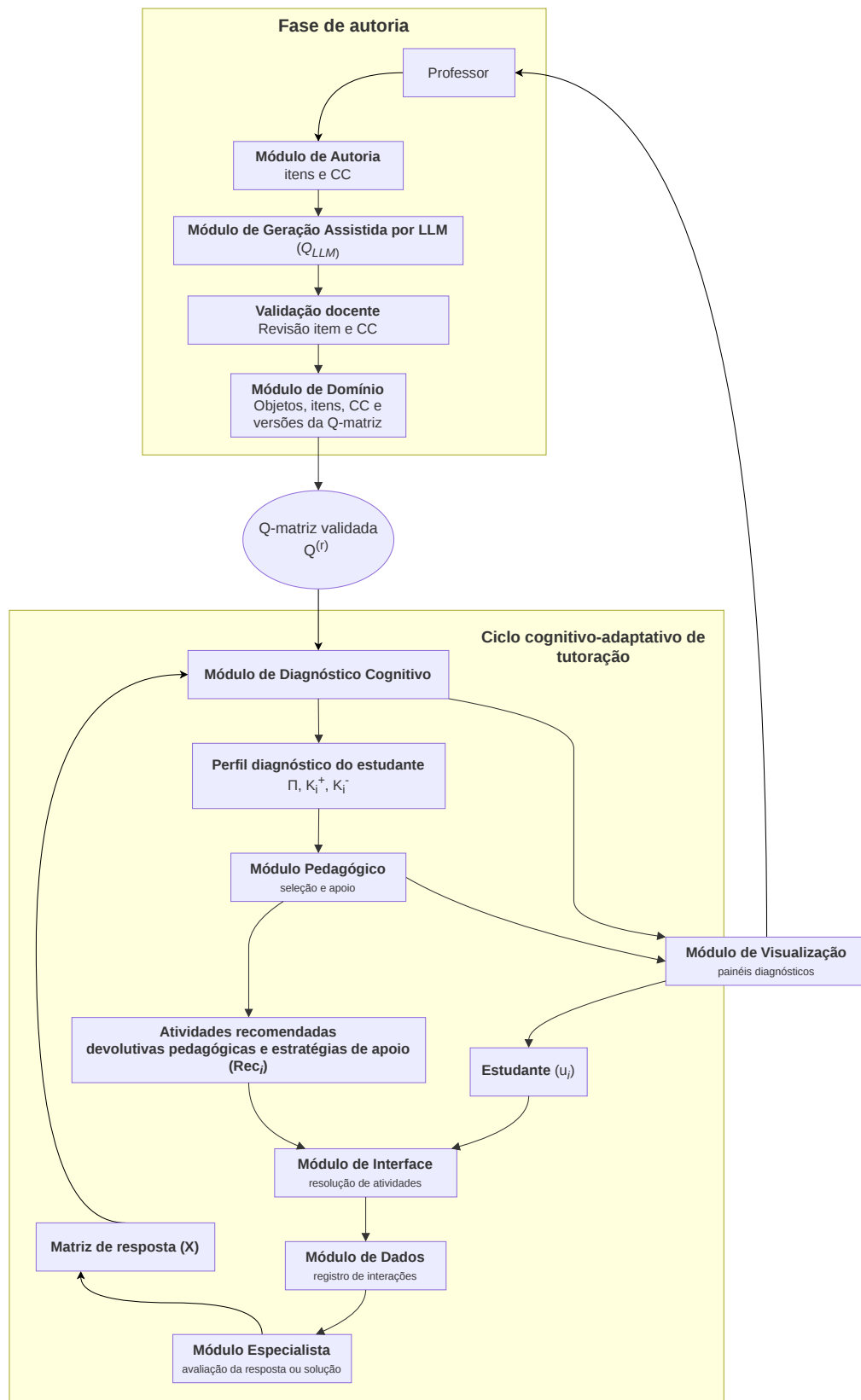
O primeiro painel ($Dash_T$) apoia a tomada de decisão pedagógica em nível de turma. O segundo ($Dash_i$) apoia a autorregulação do estudante, pois apresenta informações sobre seu próprio progresso e sobre os próximos passos sugeridos.

4.13 Integração entre os módulos

A arquitetura CAASA-ITS opera como um ciclo contínuo de autoria, avaliação, diagnóstico e adaptação. A Figura 27 apresenta o ciclo cognitivo-adaptativo da arquitetura CAASA-ITS.

A representação visual organiza a proposta em camadas. A Q-matriz validada ocupa posição central por articular a autoria pedagógica, a representação item-CC, a inferência por

Figura 27 – Ciclo cognitivo-adaptativo da arquitetura CAASA-ITS.



Fonte: Elaborado pelos autores (2026).

Modelos de Diagnóstico Cognitivo e a adaptação das atividades no processo de tutoração. Na camada de autoria, o professor cadastra itens e componentes de conhecimento, recebe sugestões do módulo baseado em LLM e valida as associações item-CC. Na camada de tutoração, o estudante interage com o sistema, suas respostas são registradas e avaliadas, e o MDC estima o perfil de domínio. Com base nesse diagnóstico, o Módulo Pedagógico seleciona atividades, produz devolutivas pedagógicas e define estratégias de apoio. O Módulo de Visualização organiza esses resultados para professor e estudante. Esse fluxo pode ser sintetizado da seguinte forma:

$$(E, C) \rightarrow \hat{Q}_{LLM} \rightarrow Q^{(r)} \rightarrow X \rightarrow \Pi \rightarrow (K_i^+, K_i^-) \rightarrow Rec_i \rightarrow Feedback_i \quad (4.33)$$

Essa sequência mostra que a Q-matriz é o elo entre a autoria pedagógica, o diagnóstico cognitivo e a adaptação instrucional. A arquitetura proposta, portanto, amplia as versões anteriores ao integrar a estrutura multiagente, o diagnóstico por MDC e a geração assistida da Q-matriz em um mesmo processo cognitivo-adaptativo.

4.14 Resumo do Capítulo

Este capítulo apresentou a arquitetura CAASA-ITS como uma proposta cognitivo-adaptativa para avaliação estruturada de habilidades em STI. Essa arquitetura foi organizada a partir da evolução dos estudos desenvolvidos ao longo da pesquisa, integrando a base pedagógica e multiagente proposta por [Lopes e Netto \(2021\)](#), a incorporação do Modelos de Diagnóstico Cognitivo ao STI LEIA apresentada por [Lopes e Netto \(2023\)](#) e o mecanismo de apoio à geração e avaliação da Q-matriz por LLM investigado por [Lopes e Netto \(2026\)](#).

É importante destacar que a arquitetura CAASA-ITS apresenta aderência à perspectiva socioconstrutivista de [Vygotsky \(1978\)](#) ao estruturar o processo de tutoria como uma sequência de mediações pedagógicas orientadas pelo estado de conhecimento do estudante. Nela, o estudante interage com o sistema, e suas ações são registradas, o domínio de habilidades é estimado e novas atividades ou devolutivas são selecionadas, sob medida, com base nas lacunas identificadas. Esse funcionamento é coerente com a ideia de apoio gradual (*scaffolding*) ao estudante, pois a intervenção pedagógica passa a ser definida de acordo com o que o estudante já domina e com o que ainda precisa desenvolver, sendo assim, aplicado o modelo ZDP.

Essa dimensão pedagógica é complementada por uma estrutura formal de representação e inferência, apoiada na teoria do *Rule Space Theory* (RST) desenvolvida por [Tatsuoka, 1984](#).

No contexto da RST, a Q-matriz representa a relação entre itens de avaliação e atributos cognitivos, permitindo explicitar quais CC são requeridas em cada atividade. Essa estrutura organiza o espaço de possíveis padrões de domínio e não domínio, servindo como base para a classificação diagnóstica dos estudantes. Na CAASA-ITS, a Q-matriz, também, assume esse papel ao relacionar itens e CC, e fornecendo a estrutura necessária para que os MDCs estimem o perfil de domínio dos estudantes a partir das respostas observadas. Dessa maneira, a arquitetura articula mediação pedagógica, representação formal do conhecimento e inferência cognitiva no mesmo ciclo de avaliação e adaptação.

Sendo assim, a CAASA-ITS não se limita à organização computacional de módulos. A proposta estabelece uma relação entre fundamentos pedagógicos e mecanismos formais de avaliação, na qual a mediação inspirada em [Vygotsky \(1978\)](#) é operacionalizada por meio de agentes ativos e reativos, módulos de tutoria e estratégias de apoio, enquanto a Q-matriz e os MDCs fornecem a base diagnóstica para identificar o estado de conhecimento do estudante. Essa integração permite que a avaliação de habilidades seja tratada como um processo contínuo, interpretável e orientado à tomada de decisão pedagógica.

O próximo capítulo apresenta os resultados alcançados nos estudos que sustentaram a construção da CAASA-ITS. São discutidas as contribuições da arquitetura multiagente de [Lopes e Netto \(2021\)](#), os resultados da incorporação do MDC G-DINA em [Lopes e Netto \(2023\)](#) e a avaliação estrutural e funcional da geração de Q-matriz por LLM em [Lopes e Netto \(2026\)](#). Esses resultados mostram como cada etapa da pesquisa contribuiu para a consolidação da arquitetura proposta nesta tese.

Capítulo

5

RESULTADOS ALCANÇADOS

Este capítulo apresenta os resultados alcançados ao longo do desenvolvimento da pesquisa, organizados a partir dos três estudos que contribuíram progressivamente para a formalização da arquitetura CAASA-ITS. Cada estudo investigou uma dimensão específica do problema de pesquisa, envolvendo, respectivamente, mecanismos de adaptação pedagógica em STI, a incorporação de MDC para avaliação de habilidades e o uso de LLM como apoio à construção e à avaliação de Q-matrizes. Em conjunto, esses estudos forneceram os elementos conceituais, computacionais e experimentais utilizados na consolidação da arquitetura proposta.

As atividades de concepção, implementação, configuração experimental e análise computacional desenvolvidas especificamente no âmbito desta tese foram conduzidas pelo autor, sob orientação acadêmica. Entre essas atividades estão a adaptação dos mecanismos de avaliação baseada em habilidades no STI LEIA, a configuração e a aplicação do modelo G-DINA, a organização das Q-matrizes e dos dados utilizados nos experimentos e o desenvolvimento do *pipeline* computacional empregado no Estudo 3 para geração e avaliação das Q-matrizes produzidas com apoio de LLM. Componentes previamente existentes do STI LEIA, bem como bibliotecas, modelos e implementações de terceiros utilizados nos experimentos, são identificados e referenciados nas respectivas seções.

5.1 Estudo 1: Arquitetura Multiagente para Recomendação e Intervenção Pedagógica

O primeiro estudo desenvolvido no percurso desta tese foi o mapeamento da literatura publicado em [Lopes e Netto \(2020\)](#). O objetivo do estudo foi identificar como os STI eram modelados e aplicados em diferentes contextos educacionais, com atenção às estratégias de ensino, às formas de acompanhamento do estudante e aos mecanismos de adaptação da aprendizagem. Embora esse trabalho não seja descrito integralmente neste capítulo, ele foi importante na delimitação da pesquisa, pois permitiu identificar tendências, limitações

e lacunas relacionadas à modelagem do estudante, à organização do domínio e ao uso de mecanismos inteligentes de apoio à tomada de decisão pedagógica.

A partir desse mapeamento, foi possível observar que muitos STI ainda apresentavam limitações na representação explícita das habilidades envolvidas nas atividades, no acompanhamento individual do progresso dos estudantes e na adaptação das intervenções de acordo com o estado de conhecimento de cada aprendiz. Essas constatações orientaram o desenvolvimento do estudo seguinte, que propõe uma arquitetura multiagente para ampliar as funcionalidades do STI LEIA, que é direcionado ao ensino inicial de álgebra, com foco na resolução de equações polinomiais de primeiro grau.

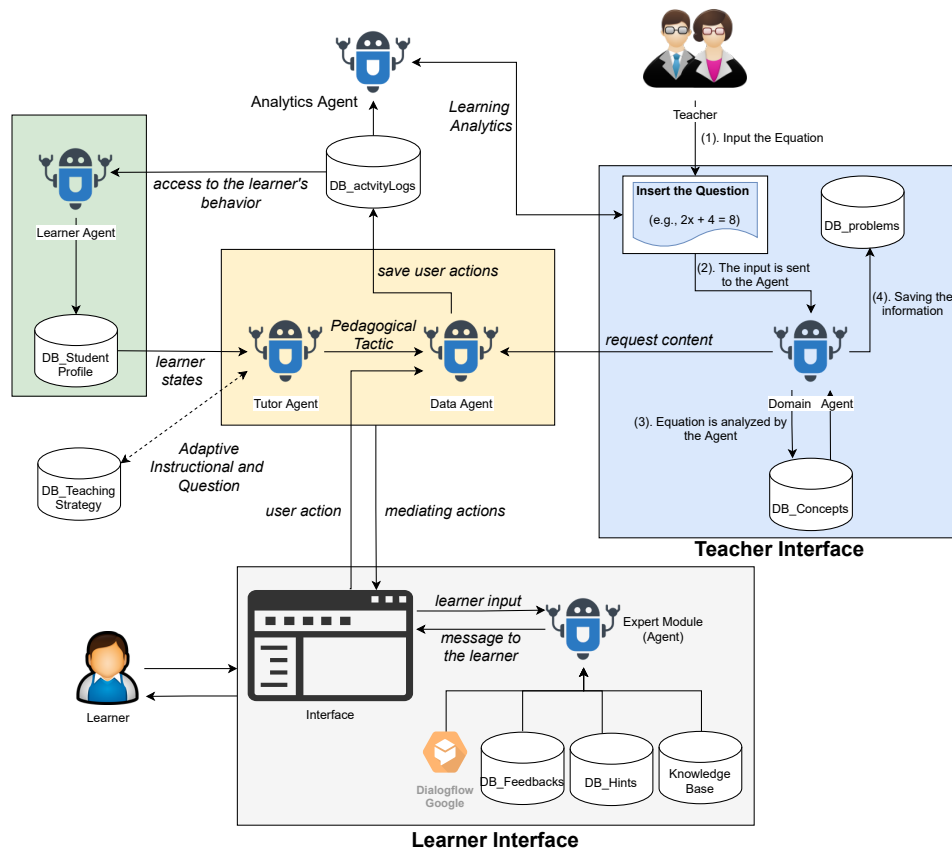
A proposta partiu da arquitetura já existente do STI (Lopes et al., 2019), que permitia a interação passo a passo do estudante durante a resolução de equações. No entanto, a sequência de problemas ainda era previamente definida, limitando a adaptação das atividades às necessidades de cada estudante. Assim, com base nas lacunas observadas no mapeamento e nas limitações identificadas no uso inicial do LEIA, Lopes e Netto (2021) propuseram a inclusão de novos agentes pedagógicos, cujas funções seriam voltadas à organização do domínio, à modelagem do estudante e à condução das estratégias tutoriais. A Figura 28 representa a arquitetura geral proposta e sintetiza a relação entre os agentes, as bases de dados, as interfaces do professor e do aluno, e os mecanismos de devolutivas, dicas e recomendações.

5.1.1 Objetivo do estudo

O objetivo deste estudo foi projetar uma arquitetura multiagente para apoiar a recomendação de atividades e a intervenção pedagógica com base nas ações dos estudantes. A proposta visou aprimorar o processo de acompanhamento do estudante, considerando a necessidade de sequenciar exercícios, identificar dificuldades durante a resolução e oferecer intervenções compatíveis com o nível de conhecimento do aprendiz.

Esse objetivo foi definido a partir de duas limitações observadas na versão anterior do sistema em (Lopes et al., 2019). A primeira estava relacionada à organização das atividades, uma vez que os problemas matemáticos e a sequência de instruções eram previamente definidos, sem adaptação direta às habilidades matemáticas de cada estudante. A segunda limitação estava associada ao trabalho do professor na construção e classificação das equações, pois a elaboração de atividades para diferentes níveis de dificuldade exige tempo e conhecimento sobre a estrutura conceitual do domínio. Assim, o estudo teve como propósito estruturar uma solução que auxiliasse tanto o estudante quanto o professor.

Figura 28 – Arquitetura Multiagente proposta por Lopes e Netto (2021).



Fonte: Lopes e Netto (2021).

Para os estudantes, a arquitetura deveria recomendar atividades compatíveis com seus perfis e oferecer apoio durante a resolução. Para o professor, a proposta deveria facilitar a organização das equações, o acompanhamento das interações e a identificação de dificuldades recorrentes. Portanto, o estudo não se concentrou apenas na automação da tutoria, mas também na criação de uma base arquitetural para integrar recomendação, monitoramento e intervenção pedagógica no mesmo ambiente.

5.1.2 Estrutura proposta

A estrutura proposta (ver Figura 28) foi organizada a partir de uma abordagem multiagente. Essa escolha permitiu distribuir as responsabilidades do sistema em componentes especializados, mantendo a possibilidade de inserção de novos agentes e de expansão de suas funcionalidades. A arquitetura inicia a partir dos agentes já existentes: o *Data Agent*, o *Expert Module* e o *Analytics Agent*. O *Data Agent* é responsável por manter o modelo do estudante, registrar o histórico de ações do aprendiz e gerenciar o ciclo de tutoria. O *Expert Module* analisa as soluções das equações passo a passo, avaliando operações como divisão,

adição de termos semelhantes e soma até a obtenção do valor da incógnita. Já o *Analytics Agent* produz relatórios de desempenho para o professor, incluindo o número de equações resolvidas, os passos trabalhados e o tempo de interação com as atividades. O *Domain Agent* foi proposto para auxiliar o professor na organização das atividades. Sua função é receber a equação inserida pelo professor, analisar sua estrutura e associá-la aos conceitos do domínio.

Além de identificar os conceitos, o *Domain Agent* organiza as equações por nível de dificuldade. As atividades do estudo, foram classificadas como fácil, médio e difícil, considerando as propriedades matemáticas envolvidas em sua resolução, tais como equivalência, associatividade, elemento inverso, elemento identidade e aplicação da ordem das operações. Essa organização possibilita que o sistema utilize as atividades de maneira mais controlada, considerando tanto a estrutura algébrica da equação quanto o perfil do estudante.

O *Learner Agent* foi projetado para modelar o estudante. Ele utiliza registros da base de interações para estimar o nível de conhecimento do estudante em relação aos conceitos do domínio. A modelagem é baseada no *Overlay Model*, no qual o conhecimento do estudante é tratado como um subconjunto do conhecimento representado pelo sistema. Dessa forma, o perfil do estudante passa a registrar acertos, erros e habilidades relacionadas aos conceitos trabalhados. Essa estrutura viabiliza para que o sistema identifique estudantes iniciantes, intermediários ou com desempenho satisfatório, segundo o domínio adquirido nas atividades.

O *Tutor Agent* foi definido como o agente responsável por conduzir o ciclo tutorial e selecionar estratégias pedagógicas de acordo com as informações produzidas pelo *Learner Agent*. Esse agente utiliza o nível de conhecimento do estudante para indicar ao *Data Agent* quais atividades e intervenções devem ser disponibilizadas. O estudo adotou a teoria sociocultural de Vygotsky como base, em particular a ideia de mediação e Zona de Desenvolvimento Proximal. Nesse sentido, o STI atua como um parceiro mais capaz, oferecendo apoio ao estudante durante a resolução das atividades.

5.1.3 Resultados alcançados

Os resultados do estudo foram de natureza arquitetural e conceitual. O principal resultado foi a definição de uma arquitetura multiagente capaz de ampliar o STI LEIA, por meio da incorporação de mecanismos de recomendação, modelagem do estudante e intervenção pedagógica. Essa arquitetura estabeleceu uma separação funcional entre os agentes responsáveis pelo domínio, pelo estudante, pela tutoria, pelos dados, pela análise das soluções e

pelos relatórios de desempenho.

Outro resultado foi a estruturação do *Domain Agent*, que permitiu a formalização de uma primeira organização de equações de primeiro grau com base em conceitos e níveis de dificuldade. Essa organização representou um avanço em relação à inserção de exercícios no sistema, pois passou a permitir que as atividades fossem tratadas como objetos associados a conceitos matemáticos específicos.

O estudo também resultou na proposição do *Learner Agent*, responsável por converter os registros de interação em informações sobre o nível de conhecimento do estudante. Essa modelagem viabiliza representar o desempenho a partir de acertos, erros e conceitos trabalhados, criando uma base para decisões pedagógicas posteriores. Dessa forma, a arquitetura passou a considerar o comportamento do estudante durante a resolução das atividades, e não apenas o resultado final da resposta.

A proposição do *Tutor Agent* também foi um resultado relevante para esta tese, pois estabeleceu a ligação entre o estado do estudante e as estratégias de ensino. Esse agente associa o perfil do aprendiz às estratégias de intervenção, como retornos pedagógicos, dicas e apoio estruturado (*scaffolding*). Além disso, o módulo de recomendação foi outro produto do estudo. A partir da relação entre a base de problemas e o perfil do estudante, o sistema adquiriu uma lógica para selecionar atividades conforme o nível de dificuldade e o desempenho observado.

Em síntese, o Estudo 1 teve caráter predominantemente conceitual e arquitetural e não envolveu, nessa etapa, uma aplicação experimental da arquitetura proposta com uma amostra definida de estudantes. A proposta partiu de componentes já existentes no STI LEIA, como o Agente de Dados, o Módulo Especialista e o Agente de Análise, e incluiu os agentes de Domínio, Estudante e Tutor para ampliar os mecanismos de organização do domínio, de modelagem do estudante e de intervenção pedagógica.

Para exemplificar a organização do domínio, foram consideradas nove equações polinomiais de primeiro grau, distribuídas em três níveis de dificuldade – fácil, médio e difícil –, envolvendo conceitos como equivalência, multiplicação, associatividade, elementos inversos, elemento identidade e ordem das operações. Essas atividades tiveram função ilustrativa na especificação da arquitetura e não constituíram um conjunto experimental aplicado a estudantes (Lopes e Netto, 2021).

5.1.4 Limitações identificadas

Apesar dos avanços arquiteturais, o Estudo 1 teve caráter predominantemente conceitual e arquitetural. Não foi conduzido, nessa etapa, um experimento formal com uma amostra definida de estudantes nem foi estabelecido um conjunto fechado de itens para avaliação experimental da arquitetura. A proposta foi desenvolvida no domínio de equações polinomiais de primeiro grau, considerando atividades organizadas em diferentes níveis de dificuldade e conceitos matemáticos associados à sua resolução. Dessa forma, os resultados desse estudo devem ser interpretados como evidências da viabilidade arquitetural da proposta, e não como uma avaliação empírica de seus efeitos sobre a aprendizagem.

Outra limitação está relacionada à modelagem do estudante. A proposta considerou níveis gerais de conhecimento e assumiu que os estudantes poderiam ser tratados segundo uma estrutura comum de aprendizagem. No entanto, estilos de aprendizagem, diferenças individuais, conhecimentos prévios e dificuldades específicas podem variar entre os alunos. Essa limitação indicou a necessidade de modelos mais aprimorados para representar o domínio de habilidades, motivando estudos posteriores com ADC.

Foi identificada também uma limitação relacionada à formalização das habilidades matemáticas. A estrutura dos conceitos e os níveis de dificuldade dependem da forma como o conteúdo é organizado, podendo variar de acordo com o currículo, o país, a série escolar e a interpretação dos especialistas. Por isso, é necessário um processo mais sistemático para definir habilidades, relacioná-las aos itens e validar se essas relações representam adequadamente as operações cognitivas exigidas pelas atividades.

Além disso, a infraestrutura tecnológica e as condições de aplicação em contexto remoto também foram apontadas como fatores limitantes. Dificuldades de acesso, disponibilidade de equipamentos e condições de uso podem restringir a aplicação do sistema e a realização de testes com um número maior de estudantes. Essas limitações são relevantes, pois o estudo foi motivado por desafios observados no ensino remoto e pela necessidade de apoiar professores e os alunos em ambientes mediados por tecnologia.

