



**Instituto Federal de Educação, Ciência e Tecnologia da
Paraíba Campus João Pessoa**

Programa de Pós-Graduação em Tecnologia da Informação

Nível Mestrado Profissional

LIANA AMÂNCIO DOS SANTOS GONÇALVES

**Uma Abordagem para Geração de Avaliações de
Aprendizagem Adaptadas a partir de Prática Inclusiva e
Preferências Discentes**

DISSERTAÇÃO DE MESTRADO

JOÃO PESSOA – PB

2026

Uma Abordagem para Geração de Avaliações de Aprendizagem Adaptadas a partir de Prática Inclusiva e Preferências Discentes

Dissertação de Mestrado apresentada como requisito final para obtenção do título de Mestre em Tecnologia da Informação pelo Programa de Pós-Graduação em Tecnologia da Informação do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba – IFPB.

Orientadora: Profa. Dra. Damires Souza

João Pessoa – PB 2026

Dados Internacionais de Catalogação na Publicação (CIP)
Biblioteca Nilo Peçanha - *campus* João Pessoa, PB.

G635a Gonçalves, Liana Amâncio dos Santos.

Uma abordagem para geração de avaliações de aprendizagem adaptadas a partir de prática inclusiva e preferências discentes / Liana Amâncio dos Santos Gonçalves. – 2026.

114 f. : il.

Dissertação (Mestrado em Tecnologia da Informação) – Instituto Federal de Educação da Paraíba / Programa de Pós-Graduação em Tecnologia da Informação (PPGTI), 2026.

Orientação: Profa. Dra. Damires Yluska de Souza Fernandes.

1.Prática inclusiva. 2. TEA. 3. Avaliação de aprendizagem adaptativa. 4. Avaliação formativa. 5. Ensino médio. I. Título.

CDU

37.091.26:376(043)

LIANA AMÂNCIO DOS SANTOS GONÇALVES

**UMA ABORDAGEM PARA GERAÇÃO DE AVALIAÇÕES DE APRENDIZAGEM ADAPTADAS
A PARTIR DE PRÁTICA INCLUSIVA E PREFERÊNCIAS DISCENTES**

DISSERTAÇÃO submetida Programa de Pós-Graduação em Tecnologia da Informação do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba - IFPB, Campus João Pessoa, como requisito para obtenção do título de **Mestre em Tecnologia da Informação**.

Aprovado em 06/05/2026 14:00.

Membros da Banca Examinadora:

Profa. Dra. DAMIRES YLUSKA DE SOUZA FERNANDES

Instituto Federal da Paraíba (IFPB)

Orientadora

Prof. Dr. FRANCISCO PETRÔNIO ALENCAR DE MEDEIROS

Instituto Federal da Paraíba (IFPB)

Examinador

Profa. Dra. VALÉRIA FARINAZZO MARTINS

Universidade Presbiteriana Mackenzie (UPM)

Examinadora

Documento assinado eletronicamente por:

- **Damires Yluska de Souza Fernandes, PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 07/05/2026 11:18:10.
- **Francisco Petronio Alencar de Medeiros, PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 11/05/2026 08:17:55.
- **VALÉRIA FARINAZZO MARTINS, PRESTADOR DE SERVIÇO**, em 31/07/2026 13:10:33.

Este documento foi emitido pelo SUAP em 24/04/2026. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse <https://suap.ifpb.edu.br/autenticar-documento/> e forneça os dados abaixo:

Código 870084
Verificador: dd2cececf4
Código de Autenticação:



Este trabalho é dedicado aos professores que acreditam na educação inclusiva e às pessoas com Transtorno do Espectro Autista, que nos ensinam, diariamente, novas formas de aprender e compreender o mundo.

.

AGRADECIMENTOS

Agradeço a toda minha família, meus amigos e colegas. A todos os professores do Programa de Mestrado em Tecnologia da Informação - PPGTI do IFPB, em especial à minha professora e orientadora, Dr. Damires Yluska de Souza Fernandes e aos alunos que participaram do Projeto de Extensão no desenvolvimento do MeuAvalia. O presente trabalho foi realizado com apoio do Instituto Federal da Paraíba (IFPB).