Por fim, o estudo concentrou-se no domínio de equações polinomiais de primeiro grau. Embora esse recorte tenha oportunizado uma modelagem mais controlada, a arquitetura ainda precisava ser investigada em outros tipos de problemas, tais como os problemas contextualizados e os de domínios com maior complexidade conceitual. Essa limitação reforçou a necessidade de avançar para uma estrutura mais geral de modelagem de habilidades, capaz

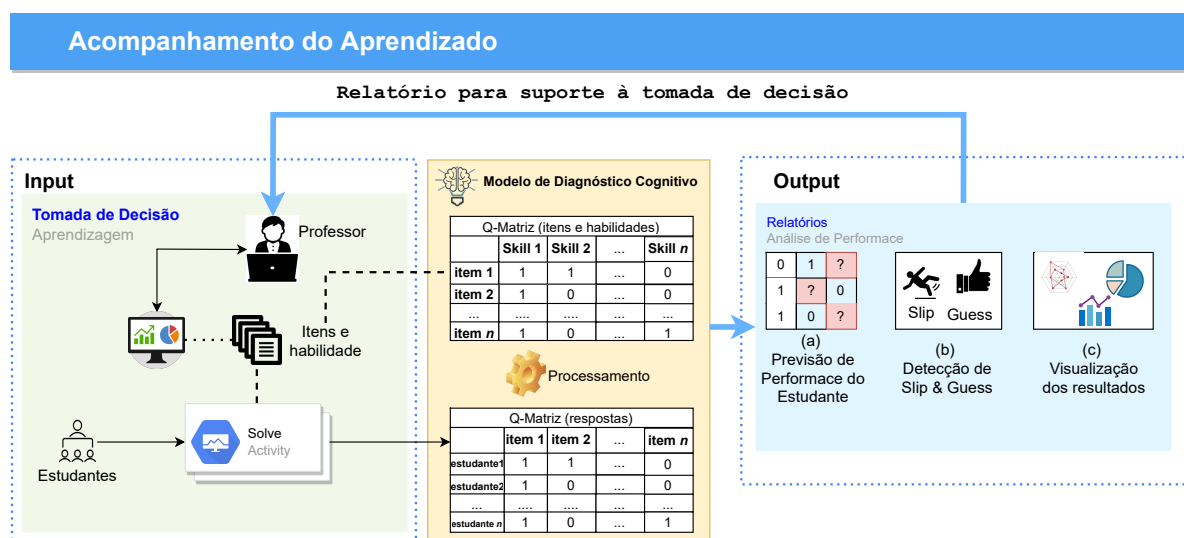
de apoiar a identificação de lacunas cognitivas e a recomendação de atividades em diferentes contextos educacionais.

5.2 Estudo 2: Incorporação do MDC G-DINA em um STI

O segundo resultado desta tese corresponde à incorporação de uma abordagem de avaliação diagnóstica cognitiva ao STI LEIA, com uso do MDC G-DINA. Diferentemente do estudo anterior, que se concentrou na proposição arquitetural de agentes pedagógicos para recomendação e intervenção, este estudo avançou na avaliação das habilidades cognitivas dos estudantes a partir das respostas registradas durante a resolução de equações polinomiais de primeiro grau.

A proposta teve como foco ampliar a capacidade diagnóstica do STI, de modo que o sistema pudesse não apenas registrar os acertos e erros, mas também estimar o domínio de habilidades específicas envolvidas nas atividades. Para tanto, foram definidos atributos de habilidades; depois estruturada uma Q-matriz e incorporado o MDC G-DINA ao fluxo de avaliação do STI LEIA. A Figura 29 apresenta o modelo de gerenciamento de atividades e o diagnóstico cognitivo incorporado ao sistema, articulando a entrada de dados, seu processamento pelo Modelo de Diagnóstico e a saída de resultados em formato de relatórios e visualizações (*dashboards*).

Figura 29 – Modelo de gerenciamento de atividades e diagnóstico cognitivo para uso no Sistema Tutor Inteligente.



Fonte: Lopes e Netto (2023).

5.2.1 Objetivo do estudo

O objetivo deste estudo foi incorporar uma abordagem baseada no MDC em um STI, com o propósito de avaliar o domínio de habilidades cognitivas dos estudantes. A pesquisa buscou fornecer aos professores informações mais detalhadas sobre o estado de conhecimento dos estudantes, possibilitando a identificação de pontos fortes e lacunas de aprendizagem.

A proposta parte do entendimento de que as habilidades cognitivas são construtos latentes, ou seja, não podem ser observadas diretamente. Por isso, a simples análise de respostas corretas ou incorretas não permite compreender com precisão quais habilidades foram dominadas ou quais ainda precisam ser aprimoradas. Nesse contexto, o uso de um MDC foi proposto como mecanismo de inferência para estimar probabilisticamente o domínio para cada habilidade associada aos itens avaliativos.

Diante disso, o estudo teve uma dupla finalidade. Do ponto de vista computacional, buscou-se adaptar o STI LEIA para receber uma estrutura de avaliação baseada em habilidades e apresentar os resultados em um *dashboard*. Do ponto de vista pedagógico, o objetivo foi fornecer subsídios para que o professor pudesse planejar intervenções com base nas habilidades com menor percentual de domínio.

5.2.2 Definição dos CC e da Q-matriz

A definição dos CC foi realizada por meio de uma análise qualitativa envolvendo documentos da literatura e a colaboração de professores de Matemática. O domínio escolhido foi a resolução de equações polinomiais de primeiro grau, por ser um conteúdo recorrente no Ensino Fundamental e relevante para a iniciação do estudo de Álgebra.

Seguindo o modelo apresentado na Figura 29 foram definidos cinco atributos de habilidades, representados por α_1 , α_2 , α_3 , α_4 e α_5 . O atributo α_1 corresponde às propriedades de divisão e multiplicação da igualdade. O atributo α_2 refere-se à simplificação de fração ao menor termo. O atributo α_3 envolve a solução de equações com constantes em ambos os lados. O atributo α_4 corresponde à solução de equações com variáveis e constantes em ambos os lados da igualdade.

Por fim, o atributo α_5 refere-se à aplicação da propriedade distributiva. A Tabela 1 apresenta a relação entre os dez itens avaliativos e os cinco atributos de habilidades definidos. Essa tabela funciona como a base da Q-matriz do estudo, indicando quais habilidades são exigidas em cada equação. Por exemplo, itens mais simples, como $2x = 8$, demandam apenas o atributo α_1 , enquanto itens mais complexos, como $7(8x - 6) = 8(8x - 5)$, exigem o conjunto

completo de habilidades avaliadas.

Tabela 1 – Atributos de habilidades relacionados às operações de resolução de equação polinomial de primeiro grau.

Itens		Atributos de habilidades					Itens		Atributos de habilidades				
		$\alpha 1$	$\alpha 2$	$\alpha 3$	$\alpha 4$	$\alpha 5$			$\alpha 1$	$\alpha 2$	$\alpha 3$	$\alpha 4$	$\alpha 5$
e1	$2x = 8$	1	0	0	0	0	e6	$5x + 6 = 46$	1	1	1	0	0
e2	$3x = 99$	1	1	0	0	0	e7	$2x - 1 = 65 - 4x$	1	1	1	1	0
e3	$6x = 92$	1	1	0	0	0	e8	$6x + 7 = 21 - 4x$	1	1	1	1	0
e4	$7x = 55$	1	0	0	0	0	e9	$6(9x - 9) = 4(8x - 8)$	1	1	1	1	1
e5	$2x - 62 = 28$	1	1	1	0	0	e10	$7(8x - 6) = 8(8x - 5)$	1	1	1	1	1

* (1 – contém atributo de habilidade, 0 – não contém)

Fonte: [Lopes e Netto \(2023\)](#).

A metodologia de avaliação dos estudantes no STI seguiu o mesmo processo descrito na Figura 6 que ilustra a metodologia empregada na avaliação por MDC. Nesse tipo de avaliação, os itens são correlacionados aos atributos de habilidades, e as respostas dos estudantes são usadas para estimar o domínio. No estudo, foram utilizadas duas matrizes: a primeira é a Q-matriz que representa a relação entre itens e os atributos, e a segunda matriz armazena as respostas dos estudantes aos itens avaliativos.

Para garantir o gerenciamento dessa estrutura, no STI, foi desenvolvida uma interface para que o professor pudesse inserir atividades e selecionar as habilidades associadas a cada uma. Essa funcionalidade é apresentada na Figura 30 a seguir.

Figura 30 – Interface de autoria do professor, voltada ao cadastro de atividades no STI e à associação dos CC a serem avaliadas pelo MDC.

Fonte: [Lopes e Netto \(2023\)](#).

As funcionalidades mostradas na Figura 30 foram implementadas como uma extensão do STI LEIA, permitindo ao professor cadastrar as atividades e explicitar as relações entre cada item e os CC avaliados.

5.2.3 Aplicação com os estudantes

A aplicação do estudo foi realizada com uma turma de Matemática do 9º ano do Ensino Fundamental da Escola Estadual Castelo Branco, localizada no município de Manacapuru, no estado Amazonas.

O experimento foi conduzido em 2022 e contou com 19 estudantes do 9º ano do Ensino Fundamental e com a participação de um professor de Matemática. Foram utilizados dez itens avaliativos relacionados a cinco atributos de habilidades no domínio de equações polinomiais de primeiro grau. O STI LEIA foi utilizado para aplicação das atividades e registro das interações dos estudantes. A versão utilizada do sistema foi desenvolvida com o *framework* Laravel¹, executada em ambiente Docker e utilizou MySQL para persistência dos dados.

A análise quantitativa foi realizada por meio do modelo G-DINA, utilizando o pacote *G-DINA*² na linguagem R, versão 4.1.2, em ambiente RStudio³. O processamento foi executado em um notebook Lenovo IdeaPad 330S, equipado com processador Intel Core i7 e 8 GB de memória RAM. As respostas registradas pelo STI foram utilizadas na composição da matriz de respostas e, juntamente com a Q-matriz previamente definida, constituíram as entradas para a estimação dos perfis de domínio das habilidades.

O procedimento iniciou com a aplicação de uma avaliação de pré-teste, composta por questões semelhantes às atividades organizadas na Tabela 1. Em seguida, os estudantes utilizaram o STI para solucionar as atividades propostas. Durante essa interação, o sistema registrou as respostas e, posteriormente, foram processadas pelo Módulo de Avaliação Cognitiva. Por conseguinte, os dados são enviados para análise do MDC G-DINA, que dá como saída os percentuais de domínio individual de habilidades para os estudantes.

Após a análise diagnóstica, o professor usou as informações produzidas pelo sistema para planejar uma aula de revisão. Essa aula foi direcionada aos estudantes com menor desempenho e às habilidades que apresentaram menor percentual de domínio. Após essa etapa de intervenção, foi aplicada uma avaliação pós-teste (ver Apêndice A), com o nível de dificuldade semelhante ao pré-teste.

¹ <https://laravel.com/>

² <https://wenchao-ma.github.io/GDINA/>

³ <https://docs.posit.co/ide/user/#rstudio-ide-oss-downloads>

Esse estudo permitiu observar o uso do MDC em um fluxo pedagógico completo: definição das habilidades, aplicação das atividades, processamento diagnóstico, visualização dos resultados, intervenção docente e comparação entre pré- e pós-teste.

5.2.4 Resultados

Os resultados diagnósticos são apresentados em percentuais de domínio das habilidades avaliadas. A Tabela 2 do artigo [Lopes e Netto \(2023\)](#) se destaca, pois apresenta as estimativas individuais geradas por inferência do MDC G-DINA. A Tabela mostra que o modelo permitiu diferenciar perfis de domínio entre os participantes, evidenciando quais habilidades estavam consolidadas e quais exigiam maior atenção pedagógica.

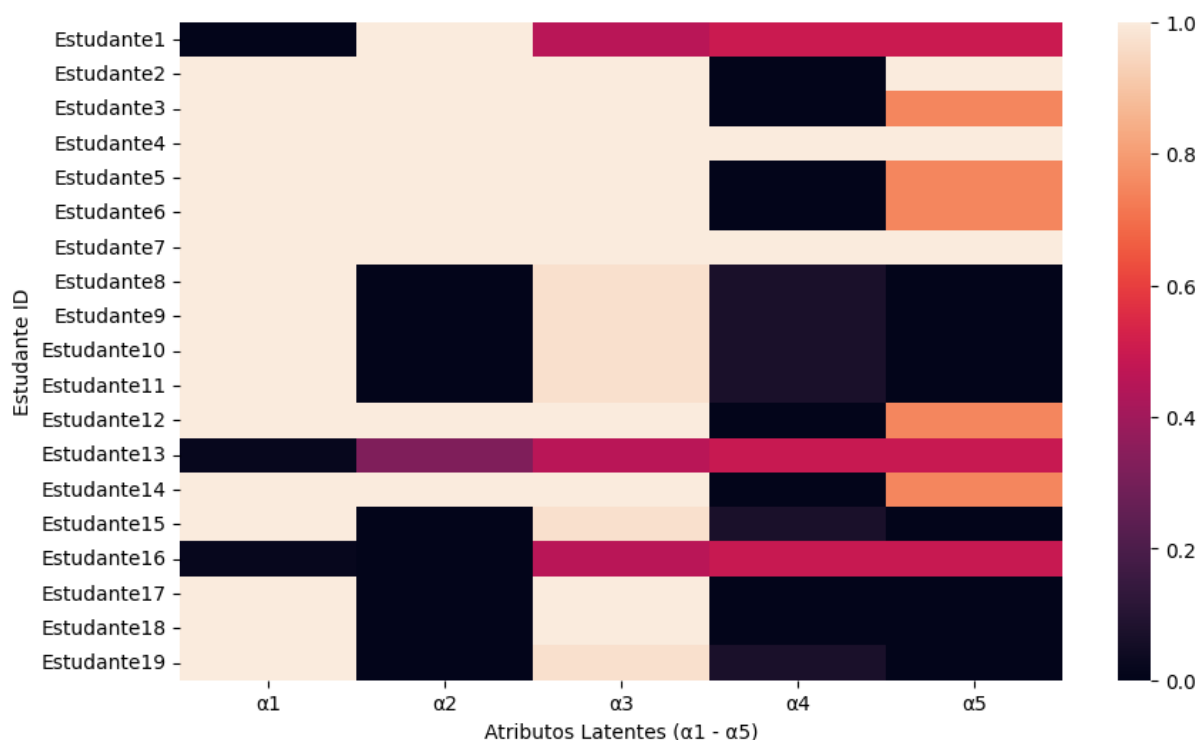
Em geral, aproximadamente 90% do grupo alcançou o domínio do atributo α_1 ; 47% dominaram a habilidade α_2 ; 94% dominaram o α_3 , 36% dominaram α_4 e 57, e 89% dominaram α_5 . Esses resultados indicam que as maiores dificuldades do grupo estavam relacionadas à habilidade α_4 , que envolve a solução de equações com variáveis e constantes em ambos os lados da igualdade, e à habilidade α_2 , relacionada a simplificação de frações.

A Figura 31 apresenta o mapa de calor das estimativas de domínio produzidas pelo MDC G-DINA para cada estudante em relação aos atributos latentes avaliados. As linhas representam os estudantes e as colunas representam os atributos α_1 a α_5 . A escala varia de 0 a 1, sendo que valores mais próximos a 1 indicam maior probabilidade de domínio da habilidade, enquanto valores próximos de 0 indicam baixa estimativa de domínio.

A Figura 31 mostra a distribuição dos perfis de conhecimento no grupo. Os atributos α_1 e α_3 apresentam maior concentração de valores elevados, indicando que a maioria dos estudantes demonstrou domínio dessas habilidades. Em contraste, o atributo α_4 apresenta maior ocorrência de valores baixos, confirmando a dificuldade do grupo em resolver equações com variáveis e constantes em ambos os lados da igualdade. O atributo α_2 também apresenta variação relevante entre os estudantes, com um grupo demonstrando domínio elevado e outro grupo com baixa estimativa, o que sugere dificuldades na simplificação de frações ao menor termo.

A visualização também evidencia diferenças individuais. Os estudantes identificados com *Estudante4* e o *Estudante7* apresentam domínio elevado em praticamente todos os atributos, enquanto os identificados *Estudante13* e o *Estudante16* apresentam estimativas mais baixas e distribuídas de forma irregular. Esses resultados reforçam a utilidade do diagnóstico por habilidade, pois permitem identificar não apenas o desempenho geral da turma, mas tam-

Figura 31 – Pontuação das habilidades dominadas pelos alunos participantes medida pelo MDC G-DINA.



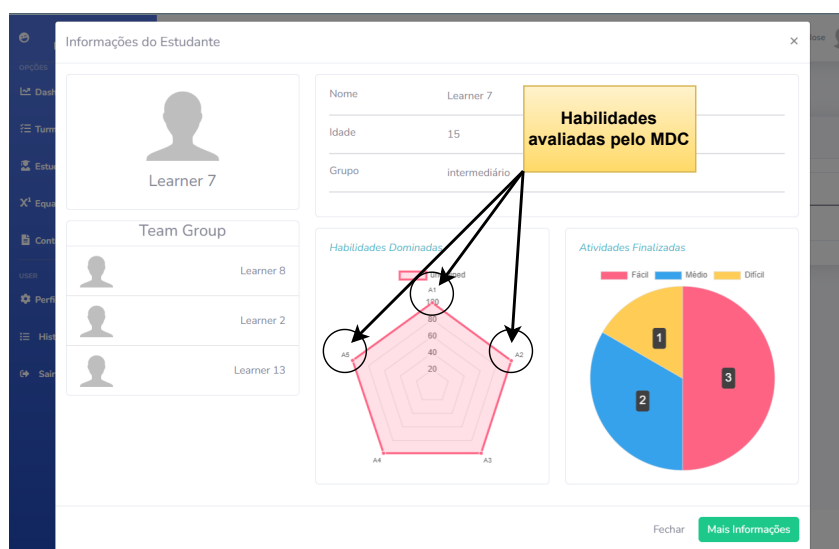
Fonte: Elaborado pelos autores (2026).

bém os estudantes e os atributos que exige maior atenção no planejamento de intervenções pedagógicas.

Com base nos dados apresentados na Figura 31, é possível observar que a maioria dos estudantes dominou com sucesso a maior parte das habilidades propostas. No entanto, observa-se uma tendência a um coeficiente de domínio relativamente baixo para o atributo α_4 , o que indica que muitos aprendizes têm dificuldade em compreender o conceito de distribuição de termos semelhantes em ambos os lados da equação. Os resultados foram plotados nos perfis individuais dentro do STI. Para apresentá-los de maneira didática, os resultados foram plotados usando um gráfico de radar, conforme mostra o exemplo da Figura 32.

Após a realização de toda a intervenção, os resultados do pré-teste e do pós-teste foram comparados por meio do *t*-teste pareado⁴. Em Lopes e Netto (2023), a média do pré-teste foi 3,10, indicando o desempenho médio inicial da turma antes da intervenção. O desvio padrão foi 3,05, o que mostra uma dispersão elevada das notas em relação à média, sugerindo que

⁴ https://rpubs.com/henriquealvarenga/t_test

Figura 32 – Painel do perfil de habilidades do estudante *Estudante7* inferido pelo MDC.

Fonte: [Lopes e Netto \(2023\)](#).

havia estudantes com desempenhos bastante diferentes no início da aplicação.

No pós-teste, a média aumentou para 5,47, indicando melhora no desempenho médio após o uso do STI e a intervenção pedagógica orientada pelos resultados diagnósticos. O desvio padrão foi 3,57, também elevado, mostrando que, embora a média da turma tenha aumentado, os desempenhos continuaram variando entre os estudantes. Assim, o resultado sugere avanço geral no grupo, mas com diferenças individuais ainda presentes.

O *t*-teste pareado indicou diferença estatisticamente significativa entre as médias, com $t(19) = -3,41$ e $p = 0,0031143$. Como o valor de p foi inferior a 0,05, rejeita-se a hipótese de que não houve diferença entre pré-teste e pós-teste. Portanto, os dados indicam que a melhora observada no pós-teste não ocorreu apenas por variação aleatória, mas está associada ao processo de intervenção realizado.

Estes resultados indicam que a incorporação do MDC G-DINA ao STI LEIA permitiu identificar lacunas de conhecimento e orientar uma intervenção pedagógica mais direcionada. O estudo também mostrou que os dados diagnósticos puderam ser integrados ao painel do sistema, facilitando a visualização do desempenho dos estudantes pelo professor.

5.2.5 Limitações identificadas

No entanto, apesar dos resultados obtidos, foram identificadas algumas limitações. A primeira está relacionada com o número de itens avaliativos. Conforme investigado por [Lopes e Netto \(2023\)](#), os MDC requerem uma quantidade maior de itens avaliativos para um melhor ajuste do modelo G-DINA. Neste estudo, foram utilizados dez itens, devido a restrições de

tempo e de recursos de *hardware*, bem como à disponibilidade dos estudantes.

Outra limitação está associada ao quantitativo de participantes. O estudo foi realizado com 19 estudantes de uma única turma. Embora os resultados tenham sido compatíveis com o comportamento observado durante a aplicação dos testes, a quantidade de alunos limita a generalização dos achados para outros contextos escolares ou domínios de conhecimento.

É também necessário ter em conta que o estudo foi aplicado apenas ao domínio de equações polinomiais de primeiro grau. Esta seleção favoreceu o controle da modelagem das habilidades, mas ainda não permite afirmar que a abordagem produzirá o mesmo tipo de resultado em conteúdos com estruturas conceituais mais amplas ou menos lineares.

Por fim, [Lopes e Netto \(2023\)](#) indicam a necessidade de investigar outros MDC e novas abordagens baseadas em atividades orientadas a habilidades, incluindo modelos híbridos entre métodos estatísticos e técnicas de aprendizagem de máquina. Essa continuidade é necessária para melhorar a precisão do diagnóstico e ampliar a autonomia do STI na avaliação e na recomendação pedagógica.

5.3 Estudo 3: Geração e avaliação de matriz-Q por LLMs

O terceiro estudo desta tese investigou o uso de LLM para criação automática de Q-matriz, com foco na associação entre itens avaliativos e os CC. O estudo parte da limitação recorrente na construção de Q-matrizes que está associada à necessidade de se consultar especialistas para definir quais habilidades ou conceitos são exigidos por cada item. Embora esse processo seja necessário para a aplicação do MDC, esse processo pode ser demorado, suscetível a divergências entre os avaliadores e pouco escalável em caso de grande volume de itens.

Neste estudo, a Q-matriz é tratada como uma representação interpretável da relação entre item e CC. Dessa forma, a criação automática por LLM foi formulada como uma tarefa de classificação multirótulo, em que cada item pode ser associado a um ou mais CC. A contribuição do estudo está na avaliação em dois níveis: a avaliação estrutural, que compara a Q-matriz gerada pelo LLM com uma Q-matriz definida por especialistas, e a funcional, que se verifica o efeito dessas matrizes no desempenho preditivo baseado em um MDC.

5.3.1 Objetivo do estudo

O objetivo foi avaliar se as LLM podem gerar Q-matrizes estruturalmente coerentes e funcionalmente úteis para aplicações de diagnóstico cognitivo. A proposta buscou responder

se as associações item–CC produzidas por modelos de linguagem se aproximam das associações definidas por especialistas, bem como se essas matrizes preservam o desempenho preditivo de um MDC.

O estudo foi orientado por três questões de pesquisa. A primeira investigou se os LLM conseguem gerar Q-matrizes compatíveis com os padrões definidos por especialistas. A segunda avaliou se as Q-matrizes geradas automaticamente preservam ou melhoram a acurácia preditiva de um MDC quando comparadas à Q-matriz dos especialistas. A terceira analisou quais estratégias de *prompting* produzem resultados mais confiáveis para a classificação automática de habilidades.

5.3.2 Configuração experimental

A configuração do experimento utilizou itens de Matemática do TIMSS 2011⁵ (Sun et al., 2025), referentes ao 8º ano, totalizando 23 (vinte e três) itens avaliativos e 7 (sete) CC. O estudo não teve a finalidade de descobrir novos CC, mas de associar cada item a um conjunto previamente definido. Assim, a tarefa atribuída aos LLM foi inferir quais CC estavam relacionados a cada item, gerando uma Q-matriz binária. O processo seguiu o roteiro ilustrado na Figura 33.

Conforme demonstrado na Figura 33, na primeira etapa, foram definidos os itens avaliativos (questões do TIMSS) (*E*) e o conjunto fixo de CC, com base em uma Q-matriz de especialista. Em seguida, as LLM receberam os itens e os CC por meio de diferentes estratégias de *prompting*. Logo em seguida, as respostas dos modelos foram convertidas em linhas binárias da Q-matriz gerada. Na quarta etapa, cada Q-matriz foi avaliada estruturalmente e funcionalmente. E na última, os resultados são apresentados. O apêndice B descreve o algoritmo aplicado para este experimento.

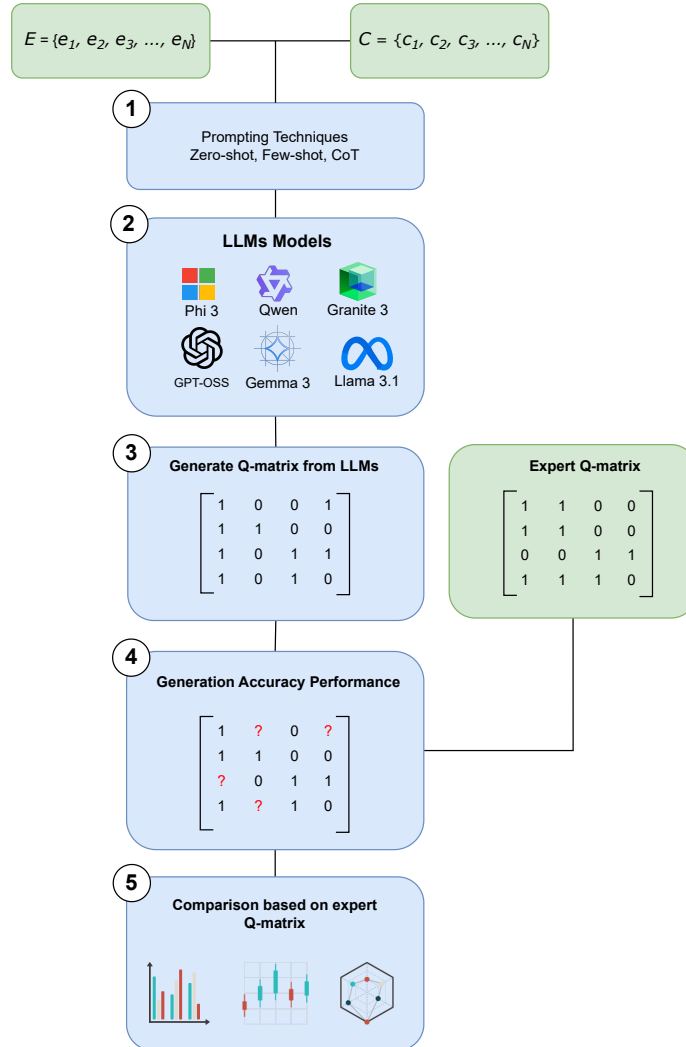
No trabalho de Lopes e Netto (2026) foram avaliadas três estratégias de *prompting*: *zero-shot*, *few-shot* e *few-shot* com CoT. Na estratégia *zero-shot*, o modelo recebeu apenas a instrução de classificação, os itens e a lista de CC. Na estratégia *few-shot*, foram incluídos descrições e exemplos para orientar a associação entre o item e os CC. Já na estratégia de *few-shot* com CoT, além dos exemplos, o modelo foi orientado a raciocinar passo a passo antes de produzir a lista final de CC.

Os experimentos com os LLM foram executados localmente por meio do Ollama Server⁶ e de sua API, em um servidor equipado com processador Intel Core i9, 32 GB de memória RAM

⁵ <https://timssandpirls.bc.edu/timss2011/international-results-mathematics.html>

⁶ <https://ollama.com/>

Figura 33 – Visão geral do processo experimental, desde os itens de avaliação (E) e CC (C) até as Q-matrizes geradas pelo LLM e as subsequentes análises estruturais e estatísticas.



Fonte: [Lopes e Netto \(2026\)](#).

e GPU NVIDIA GeForce RTX 4070 com 8 GB de memória dedicada. Foram avaliados os modelos `phi3:3.8b`, `qwen3:8b`, `gemma3:12b`, `llama3.1:8b`, `granite3-dense:8b` e `gpt-oss:20b`. Os modelos foram utilizados em suas configurações originais, sem ajuste fino específico para o domínio, com temperatura igual a 0,1 visando reduzir a variabilidade das respostas e favorecer saídas mais determinísticas.

A avaliação funcional foi realizada com o modelo DINA por meio da biblioteca EduCDM⁷. Como a matriz de respostas reais correspondente aos itens utilizados não estava disponível,

⁷ <https://github.com/bigdata-ustc/EduCDM>

foram empregadas matrizes binárias simuladas com $N = 50$, $N = 100$ e $N = 200$ estudantes conforme a metodologia orientada para avaliação diagnóstica (Ravand e Baghaei, 2020).

Em cada configuração, os dados foram divididos em 80% para treinamento e 20% para teste, utilizando semente pseudoaleatória fixa (`seed = 42`). O treinamento do DINA foi realizado por cinco épocas. Para reduzir a influência da variabilidade inerente à geração por LLM, cada configuração experimental foi executada dez vezes, sendo calculadas as médias, os desvios-padrão e os intervalos de confiança de 95% das métricas obtidas.

Para fins de reprodutibilidade e verificação dos procedimentos adotados, o código-fonte, os *prompts*, os arquivos de configuração e os demais materiais utilizados na execução dos experimentos estão disponíveis no repositório permanente indicado no Apêndice B.

5.3.3 Avaliação estrutural da Q-matriz

A avaliação estrutural verificou o grau de alinhamento entre as Q-matrizes geradas pelas LLM e a Q-matriz definida por especialistas. Para isso, foram utilizadas duas métricas complementares: *Hamming Loss*⁸ (HL) e *Jaccard Index*⁹ (JSI). O HL mede a proporção de divergências entre as duas matrizes, considerando associações adicionadas indevidamente ou omitidas. Valores menores indicam maior proximidade estrutural. Já o JSI, por sua vez, mede a sobreposição entre associações positivas, considerando apenas os vínculos ativos entre itens e CC. Valores maiores indicam maior similaridade com a Q-matriz do especialista. Os resultados estruturais são apresentados na Figura 34.

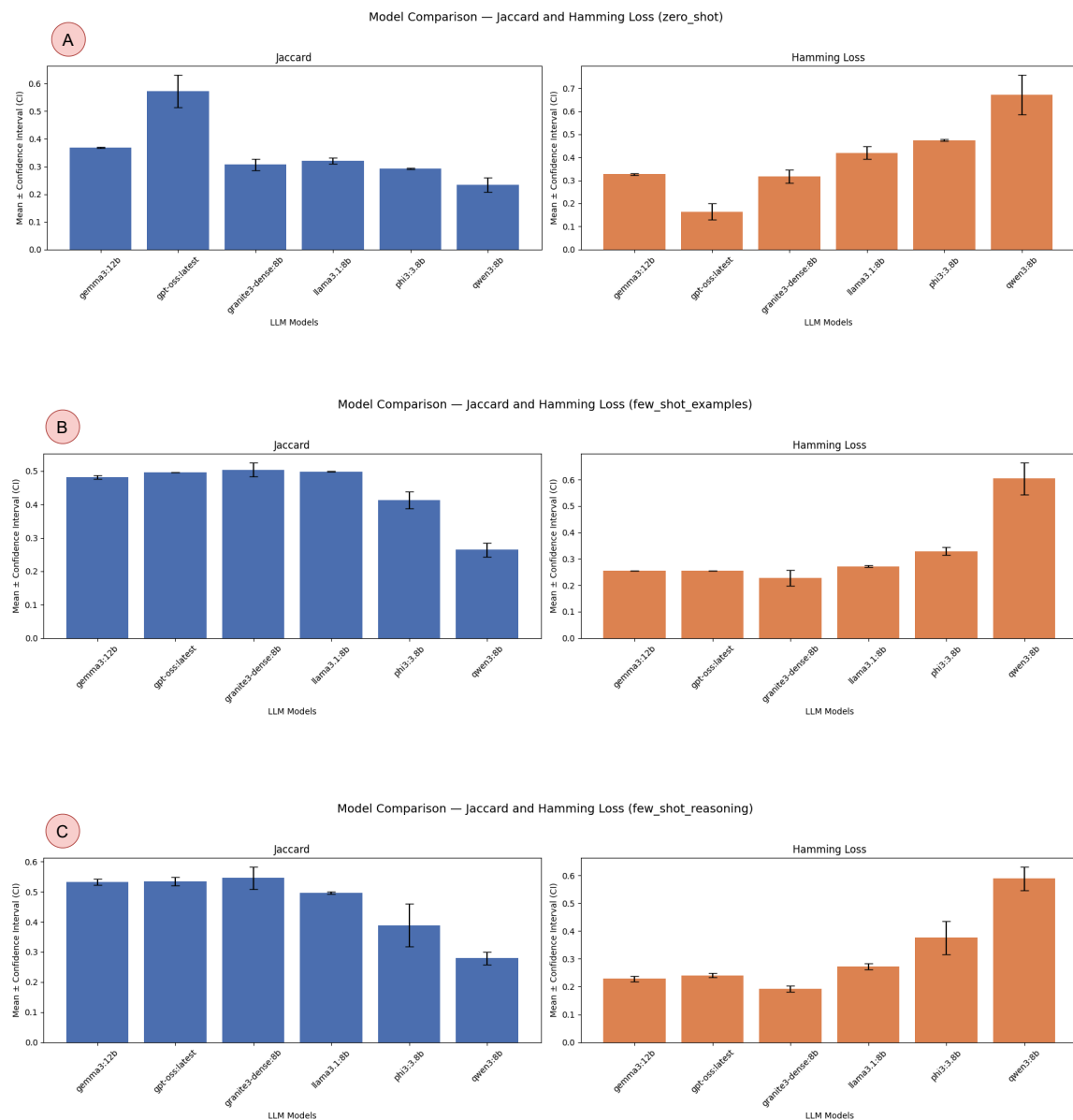
A Figura 34 apresenta os painéis (A), (B) e (C). O painel (A) refere-se à configuração sem exemplos prévios (*zero-shot*); o painel (B), à configuração com poucos exemplos (*few-shot*); e o painel (C), à configuração com poucos exemplos associada ao raciocínio explícito (CoT). Em cada um, o gráfico à esquerda apresenta o JSI, que mede o grau de sobreposição entre as associações item-CC nas Q-matrizes geradas pelo LLM e nas definidas por especialistas. O gráfico à direita apresenta a distância de HL, que quantifica as divergências estruturais entre essas matrizes. As barras representam os valores médios obtidos entre as execuções, e as barras de erro indicam os intervalos de confiança de 95%. Valores mais altos de JSI e mais baixos de HL indicam maior alinhamento estrutural com a Q-matriz definida por especialistas.

No cenário *zero-shot* (A), observou-se maior variabilidade entre os modelos, o que indica que a ausência de exemplos leva os LLM a produzir associações mais instáveis. Nesse cenário,

⁸ https://scikit-learn.org/stable/modules/generated/sklearn.metrics.hamming_loss.html

⁹ https://scikit-learn.org/stable/modules/generated/sklearn.metrics.jaccard_score.html

Figura 34 – Comparação estrutural das Q-matrizes geradas por LLM sob diferentes estratégias de *prompts*.



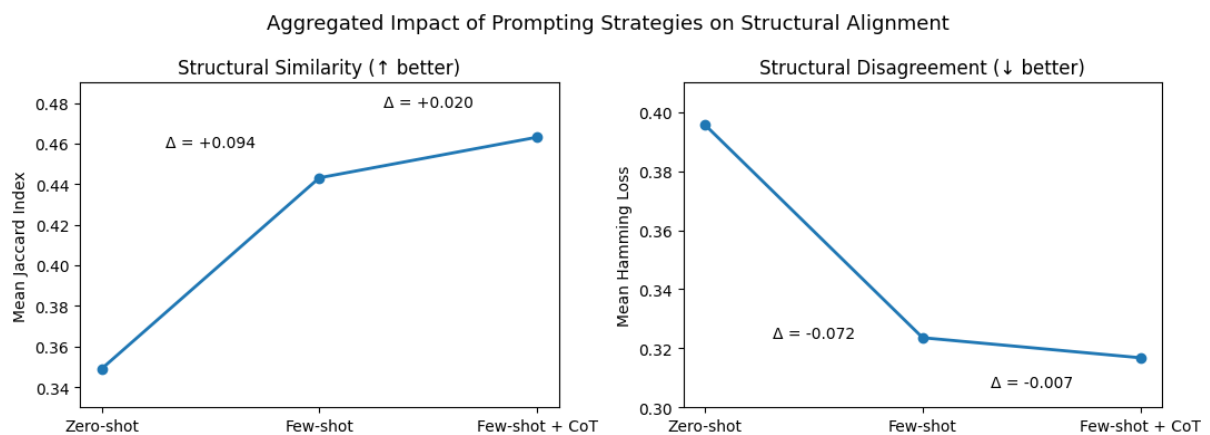
Fonte: [Lopes e Netto \(2026\)](#).

o modelo GPT-OSS apresentou o melhor alinhamento estrutural, com maior JSI e menor HL entre os modelos avaliados.

A transição para a estratégia *few-shot* (B) produziu melhoras estruturais para a maioria dos modelos. A inclusão de exemplos reduziu a quantidade de associações divergentes e aumentou a sobreposição com a Q-matriz do especialista. Esse resultado indica que os exemplos fornecidos no *prompt* funcionaram como orientação estrutural para a classificação dos CC.

A estratégia *few-shot* com CoT mostrada no painel (C) trouxe ganhos adicionais, porém menores que os observados na transição de *zero-shot* para *few-shot*. A média do JSI passou de 0,3490 no *zero-shot* para 0,4431 no *few-shot* e 0,4631 no *few-shot* com CoT. O HL foi reduzido de 0,3957 para 0,3236 e, posteriormente, para 0,3168. A Figura 35 mostra essa evolução, apresentando o ganho progressivo de alinhamento estrutural entre as estratégias de *prompt*.

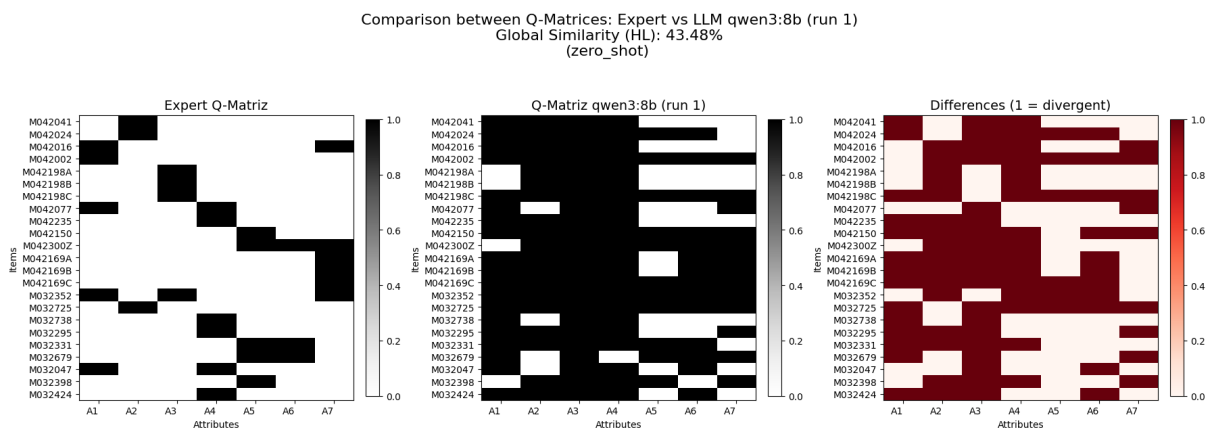
Figura 35 – Melhoria progressiva no alinhamento estrutural das Q-matrizes geradas por LLM com diferentes estratégias de *prompts*.



Fonte: Lopes e Netto (2026).

A Figura 36 complementa a análise estrutural, apresentando um exemplo visual da divergência entre a Q-matriz do especialista e a Q-matriz gerada pela LLM no cenário *zero-shot*.

Figura 36 – Comparação ilustrativa entre uma Q-matriz definida por especialistas (à esquerda) e uma Q-matriz gerada por LLM sem nenhum exemplo, produzida por *qwen3 : 8b* (ao centro), juntamente com uma matriz de diferença binária (à direita), onde o valor 1 indica associações divergentes entre item e atributo.



Fonte: Lopes e Netto (2026).

A Figura 36 mostra a matriz especialista, a matriz gerada e a matriz de diferenças, evi-

denciando associações divergentes entre itens e atributos. Essa figura é importante para demonstrar que a divergência estrutural pode ser resultado da omissão de atributos ou da atribuição de CC aos itens não previstos pelos especialistas.

5.3.4 Avaliação funcional

A avaliação funcional investigou se as Q-matrizes geradas por LLM mantinham desempenho preditivo adequado quando utilizadas em um MDC. Para tanto, cada Q-matriz foi utilizada como entrada no modelo DINA. O desempenho foi avaliado por meio das métricas RMSE e MAE, calculadas a partir da diferença entre as respostas observadas e as probabilidades previstas pelo modelo. Valores menores indicam melhor desempenho preditivo.

A Tabela 2 apresenta a avaliação funcional das matrizes-Q especialistas e das Q-matrizes geradas por LLM, considerando diferentes modelos, estratégias de *prompting* e tamanhos de amostra simulada. O melhor resultado global está destacado em vermelho, enquanto o melhor resultado dentro de cada estratégia de estímulo está em negrito.

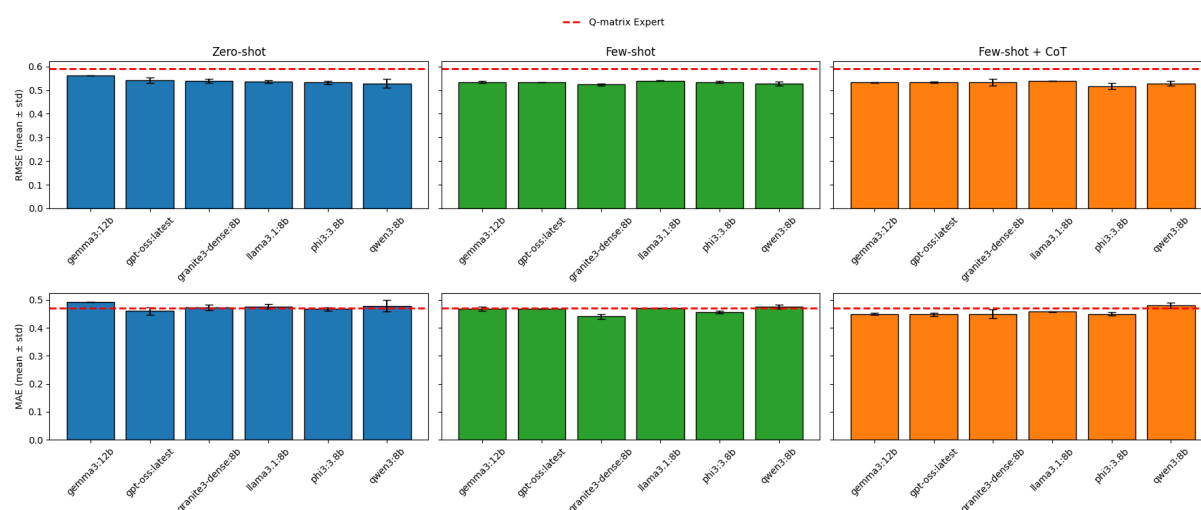
Tabela 2 – Avaliação funcional de Q-matrizes definidas por especialistas e geradas por LLM usando o modelo DINA. Os valores representam a média $\pm$ desvio padrão entre as execuções.

Prompt	Q-matrix	N = 50		N = 100		N = 200	
		RMSE	MAE	RMSE	MAE	RMSE	MAE
Zero-shot	Expert	0.5462	0.4188	0.5914	0.4797	0.5551	0.4242
	Gemma3-12B	0.5280 $\pm$ 0.0003	0.4535 $\pm$ 0.0002	0.5170 $\pm$ 0.0016	0.4730 $\pm$ 0.0020	0.5125 $\pm$ 0.0031	0.4395 $\pm$ 0.0024
	GPT-OSS-20B	0.5135 $\pm$ 0.0043	0.4257 $\pm$ 0.0054	0.5523 $\pm$ 0.0105	0.4564 $\pm$ 0.0074	0.5278 $\pm$ 0.0076	0.4395 $\pm$ 0.0063
	Granite3-Dense-8B	0.5088 $\pm$ 0.0034	0.4321 $\pm$ 0.0026	0.5254 $\pm$ 0.0067	0.4661 $\pm$ 0.0048	0.5052 $\pm$ 0.0059	0.4371 $\pm$ 0.0039
	LLaMA3.1-8B	0.5043 $\pm$ 0.0110	0.4298 $\pm$ 0.0109	0.5158 $\pm$ 0.0063	0.4470 $\pm$ 0.0061	0.5028 $\pm$ 0.0064	0.4365 $\pm$ 0.0060
	Phi3-3.8B	0.4960 $\pm$ 0.0070	0.4455 $\pm$ 0.0061	0.5126 $\pm$ 0.0047	0.4658 $\pm$ 0.0044	0.5032 $\pm$ 0.0049	0.4396 $\pm$ 0.0043
	Qwen3-8B	0.5075 $\pm$ 0.0040	0.4429 $\pm$ 0.0038	0.5169 $\pm$ 0.0051	0.4620 $\pm$ 0.0047	0.4954 $\pm$ 0.0042	0.4492 $\pm$ 0.0040
Few-shot example	Expert	0.5462	0.4188	0.5914	0.4797	0.5551	0.4242
	Gemma3-12B	0.5213 $\pm$ 0.0011	0.4345 $\pm$ 0.0010	0.5326 $\pm$ 0.0035	0.4596 $\pm$ 0.0032	0.5202 $\pm$ 0.0041	0.4400 $\pm$ 0.0038
	GPT-OSS-20B	0.5214 $\pm$ 0.0039	0.4239 $\pm$ 0.0041	0.5361 $\pm$ 0.0068	0.4597 $\pm$ 0.0059	0.5100 $\pm$ 0.0067	0.4402 $\pm$ 0.0058
	Granite3-Dense-8B	0.5142 $\pm$ 0.0036	0.4317 $\pm$ 0.0031	0.5355 $\pm$ 0.0059	0.4617 $\pm$ 0.0051	0.5195 $\pm$ 0.0058	0.4370 $\pm$ 0.0046
	LLaMA3.1-8B	0.5076 $\pm$ 0.0081	0.4239 $\pm$ 0.0075	0.5250 $\pm$ 0.0061	0.4600 $\pm$ 0.0058	0.5136 $\pm$ 0.0060	0.4352 $\pm$ 0.0054
	Phi3-3.8B	0.4960 $\pm$ 0.0047	0.4396 $\pm$ 0.0043	0.5165 $\pm$ 0.0045	0.4470 $\pm$ 0.0041	0.5165 $\pm$ 0.0049	0.4406 $\pm$ 0.0044
	Qwen3-8B	0.5037 $\pm$ 0.0045	0.4429 $\pm$ 0.0041	0.5119 $\pm$ 0.0054	0.4648 $\pm$ 0.0049	0.5031 $\pm$ 0.0048	0.4524 $\pm$ 0.0042
Few-shot reasoning	Expert	0.5462	0.4188	0.5914	0.4797	0.5551	0.4242
	Gemma3-12B	0.5213 $\pm$ 0.0012	0.4335 $\pm$ 0.0011	0.5326 $\pm$ 0.0032	0.4561 $\pm$ 0.0030	0.5202 $\pm$ 0.0040	0.4364 $\pm$ 0.0036
	GPT-OSS-20B	0.5214 $\pm$ 0.0041	0.4338 $\pm$ 0.0042	0.5361 $\pm$ 0.0066	0.4593 $\pm$ 0.0057	0.5100 $\pm$ 0.0064	0.4246 $\pm$ 0.0056
	Granite3-Dense-8B	0.5142 $\pm$ 0.0035	0.4262 $\pm$ 0.0030	0.5355 $\pm$ 0.0056	0.4596 $\pm$ 0.0049	0.5195 $\pm$ 0.0055	0.4339 $\pm$ 0.0044
	LLaMA3.1-8B	0.5076 $\pm$ 0.0078	0.4248 $\pm$ 0.0072	0.5250 $\pm$ 0.0058	0.4535 $\pm$ 0.0055	0.5192 $\pm$ 0.0059	0.4378 $\pm$ 0.0052
	Phi3-3.8B	0.4960 $\pm$ 0.0045	0.4255 $\pm$ 0.0041	0.5165 $\pm$ 0.0043	0.4512 $\pm$ 0.0040	0.5126 $\pm$ 0.0046	0.4422 $\pm$ 0.0041
	Qwen3-8B	0.5037 $\pm$ 0.0046	0.4329 $\pm$ 0.0042	0.5218 $\pm$ 0.0052	0.4655 $\pm$ 0.0048	0.4948 $\pm$ 0.0041	0.4476 $\pm$ 0.0039

Fonte: Lopes e Netto (2026).

A Figura 37 mostra a comparação entre os RMSE e MAE para as diferentes estratégias de *prompting*.

Os resultados obtidos mostraram que as diferenças funcionais entre a Q-matriz do especialista e as Q-matrizes geradas automaticamente foram moderadas. Mesmo quando havia diferenças estruturais em relação à matriz do especialista, o desempenho preditivo do DINA permaneceu próximo ao obtido com a Q-matriz de referência. Esse resultado sugere

Figura 37 – Comparação de RMSE e MAE entre diferentes estratégias de *promptings*.

Fonte: [Lopes e Netto \(2026\)](#).

que o modelo DINA apresentou robustez diante de variações moderadas na especificação da Q-matriz.

No cenário *zero-shot*, o menor RMSE global foi obtido pelo modelo *Qwen3:8b* no tamanho amostral $N = 200$, com valor médio aproximado de 0,495. Esse resultado representou uma redução relativa de cerca de 9% em relação ao RMSE obtido com a Q-matriz especialista nas mesmas condições. No entanto, os valores de MAE apresentaram menor dispersão entre os modelos, e a Q-matriz especialista manteve os menores valores em alguns cenários, indicando maior precisão média.