RESUMO

Diante dos desafios do cenário educacional da Educação Básica, marcado pela predominância de práticas avaliativas tradicionais e pela dificuldade em contemplar a diversidade de perfis dos estudantes, especialmente em contextos inclusivos, como aqueles associados ao Transtorno do Espectro Autista (TEA), abordagens baseadas em ferramentas tecnológicas têm se revelado alternativas potenciais para a adaptação da aprendizagem. Nesse contexto, este trabalho apresenta uma abordagem denominada MeuAvalia, que busca apoiar docentes na geração de avaliações de aprendizagem formativas adaptadas no Ensino Médio. Para isso, considera a prática inclusiva do Ensino Estruturado, tópicos de conteúdo presentes no plano de ensino da disciplina, e as preferências avaliativas dos estudantes. A abordagem foi implementada por meio de uma aplicação que faz uso de um modelo de linguagem amplo (LLM) para gerar as avaliações de modo automatizado. Uma avaliação experimental foi conduzida com base no *ground truth* obtido e na métrica de similaridade textual BERTScore. Os resultados obtidos indicaram que a abordagem é promissora, tendo em vista um nível moderado de similaridade semântica entre os textos gerados com respeito às avaliações originais. Mesmo com a preservação semântica, observa-se que as questões geradas tendem a ser mais objetivas e diretas, mantendo o tema central, porém com menor nível de aprofundamento do conteúdo.

Palavras-chaves: Prática Inclusiva; TEA; Avaliação de aprendizagem adaptada; Avaliação formativa; LLM; Ensino Médio.

ABSTRACT

Given the challenges of the Basic Education scenario, marked by the predominance of traditional assessment practices and difficulties in addressing the diversity of student profiles, especially in inclusive contexts such as those related to the Autism Spectrum Disorder (ASD), technology-based approaches have emerged as promising alternatives for adapting learning processes. In this light, this work presents an approach named *MeuAvalia*, which aims to support teachers in generating adaptive formative learning assessments for High School. The proposal is based on three main elements: inclusive practices grounded in Structured Teaching, content topics defined in the course syllabus, and students' assessment preferences. The approach was implemented through an application that uses a Large Language Model (LLM) to automatically generate assessments. An experimental evaluation was conducted based on the ground truth data and the BERTScore textual similarity metric. The results indicate that the approach is promising, since evidences of a moderate level of semantic similarity between the generated and reference assessments have been obtained. Although semantic preservation is achieved, it is also observed that the generated questions tend to be more objective and concise, presenting a lower level of depth compared to the original assessments

Keywords: Inclusive practice; ASD; Adaptive assessment; Formative assessment; LLM; High School.

LISTA DE FIGURAS

Figura 1 - Exemplo de questão objetiva em uma avaliação de aprendizagem.	26
Figura 2-Exemplo de questão discursiva.....	27
Figura 3 - Elementos que compõem a estrutura de um prompt.....	30
Figura 4- Etapas e arquitetura do sistema QGen	37
Figura 5- Módulos da arquitetura da plataforma Mind Bridge	41
Figura 6- O processo metodológico da plataforma	42
Figura 7 - Prompt para geração dos itens matemáticos	45
Figura 8- Exemplo da questão de múltipla escolha gerada	46
Figura 9 - Prompt de geração de perguntas.	48
Figura 10 – Visão geral da abordagem MeuAvalia	53
Figura 11 – Algoritmo 1: Lógica do Motor Gerador de Avaliação Adaptada	58
Figura 12 - Fragmento de uma avaliação tradicional.	64
Figura 13 - Fragmento da AA.....	64
Figura 14- Diagrama de Casos de Uso do MeuAvalia	66
Figura 15 – Diagrama de atividades representando o funcionamento geral da ferramenta MeuAvalia.	67
Figura 16 - Arquitetura de Componentes da Implementação do MeuAvalia.....	69
Figura 17 - Registro das preferências avaliativas na visão do aluno.....	71
Figura 18 - Geração da avaliação na visão do docente.....	71
Figura 19 - Estrutura do template de prompt parametrizado utilizado no MeuAvalia.....	73
Figura 20- Instanciação do prompt para avaliação objetiva.	74
Figura 21 - Instanciação do prompt para avaliações dos tipos discursiva, prática e oral.....	74
Figura 22 – Exemplo de Avaliação Adaptada	75
Figura 23 - Médias de precisão, revocação e F1-score por tipo de AA.	79
Figura 24: Exemplo da AA oral de Português gerada pelo MeuAvalia.	80
Figura 25 – Exemplo de avaliação objetiva de Português do 3º ano gerada pela abordagem MeuAvalia.	81
Figura 26- Média F1 por Disciplina	84