A estratégia *few-shot* apresentou melhorias pontuais, principalmente em MAE e em amostras menores. No entanto, esses ganhos não se repetiram de forma uniforme no RMSE. A estratégia com CoT produziu resultados heterogêneos: alguns modelos apresentaram pequenas reduções em MAE, mas sem ganhos sistemáticos em RMSE. Dessa forma, a avaliação funcional indica que melhorias estruturais não necessariamente se traduziram em ganhos proporcionais no desempenho preditivo.

Esses resultados são relevantes para a tese, pois mostram que a avaliação de Q-matrizes geradas por LLM não deve depender apenas da comparação estrutural com especialistas. Uma Q-matriz pode apresentar diferenças em relação à referência e, mesmo assim, produzir um desempenho funcional adequado no modelo diagnóstico. Por isso, a combinação de avaliações estruturais e funcionais oferece uma forma mais completa de analisar a qualidade das Q-matrizes geradas automaticamente.

5.3.5 Limitações identificadas

O Estudo 3 apresentou limitações que devem ser consideradas ao se interpretar os resultados. A primeira limitação diz respeito à generalização externa. Todos os experimentos foram conduzidos utilizando um único conjunto de dados, um conjunto fixo de itens e CC, bem como uma Q-matriz. Embora essa configuração tenha permitido comparações controladas, ela restringe a generalização para outros domínios, níveis educacionais e estruturas cognitivas mais complexas.

A segunda limitação está relacionada ao uso de respostas simuladas de estudantes na avaliação funcional. Essa escolha possibilitou o isolamento do efeito das variações da Q-matriz sob condições controladas, reduzindo interferências, como o comportamento de chute (*guessing*), o ruído de medida e a heterogeneidade da população. Por isso, os dados simulados podem não capturar efeitos presentes em contextos reais de aprendizagem. Assim, os resultados funcionais devem ser interpretados com cautela quando transferidos para ambientes educacionais reais.

Outra limitação foi a ausência de comparações com métodos tradicionais de estimativa ou de refinamento da Q-matriz, como abordagens estatísticas ou baseadas em otimização. A inclusão dessas linhas de base permitiria uma avaliação mais abrangente das vantagens e limitações da geração por LLM em relação aos métodos já utilizados na literatura.

O escopo das estratégias de *prompting* também foi limitado. Embora o estudo tenha avaliado as estratégias de *zero-shot*, *few-shot* e *few-shot* com CoT, não foram exploradas estratégias como *prompting* iterativo, autorrefinamento, validação automática por outro modelo ou fluxos com especialista no processo. Essas alternativas podem produzir resultados diferentes em termos de alinhamento estrutural, desempenho funcional e custo computacional.

Por fim, a avaliação utilizou métricas quantitativas estruturais e preditivas. Embora sejam apropriadas para comparações computacionais, elas não capturam completamente as dimensões qualitativas da Q-matriz, tais como plausibilidade pedagógica, o alinhamento curricular e a interpretabilidade para os especialistas. Essa limitação indica a necessidade de estudos futuros envolvendo os especialistas, uso de dados reais de resposta dos estudantes e validação em ambientes como STI e outras plataformas de aprendizagem adaptativa.

5.4 Resumo do Capítulo

Esse capítulo apresentou os principais estudos desenvolvidos ao longo da construção da tese, organizados em três etapas complementares. O primeiro estudo (ver seção 5.1) abordou a proposição de uma arquitetura para recomendação de atividades e intervenção pedagógica em um STI. Esse estudo estabeleceu uma base arquitetural para ampliação do sistema por meio de agentes pedagógicos, responsáveis pela organização do domínio, modelagem do estudante e condução de estratégias tutoriais. A proposta indicou que a adaptação das atividades não deve depender apenas de sequências fixas de exercícios, e sim de uma estrutura capaz de registrar ações dos estudantes, relacioná-las aos conceitos do domínio e apoiar intervenções pedagógicas durante a resolução.

O segundo estudo (ver seção 5.2) avançou sobre essa base ao incorporar MDC G-DINA. Com a definição dos atributos de habilidades e da construção de uma Q-matriz para atividades envolvendo equações polinomiais de primeiro grau, o sistema passou a estimar o domínio das habilidades cognitivas dos estudantes. A aplicação com os estudantes mostrou que a ADC pode fornecer aos professores informações mais detalhadas sobre as lacunas de aprendizagem, permitindo o planejamento de intervenções direcionadas. Os resultados indicaram diferenças entre as habilidades avaliadas, sendo os atributos relacionados à solução de equações com variáveis e constantes em ambos os lados da igualdade e à simplificação de frações os que apresentaram maior dificuldade.

O terceiro estudo (ver seção 5.3) ampliou a investigação ao analisar o uso de LLM para geração automatizada da Q-matriz. Esse estudo transferiu parte do processo de associação entre itens e CC para um procedimento computacional assistido por modelos de linguagem. A avaliação foi realizada em dois níveis: estrutural, por meio da comparação entre as Q-matrizes geradas e a Q-matriz dos especialistas, e o funcional, por meio do impacto dessas matrizes no desempenho preditivo do modelo DINA. Os resultados mostraram que estratégias de *prompting* com exemplos melhoraram o alinhamento estrutural das matrizes geradas. A avaliação funcional, por sua vez, indicou que o desempenho preditivo permaneceu próximo ao obtido com a Q-matriz dos especialistas, mesmo diante de diferenças estruturais moderadas.

Juntos, os três estudos mostram a evolução da proposta da tese. A sequência mostra que a tese não se limita à implementação de um único módulo avaliativo, mas apresenta uma linha de desenvolvimento para sistemas educacionais inteligentes focados no diagnóstico de habilidades. Com base nos estudos realizados, a avaliação do estudante passa a ser tratada

como um processo composto por três elementos: a organização formal do domínio, a inferência diagnóstica do domínio de habilidades e a possibilidade de automatização de parte da construção da Q-matriz. Esses elementos sustentam a proposta de um modelo de avaliação cognitiva para STI, que tem o potencial de apoiar os professores na identificação de lacunas de aprendizagem e na tomada de decisões pedagógicas.

As limitações observadas nos estudos também orientam os desdobramentos da pesquisa. A arquitetura multiagente ainda requer validações mais abrangentes em diferentes contextos educacionais. A aplicação do G-DINA mostrou a necessidade de maior quantidade de itens, mais estudantes, mais CC e testes em outros domínios. A geração da Q-matriz por LLM, por sua vez, indicou a necessidade de validação com especialistas, de dados de respostas reais dos estudantes e de comparação com outros métodos automáticos de criação ou refinamento da Q-matriz.

As limitações descritas não anulam os resultados obtidos, mas delimitam o alcance das conclusões e indicam os requisitos para a consolidação do modelo proposto. Portanto, os resultados apresentados neste capítulo constituem a base empírica e metodológica para a formulação da proposta final da tese, descrita no capítulo [4](#).

Capítulo

6

CONCLUSÕES FINAIS

Esta tese teve como objetivo geral formalizar, instanciar e avaliar a arquitetura cognitivo-adaptativa CAASA-ITS para avaliação estruturada de habilidades em Sistema Tutor Inteligente (STI). A proposta integrou a representação item–Componentes de Conhecimento (CC), inferência por Modelos de Diagnóstico Cognitivo (MDC), modelagem do estudante, visualização diagnóstica, adaptação pedagógica e apoio de Modelo de Linguagem de Grande Escala (LLM) para gerar Q-matrizes.

A arquitetura proposta não surgiu de forma isolada, mas como uma síntese dos estudos desenvolvidos durante o doutorado. O primeiro estudo (ver seção 5.1) estabeleceu a base pedagógico-computacional relacionada aos STI, ao acompanhamento das ações dos estudantes e à necessidade de organizar intervenções pedagógicas com base em evidências de desempenho. O segundo estudo (ver seção 5.2) incorporou a dimensão psicométrica da avaliação diagnóstica, por meio da aplicação do G-DINA e da análise do domínio de habilidades com base na Q-matriz. O terceiro estudo (ver seção 5.3) acrescentou a dimensão da autoria assistida por IA, investigando o uso de LLM na geração e avaliação de Q-matrizes.

Esses estudos, em conjunto, possibilitaram a estruturação de uma arquitetura que contempla todo o ciclo da avaliação cognitiva em STI, sendo: a definição dos CC, associação entre itens e CC, construção da Q-matriz, aplicação das atividades, coleta de respostas, inferência do domínio cognitivo, atualização do modelo do estudante, visualização diagnóstica e apoio à tomada de decisão pedagógica. Assim, a principal contribuição da tese está em integrar esses componentes em uma arquitetura voltada à avaliação estruturada de habilidades.

Dessa forma, a tese alcançou seu objetivo geral. Além disso, os objetivos específicos também foram alcançados. A formalização dos elementos conceituais e computacionais foi realizada por meio da definição das relações entre estudantes, itens, CC, respostas observadas, Q-matriz e estados de domínio. A instanciação modular foi consolidada na arquitetura CAASA-ITS, que organiza os componentes de autoria, domínio, estudante, avaliação diagnóstica, tutoria, adaptação pedagógica e visualização. A aplicação de MDC foi avaliada com

a incorporação do G-DINA ao STI LEIA, permitindo gerar perfis diagnósticos individuais e coletivos. O uso de LLM foi investigado na geração de Q-matrizes, considerando tanto o alinhamento estrutural com matrizes de referência quanto o impacto funcional nos modelos diagnósticos.

As questões norteadoras também foram respondidas ao longo dos estudos. A pesquisa mostrou que os MDC podem ser integrados aos STI quando a relação entre itens e CC é formalizada por meio da Q-matriz. Também evidenciou que a construção manual da Q-matriz apresenta limitações relacionadas ao esforço de autoria, à dependência de especialistas e à definição da granularidade dos CC. Os resultados também indicaram que os modelos LLM podem apoiar a geração inicial da Q-matriz, desde que suas sugestões sejam revisadas e avaliadas estrutural e funcionalmente. Por fim, a arquitetura CAASA-ITS mostrou como esses elementos podem ser integrados em uma proposta única para avaliação cognitiva baseada em habilidades.

A contribuição central da tese está na consolidação da CAASA-ITS como uma arquitetura integrada para avaliação de habilidades em STI. A proposta articula representação, inferência e automação em uma estrutura voltada à produção de diagnósticos interpretáveis, ao apoio à tomada de decisão pedagógica e ao desenvolvimento de sistemas educacionais orientados por evidências cognitivas.

Os resultados obtidos demonstraram que a tese avançou na estruturação do processo de avaliação de habilidades em STI, ao organizar a avaliação como um fluxo composto por autoria, representação, aplicação, inferência, interpretação e tomada de decisão pedagógica. Esse fluxo permitiu superar uma visão fragmentada da avaliação, na qual a modelagem do estudante, a Q-matriz, os modelos diagnósticos e os mecanismos de apoio ao professor eram tratados de maneira isolada.

A arquitetura CAASA-ITS também consolidou uma abordagem modular para avaliação cognitiva baseada em habilidades. A separação entre os módulos de domínio, autoria, estudante, avaliação diagnóstica, tutoria, adaptação e visualização permitiu explicitar as responsabilidades de cada componente e definir como eles se articulam no processo de diagnóstico. Essa organização favorece a rastreabilidade das decisões avaliativas e oferece uma base conceitual e computacional para futuras instanciações em outros STI.

A integração do G-DINA ao STI LEIA demonstrou a viabilidade de se utilizar MDC em um ambiente educacional real para estimar perfis de domínio de habilidades. Com isso, a avaliação deixou de ser interpretada apenas por acertos, erros ou notas totais, passando,

também, a considerar as habilidades cognitivas mobilizadas na resolução das atividades. Essa mudança reforça a contribuição da tese para modelos de avaliação mais interpretáveis e orientados à identificação de lacunas de aprendizagem.

A investigação com LLM ampliou essa contribuição ao tratar a construção da Q-matriz como uma etapa passível de apoio computacional. Os resultados indicaram que as Q-matrizes geradas com apoio de LLM podem apresentar desempenho funcional comparável ao das matrizes construídas por especialistas em cenários controlados, embora ainda exijam revisão humana, bem como validação estrutural e funcional. Essa constatação contribui para reduzir o esforço de autoria, sem transferir a responsabilidade pedagógica e psicométrica para o especialista.

Essa tese mostrou que a qualidade estrutural de uma Q-matriz e seu desempenho funcional em MDC não são dimensões equivalentes. Essa distinção é relevante para a área de Informática na Educação, pois mostra que a avaliação de representações item-atributo deve considerar tanto sua proximidade com uma referência especializada quanto seu impacto no desempenho dos modelos de inferência.

6.1 Contribuições da Tese

A partir da contribuição central, podem ser destacadas contribuições específicas que constituem a arquitetura CAASA-ITS e fornecem evidências para sua proposição. As principais contribuições são descritas a seguir:

- (C1) **Estruturação do processo de avaliação de habilidades em STI:** Esta tese propõe um modelo conceitual que organiza a avaliação de habilidades em três dimensões interdependentes: (i) representação das relações entre itens e CC, (ii) inferência do domínio de habilidades por meio do MDC e (iii) uso dos resultados para interpretação no contexto educacional. Essa estrutura fornece uma visão unificada do processo de avaliação, superando abordagens fragmentadas encontradas na literatura.
- (C2) **Formalização e organização modular da CAASA-ITS:** Com base no modelo conceitual proposto, esta pesquisa formaliza uma arquitetura modular para avaliação de habilidades em STI, organizando os componentes responsáveis pela representação do conhecimento, autoria e gerenciamento da Q-matriz, modelagem do estudante, inferência cognitiva, adaptação pedagógica e visualização dos resultados. A separação

dessas responsabilidades favorece a extensibilidade, a reutilização e a rastreabilidade dos processos envolvidos.

- (C3) **Integração de MDC em um STI real:** A tese demonstra a viabilidade da aplicação do MDC para inferência de habilidades em um ambiente educacional real. Essa contribuição evidencia a aplicabilidade prática dos modelos cognitivos no contexto de STI, preservando a interpretabilidade dos resultados.
- (C4) **Investigação do apoio de LLM à construção da Q-matriz:** Esta pesquisa investiga o uso de LLM como mecanismo de apoio à construção e à revisão da representação item-CC, tradicionalmente realizada de forma manual. Os resultados indicam que Q-matrizes produzidas com esse apoio podem apresentar desempenho funcional comparável ao das matrizes construídas por especialistas em cenários controlados, contribuindo para reduzir o esforço de autoria sem eliminar a necessidade de avaliação e validação especializada.
- (C5) **Análise da relação entre qualidade estrutural e desempenho funcional:** A tese evidencia que a similaridade estrutural entre representações de CC não implica necessariamente na equivalência funcional no desempenho do MDC. Essa constatação contribui para a compreensão crítica da avaliação de representações em contextos educacionais.

Em conjunto, essas contribuições constituem e fornecem evidências para a CAASA-ITS. A contribuição da tese não está relacionada à proposição de um novo MDC ou de um novo LLM, mas à organização e integração desses recursos com os demais componentes necessários à avaliação estruturada de habilidades em STI. A arquitetura estabelece uma separação explícita de responsabilidades entre autoria e representação do domínio, geração assistida da Q-matriz, coleta das evidências de desempenho, inferência cognitiva, Modelagem do Estudante, visualização dos resultados e apoio à tomada de decisão pedagógica.

6.2 Limitações

Apesar dos resultados alcançados, algumas limitações devem ser consideradas. A primeira está relacionada ao tamanho das amostras utilizadas nos estudos empíricos. A aplicação com estudantes ocorreu em um contexto específico, com um número reduzido de participantes, abrangendo um domínio delimitado da Matemática. Embora esses resultados tenham permitido observar o funcionamento da proposta, são necessários estudos com

amostras maiores para avaliar a estabilidade das estimativas e a possibilidade de generalizar os resultados.

Outra limitação refere-se ao número de itens avaliativos utilizados. Os MDC dependem de instrumentos bem planejados, com quantidade suficiente de itens por CC. Quando há poucos itens ou quando alguns CC aparecem de forma desigual na Q-matriz, a estimativa de domínio pode se tornar menos estável. Essa limitação reforça a necessidade de um planejamento cuidadoso da avaliação e de validação da Q-matriz antes da aplicação do MDC.

A definição dos CC também representa uma limitação. A granularidade interfere diretamente no diagnóstico. CC muito amplos reduzem a capacidade de identificar lacunas específicas. CC muito fragmentados podem dificultar a construção dos itens, aumentar a complexidade da Q-matriz e gerar interpretações pouco úteis para o professor. A tese avançou nessa discussão, mas critérios mais objetivos para a definição da granularidade ainda precisam ser investigados.

No estudo com LLM, a principal limitação está relacionada à confiabilidade das Q-matrizes geradas automaticamente. Embora os modelos de linguagem possam sugerir relações plausíveis entre itens e CC, essas sugestões não garantem validade pedagógica ou psicométrica. Os LLM podem produzir respostas coerentes em linguagem natural, mas inconsistentes do ponto de vista do modelo diagnóstico. Por isso, os resultados indicam que sua atuação deve ser entendida como apoio à autoria, e não como substituição do especialista.

Também se reconhece como limitação a concentração dos estudos em domínios matemáticos. O uso de equações de primeiro grau permitiu testar a proposta em contextos com habilidades relativamente delimitadas. No entanto, a aplicação em outros domínios, como leitura, programação, ciências naturais ou competências transversais, pode exigir ajustes na definição dos CC, na construção dos itens e na interpretação dos perfis de domínio.

Uma limitação central desta tese refere-se ao fato de a CAASA-ITS não ter sido implementada e avaliada integralmente como uma única solução computacional. Os principais componentes da arquitetura foram investigados separadamente nos estudos que compõem a pesquisa, fornecendo evidências sobre sua viabilidade individual. Entretanto, não foi realizado um experimento no qual o fluxo completo de autoria, geração assistida da Q-matriz, validação, aplicação das atividades, registro das respostas, diagnóstico cognitivo, atualização do Modelo do Estudante, adaptação pedagógica e visualização diagnóstica operasse de maneira integrada. Dessa forma, os resultados desta tese sustentam a formalização da arquitetura e seus principais mecanismos.

Também não foi avaliado, nesta tese, o desempenho computacional da CAASA-ITS em uma implementação integrada. Essa questão é relevante porque alguns de seus componentes possuem exigências computacionais próprias. A aplicação de MDC pode ser afetada pela quantidade de estudantes, itens e CC, enquanto a execução de LLM depende, entre outros fatores, do tamanho do modelo, da infraestrutura disponível e da quantidade de solicitações processadas. Como esses componentes não foram avaliados conjuntamente em um único ambiente, ainda não é possível determinar o impacto da integração sobre tempo de processamento, consumo de memória, capacidade de atendimento simultâneo ou escalabilidade da arquitetura.

Por fim, a avaliação funcional das Q-matrizes geradas por LLM utilizou dados simulados. Essa escolha permitiu controlar as condições experimentais, mas não substitui estudos com dados reais de estudantes. Aplicações em ambientes educacionais autênticos são necessárias para verificar como as matrizes geradas se comportam diante de respostas reais, ruídos de aprendizagem, estratégias de resolução e diferenças individuais.

6.3 Trabalhos Futuros

Como principal continuidade desta pesquisa, propõe-se implementar e avaliar a CAASA-ITS de forma integrada, reunindo em uma única aplicação os módulos de autoria, domínio, geração assistida da Q-matriz, validação docente, registro das interações, diagnóstico cognitivo, Modelagem do Estudante, adaptação pedagógica e visualização dos resultados. Essa implementação permitirá verificar o comportamento do ciclo cognitivo-adaptativo completo e avaliar se as evidências obtidas individualmente nos estudos desta tese permanecem consistentes quando os componentes operam de forma articulada.

Essa avaliação deverá considerar não apenas aspectos pedagógicos e diagnósticos, mas também requisitos computacionais da arquitetura. Entre os aspectos a serem investigados estão o tempo de processamento das etapas de diagnóstico e geração assistida, o consumo de memória e recursos computacionais, o impacto do aumento do número de estudantes, itens e CC, a execução concorrente dos módulos e a escalabilidade do sistema em diferentes configurações de *hardware*. Esse tipo de análise permitirá identificar possíveis gargalos associados tanto aos MDC quanto aos LLM e avaliar estratégias de otimização ou distribuição dos componentes.

Para outros trabalhos futuros, recomenda-se ampliar a aplicação da proposta CAASA-ITS em turmas maiores e em diferentes níveis de ensino. Essa ampliação permitirá avaliar

a estabilidade das estimativas produzidas pelos MDC e verificar se os perfis de domínio correspondem às dificuldades observadas pelos professores na sala de aula.

Outra possibilidade é realizar estudos longitudinais para acompanhar a evolução do estado de conhecimento dos estudantes ao longo de várias atividades, unidades ou períodos letivos. Essa análise pode mostrar como os perfis de domínio se modificam com o tempo e como a arquitetura pode apoiar intervenções pedagógicas contínuas.

Também se recomenda comparar diferentes MDC no mesmo conjunto de dados. A tese utilizou o G-DINA, mas estudos futuros podem avaliar DINA, DINO, G-DINA, RCD ([Gao et al., 2021](#)), NeuralCDM ([Wang et al., 2020](#)) e outros modelos. Essa comparação pode indicar quais os modelos são mais adequados para diferentes tipos de domínio, quantidade de dados, estrutura de habilidades e objetivos pedagógicos.

A integração entre diagnóstico e recomendação pedagógica também deve ser aprofundada. Os estudos realizados indicam que o diagnóstico por habilidades pode apoiar a tomada de decisão, mas trabalhos futuros podem investigar como transformar automaticamente os perfis de domínio em recomendações de atividades, trilhas de aprendizagem, devolutivas personalizadas e em sugestões de intervenções do professor.

Outro caminho possível consiste em investigar fluxos híbridos de autoria da Q-matriz, combinando sugestões de LLM, revisão de professores, validação psicométrica e avaliação funcional. Esse tipo de fluxo pode reduzir o esforço inicial de construção da Q-matriz, mantendo a validação humana e a análise estatística como etapas obrigatórias.

Por fim, trabalhos futuros podem explorar o uso de dados processuais além das respostas corretas ou incorretas, como o tempo de resolução, o número de tentativas, a sequência de passos, os erros intermediários e os padrões de interação que podem enriquecer a modelagem do estudante. A incorporação desses dados pode ser outra maneira de refinar a identificação de habilidades e oferecer uma visão mais detalhada do processo de aprendizagem.

Dessa forma, conclui-se que os modelos de avaliação baseados em habilidades representam uma abordagem adequada para a área de Informática na Educação, pois permitem interpretar o desempenho do estudante em termos de domínio cognitivo e não apenas de pontuação final. A integração do STI, MDC e do apoio de LLMs para autoria da Q-matriz constitui uma direção viável para o desenvolvimento de ambientes educacionais com diagnósticos interpretáveis e orientados à tomada de decisão pedagógica. Os resultados obtidos oferecem uma base conceitual, computacional e experimental para novas pesquisas sobre avaliação cognitiva para os sistemas educacionais inteligentes.

REFERÊNCIAS

- ABDELAZIZ, I.; BASU, K.; AGARWAL, M.; KUMARAVEL, S.; STALLONE, M.; PANDA, R.; RIZK, Y.; BHARGAV, G. P. S.; CROUSE, M.; GUNASEKARA, C.; IKBAL, S.; JOSHI, S.; KARANAM, H.; KUMAR, V.; MUNAWAR, A.; NEELAM, S.; RAGHU, D.; SHARMA, U.; SORIA, A. M.; SREEDHAR, D.; VENKATESWARAN, P.; UNUVAR, M.; COX, D. D.; ROUKOS, S.; LASTRAS, L. A.; KAPANIPATHI, P. Granite-Function Calling Model: Introducing Function Calling Abilities via Multi-task Learning of Granular Tasks. In: Dernoncourt, Franck and Preoțiuc-Pietro, Daniel and Shimorina, Anastasia (Ed.). *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Miami, Florida, US: Association for Computational Linguistics, 2024. p. 1131–1139. DOI: <https://doi.org/10.18653/v1/2024.emnlp-industry.85>. 68
- ABDELRAHMAN, G.; WANG, Q.; NUNES, B. Knowledge Tracing: A Survey. *ACM Computing Surveys*, ACM New York, NY, v. 55, n. 11, p. 1–37, 2023. DOI: <https://doi.org/10.1145/3569576.1>, 36, 41
- ABDELSHIHEED, M.; BARNES, T.; CHI, M. How and When: the Impact of Metacognitive Knowledge Instruction and Motivation on transfer across Intelligent Tutoring Systems. *International Journal of Artificial Intelligence in Education – (IJAIED)*, Springer, v. 34, n. 3, p. 974–1007, 2024. DOI: <https://doi.org/10.1007/s40593-023-00371-0>. 19, 20
- AHMAD, K.; IQBAL, W.; EL-HASSAN, A.; QADIR, J.; BENHADDOU, D.; AYYASH, M.; AL-FUQAHA, A. Data-Driven Artificial Intelligence in Education: A Comprehensive Review. *IEEE Transactions on Learning Technologies*, IEEE, v. 17, p. 12–31, 2024. DOI: <https://doi.org/10.1109/TLT.2023.3314610>. 4
- AKYILDIZ, A.; EKEN, S. From Raw Data to Intelligent Questions: LLM-Powered Generation with Human Validation. *The Age of Generative Artificial Intelligence*, Izmir Academy Association, p. 90–142, 2025. DOI: <https://doi.org/10.5281/zenodo.16009204>. 19
- AKYUZ, Y. Effects of Intelligent Tutoring Systems on Personalized Learning. *Creative Education*, Scientific Research Publishing, v. 11, n. 06, p. 953 – 978, 2020. DOI: <https://doi.org/10.4236/ce.2020.116069>. 1
- ALABDULHADI, A.; FAISAL, M. Systematic Literature Review of STEM Self-Study Related ITS. *Education and Information Technologies*, Springer, v. 26, n. 2, p. 1549–1588, 2021. DOI: <https://doi.org/10.1007/s10639-020-10315-z>. 5
- ALBANO, G. Knowledge, Skills, Competencies: A Model for Mathematics E-Learning. In: Kwan, Reggie and McNaught, Carmel and Tsang, Philip and Wang, Fu Lee and Li, Kam Cheong (Ed.). *Enhancing Learning Through Technology. Education Unplugged: Mobile Technologies and Web 2.0*. Berlin, Heidelberg: Springer, 2011. p. 214–225. DOI: https://doi.org/10.1007/978-3-642-22383-9_18. 76, 77
- ALEVEN, V. Rule-based Cognitive Modeling for Intelligent Tutoring Systems. *Advances in Intelligent Tutoring Systems*, Springer, Berlin, p. 33–62, 2010. DOI: https://doi.org/10.1007/978-3-642-14363-2_3. 18

- ALMEIDA, T. O.; NETTO, J. F. D. M.; LOPES, A. M. M. Multi-agent System for Recommending Learning Objects in E-learning Environments. In: IEEE. *2021 IEEE Frontiers in Education Conference (FIE)*. Lincoln, NE, USA: IEEE, 2021. p. 1–5. DOI: <https://doi.org/10.1109/FIE49875.2021.9637258>. 14
- ALMEIDA, T. O.; NETTO, J. F. de M.; LOPES, A. M. M. Recommendation Model of Personalized Teaching Materials in E-Learning Environments. In: IEEE. *2021 IEEE Frontiers in Education Conference (FIE)*. Lincoln, NE, USA: IEEE, 2021. p. 1–7. DOI: <https://doi.org/10.1109/FIE49875.2021.9637286>. 14
- ALSHAIKH, F.; HEWAHI, N. AI and Machine Learning Techniques in the Development of Intelligent Tutoring System: A Review. In: IEEE. *2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*. Zallaq, Bahrain: IEEE, 2021. p. 403–410. DOI: <https://doi.org/10.1109/3ICT53449.2021.9582029>. 17
- ANDRADE, A. F. *Uma Aplicação da Abordagem Sociointeracionista de Vygotsky para a Construção de um Ambiente Computacional de Aprendizagem*. Tese (Tese (Doutorado)) — Programa de Pós-Graduação em Informática na Educação, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, 2003. 23
- ANDRADE, D. F. d.; TAVARES, H. R.; VALLE, R. d. C. *Teoria da Resposta ao Item: conceitos e aplicações*. São Paulo: Associação Brasileira de Estatística – ABE, 2000. Disponível em: <https://www.heliton.ufpa.br/arquivos/tri/LivroTRI.pdf>. Acesso em: 18 de setembro de 2026. 34, 35, 36
- ANDRADE, D. F. de; VALLE, R. da C. Introdução à Teoria da Resposta ao Item. *Estudos em Avaliação Educacional*, n. 18, p. 13–32, 1998. DOI: <https://doi.org/10.18222/eae01819982250>. 34
- ANZIEU, P.; BUHAI, B. ANNETTE Project: Advanced Networking for Nuclear Education and Training and Transfer of Expertise. *The EU Framework Programme for Research and Innovation*, v. 4, 2017. Disponível em: <http://www.annette.eu/wp-content/uploads/2021/02/D4.2-Software-Support-for-Competence-Management.pdf>. 82, 83
- ARYADOUST, V. A cognitive diagnostic assessment study of the listening test of the Singapore–Cambridge general certificate of education O-level: Application of DINA, DINO, G-DINA, HO-DINA, and RRUM. *International journal of Listening*, Taylor & Francis, v. 35, n. 1, p. 29–52, 2021. DOI: <https://doi.org/10.1080/10904018.2018.1500915>. 45
- ASCARI, S. R.; GOTTARDO, E.; PIMENTEL, A. R. Mafint: Modelo Afetivo de Intervenção Tutorial para Ambientes Virtuais de Aprendizagem. In: XXXI *Simpósio Brasileiro de Informática na Educação – (SBIE)*. formato Virtual, Natal, RN: SBC, 2020. p. 832–841. DOI: <https://doi.org/10.5753/cbie.sbie.2020.832>. 22
- BAI, J.; BAI, S.; CHU, Y.; CUI, Z.; DANG, K.; DENG, X.; FAN, Y.; GE, W.; HAN, Y.; HUANG, F.; HUI, B.; JI, L.; LI, M.; LIN, J.; LIN, R.; LIU, D.; LIU, G.; LU, C.; LU, K.; MA, J.; MEN, R.; REN, X.; REN, X.; TAN, C.; TAN, S.; TU, J.; WANG, P.; WANG, S.; WANG, W.; WU, S.; XU, B.; XU, J.; YANG, A.; YANG, H.; YANG, J.; YANG, S.; YAO, Y.; YU, B.; YUAN, H.; YUAN, Z.; ZHANG, J.; ZHANG, X.; ZHANG, Y.; ZHANG, Z.; ZHOU, C.; ZHOU, J.; ZHOU, X.; ZHU, T. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. DOI: <https://doi.org/10.48550/arXiv.2309.16609>. 65, 66

- BAKER, R. S. J. d.; CORBETT, A. T.; ALEVEN, V. More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. In: *International Conference on Intelligent Tutoring Systems*. Berlin, Heidelberg: Springer, 2008. p. 406–415. DOI: https://doi.org/10.1007/978-3-540-69132-7_44. 39
- BARBOSA, F. M.; NETTO, J. F. de M.; LOPES, A. M. M. Uma Arquitetura de STI baseada nas Experiências de Aprendizagem Mediadas para o Ensino de Trigonometria. In: *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*. Manaus, Amazonas: SBC, 2022. p. 1066–1076. DOI: <https://doi.org/10.5753/sbie.2022.225210>. 14, 23
- BARNES, T. M. The Q-matrix Method: Mining Student Response Data for Knowledge. In: *American Association for Artificial Intelligence 2005 Educational Data Mining Workshop*. Pittsburgh, PA, USA: AAI Press, 2005. p. 1–8. Disponível em: <https://cdn.aaai.org/Workshops/2005/WS-05-02/WS05-02-006.pdf>. Acesso em: 22 de jun. de 2026. 54
- BARNES, T. M.; BITZER, D. L.; BARNES, T.; BITZER, D.; VOUK, M. *Comparison of the Q-Matrix Clustering Method with Factor Analysis and K-Means Cluster Analysis*. [S.l.], 2005. Disponível em: <https://www.researchgate.net/publication/249871175>. Acesso em: 04 de maio de 2026. 47
- BARREIRA, C.; BOAVIDA, J.; ARAÚJO, N. Avaliação Formativa: Novas Formas de Ensinar e Aprender. *Revista Portuguesa de Pedagogia*, p. 95–133, 2006. DOI: https://doi.org/10.14195/1647-8614_40-3_4. 25
- BAUMEISTER, R. F.; LEARY, M. R. Writing narrative literature reviews. *Review of General Psychology*, SAGE Publications Sage CA: Los Angeles, CA, v. 1, n. 3, p. 311–320, 1997. DOI: <https://doi.org/10.1037/1089-2680.1.3.311>. 75
- BENGESI, S.; EL-SAYED, H.; SARKER, M. K.; HOUKPATI, Y.; IRUNGU, J.; OLADUNNI, T. Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion model, and Transformers. *IEEE Access*, IEEE, v. 12, p. 69812–69837, 2024. DOI: <https://doi.org/10.1109/ACCESS.2024.3397775>. 62
- BLACK, P.; WILIAM, D. Assessment and Classroom Learning. *Assessment in Education: principles, policy & practice*, Taylor & Francis, v. 5, n. 1, p. 7–74, 1998. DOI: <https://doi.org/10.1080/0969595980050102>. 25
- BLOOM, B. S.; MADAUS, G. F.; HASTINGS, J. T. *Handbook on Formative and Summative Evaluation of Student Learning*. New York: McGraw-Hill, 1971. 25, 26
- BOLLELA, V. R.; BORGES, M. d. C.; TRONCON, L. E. d. A. Avaliação Somativa de Habilidades Cognitivas: Experiência envolvendo boas Práticas para a Elaboração de Testes de Múltipla Escolha e a Composição de Exames. *Revista Brasileira de Educação Médica*, SciELO Brasil, v. 42, p. 74–85, 2018. DOI: <https://doi.org/10.1590/1981-52712015v42n4RB20160065>. 31, 32
- BREMGARTNER, V. *Arcabouço Conceitual de Adaptação de Recursos Educacionais*. Tese (Doutorado em Informática) — Universidade Federal do Amazonas, Manaus, AM, 2017. Disponível em: <https://tede.ufam.edu.br/handle/tede/6216>. Acesso em: 26 de junho 2026. 2, 22

BREMGARTNER, V.; NETTO, J. F. de M.; MENEZES, C. S. de. Conceptual framework of educational resources adaptation for improve collaborative learning in virtual learning environments. In: IEEE. *2017 IEEE Frontiers in Education Conference – (FIE)*. Indianapolis, IN, USA: IEEE, 2017. p. 1–9. DOI: <https://doi.org/10.1109/FIE.2017.8190707>. 84, 85, 86

BRIGGS DEREK C. ALONZO, A. C.; SCHWAB, C.; WILSON, M. Diagnostic Assessment with Ordered Multiple-choice Items. *Educational Assessment*, Taylor & Francis, v. 11, n. 1, p. 33–63, 2006. DOI: https://doi.org/10.1207/s15326977ea1101_2. 26, 31, 70

BROWN, T. B.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J.; DHARIWAL, P.; NEELAKANTAN, A.; SHYAM, P.; SASTRY, G.; ASKELL, A.; AGARWAL, S.; HERBERT-VOSS, A.; KRUEGER, G.; HENIGHAN, T.; CHILD, R.; RAMESH, A.; ZIEGLER, D. M.; WU, J.; WINTER, C.; HESSE, C.; CHEN, M.; SIGLER, E.; LITWIN, M.; GRAY, S.; CHESS, B.; CLARK, J.; BERNER, C.; MCCANDLISH, S.; RADFORD, A.; SUTSKEVER, I.; AMODEI, D. Language Models are Few-shot Learners. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2020. (NIPS '20, v. 33), p. 1877–1901. DOI: <https://doi.org/10.48550/arXiv.2005.14165>. 55, 60

CARLE, S. M. *Student-Held Misconceptions Regarding Area and Perimeter of Rectangles*. Dissertação (Master of Arts Thesis) — University of Massachusetts Boston, Boston, 1993. Critical and Creative Thinking Program. Disponível em: https://scholarworks.umb.edu/cct_capstone/46/. Acesso em: 22 jun. 2026. 45

CASTILHOS, F. M. de; JAQUES, P. A. Deep Learning para Autoria Automatizada de Modelos de Domínios de Sistemas Tutores Baseados em Passos. *Revista Brasileira de Informática na Educação – RBIE*, CEIE, SBC, Curitiba, PR, v. 32, p. 871–894, 2024. DOI: <https://doi.org/10.5753/rbie.2024.3702>. 19

CHICO, V. J. S.; REGINO, A. G.; BONACIN, R.; REIS, J. C. dos. LLM-Based Automatic Generation of Multiple-Choice Questions With Meaningful Distractors. In: *Anais XXXVI Simpósio Brasileiro de Informática na Educação – SBIE*. Curitiba, PR: SBC, 2025. p. 813–827. DOI: <https://doi.org/10.5753/sbie.2025.12675>. 69, 73

CHIUSOLE, D. de; STEFANUTTI, L.; ANSELMINI, P.; ROBUSTO, E. Stat-Knowlab. Assessment and Learning of Statistics with Competence-based Knowledge Space Theory. *International Journal of Artificial Intelligence in Education*, Springer, v. 30, n. 4, p. 668–700, 2020. DOI: <https://doi.org/10.1007/s40593-020-00223-1>. 95, 96, 100

CORBETT, A. T.; ANDERSON, J. R. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*, v. 4, p. 253–278, 1994. DOI: <https://doi.org/10.1007/BF01099821>. 36

CRUZ-CASTAÑEDA, W. A.; AMADEUS, M. Large Languages Models in Brazilian Portuguese: A Chronological Survey. *Journal of the Brazilian Computer Society*, v. 31, n. 1, p. 1168–1187, 2025. DOI: <https://doi.org/10.5753/jbcs.2025.5789>. 4, 62, 65, 67

DENG, C.; YUAN, B. Research on an Intelligent Tutoring System Based on Automatic Construction of Multimodal Knowledge Graphs and Retrieval-Augmented Generation. *Frontiers in Computer Science*, Frontiers, v. 8, p. 1777749, 2026. DOI: <https://doi.org/10.3389/fcomp.2026.1777749>. 19

- DESMARAIS, M. C.; BEHESHTI, B.; NACEUR, R. Item to Skills Mapping: Deriving a Conjunctive Q-matrix from Data. In: Cerri, S. A. and Clancey, W. J. and Papadourakis, G. and Panourgia, K. (Ed.). *International Conference on Intelligent Tutoring Systems – ITS 2012*. Berlin, Heidelberg, 2012. p. 454–463. DOI: https://doi.org/10.1007/978-3-642-30950-2_58. 28, 47, 54, 70
- DING, Z.; WANG, X.; WU, Y.; CAO, G.; CHEN, L. Tagging Knowledge Concepts for Math Problems Based on Multi-Label Text Classification. v. 267, p. 126232, 2025. DOI: <https://doi.org/10.1016/j.eswa.2024.126232>. 3, 54, 61
- DÜNTSCH, I.; GEDIGA, G. Skills and Knowledge Structures. *British Journal of Mathematical and Statistical Psychology*, Wiley Online Library, v. 48, n. 1, p. 9–27, 1995. DOI: <https://doi.org/10.1111/j.2044-8317.1995.tb01047.x>. 28
- DURAND, G.; BELACEL, N.; GOUTTE, C. Evaluation of Expert-based Q-matrices Predictive Quality in Matrix Factorization Models. In: SPRINGER. *European Conference on Technology Enhanced Learning*. [S.l.]: Springer, 2015. p. 56–69. DOI: https://doi.org/10.1007/978-3-319-24258-3_5. 47
- EMOND, B. Learning traces, Measurement and Assessment Templates for AIS Interoperability. In: Sottolare, R. A. and Schwarz, J. (Ed.). *International Conference on Human-Computer Interaction*. [S.l.]: Springer, Cham, 2020. v. 12214, p. 71–87. DOI: https://doi.org/10.1007/978-3-030-50788-6_6. 28
- ESSA, A. A Possible Future for Next Generation Adaptive Learning Systems. *Smart Learning Environments*, Springer, v. 3, n. 1, p. 16–40, 2016. DOI: <https://doi.org/10.1186/s40561-016-0038-y>. 24
- FALMAGNE, J.-C.; DOBLE, C. Overview. In: _____. *Knowledge Spaces: Applications in Education*. Berlin, Heidelberg: Springer, 2013. p. 3–26. ISBN 978-3-642-35329-1. DOI: https://doi.org/10.1007/978-3-642-35329-1_1. 28
- FARID, M.; ELGOHARY, R.; MOAWAD, I.; ROUSHDY, M. User Profiling Approaches, Modeling, and Personalization. In: *Proceedings of the 11th International Conference on Informatics & Systems – (INFOS 2018)*. Cairo, Egypt: [s.n.], 2018. p. 7. DOI: <http://dx.doi.org/10.2139/ssrn.3389811>. 20
- FERNANDES, R. d. S. *Contribuições para Modelos Gerais de Diagnóstico Cognitivo*. Tese (Doutorado em Estatística) — Estatística Interinstitucional do ICMC e UFSCarr, Universidade de São Paulo, São Carlos, 2024. DOI: <https://doi.org/10.11606/T.104.2024.tde-05082024-095246>. 2, 5, 50
- FERREIRA, H. N. M. *Uma Abordagem Híbrida Baseada em Redes Bayesianas e Ontologias para Modelagem do Estudante em Sistemas Adaptativos e Inteligentes para Educação*. Tese (Doutorado em Ciência da Computação) — Universidade Federal de Uberlândia (UFU), 2018. DOI: <http://dx.doi.org/10.14393/ufu.te.2018.786>. 21
- FREITAS, S. L.; COSTA, M. G. N. da; MIRANDA, F. A. de. Avaliação Educacional: Formas de Uso na Prática Pedagógica. *Revista Meta: Avaliação*, v. 6, n. 16, p. 85–98, 2014. DOI: <http://dx.doi.org/10.22347/2175-2753v6i16.217>. 25, 26

- FU, D.; QIN, C.; LUO, Z.; LI, Y.; YU, X.; YE, Z. Using Regularized Methods to Validate Q-matrix in Cognitive Diagnostic Assessment. *Journal of Educational and Behavioral Statistics*, SAGE Publications, Los Angeles, CA, v. 50, n. 1, p. 149–179, 2024. DOI: <https://doi.org/10.3102/10769986241240084>. 53
- GAO, W.; LIU, Q.; HUANG, Z.; YIN, Y.; BI, H.; WANG, M.-C.; MA, J.; WANG, S.; SU, Y. RCD: Relation Map Driven Cognitive Diagnosis for Intelligent Education Systems. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Evento Virtual, Canada: Association for Computing Machinery, New York, NY, United States, 2021. p. 501–510. DOI: <https://doi.org/10.1145/3404835.3462932>. 151
- GEORGE, A. C.; ROBITZSCH, A. Cognitive Diagnosis Models in R: A Didactic. *The Quantitative Methods for Psychology*, v. 11, n. 3, p. 189–205, 2015. DOI: <https://doi.org/1020982/tqmp.11.3.p189>. 45
- GERHARDT, T. E.; SILVEIRA, D. T. (Ed.). *Métodos de Pesquisa*. Porto Alegre: Editora da UFRGS, 2009. 120 p. (Educação a Distância – EaD). 8, 9
- GIRAFFA, L.; KOHLS-SANTOS, P. Inteligência Artificial e Educação: Conceitos, Aplicações e Implicações no fazer docente. *Educação em análise*, Londrina, PR, v. 8, n. 1, p. 116–134, 2023. DOI: <https://doi.org/10.5433/1984-7939.2023v8n1p116>. 17
- GIRAFFA, L. M. M. *Uma Arquitetura de Tutor Utilizando Estados Mentais*. Tese (Doutorado em Informática) — Universidade Federal do Rio Grande do Sul, 1999. Disponível em: <http://hdl.handle.net/10183/17620>. Acesso em: 26 de junho 2026. 22
- GONG, Y.; BECK, J. E.; HEFFERNAN, N. T. How to Construct More Accurate Student Models: Comparing and Optimizing Knowledge Tracing and Performance Factor Analysis. *International Journal of Artificial Intelligence in Education*, IOS Press, v. 21, n. 1-2, p. 27–46, 2011. DOI: <https://doi.org/10.3233/JAI-2011-016>. 39
- GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. Cambridge, MA: MIT Press, 2016. Acessado em: 25 de maio de 2025. Disponível em: <https://www.deeplearningbook.org>. 59
- GRAESSER, A. C.; CONLEY, M. W.; OLNEY, A. Intelligent Tutoring Systems. In: Karen R. Harris and Steve Graham and Tim Urdan and Anna G. Bus and Stephen Major and H. Lee Swanson (Ed.). *APA Educational Psychology Handbook, Vol. 3: Application to Learning and Teaching*. [S.l.]: American Psychological Association, 2012. p. 451–473. DOI: <https://doi.org/10.1037/13275-018>. 2
- Granite Team. *Granite 3.0 Language Models*. Armonk, NY, 2024. Disponível em: <https://www.rivista.ai/wp-content/uploads/2024/10/paper-1.pdf>. Acesso em: 22 maio 2026. 67
- Granite Vision Team; KARLINSKY, L.; ARBELLE, A.; DANIELS, A.; NASSAR, A.; ALFASSI, A.; WU, B.; SCHWARTZ, E.; JOSHI, D.; KONDIC, J.; SHABTAY, N.; LI, P.; HERZIG, R.; ABEDIN, S.; PEREK, S.; HARARY, S.; BARZELAY, U.; GOLDFARB, A. R.; OLIVA, A.; WIELES, B.; BHATTACHARJEE, B.; HUANG, B.; AUER, C.; GUTFREUND, D.; BEYMER, D.; WOOD, D.; KUEHNE, H.; HANSEN, J.; SHTOK, J.; WONG, K.; BATHEN, L. A.; MISHRA, M.; LYSAK, M.; DOLFI, M.; YUROCHKIN, M.; LIVATHINOS, N.; HAREL, N.; AZULAI, O.; NAPARSTEK, O.; LIMA, R. T. de; PANDA, R.; DOVEH, S.; GUPTA, S.; DAS, S.; ZAWAD, S.; KIM, Y.; HE, Z.; BROOKS, A.; GOODHART, G.;

- GOVINDJEE, A.; LEIST, D.; IBRAHIM, I.; SOFFER, A.; COX, D.; SOULE, K.; LASTRAS, L.; DESAI, N.; OFEK-KOIFMAN, S.; RAGHAVAN, S.; SYEDA-MAHMOOD, T.; STAAR, P.; DRORY, T.; FERIS, R. Granite Vision: a Lightweight, Open-Source Multimodal Model for Enterprise Intelligence. *arXiv preprint arXiv:2502.09927*, 2025. DOI: <https://doi.org/10.48550/arXiv.2502.09927>. 68
- GRANT, M. J.; BOOTH, A. A Typology of Reviews: an Analysis of 14 Review Types and Associated Methodologies. *Health information & libraries journal*, Wiley Online Library, v. 26, n. 2, p. 91–108, 2009. <https://doi.org/10.1111/j.1471-1842.2009.00848.x>. 75
- GUSMÃO, F. A. F.; AMORIM, S. S. Revisão Sistemática: Pesquisas empíricas acerca do Saeb (1995-2021). *Estudos em Avaliação Educacional*, Fundação Carlos Chagas, v. 34, 2023. DOI: <https://doi.org/10.18222/eae.v34.9051>. 32
- HAMBLETON, R. K.; SWAMINATHAN, H. Item Response Theory. In: _____. *Item Response Theory: Principles and Applications*. Dordrecht: Springer Netherlands, 1985. p. 33–52. DOI: https://doi.org/10.1007/978-94-017-1988-9_3. 34, 35, 36
- HAN, J.; ZHAO, W.; JIANG, Q.; OUBIBI, M.; HU, X. Intelligent Tutoring System Trends 2006-2018: A Literature Review. In: IEEE. *2019 Eighth International Conference on Educational Innovation Through Technology – (EITT)*. Biloxi, MS, USA: IEEE, 2019. p. 153–159. DOI: <https://doi.org/10.1109/EITT.2019.00037>. 4, 5
- HARLEN, W.; JAMES, M. Assessment and Learning: Differences and Relationships Between Formative and Summative Assessment. *Assessment in Education: Principles, Policy & Practice*, Taylor & Francis, v. 4, n. 3, p. 365–379, 1997. DOI: <https://doi.org/10.1080/0969594970040304>. 26
- HEEREN, B.; JEURING, J. Feedback Services for Stepwise Exercises. *Science of Computer Programming*, Elsevier, v. 88, p. 110–129, 2014. DOI: <https://doi.org/10.1016/j.scico.2014.02.021>. 20
- HELLER, J.; ÜNLÜ, A.; ALBERT, D. Skills, Competencies and Knowledge Structures. In: Falmagne, J. C. and Albert, D. and Doble, C. and Eppstein, D. and Hu, X. (Ed.). *Knowledge spaces*. Berlin, Heidelberg: Springer, 2013. p. 229–242. DOI: https://doi.org/10.1007/978-3-642-35329-1_11. 29, 30
- HENSON, R. A.; TEMPLIN, J. L.; WILLSE, J. T. Defining a Family of Cognitive Diagnosis Models using Log-linear Models with Latent Variables. *Psychometrika*, Springer-Verlag, v. 74, n. 2, p. 191–210, 2009. DOI: <https://doi.org/10.1007/s11336-008-9089-5>. 51
- HERLIANA, F.; HALIM, A.; ANDRIANI, R.; YASIN, B. et al. Cognitive Diagnostic Assessment in Educational Contexts: Global Trends, Methodological Innovations, and Emerging Direction. *Eurasia Journal of Mathematics, Science and Technology Education*, Modestum, v. 22, n. 2, p. 2775, 2026. DOI: <https://doi.org/10.29333/ejmste/17827>. 2, 5
- HNIDA, M.; IDRISSE, M. K.; BENNANI, S. Overview of CALEP: A Competency Based Learning Path Generation System. In: *2018 17th International Conference on Information Technology Based Higher Education and Training – (ITHET)*. [S.l.]: IEEE, 2018. p. 1–6. DOI: <https://doi.org/10.1109/ITHET.2018.8424772>. 87, 88, 89, 101
- HODSON, T. O. Root Mean Square Error (RMSE) or Mean Absolute Error (MAE): When to use them or not. *Geoscientific Model Development Discussions*, Göttingen, Germany, v. 2022, p. 1–10, 2022. DOI: <https://doi.org/10.5194/gmd-15-5481-2022>. 51