LISTA DE TABELAS

Tabela 1 – Nível de suporte versus alguns aspectos.....	20
Tabela 2- Caracterização quantitativa do conjunto de dados de avaliações.....	78
Tabela 3 - Comparação entre questões da avaliação original e da avaliação gerada.....	79
Tabela 4- Estatísticas descritivas conforme tipo de método avaliativo.....	79
Tabela 5- Exemplos de questões das avaliações docentes e das avaliações geradas pelo MeuAvalia.....	83

LISTA DE QUADROS

Quadro 1 - Síntese dos estudos relacionados à pesquisa.....	51
Quadro 2 - Preferências Avaliativas associadas aos Métodos Avaliativos.....	57
Quadro 3- Preferências Instrucionais associadas aos Métodos Avaliativos.....	57
Quadro 4- Exemplo de preferências com pesos.....	61
Quadro 5 – Exemplo do prompt utilizado pelo motor gerador de avaliações adaptadas.....	63

LISTA DE ABREVIATURAS E SIGLAS

IA Inteligência Artificial

LLM Large Language Model

TEA Transtorno do Espectro Autista

SUMÁRIO

1. INTRODUÇÃO.....	15
1.1 Motivação e Definição do Problema.....	15
1.2 Objetivos.....	17
1.3 Contribuições esperadas	18
1.4 Estrutura do Documento.....	18
2. FUNDAMENTAÇÃO TEÓRICA.....	19
2.1 Transtorno do Espectro Autista (TEA).....	19
2.2 Processo de ensino e aprendizagem na Educação inclusiva.....	21
2.3 Ensino Estruturado como Prática Inclusiva.....	22
2.4 Avaliação da Aprendizagem.....	24
2.5 Modelos de Linguagem.....	28
2.5.1 Engenharia de Prompts.....	29
2.5.2 Avaliação em LLMs.....	31
2.5.3 Métrica de Avaliação	32
2.6 Uso de LLMs para geração da avaliação de aprendizagem.....	33
2.7 Considerações.....	35
3. TRABALHOS RELACIONADOS.....	37
3.1 Bhowmick et al., (2023).....	37
3.2 Silva et al. (2024).....	41
3.3 Meissner et al., (2024).....	44
3.4 Wang et al., (2025).....	47
3.5 Diyab et al., (2025).....	48
3.6 Síntese dos trabalhos relacionados	51
4. Abordagem MeuAvalia.....	54
4.1 Visão geral da abordagem MeuAvalia.....	54
4.2 Motor gerador de Avaliação Adaptada.....	58
4.3 Exemplo de uso.....	63
4.4 Considerações do capítulo.....	65
5. Implementação e Experimentos.....	66
5.1 Ferramenta MeuAvalia.....	66
5.2 Implementação do Motor gerador de Avaliação Adaptada.....	72
5.2.1 Modelo de Linguagem e Estratégia de Prompting.....	72
5.3 Exemplo de uma avaliação gerada pela ferramenta.....	75
5.4 Avaliação experimental	76
5.4.1 Ambiente Experimental.....	77
5.4.2 Conjunto de dados	77
5.4.2 Resultados e Discussão	78
5.5 Considerações sobre a QP1.....	87