- HOLDEN, H. K.; SINATRA, A. M. A Guide to Scaffolding and Guided Instructional Strategies for ITSs. In: Sottolare, Robert A. and Graesser, Arthur C. and Hu, Xiangen and Goldberg, Benjamin S. (Ed.). *Design Recommendations for Intelligent Tutoring Systems – Vol. 2: Instructional Management*. 1. ed. Orlando, FL: U.S. Army Research Laboratory, 2014. v. 2, p. 265–281. ISBN 978-0-9893923-2-7. [20](#)
- HUANG, X.; XU, W.; LIU, R. Effects of Intelligent Tutoring Systems on Educational Outcomes. *International Journal of Distance Education Technologies*, v. 23, p. 1–25, 2 2025. ISSN 1539-3100. DOI: <https://doi.org/10.4018/IJDET.368420>. [1](#), [4](#), [5](#)
- INEP. ENEM. 2026. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/saeb>. Acesso em: 21 junho 2026. [32](#)
- INEP. Saeb. 2026. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/saeb>. Acesso em: 21 junho 2026. [32](#)
- ION, G.; CANO, E.; CABRERA, N. Competency Assessment Tool (CAT). The Evaluation of an Innovative Competency-based Assessment Experience in Higher Education. *Technology, Pedagogy and Education*, Taylor & Francis, v. 25, n. 5, p. 631–648, 2016. DOI: <https://doi.org/10.1080/1475939X.2015.1134635>. [80](#), [81](#), [82](#)
- JAIQUES, P. A.; SEFFRIN, H.; RUBI, G.; MORAIS, F. de; GHILARDI, C.; BITTENCOURT, I. I.; ISOTANI, S. Rule-based Expert Systems to Support Step-by-Step Guidance in Algebraic Problem Solving: The Case of the Tutor PAT2Math. *Expert Systems with Applications*, Elsevier, v. 40, n. 14, p. 5456–5465, 2013. DOI: <https://doi.org/10.1016/j.eswa.2013.04.004>. [18](#), [37](#)
- JAVAHERIPI, M.; BUBECK, S.; ABDIN, M.; ANEJA, J.; MENDES, C. C. T.; CHEN, W.; GIORNO, A. D.; ELDAN, R.; GOPI, S.; GUNASEKAR, S.; KAUFFMANN, P.; LEE, Y. T.; LI, Y.; NGUYEN, A.; ROSA, G. de; SAARIKIVI, O.; SALIM, A.; SHAH, S.; SANTACROCE, M.; BEHL, H. S.; KALAI, A. T.; WANG, X.; WARD, R.; WITTE, P.; ZHANG, C.; ZHANG, Y. Phi-2: The Surprising Power of Small Language Models. *Microsoft Research Blog*, v. 1, n. 3, p. 3, 2023. Disponível em: https://nips.cc/media/neurips-2023/Slides/83968_5GxuY2z.pdf. Acesso em: 26 de junho de 2026. [64](#)
- JAVIDANMEHR, Z.; SARAB, M. R. A. Cognitive Diagnostic Assessment: Issues and Considerations. *International Journal of Language Testing*, ERIC, v. 7, n. 2, p. 73–98, 2017. [31](#), [33](#), [36](#), [101](#)
- JESUS, A. Sistemas Tutores Inteligentes uma Visão Geral. *Revista Eletrônica de Sistemas de Informação*, v. 2, n. 2, 2003. DOI: <https://doi.org/10.21529/RESI.2003.0202006>. [23](#)
- JL, Y.; ZHAN, Z.; LI, T.; ZOU, X.; LYU, S. Human–Machine Cocreation: The Effects of ChatGPT on Students’ Learning Performance, AI Awareness, Critical Thinking, and Cognitive Load in a STEM Course Toward Entrepreneurship. *IEEE Transactions on Learning Technologies*, v. 18, p. 402–415, 2025. DOI: <https://doi.org/10.1109/TLT.2025.3554584>. [4](#)
- JIANG, A. Q.; SABLAYROLLES, A.; ROUX, A.; MENSCH, A.; SAVARY, B.; BAMFORD, C.; CHAPLOT, D. S.; CASAS, D. de las; HANNA, E. B.; BRESSAND, F.; LENGUEL, G.; BOUR, G.; LAMPLE, G.; LAVAUD, L. R.; SAULNIER, L.; LACHAUX, M.-A.; STOCK, P.; SUBRAMANIAN, S.; YANG, S.; ANTONIAK, S.; SCAO, T. L.; GERVET, T.; LAVRIL, T.; WANG, T.; LACROIX, T.; SAYED, W. E. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. DOI: <https://doi.org/10.48550/arXiv.2401.04088>. Disponível em: <https://arxiv.org/abs/2401.04088>. [65](#)

- JOHNSON, R. B.; CHRISTENSEN, L. *Educational research: Quantitative, qualitative, and mixed approaches*. [S.l.]: Sage publications, 2014. v. 11. 9
- JOSÉ, A. S. C. *Creating a Knowledge Space for the SQL Domain*. Dissertação (Master's dissertation) — NOVA Information Management School, Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa, Lisbon, Portugal, nov. 2022. Disponível em: <<https://run.unl.pt/bitstream/10362/140846/1/TAA0155.pdf>>. Acesso em: 08 de maio de 2026. 27
- JÚNIOR, A. B. d. L.; SILVA, L. T. G. Os Sistemas Tutores Inteligentes e a Adaptação do Ensino aos Perfis de Aprendizagem do Usuário. *Educação Temática Digital – (ETD)*, UNICAMP, v. 24, n. 3, p. 618–632, 2022. DOI: <https://doi.org/10.20396/etd.v24i3.8663707>. 17, 23
- KANG, C.; YANG, Y.; ZENG, P. Q-matrix Refinement Based on Item Fit Statistic RMSEA. *Applied Psychological Measurement*, SAGE Publications: Los Angeles, CA, v. 43, n. 7, p. 527–542, 2019. DOI: <https://doi.org/10.1177/0146621618813104>. 52
- KAUTZMANN, T. R.; JAQUES, P. A. Treinamento da Habilidade Metacognitiva de Monitoramento do Conhecimento em Sistemas Tutores. *Revista Brasileira de Informática na Educação – (RBIE)*, SBC, v. 24, n. 2, p. 22–43, 2016. DOI: <https://doi.org/10.5753/rbie.2016.24.02.22>. 20
- KÖHN, H.-F.; CHIU, C.-Y.; OLUWALANA, O.; KIM, H.; WANG, J. A two-step q-matrix estimation method. *Applied Psychological Measurement*, SAGE Publications: Los Angeles, CA, v. 49, n. 1-2, p. 3–28, 2025. DOI: <https://doi.org/10.1177/01466216241284418>. 54
- KUO, B.-C.; CHEN, C.-H.; TORRE, J. de la. A Cognitive Diagnosis Model for Identifying Coexisting Skills and Misconceptions. *Applied Psychological Measurement*, SAGE Publications: Los Angeles, CA, v. 42, n. 3, p. 179–191, 2018. DOI: <https://doi.org/10.1177/0146621617722791>. 45
- LAINE, J.; LINDQVIST, T.; KORHONEN, T.; HAKKARAINEN, K. Systematic Review of Intelligent Tutoring Systems for Hard Skills Training in Virtual Reality Environments. *International Journal of Technology in Education and Science*, ISTES Organization, v. 6, p. 178–203, 5 2022. DOI: <https://doi.org/10.46328/ijtes.348>. 1
- LARA, J. A.; ALJAWARNEH, S.; PAMPLONA, S. Special issue on the current trends in E-learning Assessment. *Journal of Computing in Higher Education*, Springer, v. 32, n. 1, p. 1–8, 2020. DOI: <https://doi.org/10.1007/s12528-019-09235-w>. 24, 26
- LEFEBVRE, B.; GAUTHIER, G.; TADIÉ, S.; DUC, T. H.; ACHABA, H. Competence Ontology for Domain Knowledge Dissemination and Retrieval. *Applied Artificial Intelligence*, Taylor & Francis, v. 19, n. 9-10, p. 845–859, 2005. DOI: <https://doi.org/10.1080/08839510500234222>. 18
- LI, H.; XU, T.; TANG, J.; WEN, Q. Automate Knowledge Concept Tagging on Math Questions with LLMs. *arXiv e-prints*, p. arXiv:2403.17281, 2024. DOI: <https://doi.org/10.48550/arXiv.2403.17281>. 61, 65, 67, 74
- LI, H.; XU, T.; TANG, J.; WEN, Q. Knowledge Tagging System on Math Questions via LLMs with Flexible Demonstration Retriever. *arXiv preprint arXiv:2406.13885*, 2024. DOI: <https://doi.org/10.48550/arXiv.2406.13885>. 54
- LI, J.; CHEN, P. A new Q-matrix Validation Method Based on Signal Detection Theory. *British Journal of Mathematical and Statistical Psychology*, Wiley Online Library, v. 78, n. 2, p. 522–554, 2025. DOI: <https://doi.org/10.1111/bmsp.12371>. 54

- LIU, H.; ZHANG, Y.; JIA, J. The Design of Guiding and Adaptive Prompts for Intelligent Tutoring Systems and Its Effect on Students' Mathematics Learning. *IEEE Transactions on Learning Technologies*, IEEE, 2024. DOI: <https://doi.org/10.1109/TLT.2024.3382000>. 1
- LIU, J.; TANG, W.; HE, X.; YANG, B.; WANG, S. Research on DINA Model in Online Education. In: LIU, S.; SUN, G.; FU, W. (Ed.). *International Conference on E-Learning, E-Education, and Online Training*. [S.l.]: Springer, Cham, 2020. v. 340, p. 279–291. DOI: https://doi.org/10.1007/978-3-030-63955-6_24. 45, 49
- LIU, R. A Diagnostic Measurement Analysis of the Subskill Mastery on the Michigan English Test. *SPAAN Research Grant Working Paper*, 2019. Disponível em: https://michiganassessment.org/wp-content/uploads/2021/02/SPAAN_2019-01_Liu-2.pdf. Acesso em: 26 de junho de 2026. 45
- LOPES, A. M. M.; NETTO, J. F.; LIMA, D. P. R. de. O Uso de Agentes Conversacionais no Apoio do Ensino de Resolução de Problemas Matemáticos: Uma Revisão Sistemática da Literatura. In: *Simpósio Brasileiro de Informática na Educação – (SBIE)*. [S.l.]: SBC, 2018. v. 29, n. 1, p. 1403–1412. DOI: <https://doi.org/10.5753/cbie.sbie.2018.1403>. 23
- LOPES, A. M. M.; NETTO, J. F. D. M. Designing Pedagogical Agents Toward the Recommendation and Intervention Based on Students' Actions in an ITS. In: IEEE. *2021 IEEE Frontiers in Education Conference (FIE)*. Lincoln, NE, USA: IEEE, 2021. p. 1–8. DOI: <https://doi.org/10.1109/FIE49875.2021.9637406>. ii, 12, 13, 20, 105, 109, 110, 113, 115, 116, 117, 119, 120, 122, 123, 125
- LOPES, A. M. M.; NETTO, J. F. d. M. Incorporando uma Abordagem para Avaliação de Habilidades Cognitivas em um Sistema Tutor Inteligente. *Revista Novas Tecnologias na Educação — RENOTE*, SBC, v. 21, p. 178–187, 2023. DOI: <https://doi.org/10.22456/1679-1916.137738>. 10, 12, 14, 46, 71, 106, 110, 114, 117, 119, 120, 127, 129, 131, 132, 133, 134
- LOPES, A. M. M.; NETTO, J. F. D. M. Large Language Models for Learning Technologies: Structural and Functional Evaluation of Automated Q-Matrix Generation for Educational Assessment. *IEEE Access*, IEEE, v. 14, p. 58466–58483, 2026. DOI: <https://doi.org/10.1109/ACCESS.2026.3683811>. 10, 12, 14, 28, 53, 72, 74, 111, 119, 120, 135, 136, 138, 139, 140, 141, 178
- LOPES, A. M. M.; NETTO, J. F. de M. Sistemas Tutores Inteligentes: um Mapeamento das Produções Brasileiras. In: SBC. *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. Evento Virtual, Natal, RN, Brasil: SBC, 2020. p. 1313–1322. DOI: <https://doi.org/10.5753/cbie.sbie.2020.1313>. 4, 5, 10, 11, 13, 18, 75, 121
- LOPES, A. M. M.; NETTO, J. F. de M.; SOUZA, R. A. de; MOURÃO, A. B.; ALMEIDA, T. O.; LIMA, D. P. de. Improving Students Skills to Solve Elementary Equations in K-12 Programs Using an Intelligent Tutoring System. In: *2019 IEEE Frontiers in Education Conference – (FIE)*. Cincinnati, Ohio: IEEE, 2019. p. 1–9. DOI: <https://doi.org/10.1109/FIE43999.2019.9028363>. 18, 23, 110, 122
- LU, Z.; LI, X.; CAI, D.; YI, R.; LIU, F.; ZHANG, X.; LANE, N. D.; XU, M. Small language models: Survey, measurements, and insights. *arXiv preprint arXiv:2409.15790*, 2024. DOI: <https://doi.org/10.48550/arXiv.2409.15790>. 64

- LUISE, R. S. A.-D.; AREVALILLO-HERRÁEZ, M.; ARNAU, D. On Using Conversational Frameworks to Support Natural Language Interaction in Intelligent Tutoring Systems. *IEEE Transactions on Learning Technologies*, IEEE, v. 16, n. 5, p. 722–735, 2023. DOI: <https://doi.org/10.1109/TLT.2023.3245121>. 19
- MA, C.; OUYANG, J.; XU, G. Learning Latent and Hierarchical Structures in Cognitive Diagnosis Models. *Psychometrika*, Springer, p. 1–33, 2022. DOI: <https://doi.org/10.1007/s11336-022-09867-5>. 44, 45
- MA, W.; MINCHEN, N.; TORRE, J. de la. Choosing between CDM and unidimensional IRT: The proportional reasoning test case. *Measurement: Interdisciplinary Research and Perspectives*, Taylor & Francis, v. 18, n. 2, p. 87–96, 2020. DOI: <https://doi.org/10.1080/15366367.2019.1697122>. 44, 101
- MA, W.; TORRE, J. de la. G-DINA: An R Package for Cognitive Diagnosis Modeling. *Journal of Statistical Software*, v. 93, p. 1–26, 2020. DOI: <https://doi.org/10.18637/jss.v093.i14>. 51
- MAITY, S.; DEROY, A. The Future of Learning in the Age of Generative AI: Automated Question Generation and Assessment with Large Language Models. *arXiv preprint arXiv:2410.09576*, 2024. DOI: <https://doi.org/10.48550/arXiv.2410.09576>. 61, 68, 73
- MALDANER, N.; POZZEBON, E.; SANTOS, T. N. dos. Proposta de Implementação da Computação Afetiva no Sistema Tutor Inteligente MAZK: Conciliando Emoções com o Processo de Aprendizagem. *Revista Novas Tecnologias na Educação – RENOTE*, SBC, v. 20, n. 2, 2022. DOI: <https://doi.org/10.22456/1679-1916.129153>. 22
- MAN, X. Reform Method of Music Teaching Theory System in Colleges and Universities Based on Deep Learning Analysis Technology. *Applied Mathematics and Nonlinear Sciences*, v. 9, 2024. DOI: <https://doi.org/10.2478/amns-2024-0198>. 42
- MANCHANDA, J.; BOETTCHER, L.; WESTPHALEN, M.; JASSER, J. The Open Source Advantage in Large Language Models (LLMs). *arXiv preprint arXiv:2412.12004*, 2024. DOI: <https://doi.org/10.48550/arXiv.2412.12004>. 61, 62, 66
- MARINHO, W.; MARTI, L. Estudo de Técnicas de Modelagem de Proficiência de Alunos para emprego em Sistemas Tutores Inteligentes. In: INSTITUTO DE COMPUTAÇÃO – UFF *Workshop de Pesquisas e Desenvolvimento em Inteligência Artificial, Inteligência Coletiva e Ciência de Dados - WORKPEDIA*. [S.l.], 2016. 33, 34
- MAROUF, A.; AL-DAHDOOH, R.; GHALI, M. J. A.; MAHDI, A. O.; ABUNASSER, B. S.; ABUNASER, S. S. Enhancing Education with Artificial Intelligence: The Role of Intelligent Tutoring Systems. *International Journal of Engineering and Information Systems (IJEAIS)*, v. 8, n. 8, p. 10–16, 2024. Disponível em: <http://ijeais.org/wp-content/uploads/2024/8/IJEAIS240803.pdf>. Acesso em: 26 de junho de 2026. 2
- MATSUDA, N.; WOOD, J.; SHRIVASTAVA, R.; SHIMMEI, M.; BIER, N. Latent Skill Mining and Labeling from Courseware Content. *Journal of Educational Data Mining*, v. 14, 8 2022. DOI: <https://doi.org/10.5281/zenodo.7086211>. 3, 54
- MEDRANO, Á. A. Utilización del modelo G-DINA para promover la evaluación como forma de aprendizaje. *Revista de Estudios y Experiencias en Educación – REXE*, v. 15, n. 28, p. 15–28, 2016. DOI: <https://doi.org/10.21703/rexe.20162815281>. 45

- MENEZES, C. S. d.; JUNIOR, A. N. d. C.; ARAGÓN, R. Arquiteturas pedagógicas para aprendizagem em rede. *Série de livros texto da CEIE-SBC*, v. 7, p. 21–08, 2020. Disponível em: <<https://ieducacao.ceie-br.org/arquiteturas-pedagogicas/>>. Acesso em: 2 de setembro de 2026. 5
- MENEZES, M. F. D.; NETTO, J. F. D. M.; LOPES, A. M. M. Educational Data Mining: Analysis Based On an Intelligent Tutoring System for Teaching Algebra. In: IEEE. 2022 *IEEE Frontiers in Education Conference – (FIE)*. Uppsala, Sweden: IEEE, 2022. p. 1–7. DOI: <https://doi.org/10.1109/FIE56618.2022.9962484>. 15
- MENEZES, M. F. D.; NETTO, J. F. d. M.; LOPES, A. M. M.; BARBOSA, F. M. D. Applying Educational Data Mining to Classify Students in an Intelligent Tutoring System for Algebra Instruction. In: IEEE. 2023 *IEEE Frontiers in Education Conference – (FIE)*. College Station, TX, USA: IEEE, 2023. p. 1–8. DOI: <https://doi.org/10.1109/FIE58773.2023.10342993>. 15
- MENEZES, M. F. D.; NETTO, J. F. d. M.; LOPES, A. M. M.; SILVA, R. C.; MENEZES, R. A. The Proposal of a Dashboard for Analysis and Visualization of Educational Data of Intelligent Tutoring Systems. In: IEEE. 2023 *IEEE Frontiers in Education Conference – (FIE)*. College Station, TX, USA: IEEE, 2023. p. 1–8. DOI: <http://doi.org/10.1109/FIE58773.2023.10343000>. 14
- MENEZES, M. F. de; NETTO, J. F. de M.; LOPES, A. M. M. Visualização de dados educacionais: Desenvolvendo uma ferramenta para análise de dados de sistemas tutores inteligentes. *Revista Novas Tecnologias na Educação – RENOTE*, SBC, v. 22, n. 1, p. 207–216, 2024. DOI: <https://doi.org/10.22456/1679-1916.141548>. 14
- MIN, S.; HE, L. Developing Individualized Feedback for Listening Assessment: Combining Standard Setting and Cognitive Diagnostic Assessment Approaches. *Language Testing*, SAGE Publications Sage UK: London, England, v. 39, n. 1, p. 90–116, 2022. DOI: <https://doi.org/10.1177/0265532221995475>. 45
- MINN, S.; YU, Y.; DESMARAIS, M. C.; ZHU, F.; VIE, J.-J. Deep Knowledge Tracing and Dynamic Student Classification for Knowledge Tracing. In: IEEE. 2018 *IEEE International Conference on Data Mining (ICDM)*. [S.l.]: IEEE, 2018. p. 1182–1187. DOI: <https://doi.org/10.1109/ICDM.2018.00156>. 41
- MONTEIRO, E. J. L. B.; LEITÃO, G. de S.; BARRETO, R. da S. Um método baseado na teoria da resposta ao item para avaliação e feedback automático no contexto de educação digital. *Revista Brasileira de Informática na Educação – RBIE*, v. 29, p. 746–774, 2021. DOI: <https://doi.org/10.5753/rbie.2021.29.0.746>. Disponível em: <<https://journals-sol.sbc.org.br/index.php/rbie/article/view/2969>>. Acesso em: 27 de junho de 2026. 35
- MOORE, S.; SCHMUCKER, R.; MITCHELL, T.; STAMPER, J. Automated Generation and Tagging of Knowledge Components from Multiple-choice Questions. In: *Proceedings of the Eleventh ACM Conference on Learning@ Scale*. Online publication: ACM, 2024. p. 122–133. DOI: <https://doi.org/10.1145/3657604.3662030>. 3, 61, 74
- MORAIS, F. de; JAQUES, P. A. Dinâmica de Afetos em um Sistema Tutor Inteligente de Matemática no Contexto Brasileiro: uma Análise da Transição de Emoções Acadêmicas. *Revista Brasileira de Informática na Educação – RENOTE*, SBC, v. 30, p. 519–541, 2022. DOI: <https://doi.org/10.5753/rbie.2022.2577>. 22