5.6 Considerações sobre a QP2.....	87
6. Considerações Finais.....	88
6.1 Trabalhos Futuros.....	88
6.2 Cuidados éticos.....	88
6.3 Declaração sobre uso de Inteligência Artificial.....	89
REFERÊNCIAS BIBLIOGRÁFICAS.....	90
APÊNDICES.....	98
APÊNDICE A – Questionários utilizados na elicitação de requisitos.....	98
APÊNDICE B – Experimentos Para Escolha Do LLM.....	104
APÊNDICE C – Resultados da similaridade semântica por algumas questões.....	105
ANEXO A -Avaliação gerada pelos Modelos.....	107

1. INTRODUÇÃO

Este capítulo aborda uma visão geral desta pesquisa e descreve o contexto no qual o trabalho de mestrado está inserido. Inicialmente, são expostas as motivações que impulsionaram o desenvolvimento do trabalho proposto. A partir dessa motivação, é definido o problema de pesquisa que orienta este estudo. Em seguida, são apresentados os objetivos pretendidos com a pesquisa, bem como as contribuições esperadas, tanto no âmbito acadêmico quanto no contexto aplicado educacional. Por fim, descreve-se a estrutura geral da dissertação.

1.1 Motivação e Definição do Problema

O ensino médio é uma etapa fundamental da educação básica, pois funciona como uma ponte entre a formação geral do estudante e suas escolhas futuras, sendo um período em que diferentes perfis de estudantes, geralmente na adolescência e juventude inicial, se manifestam de forma mais evidente, o que intensifica a necessidade de práticas educacionais inclusivas.

A educação inclusiva no ensino médio deve ser compreendida como um conjunto de princípios que orienta práticas pedagógicas voltadas à valorização da diversidade e à garantia do direito de aprendizagem para todos os estudantes, uma vez que as barreiras à aprendizagem podem estar nos métodos e materiais educacionais, e não nas capacidades dos estudantes (Rose, 2002). Exemplos de práticas pedagógicas inclusivas incluem o uso de instruções claras e objetivas, a organização sequencial das atividades, o emprego de recursos visuais para apoiar a compreensão dos conteúdos, a previsibilidade das tarefas por meio de rotinas estruturadas e a adaptação dos instrumentos de avaliação às necessidades dos estudantes (Schopler et al., 1995; Silva, 2022). Tais estratégias podem ser observadas em situações como a apresentação de roteiros visuais para execução de atividades, a divisão de tarefas complexas em etapas menores, o uso de cronogramas previamente definidos e a oferta de diferentes formatos avaliativos para que os estudantes demonstrem seus conhecimentos (Rasmussen et al., 2021; Tanaka et al., 2024).

Nesse sentido, segundo Paulo Freire (Freire, 1996), o processo educativo deve reconhecer e valorizar as diferenças entre os sujeitos, promovendo uma aprendizagem baseada no diálogo e no respeito às singularidades. De forma complementar, Montessori (2013) destaca que a educação deve considerar as potencialidades individuais dos estudantes. Esse princípio estende-se também ao processo avaliativo, uma vez que a avaliação constitui parte integrante do ensino e da aprendizagem, reforçando a necessidade de práticas avaliativas mais flexíveis e adaptáveis no ensino médio.

As práticas avaliativas compreendem os instrumentos utilizados para acompanhar, orientar e melhorar o processo de ensino e aprendizagem, tanto sob a perspectiva do estudante quanto do docente (Almeida et al., 2023; Sousa, 2022). Nesse contexto, observa-se um movimento crescente em direção ao desenvolvimento de avaliações mais inclusivas e adaptadas, impulsionado pela diversidade de perfis presentes nas salas de aula contemporâneas (Adlit et al., 2023).