- NAVEED, H.; KHAN, A. U.; QIU, S.; SAQIB, M.; ANWAR, S.; USMAN, M.; AKHTAR, N.; BARNES, N.; MIAN, A. A Comprehensive Overview of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, ACM New York, NY, v. 16, n. 5, p. 1–72, 2025. DOI: <https://doi.org/10.1145/3744746>. 55, 57, 58, 59, 60, 62, 63, 64, 65, 66
- NOGUEIRA, J. V. d. C. M. F.; CASTRO, J. A. P. de; LARCHER, L. O.; FERREIRA, R. V. J.; BARBOSA, B. T.; SOUZA, J. F. de. Uma Análise de Qualidade do uso de Grandes Modelos de Linguagem para Geração Automática de Itens Avaliativos em Português. In: *Anais do XXXVI Simpósio Brasileiro de Informática na Educação (SBIE)*. Curitiba, PR: SBC, 2025. p. 234–247. DOI: <https://doi.org/10.5753/sbie.2025.12210>. 69, 71, 73, 74
- NYE, B. D.; GRAESSER, A. C.; HU, X. AutoTutor and Family: A Review of 17 years of Natural Language Tutoring. *International Journal of Artificial Intelligence in Education*, Springer, v. 24, n. 4, p. 427–469, 2014. DOI: <https://doi.org/10.1007/s40593-014-0029-5>. 23
- OEDA, S.; IIZUMI, S. Data Preprocessing with IRM for Skill Extraction Using NMF. *Procedia Computer Science*, Elsevier, v. 207, p. 2283–2290, 2022. DOI: <https://doi.org/10.1016/j.procs.2022.09.287>. 47
- OLIVEIRA, A. S. F.; ISOTANI, S.; BITTENCOURT, I. I. Tecnologia Educacional na era da IA: explorando a Aplicação de Sistemas Tutores Inteligentes nas Diferentes Áreas do Saber. *Tecnologias, Sociedade e Conhecimento*, Campinas, SP, v. 10, n. 2, p. 68–101, 2023. DOI: <https://doi.org/10.20396/tsc.v10i2.18367>. 4, 17
- OLIVEIRA, G. P. d. Avaliação formativa nos cursos superiores: Verificações qualitativas no processo de ensino-aprendizagem e a autonomia dos educandos. *Revista Iberoamericana de Educación*, v. 15, p. 1–6, 2002. 25, 26
- PAIVA, R. C.; FERREIRA, M. S.; FRADE, M. M. Intelligent Tutorial System Based on Personalized System of Instruction to Teach or Remind Mathematical Concepts. *Journal of Computer Assisted Learning*, Wiley Online Library, v. 33, n. 4, p. 370–381, 2017. DOI: <https://doi.org/10.1111/jcal.12186>. 20
- PALADINES, J.; RAMIREZ, J. A Systematic Literature Review of Intelligent Tutoring Systems with Dialogue in Natural Language. *IEEE Access*, IEEE, v. 8, p. 164246–164267, 2020. DOI: <https://doi.org/10.1109/ACCESS.2020.3021383>. 23
- PAPAMITSIOU, Z.; SHARMA, K.; GIANNAKOS, M. N. On the Dependence Structure Between Learners' Response-time and Knowledge Mastery: If Not Linear, Then What? In: *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. New York, NY, USA: Association for Computing Machinery, 2020. (UMAP '20), p. 201–210. DOI: <https://doi.org/10.1145/3340631.3394865>. 5
- PARAKHIYA, V. *Decoding Large Language Models: A Deep Dive into the Self-Attention Mechanism*. 2023. Medium. Acesso em: 25 maio 2026. Disponível em: <https://medium.com/@vimal.parakhiya/decoding-large-language-models-a-deep-dive-into-the-self-attention-mechanism-967d6207d4ae>. 58
- PARDOS, Z. A.; HEFFERNAN, N. T. Modeling Individualization in a Bayesian Networks Implementation of Knowledge Tracing. In: De Bra, P. and Kobsa, A. and Chin, D. (Ed.). *User Modeling, Adaptation, and Personalization. UMAP 2010*. [S.l.]: Springer Berlin, Heidelberg, 2010. p. 255–266. DOI: https://doi.org/10.1007/978-3-642-13470-8_24. 39

- PARDOS, Z. A.; HEFFERNAN, N. T. Tutor Modeling versus Student Modeling. In: *Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference*. Marco Island, Florida, USA: [s.n.], 2012. p. 420–425. Disponível em: <<https://cdn.aaai.org/ocs/4483/4483-21568-1-PB.pdf>>. Acesso em: 26 de junho de 2026. 39
- PASQUALI, L. Psychometrics. *Revista da Escola de Enfermagem da USP*, SciELO Brasil, v. 43, p. 992–999, 2009. DOI: <https://doi.org/10.1590/S0080-62342009000500002>. 33, 34
- PATEL, A.; KINSHUK; RUSSELL, D. Intelligent tutoring tools for cognitive skill acquisition in lifelong learning. *Journal of Educational Technology & Society*, JSTOR, v. 3, n. 1, p. 32–40, 2000. Disponível em: <<http://www.jstor.org/stable/jeductechsoci.3.1.32>>. Acesso em: 26 de junho de 2026. 28
- PAVLIK, P. I.; CEN, H.; KOEDINGER, K. R. Performance Factors Analysis – A New Alternative to Knowledge Tracing. *Online submission*, ERIC, 2009. DOI: <https://doi.org/10.3233/978-1-60750-028-5-531>. 39, 40
- PEDASTE, M.; MÄEOTS, M.; SIIMAN, L. A.; JONG, T. D.; RIESEN, S. A. V.; KAMP, E. T.; MANOLI, C. C.; ZACHARIA, Z. C.; TSOURLIDAKI, E. Phases of Inquiry-based Learning: Definitions and the Inquiry Cycle. *Educational research review*, Elsevier, v. 14, p. 47–61, 2015. DOI: <https://doi.org/10.1016/j.edurev.2015.02.003>. 9, 10
- PEIKOS, G.; STAVROU, D. ChatGPT for Science Lesson Planning: An Exploratory Study Based on Pedagogical Content Knowledge. *Education Sciences*, v. 15, n. 3, p. 338, 2025. ISSN 2227-7102. DOI: <https://doi.org/10.3390/educsci15030338>. 4
- PELÁNEK, R. Bayesian Knowledge Tracing, Logistic Models, and Beyond: an Overview of Learner Modeling Techniques. *User Modeling and User-Adapted Interaction*, Springer Netherlands, v. 27, p. 313–350, 12 2017. ISSN 15731391. DOI: <https://doi.org/10.1007/s11257-017-9193-2>. 21, 24
- PELÁNEK, R. Managing Items and Knowledge Components: Domain Modeling in Practice. *Educational Technology Research and Development*, Springer, v. 68, p. 529–550, 2020. DOI: <https://doi.org/10.1007/s11423-019-09716-w>. 27, 28, 30, 54
- PELÁNEK, R. Adaptive Learning is Hard: Challenges, Nuances, and Trade-offs in Modeling. *International Journal of Artificial Intelligence in Education*, Springer, v. 35, p. 304–329, 2025. DOI: <https://doi.org/10.1007/s40593-024-00400-6>. 3, 24
- PELÁNEK, R.; EFFENBERGER, T.; ČECHÁK, J. Complexity and difficulty of items in learning systems. *International Journal of Artificial Intelligence in Education*, Springer, v. 32, n. 1, p. 196–232, 2022. DOI: <https://doi.org/10.1007/s40593-021-00252-4>. 3, 31
- PEREIRA, A. J.; GOMES, A. S.; PRIMO, T. T. Evaluation of Human-Artificial Hybrid Tutoring System for Mediation of Engagement in E-Learning. *Revista Brasileira de Informática na Educação – RBIE*, SBC, v. 33, p. 216–246, 2025. DOI: <https://doi.org/10.5753/rbie.2025.4988>. 4, 5
- PEREIRA, D. N.; DEUS, G. B. d.; MENCHIK, J. d. C.; AMESTOY, M. B. ENEM como Política de Avaliação em Larga Escala: uma Revisão das Dissertações e Teses Brasileiras (2011-2021). *Eccos Revista Científica*, Universidade Nove de Julho-UNINOVE, n. 65, 2023. DOI: <https://doi.org/10.5585/eccos.n65.23607>. 32

- PIECH, C.; BASSEN, J.; HUANG, J.; GANGULI, S.; SAHAMI, M.; GUIBAS, L. J.; SOHL-DICKSTEIN, J. Deep Knowledge Tracing. *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, v. 28, 2015. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2015/file/bac9162b47c56fc8a4d2a519803d51b3-Paper.pdf>. Acesso em: 26 de junho de 2026. 41
- PINHEIRO, D. P.; REBOUÇAS, J. A Importância da Avaliação Diagnóstica no Projeto de Nivelamento Matemático com Discentes do Ensino Médio Integrado. In: *V Congresso Nacional de Educação. Brazil*. [s.n.], 2018. Disponível em: <https://www.editorarealize.com.br/editora/anais/conedu/2018/TRABALHO_EV117_MD1_SA13_ID1179_17092018235053.pdf>. Acesso em: 26 de junho de 2026. 1, 26
- PRIMI, R.; SANTOS, A. A.; VENDRAMINI, C. M.; TAXA, F.; MULLER, F. A.; LUKJANENKO, M. d. F.; SAMPAIO, I. S. Competências e Habilidades Cognitivas: Diferentes Definições dos Mesmos Construtos. *Psicologia: Teoria e Pesquisa*, SciELO Brasil, v. 17, p. 151–159, 2001. DOI: <https://doi.org/10.1590/S0102-37722001000200007>. 31
- PRODANOV, C. C.; FREITAS, E. C. D. *Metodologia do Trabalho Científico: Métodos e Técnicas da Pesquisa e do Trabalho Acadêmico-2ª Edição*. Novo Hamburgo, RS: Editora Feevale, 2013. Disponível em: <<https://encurtador.com.br/oZvr>>. Acesso em: 26 de junho de 2026. 8, 9
- PU, Y.; WU, W.; PENG, T.; LIU, F.; LIANG, Y.; YU, X.; CHEN, R.; FENG, P. Embedding Cognitive Framework with Self-attention for Interpretable Knowledge Tracing. *Scientific Reports*, Nature Publishing Group UK London, v. 12, n. 1, p. 17536, 2022. DOI: <https://doi.org/10.1038/s41598-022-22539-9>. 4
- QUINCOZES, C. B.; MOLINOS, D.; ARAÚJO, R. D.; QUINCOZES, S.; GUEDES, G. T. Engenharia de Prompt para a Geração Automatizada de Questões Assistida por LLMs: Uma Análise Comparativa. In: *Anais do XXXVI Simpósio Brasileiro de Informática na Educação (SBIE)*. Curitiba, PR: SBC, 2025. p. 1347–1360. DOI: <https://doi.org/10.5753/sbie.2025.12891>. 61, 69, 72, 73
- RAABE, A. L. A. *Uma Proposta de Arquitetura de Sistema Tutor Inteligente Baseada na Teoria das Experiências de Aprendizagem Mediadas*. Tese (Doutorado) — Universidade Federal do Rio Grande do Sul, 2005. Disponível em: <<https://repositorio.minedu.gob.pe/handle/20.500.12799/156>>. Acesso em: 26 de junho de 2026. 18, 19, 24
- RAHMAN, M. S.; BORERA, E. Variations of Large Language Models: An Overview. *Preprints*, Preprints, December 2024. DOI: <https://doi.org/10.20944/preprints202412.0811.v1>. 59
- RAHMOUNI, N.; LOURDEAUX, D.; BENABBOU, A.; BENSEBAA, T. Diag-Skills: A Diagnosis System Using Belief Functions and Semantic Models in ITS. *Applied Sciences*, MDPI, v. 11, n. 23, p. 11326, 2021. DOI: <https://doi.org/10.3390/app112311326>. 97, 98, 99, 100
- RAIAAN, M. A. K.; MUKTA, M. S. H.; FATEMA, K.; FAHAD, N. M.; SAKIB, S.; MIM, M. M. J.; AHMAD, J.; ALI, M. E.; AZAM, S. A Review on Large Language Models: Architectures, Applications, Taxonomies, Open issues and Challenges. *IEEE access*, IEEE, v. 12, p. 26839–26874, 2024. DOI: <https://doi.org/10.1109/ACCESS.2024.3365742>. 62, 63, 64, 66
- RAMBE, S. L. Creative Techniques for Online Learning Assessment. *Englisia: Journal of Language, Education, and Humanities*, v. 8, n. 2, p. 138–147, 2021. DOI: <https://doi.org/10.22373/ej.v8i2.8111>. 3

- RAMIREZ, C.; SANCHEZ, E. Competences Memory Map: A Model for the Representation of Competences Applied in Education. In: *2012 IEEE 11th International Conference on Cognitive Informatics and Cognitive Computing*. [S.l.]: IEEE, 2012. p. 363–371. DOI: <https://doi.org/10.1109/icci-cc.2012.6311175>. 77, 78, 79
- RAVAND, H.; BAGHAEI, P. Diagnostic Classification Models: Recent developments, Practical issues, and Prospects. *International Journal of Testing*, Taylor & Francis, v. 20, n. 1, p. 24–56, 2020. DOI: <https://doi.org/10.1080/15305058.2019.1588278>. 2, 5, 27, 31, 33, 36, 44, 47, 53, 137
- RAVAND, H.; ROBITZSCH, A. Cognitive Diagnostic Modeling using R. *Practical Assessment, Research, and Evaluation*, v. 20, n. 1, p. 11, 2015. DOI: <https://doi.org/10.7275/5g6f-ak15>. 45, 100
- REDDY, S.; LABUTOV, I.; JOACHIMS, T. Latent Skill Embedding for Personalized Lesson Sequence Recommendation. *CoRR*, abs/1602.07029, 2016. Disponível em: <http://arxiv.org/abs/1602.07029>. Acesso em: 26 de junho de 2026. 3
- REEVE, B. B.; FAYERS, P. Applying Item Response Theory Modelling for Evaluating Questionnaire Item and Scale Properties. In: FAYERS, P. M.; HAYS, R. D. (Ed.). *Assessing Quality of Life in Clinical Trials: Methods and Practice*. 2. ed. Oxford: Oxford University Press, 2005. p. 55–74. DOI: <https://doi.org/10.1093/oso/9780198527695.003.0005>. 35, 36
- RIBEIRO, C. R. A Concepção dos Professores que Atuam nas Séries Iniciais do Ensino Fundamental Sobre Avaliação Diagnóstica, Avaliação Formativa e Avaliação Somativa. *Consult. A*, v. 23, 2014. Disponível em: <https://revista.fait.edu.br/pub/1629>. Acesso em: 03 de maio de 2026. 26
- ROBSON, R. Assessing Competency Instead of Learning Outcomes. In: *International Conference on Artificial Intelligence in Adaptive Education (AIED)*. Beijing, China: [s.n.], 2019. 92, 93, 94
- ROBSON, R.; HAVAS, K.; RAY, R.; SCHATZ, S.; STAFFORD, M.; ROBSON, E.; BREWER, V.; HERNANDEZ, M.; SMITH, B.; GORDON, J. *Competency Framework Development Process Report*. [S.l.], 2020. Disponível em: <https://apps.dtic.mil/sti/tr/pdf/AD1094921.pdf>. Acesso em: 26 de junho de 2026. 28, 92, 93
- ROBSON, R.; POLTRACK, J. Using Competencies to Map Performance Across Multiple Activities. In: *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (IIITSEC)*. [S.l.: s.n.], 2017. v. 17139, p. 1–12. 94
- ROTHER, E. T. Revisão sistemática x revisão narrativa. *Acta Paulista de Enfermagem*, v. 20, n. 2, p. v–vi, 2007. DOI: <https://doi.org/10.1590/S0103-21002007000200001>. 75
- RUPP, A. A.; TEMPLIN, J.; HENSON, R. A. *Diagnostic Measurement: Theory, Methods, and Applications*. New York: Guilford Press, 2010. (Methodology in the Social Sciences). ISBN 9781606235287. 47, 53
- SADLER, D. R. Formative Assessment and the Design of Instructional Systems. *Instructional science*, Springer, v. 18, n. 2, p. 119–144, 1989. DOI: <https://doi.org/10.1007/BF00117714>. 25
- SANI, S. M.; BICHI, A. B.; AYUBA, S. Artificial Intelligence Approaches in Student Modeling: Half Decade Review (2010-2015). *International Journal of Computer Science and Network – (IJCSN)*, v. 5, n. 5, 2016. Disponível em: <https://eprints.rclis.org/30166/>. Acesso em: 19 de setembro de 2026. 21, 22

- SANTOS, M. M.; BARROS, A. P.; SANTOS, E.; SILVA, J. G. da; ISOTANI, S.; BITTENCOURT, I. I.; MACARIO, V.; RODRIGUES, L.; DERMEVAL, D. Geração de Questões com LLMs Leves: Um Estudo Inicial sobre a Percepção de Educadores. In: *Anais XXXVI Simpósio Brasileiro de Informática na Educação (SBIE)*. Curitiba, PR: SBC, 2025. p. 1635–1646. DOI: <https://doi.org/10.5753/sbie.2025.12574>. 70, 72, 74
- SAPOUNTZI, A.; BHULAI, S.; CORNELISZ, I.; KLAVEREN, C. van. Dynamic Models for Knowledge Tracing & Prediction of Future Performance. In: BHULAI, S.; KARDARAS, D.; SEMANJSKI, I. (Ed.). *The Seventh International Conference on Data Analytics – DATA ANALYTICS 2018*. Athens, Greece: IARIA, 2018. p. 121–129. 37, 39
- SARTES, L. M. A.; SOUZA-FORMIGONI, M. L. O. d. Avanços na Psicometria: da Teoria Clássica dos Testes à Teoria de Resposta ao Item. *Psicologia: Reflexão e Crítica*, SciELO Brasil, v. 26, p. 241–250, 2013. DOI: <https://doi.org/10.1590/S0102-79722013000200004>. 33, 34, 35, 36
- SCARLATOS, A.; LIU, N.; LEE, J.; BARANIUK, R.; LAN, A. Training LLM-based Tutors to Improve Student Learning Outcomes in Dialogues. In: Cristea, Alexandra I. and Walker, Erin and Lu, Yu and Santos, Olga C. and Isotani, Seiji (Ed.). *International Conference on Artificial Intelligence in Education*. Cham: Springer Nature Switzerland, 2025. p. 251–266. DOI: https://doi.org/10.1007/978-3-031-98414-3_18. 23
- SCHERER, S. S. Estado Avaliador e Políticas de Resultado na Escola Pública no Brasil. *Interfaces da Educação*, v. 16, n. 45, p. 50–68, 2025. DOI: <https://doi.org/10.61389/inter.v16i45.6122>. 32
- SCHUNK, D. H. *Learning Theories: An Educational Perspective*. 6. ed. Boston: Pearson, 2012. ISBN 978-0-13-707195-1. 28, 32
- SHEMSHACK, A.; SPECTOR, J. M. A Systematic Literature Review of Personalized Learning Terms. *Smart Learning Environments*, Springer, v. 7, n. 1, p. 33, 2020. DOI: <https://doi.org/10.1186/s40561-020-00140-9>. 1
- SHI, Q.; MA, W.; ROBITZSCH, A.; SORREL, M. A.; MAN, K. Cognitively Diagnostic Analysis Using the G-DINA Model in R. *Psych*, Multidisciplinary Digital Publishing Institute, v. 3, n. 4, p. 812–835, 2021. DOI: <https://doi.org/10.3390/psych3040052>. 2, 5, 44, 45, 49
- SHIRKAR, S.; LANGHI, D.; AGRAWAL, S. LLM-based Quiz Generation for Assessments in Learning Management System. In: *2025 Second International Conference on Cognitive Robotics and Intelligent Systems (ICC-ROBINS)*. [S.l.]: IEEE, 2025. p. 464–470. DOI: <https://doi.org/10.1109/ICC-ROBINS64345.2025.11086273>. 69, 73
- SHUTE, V. J.; LEIGHTON, J. P.; JANG, E. E.; CHU, M. Advances in the Science of Assessment. *Educational Assessment*, Taylor & Francis, v. 21, n. 1, p. 34–59, 2016. DOI: <https://doi.org/10.1080/10627197.2015.1127752>. 1
- SILVA, C.; MOREIRA, T.; FERNANDES, I.; PASSOS, C.; DUARTE, J.; GOLDSCHMIDT, R. Sistemas Tutores Inteligentes na Aprendizagem por Competências: Uma Revisão Sistemática da Literatura. In: *Anais do XXXIV Simpósio Brasileiro de Informática na Educação – SBIE*. Porto Alegre, RS, Brasil: SBC, 2023. p. 1120–1132. DOI: <https://doi.org/10.5753/sbie.2023.233628>. 2
- SILVA, J. V. da; ANDRADE, T. L. de. A Efetividade de Sistemas Tutores Inteligentes no ensino: Um Estudo Baseado no Mapeamento Sistemático da Literatura. *Simpósio Brasileiro de Informática na Educação – SBIE*, SBC, Curitiba, PR, p. 483–494, 2025. DOI: <https://doi.org/10.5753/sbie.2025.12468>. 4, 5

SILVA, L. S. P. da; AZEVEDO, D. S. C.; COLVERO, R. B. et al. A implementação das Avaliações em Larga Escala no Brasil: a Trajetória das Provas Saeb e Saers e suas Implicações como Política Pública no Âmbito Escolar. *Revista de Estudos Interdisciplinares*, v. 7, n. 1, p. 1–16, 2025. DOI: <https://doi.org/10.56579/rei.v7i1.1545>. 32

SILVA, R. C. da; NETTO, J. F. de M.; LOPES, A. M. M.; MENEZES, M. F. D.; MENEZES, R. A. ERPLAB: Remote Laboratory for Teaching Robotics and Programming. In: *2023 IEEE Frontiers in Education Conference (FIE)*. College Station, TX, USA: IEEE, 2023. p. 1–7. DOI: <https://doi.org/10.1109/FIE58773.2023.10343379>. 15

SILVA, R. C. da; NETTO, J. F. de M.; MENEZES, M. F. de; LOPES, A. M. M. Percepções dos Estudantes sobre o Laboratório Remoto de Ensino de Robótica e Programação ERPLAB. In: *Simpósio Brasileiro de Informática na Educação (SBIE)*. Rio de Janeiro, RJ: SBC, 2024. p. 951–960. DOI: <https://doi.org/10.5753/sbie.2024.242610>. 15

SILVA, V. Tutores Inteligentes como Mediador para o Ensino e Aprendizagem. *Revista Tecnologias em Projeção*, v. 2, n. 1, 2011. Disponível em: <https://www.projecaociencia.com.br/index.php/Projecao4/article/view/78>. Acesso em: 26 de junho de 2026. 18

SINGH, S.; SINGH, V. An Architecture of Domain Independent and Extensible Intelligent Tutoring System Based on Concept Dependencies and Subject Paths. *International Journal of Advanced Computer Science and Applications*, Science and Information (SAI) Organization Limited, v. 13, n. 5, 2022. DOI: <https://doi.org/10.14569/IJACSA.2022.0130538>. 18

SKLAVAKIS, D.; REFANIDIS, I. The MATHEISIS Meta-knowledge Engineering Framework: Ontology-driven Development of Intelligent Tutoring Systems. *Applied Ontology*, SAGE Publications Sage UK: London, England, v. 9, n. 3-4, p. 237–265, 2014. DOI: <https://doi.org/10.3233/AO-140138>. 18

SNYDER, H. Literature Review as a Research Methodology: An Overview and Guidelines. *Journal of Business Research*, Elsevier, v. 104, p. 333–339, 2019. DOI: <https://doi.org/10.1016/j.jbusres.2019.07.039>. 75

SOARES, D. J. M.; SOARES, T. E. A.; SANTOS, W. D. Análise da Qualidade Psicométrica da prova de matemática do Exame Nacional do Ensino Médio Brasileiro de 2018. *Actualidades Investigativas en Educación*, SciELO Brasil, Online, v. 21, n. 1, p. 86–115, 2021. DOI: <http://dx.doi.org/10.15517/aie.v21i1.42338>. 35

SOARES, S. S. D.; NASCIMENTO, P. A. M. M. Evolução do Desempenho Cognitivo do Brasil de 2000 a 2009 face aos demais países. Instituto de Pesquisa Econômica Aplicada (Ipea), 2011. Disponível em: <http://repositorio.ipea.gov.br/handle/11058/1465>. Acesso em: 19 de setembro de 2026. 32, 34, 35

SOTTILARE, R. A.; FALCO, J. A. d.; CONNOR, J. A. A Guide to Instructional Techniques, Strategies and Tactics to Manage Learner Affect, Engagement, and Grit. In: SOTTILARE, R. A.; GRAESSER, A. C.; HU, X.; GOLDBERG, B. S. (Ed.). *Design Recommendations for Intelligent Tutoring Systems – Vol. 2: Instructional Management*. 1. ed. Orlando, FL: U.S. Army Research Laboratory, 2014. v. 2, p. 7–33. ISBN 978-0-9893923-2-7. 20

SOUZA, A. *Explicando o Funcionamento dos LLMs de uma forma simples!* 2024. Medium. Acesso em: 25 de maio de 2026. Disponível em: <https://medium.com/blog-do-zouza/explicando-o-funcionamento-dos-llm-de-uma-forma-simples-46dee59683cb>. Acesso em: 25 maio 2026. 56