Dentre esses perfis, destacam-se os estudantes com Transtorno do Espectro Autista (TEA), que frequentemente apresentam necessidades específicas relacionadas à comunicação, interação social e formas de processamento da informação. No entanto, os modelos avaliativos tradicionalmente adotados no ensino médio tendem a não considerar adequadamente essas particularidades, o que pode comprometer a participação e o desempenho acadêmico desses estudantes (Azevedo et al., 2024; Perondi et al., 2025; Freiá et al., 2023).

Embora haja reconhecimento da importância de avaliações mais inclusivas, professores do ensino médio enfrentam dificuldades práticas para adaptar seus modelos avaliativos às diferentes necessidades dos estudantes (Fathoni et al., 2024; Tai et al., 2023). Fatores como turmas numerosas, limitação de tempo para planejamento pedagógico, escassez de recursos didáticos e pressão por desempenho em avaliações externas contribuem para a manutenção de práticas avaliativas padronizadas (Casagrande et al., 2025).

Ainda que tais dificuldades sejam particularmente evidentes no contexto do TEA, elas também se estendem ao campo da avaliação educacional de forma mais ampla (Rasmussen et al., 2021). Nesse cenário, observa-se a busca por ferramentas que apoiem o docente na formulação de avaliações mais adaptadas aos diferentes perfis de estudantes (Owan et al., 2023; Mithi et al., 2024; Mahamuni et al., 2024; Missaoui et al., 2021). Apoiar esse processo implica auxiliar o docente na seleção e organização de métodos avaliativos, na adaptação das atividades às características dos estudantes e na promoção de práticas mais acessíveis e inclusivas (Almeida et al., 2023; Oliveira; Sousa, 2022; Fathoni et al., 2024).

As dificuldades relatadas na literatura também podem ser observadas em contextos educacionais específicos. Considerando cenários de escolas públicas estaduais da Paraíba, particularmente a Escola Cidadã Integral Técnica (ECIT) Liceu Paraibano e a ECIT Escritor Horácio de Almeida, observam-se desafios relacionados à adaptação de avaliações tradicionais à diversidade de estudantes. Em muitos casos, não se considera que alunos com TEA e outras necessidades educacionais podem apresentar preferências por diferentes formatos avaliativos, como avaliações práticas, discursivas, orais ou objetivas. Além disso, a centralidade na

transmissão de conteúdo durante a avaliação tende a limitar a adoção de estratégias mais flexíveis e inclusivas.

O contexto apresentado evidencia uma lacuna entre a necessidade de práticas avaliativas inclusivas, com atenção especial aos estudantes com TEA e às condições reais de implementação no cotidiano docente. Nesse panorama, originam-se as questões de pesquisa que norteiam este trabalho:

QP1: Como apoiar o docente na elaboração de avaliações de aprendizagem adaptadas, considerando prática inclusiva de aprendizagem e a diversidade de estudantes?

QP2: Como avaliar o alinhamento semântico das avaliações geradas em relação às avaliações de referência?

Para responder às questões postas, apresenta-se uma abordagem denominada de MeuAvalia, voltada à geração automatizada de Avaliações de Aprendizagem Adaptadas (AAA) para apoio a docentes. A abordagem baseia-se em processo avaliativo formativo, em uma prática inclusiva de aprendizagem e no uso de preferências avaliativas dos estudantes como forma de atendimento à diversidade.

A abordagem proposta é implementada por meio de uma ferramenta que faz uso de um modelo de linguagem grande (do inglês, *Large Language Model* – LLM) para a geração das avaliações. Uma avaliação experimental é conduzida para verificar a similaridade semântica das avaliações obtidas.

1.2 Objetivos

Este trabalho tem como objetivo geral propor e desenvolver uma abordagem computacional para a geração de avaliações de aprendizagem formativas adaptadas no Ensino Médio que possa ser usada como suporte ao docente.

Como objetivos específicos, destacam-se:

- Integrar à abordagem proposta uma prática inclusiva, o plano de ensino da disciplina em pauta e preferências que compõem o perfil avaliativo dos estudantes;
- Propor e desenvolver um motor de geração de avaliações;
- Implementar ferramenta que suporte à abordagem.