- ST-HILAIRE, F.; VU, D. D.; FRAU, A.; BURNS, N.; FARAJI, F.; POTOCHNY, J.; ROBERT, S.; ROUSSEL, A.; ZHENG, S.; GLAZIER, T.; ROMANO, J. V.; BELFER, R.; SHAYAN, M.; SMOFSKY, A.; DELAROSBIL, T.; AHN, S.; EDEN-WALKER, S.; SONY, K.; CHING, A. O.; ELKINS, S.; STEPANYAN, A.; MATAJOVA, A.; CHEN, V.; SAHRAEI, H.; LARSON, R.; MARKOVA, N.; BARKETT, A.; CHARLIN, L.; BENGIO, Y.; SERBAN, I. V.; KOCHMAR, E. *A New Era: Intelligent Tutoring Systems Will Transform Online Learning for Millions*. 2022. Disponível em: <<https://arxiv.org/abs/2203.03724>>. Acesso em: 26 de junho de 2026. 1
- STORARE, M. A.; COLETO, M. G.; JUNIOR, R. M. S. Mapeamento Sistemático Da Literatura Sobre O Uso De Sistemas Tutores Inteligentes Aplicado Ao Ensino De Computação Na Educação Brasileira. *Revista de Produtos Educacionais e Pesquisas em Ensino*, v. 9, n. 2, p. 363–375, 2025. Disponível em: <<https://periodicos.uenp.edu.br/index.php/reppe/article/view/2063>>. Acesso em: 26 de junho de 2026. 5
- SUEHIRO, A. C. B.; BENFICA, T. d. S.; CARDIM, N. A. Avaliação Cognitiva Infantil nos Periódicos Científicos Brasileiros. *Psicologia: Teoria e Pesquisa*, SciELO Brasil, v. 31, p. 25–32, 2015. DOI: <https://doi.org/10.1590/0102-37722015011755025032>. 27
- SUN, S.; WU, X.; XU, T. A Theoretical Framework for a Mathematical Cognitive Model for Adaptive Learning Systems. *Behavioral Sciences*, MDPI, v. 13, n. 5, p. 406, 2023. DOI: <https://doi.org/10.3390/bs13050406>. 5
- SUN, X.; ZHANG, T.; NIE, C.; SONG, N.; XIN, T. Combining Regularization and Logistic Regression Model to Validate the Q-matrix for Cognitive Diagnosis Model. *British Journal of Mathematical and Statistical Psychology*, Wiley Online Library, v. 78, n. 1, p. 1–21, 2025. DOI: <https://doi.org/10.1111/bmsp.12346>. 135
- SUN, Y.; SUZUKI, M. Diagnostic assessment for Improving Teaching Practice. *International Journal of Information and Education Technology*, IACSIT Press, v. 3, n. 6, p. 607, 2013. DOI: <https://doi.org/10.7763/IJiet.2013.V3.345>. 24, 26
- TANG, R.; ZHU, C.; CHEN, B.; ZHANG, W.; ZHU, M.; DAI, X.; GUO, H. LLM4Tag: Automatic Tagging System for Information Retrieval via Large Language Models. In: *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. Toronto, CA: Association for Computing Machinery, New York, NY, United States, 2025. p. 4882–4890. DOI: <https://doi.org/10.1145/3711896.3737242>. 4, 54
- TATSUOKA, K. K. Analysis of Errors in Fraction Addition and Subtraction Problems (Report No. NIE-G-81-0002). *Urbana: Computer-based Education Research Laboratory, University of Illinois*, 1984. 47, 48, 53, 119
- TEMPLIN, J. L.; HENSON, R. A. Measurement of Psychological Disorders Using Cognitive Diagnosis Models. *Psychological methods*, American Psychological Association, v. 11, n. 3, p. 287, 2006. DOI: <https://doi.org/10.1037/1082-989X.11.3.287>. 50
- TENÓRIO, T.; ISOTANI, S.; BITTENCOURT, I. I. Inteligência Coletiva como Ferramenta de apoio na Construção de Loops Internos em Sistemas Tutores Inteligentes. In: *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. Porto Alegre, RS, Brasil: SBC, 2020. p. 1233–1242. DOI: <https://doi.org/10.5753/cbie.sbie.2020.1233>. 2
- TOMI, S. H. Intelligent Learning Systems in Enhancing Education Through AI: Review Research. *Using AI Tools in Text Analysis, Simplification, Classification, and Synthesis*, IGI

Global Scientific Publishing, p. 1–36, 2025. DOI: <https://doi.org/10.4018/979-8-3693-9511-0.ch001>. 2

TORRE, J. D. L. A Cognitive Diagnosis Model for Cognitively Based Multiple-Choice Options. *Applied Psychological Measurement*, Sage Publications: Los Angeles, CA, v. 33, n. 3, p. 163–183, 2009. DOI: <https://doi.org/10.1177/0146621608320523>. 47, 53

TORRE, J. D. L. DINA Model and Parameter Estimation: A Didactic. *Journal of Educational and Behavioral Statistics*, Sage Publications, Los Angeles, CA, v. 34, n. 1, p. 115–130, 2009. DOI: <https://doi.org/10.3102/1076998607309474>. 48, 49

TORRE, J. D. L. The Generalized DINA Model Framework. *Psychometrika*, Springer-Verlag, v. 76, n. 2, p. 179–199, 2011. DOI: <https://doi.org/10.1007/s11336-011-9207-7>. 48, 49, 50

TORRE, J. D. L.; MINCHEN, N. Cognitively Diagnostic Assessments and the Cognitive Diagnosis Model Framework. *Psicología Educativa*, Elsevier, v. 20, n. 2, p. 89–97, 2014. DOI: <https://doi.org/10.1016/j.pse.2014.11.001>. 31, 33, 44, 48

TORRE, J. d. l.; MINCHEN, N. D. The G-DINA Model Framework. In: von Davier, M.; LEE, Y. S. (Ed.). *Handbook of Diagnostic Classification Models*. [S.l.]: Springer, Cham, 2019, (Methodology of Educational Measurement and Assessment). p. 155–169. DOI: https://doi.org/10.1007/978-3-030-05584-4_7. 50

TOUVRON, H.; LAVRIL, T.; IZACARD, G.; MARTINET, X.; LACHAUX, M.-A.; LACROIX, T.; ROZIÈRE, B.; GOYAL, N.; HAMBRO, E.; AZHAR, F.; RODRIGUEZ, A.; JOULIN, A.; GRAVE, E.; LAMPLE, G. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, v. 10, 2023. DOI: <https://doi.org/10.48550/arXiv.2302.13971>. 63

TOUVRON, H.; MARTIN, L.; STONE, K.; ALBERT, P.; ALMAHAIRI, A.; BABAEI, Y.; BASHLYKOV, N.; BATRA, S.; BHARGAVA, P.; BHOSALE, S.; BIKEL, D.; BLECHER, L.; FERRER, C. C.; CHEN, M.; CUCURULL, G.; ESIOBU, D.; FERNANDES, J.; FU, J.; FU, W.; FULLER, B.; GAO, C.; GOSWAMI, V.; GOYAL, N.; HARTSHORN, A.; HOSSEINI, S.; HOU, R.; INAN, H.; KARDAS, M.; KERKEZ, V.; KHABSA, M.; KLOUMANN, I.; KORENEV, A.; KOURA, P. S.; LACHAUX, M.-A.; LAVRIL, T.; LEE, J.; LISKOVICH, D.; LU, Y.; MAO, Y.; MARTINET, X.; MIHAYLOV, T.; MISHRA, P.; MOLYBOG, I.; NIE, Y.; POULTON, A.; REIZENSTEIN, J.; RUNGTA, R.; SALADI, K.; SCHELLEN, A.; SILVA, R.; SMITH, E. M.; SUBRAMANIAN, R.; TAN, X. E.; TANG, B.; TAYLOR, R.; WILLIAMS, A.; KUANG, J. X.; XU, P.; YAN, Z.; ZAROV, I.; ZHANG, Y.; FAN, A.; KAMBADUR, M.; NARANG, S.; RODRIGUEZ, A.; STOJNIC, R.; EDUNOV, S.; SCIALOM, T. Llama 2: Open Foundation and Fine-tuned Chat Models. *arXiv preprint arXiv:2307.09288*, 2023. DOI: <https://doi.org/10.48550/arXiv.2307.09288>. 63

VALERIANO, E. C. d. F.; CORRÊA, A. P.; POZZEBON, E. O Sistema Tutor Inteligente MAZK no ensino fundamental I. In: *Anais do XXX Simpósio Brasileiro de Informática na Educação (SBIE)*. Brasília, DF: SBC, 2019. p. 616–625. DOI: <https://doi.org/10.5753/cbie.sbie.2019.616>. 23

VANLEHN, K. The Behavior of Tutoring Systems. *International Journal of Artificial Intelligence in Education*, SAGE Publications Sage UK: London, England, v. 16, n. 3, p. 227–265, 2006. DOI: [https://doi.org/10.3233/IRG-2006-16\(3\)02](https://doi.org/10.3233/IRG-2006-16(3)02). 4, 5

VARMA, A. D.; ANAND, S.; THEKKEDATH, A. K.; P, K.; PRASANTH, N.; RAO, S. N. LLM for Question Generation and Validation to Enhance School Level Student Assessment. In: 2025

5th International Conference on Pervasive Computing and Social Networking (ICPCSN). [S.l.]: IEEE, 2025. p. 1609–1615. DOI: <https://doi.org/10.1109/ICPCSN65854.2025.11035415>. 68, 73

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. Attention is All You Need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems – NIPS’17*. Red Hook, NY, USA: Curran Associates Inc., 2017. v. 30, p. 6000–6010. 56, 57, 58

VAZ, R. F. N.; NASSER, L. Um Estudo sobre o Feedback Formativo na Avaliação em Matemática e sua Conexão com a Atribuição de Notas. *Bolema: Boletim de Educação Matemática*, SciELO Brasil, v. 35, p. 3–21, 2021. DOI: <https://doi.org/10.1590/1980-4415v35n69a01>. 25, 26

VEENMAN, M. V.; CLEEF, D. van. Measuring Metacognitive Skills for Mathematics: Students’ Self-Reports versus On-line Assessment Methods. *ZDM*, Springer, v. 51, n. 4, p. 691–701, 2019. DOI: <https://doi.org/10.1007/s11858-018-1006-5>. 2

VICARI, R. M.; GIRAFFA, L. M. Fundamentos dos Sistemas Tutores Inteligentes. *Sociedades Artificiais: a Nova Fronteira da Inteligência das Máquinas*. Porto Alegre: Bookman, 2003. 21, 22, 23

VIE, J.-J.; POPINEAU, F.; BRUILLARD, É.; BOURDA, Y. A Review of Recent Advances in Adaptive Assessment. In: _____. *Learning Analytics: Fundaments, Applications, and Trends. Studies in Systems, Decision and Control*. [S.l.]: Springer, Cham, 2017. p. 113–142. DOI: https://doi.org/10.1007/978-3-319-52977-6_4. 100

VILLANUEVA, N. M.; COSTAS, A. E.; HERMIDA, D. F.; RODRÍGUEZ, A. C. Simplify ITS: An Intelligent Tutoring System based on Cognitive Diagnosis Models and Spaced Learning. In: *2018 IEEE global engineering education conference (EDUCON)*. [S.l.]: IEEE, 2018. p. 1703–1712. DOI: <https://doi.org/10.1109/EDUCON.2018.8363440>. 71, 90, 91, 100, 101

VYGOTSKY, L. Mind in society: the development of higher mental processes harvard university press. *Cambridge, Mass*, 1978. 23, 119, 120

WAFI, M. N.; HUSSANI, S. A. M.; PAZHMAN, J. Evaluation of Students’ Mathematical Ability in Afghanistan’s Schools Using Cognitive Diagnosis Models. *Eurasia Journal of Mathematics, Science and Technology Education*, Modestum, v. 16, n. 6, 2020. DOI: <https://doi.org/10.29333/ejmste/7834>. 45

WANG, D.; HAN, H.; ZHAN, Z.; XU, J.; LIU, Q.; REN, G. A Problem-Solving Oriented Intelligent Tutoring System to Improve Students’ Acquisition of Basic Computer Skills. *Computers & Education*, v. 81, p. 102–112, 2015. ISSN 0360-1315. DOI: <https://doi.org/10.1016/j.compedu.2014.10.003>. 28

WANG, F.; LIU, Q.; CHEN, E.; HUANG, Z.; CHEN, Y.; YIN, Y.; HUANG, Z.; WANG, S. Neural Cognitive Diagnosis for Intelligent Education Systems. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2020. v. 34, n. 04, p. 6153–6161. DOI: <https://doi.org/10.1609/aaai.v34i04.6080>. 19, 151

WANG, H.; TLILI, A.; HUANG, R.; CAI, Z.; LI, M.; CHENG, Z.; YANG, D.; LI, M.; ZHU, X.; FEI, C. Examining the Applications of Intelligent Tutoring Systems in Real Educational Contexts: A Systematic Literature Review from the Social Experiment Perspective. *Education and Information Technologies*, Springer, v. 28, n. 7, p. 9113–9148, 2023. DOI: <https://doi.org/10.1007/s10639-022-11555-x>. 1, 2

- WEI, Y.; CARVALHO, P.; STAMPER, J. KCluster: An LLM-based Clustering Approach to Knowledge Component Discovery. In: Caitlin Mills and Giora Alexandron and Davide Taibi and Giosuè Lo Bosco and Luc Paquette (Ed.). *Proceedings of the 18th International Conference on Educational Data Mining*. Palermo, Italy: International Educational Data Mining Society, 2025. p. 228–240. DOI: <https://doi.org/10.5281/zenodo.15870197>. 61, 65
- WILLIAMSON, J. Cognitive Diagnostic Models and How They Can Be Useful? *Cambridge University Press & Assessment*, Cambridge, 2023. Disponível em: <<https://www.cambridgeassessment.org.uk/Images/701443-cognitive-diagnostic-models-and-how-they-can-be-useful.pdf>>. Acesso em: 26 de junho de 2026. 5, 31
- WONG, K.; LEUNG, K.; KWAN, R.; TSANG, P. E-learning: Developing a Simple Web-based Intelligent Tutoring System Using Cognitive Diagnostic Assessment and Adaptive Testing Technology. In: _____. *Hybrid Learning. Lecture Notes in Computer Science*. [S.l.]: Springer, Berlin, Heidelberg, 2010. v. 6248, p. 23–34. DOI: https://doi.org/10.1007/978-3-642-14657-2_4. 23
- WU, X.; XU, T.; ZHANG, Y. Research on the Data Analysis Knowledge Assessment of Pre-service Teachers from China based on Cognitive Diagnostic Assessment. *Current Psychology*, Springer, v. 42, p. 4885–4899, 2021. DOI: <https://doi.org/10.1007/s12144-021-01836-y>. 45
- XIONG, J.; LUO, Z.; LUO, G.; YU, X.; LI, Y. An exploratory Q-matrix Estimation Method based on Sparse Non-negative Matrix Factorization. *Behavior Research Methods*, Springer, v. 56, n. 7, p. 7647–7673, 2024. DOI: <https://doi.org/10.3758/s13428-024-02442-z>. 54
- YAMAGUCHI, K. Bayesian Diagnostic Classification Models for a Partially Known Q-Matrix. *Journal of Educational and Behavioral Statistics*, SAGE Publications Sage CA: Los Angeles, CA, v. 50, n. 2, p. 331–382, 2025. DOI: <https://doi.org/10.3102/10769986241301055>. 54
- YAMAGUCHI, K.; OKADA, K. Hybrid Cognitive Diagnostic Model. *Behaviormetrika*, Springer, v. 47, n. 2, p. 497–518, 2020. DOI: <https://doi.org/10.1007/s41237-020-00111-x>. 27
- YAMAGUCHI, K.; ZHANG, J. Fully Gibbs Sampling Algorithms for Bayesian Variable Selection in Latent Regression Models. *Journal of Educational Measurement*, Wiley Online Library, v. 60, n. 2, p. 202–234, 2023. DOI: <https://doi.org/10.1111/jedm.12348>. 52
- YEUNG, C.-K. Deep-IRT: Make Deep Learning Based Knowledge Tracing Explainable Using Item Response Theory. *arXiv preprint arXiv:1904.11738*, 2019. DOI: <https://doi.org/10.48550/arXiv.1904.11738>. 101
- YUAN, L.; LIU, Y.; CHEN, P.; XIN, T. Development of a New Learning Progression Verification Method based on the Hierarchical Diagnostic Classification Model: Taking Grade 5 Students' Fractional Operations as an Example. *Educational Measurement: Issues and Practice*, Wiley Online Library, 2022. DOI: <https://doi.org/10.1111/emip.12488>. 45
- ZAFFALON, F.; VARGAS, A. P.; SOUZA, R. L. de; PENNA, R.; BEZ, J.; TONIN, N.; BOTELHO, S. S. da C. Um Estudo Comparativo entre dois Modelos que estimam a Habilidade dos Estudantes: ELO e Teoria de Resposta ao Item. In: *Simpósio Brasileiro de Informática na Educação – SBIE*. Brasília, DF: SBC, 2019. v. 30, n. 1, p. 459–468. DOI: <https://doi.org/10.5753/cbie.sbie.2019.459>. 2

- ZERKOUK, M.; MIHOUBI, M.; CHIKHAOUI, B. A comprehensive review of AI-based Intelligent Tutoring Systems: Applications and Challenges. *arXiv preprint arXiv:2507.18882*, 2025. DOI: <https://doi.org/10.48550/arXiv.2507.18882>. 18, 19, 20, 23
- ZHANG, J.; SHI, X.; KING, I.; YEUNG, D.-Y. Dynamic Key-Value Memory Networks for Knowledge Tracing. In: *Proceedings of the 26th international conference on World Wide Web*. Perth, Australia: International World Wide Web Conferences Steering Committee Republic and Canton of Geneva Switzerland, 2017. p. 765–774. DOI: <https://doi.org/10.1145/3038912.3052580>. 41
- ZHANG, M.; JIAO, C.; LI, G.; ZHU, H.; ZHANG, W. Item Selection Using the Multiple-Strategy RRUM Based on CD-CAT. *Acta Psychologica Sinica*, v. 47, n. 12, p. 1511, 2015. DOI: <https://doi.org/10.3724/SPJ.1041.2015.01511>. 51
- ZHOU, X.; ZHANG, Z.; XIE, X.; ZHANG, J. Deep Learning based Knowledge Tracing in Intelligent Tutoring Systems. *Scientific Reports*, Nature Research, v. 15, 12 2025. DOI: <https://doi.org/10.1038/s41598-025-07422-7>. 5
- ZHUANG, Y.; LIU, Q.; BI, H.; HUANG, Z.; HUANG, W.; LI, J.; YU, J.; LIU, Z.; HU, Z.; HONG, Y. et al. Survey of Computerized Adaptive Testing: a Machine Learning Perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, p. 1–21, 2026. DOI: <https://doi.org/10.1109/TPAMI.2026.3672850>. 5
- ZOMPERO, A. d. F.; LABURÚ, C. E.; VILAÇA, T. Instrumento analítico para Avaliar Habilidades Cognitivas dos Estudantes da Educação básica nas Atividades de Investigação. *Investigação em Ensino de Ciências – IENCI*, Universidade Federal do Rio Grande do Sul (UFRGS), v. 24, p. 200–211, 2019. DOI: <https://doi.org/10.22600/1518-8795.ienci2019v24n2p200>. 32

APÊNDICE**A****AVALIAÇÃO PRÉ- E PÓS-TESTE**

AVALIAÇÃO PRÉ-TESTE

- 1) Nome _____
- 2) Idade _____
- 3) Gênero () Masculino () Feminino
- 4) Turma _____ Turno _____

Resolva para x, onde:

a) $8x = 16$

b) $9x + 4 = 85$

c) $5(8x + 6) = 5(6x + 8)$

AVALIAÇÃO PÓS-TESTE

- 1) Nome _____
- 2) Idade _____
- 3) Gênero () Masculino () Feminino
- 4) Turma _____ Turno _____

Resolva para x, onde:

a) $8x = 56$

b) $3x - 9 = 9$

c) $3(2x + 1) = 5(5x + 3)$

APÊNDICE

B

PIPELINE EXPERIMENTAL PARA GERAÇÃO E AVALIAÇÃO DE Q-MATRIZES

Algorithm 1 Experimental pipeline for Q-matrix generation and evaluation

Input: E : set of assessment items C : set of latent concepts/skills $Q^{\text{EXP}} \in \{0, 1\}^{|E| \times |C|}$: expert-defined Q-matrix L : set of large language models (LLM) P : set of prompting strategies R : number of independent repetitions**Output:** aggregated structural and functional evaluation metrics**GENERATE_Q-MATRIX**($E, C, Q^{\text{EXP}}, L, P, R$)**for all** $l \in L$ **do** **for all** $p \in P$ **do** **for** $r = 1$ to R **do** Initialize $Q^{\text{LLM}} \leftarrow \emptyset$ **for all** $e_j \in E$ **do** Infer $\hat{C}_j \subseteq C$ using model l with strategy p Encode $\hat{C}_j$ into row j of Q^{LLM} **end for** Align Q^{LLM} with Q^{EXP}

Compute structural metrics: HL, JSI

 Evaluate Q^{LLM} using DINA

Compute functional metrics: RMSE, MAE

end for **end for****end for****return** aggregated metrics over R

 Fonte: [Lopes e Netto \(2026\)](#).

Disponibilidade do código e materiais de reprodução

O código-fonte e os materiais utilizados na implementação do pipeline experimental apresentado neste apêndice encontram-se disponíveis em repositório público e foram

preservados em uma versão específica no Zenodo.

DOI: <<https://doi.org/10.5281/zenodo.22852027>>

O registro contém a versão do código utilizada nos experimentos do Estudo 3, incluindo as rotinas de comunicação com os modelos de linguagem, processamento das respostas, construção das Q-matrizes e execução das avaliações estrutural e funcional.

APÊNDICE

C

DECLARAÇÃO DE USO DE FERRAMENTAS DE INTELIGÊNCIA ARTIFICIAL

Eu, **Arcanjo Miguel Mota Lopes**, declaro que utilizei ferramentas de Inteligência Artificial no processo de elaboração da tese intitulada “**Uma Arquitetura Conceitual Integrando Diagnóstico Cognitivo e Modelos de Linguagem de Grande Escala para Avaliação de Habilidades**”, apresentada ao **Programa de Pós-Graduação em Informática da Universidade Federal do Amazonas – UFAM¹**, conforme detalho abaixo.

1. Ferramentas utilizadas

Foram utilizadas as seguintes ferramentas de Inteligência Artificial:

- **ChatGPT (versão paga), Gemini, Deepseek, Grammarly, Language Tools, Google Translator, Deepl Write, Consensus e Research Rabbit.**

2. Etapas em que foram utilizadas

As ferramentas de IA foram utilizadas nas seguintes etapas do trabalho:

- a) apoio à organização preliminar de ideias e estruturação;
- b) revisão de linguagem, com foco em gramática, ortografia, coesão, clareza e fluidez textual;

¹ Declaração elaborada com base no referencial divulgado pela UFAM sobre o uso ético, seguro e estratégico da Inteligência Artificial nas universidades federais, disponível em: <<https://www.ufam.edu.br/noticias/7254-universidades-federais-elaboram-referencial-para-o-uso-da-inteligencia-artificial.html>>; na Política de Integridade na Atividade Científica do CNPq, disponível em: <http://memoria2.cnpq.br/web/guest/view/-/journal_content/56_INSTANCE_0oED/10157/23142775>; e no modelo de “Declaração de Uso de Inteligência Artificial”, disponível em: <<https://revistas.marilia.unesp.br/index.php/index/declaracaoia>>

- c) auxílio na reformulação de frases e parágrafos previamente redigidos pelo autor;
- d) apoio à tradução ou revisão de termos em língua estrangeira, quando necessário;
- e) apoio à padronização textual e à conferência da formatação de referências, sem substituição da verificação manual das fontes originais.
- f) apoio no desenvolvimento do código na linguagem Python descrito no Apêndice B. As decisões de implementação, integração dos componentes, execução dos experimentos, testes, correções e validação dos resultados permaneceram sob responsabilidade do autor;
- g) no Estudo 3 (ver Seção 5.3), os LLM avaliados constituíram parte do próprio objeto experimental da pesquisa, sendo utilizados para produzir sugestões de associações item-CC e gerar Q-matrizes posteriormente submetidas às avaliações estrutural e funcional. Esse uso experimental deve ser distinguido do uso de ferramentas de IA como apoio à redação e à programação da tese.

3. Seções da tese em que foram utilizadas

O uso das ferramentas ocorreu de forma pontual e supervisionada nas seguintes seções da tese: **Resumo, Abstract, Introdução, Referencial teórico, Metodologia, Análise e discussão dos resultados, Considerações finais**. Na seção de **Referências**, as ferramentas foram utilizadas como apoio à padronização formal.

Nas seções textuais indicadas, o uso das ferramentas de IA concentrou-se no apoio redacional, na melhoria da clareza textual e na organização da exposição acadêmica. Adicionalmente, ferramentas de IA foram utilizadas de forma pontual como apoio ao desenvolvimento, à revisão e à correção de trechos do código Python empregado no pipeline computacional do Estudo 3 (ver Seção 5.3). Em todos os casos, as sugestões produzidas foram analisadas, testadas, corrigidas e validadas pelo autor, não sendo utilizadas para produzir autonomamente dados, resultados, análises ou conclusões da pesquisa.

Todo conteúdo sugerido foi analisado, revisado, corrigido e validado pelo autor antes de sua incorporação à versão final da tese. As ferramentas não foram utilizadas como autoras ou coautoras do trabalho, tampouco foram empregadas para fabricar dados, criar referências inexistentes, manipular resultados, alterar imagens, omitir fontes, plagiar conteúdos ou comprometer a integridade científica da pesquisa.

Declaro, ainda, que assumo total responsabilidade pelo conteúdo da tese, incluindo sua originalidade, precisão, validade, integridade ética, adequação metodológica, interpretação dos resultados e conformidade com as normas acadêmicas da Universidade Federal do Amazonas, do Programa de Pós-Graduação ao qual o trabalho está vinculado e das diretrizes nacionais de integridade científica.