1.3 Contribuições esperadas

As principais contribuições desta pesquisa são:

- Abordagem e a ferramenta MeuAvalia, que integra uma prática inclusiva de aprendizagem, tópicos do plano de ensino e preferências avaliativas dos estudantes no contexto do Ensino Médio;
- Consideração das preferências avaliativas dos alunos como elemento central associado à diversidade dos estudantes e na adaptação dos modelos avaliativos;
- Contribuição para a inclusão educacional, especialmente de estudantes com TEA, ao propor a prática inclusiva do Ensino Estruturado no processo adaptativo dos modelos avaliativos;
- Avaliação experimental do motor de geração de avaliações de aprendizagem adaptadas, a partir de uma métrica de similaridade textual.

1.4 Estrutura do Documento

Além do presente capítulo, este trabalho está dividido da seguinte forma: o Capítulo 2 apresenta a fundamentação teórica acerca dos principais conceitos que embasam esta pesquisa. O Capítulo 3 discute os trabalhos relacionados. O Capítulo 4 descreve a abordagem proposta. O Capítulo 5 apresenta a ferramenta desenvolvida e os experimentos realizados para avaliação das avaliações geradas. Por fim, o Capítulo 6 reúne as considerações finais da pesquisa, incluindo as contribuições alcançadas, as limitações do estudo, as sugestões para trabalhos futuros, os aspectos éticos da pesquisa e a declaração de uso da Inteligência Artificial.

2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, são introduzidos conceitos relacionados aos temas centrais pertinentes ao desenvolvimento deste trabalho. O Transtorno do Espectro Autista (TEA) é abordado juntamente com aspectos da educação inclusiva e dos processos avaliativos. Adicionalmente, são discutidos tópicos sobre Modelos de Linguagem, Engenharia de Prompts e uso de LLMs e ferramentas computacionais para a geração de avaliações educacionais.

2.1 Transtorno do Espectro Autista (TEA)

O Transtorno do Espectro Autista (TEA) é uma condição neurológica que afeta o desenvolvimento e o funcionamento da pessoa no dia a dia em diversas áreas, como interação social, comunicação e comportamentos restritivos ou repetitivos (APA, 2022). Os sintomas do TEA manifestam-se de forma mais pronunciada durante a primeira infância e nos anos iniciais da vida escolar. Nesse período, os sinais se caracterizam principalmente por atrasos no desenvolvimento da linguagem, acompanhados de falta de interesse social ou interações sociais incomuns (APA, 2022).

Conforme o APA (2022), as manifestações do TEA variam significativamente dependendo da gravidade da condição. Essa variabilidade é contemplada por meio da classificação em níveis de suporte (níveis 1, 2 e 3), os quais indicam a intensidade do apoio necessário em diferentes domínios do funcionamento (APA, 2022).

Destaca-se que o nível de suporte requerido pode variar de acordo com o contexto em que a pessoa está inserida, bem como sofrer alterações ao longo do desenvolvimento. Os níveis de suporte são descritos a seguir e sumarizados na Tabela 1 (ATAIDE et al., 2023; APA, 2022):

- **Nível 1:** A pessoa diagnosticada com TEA em nível 1 de suporte exige uma quantidade menor de apoio em aspectos de interação social, na inflexibilidade de comportamento, dificuldade em mudar de atividade e em problemas para organização e planejamento.
- **Nível 2:** Exige uma necessidade de apoio substancial, pois apresenta déficits graves nas habilidades de comunicação social verbal e não verbal, dificuldades nos aspectos de interação social mesmo com apoio, além da inflexibilidade do comportamento e dificuldade de lidar com a mudança ou outros comportamentos restritos/repetitivos que aparecem com mais frequência;

- **Nível 3:** Neste nível de suporte, à pessoa com TEA exige apoio muito substancial e apresenta além das características dos níveis anteriores, de forma mais grave, resposta mínima a aberturas sociais que partem de outras pessoas e grande dificuldade para mudar o foco ou as ações.

Tabela 1 – Nível de suporte versus alguns aspectos

Nível de suporte	Comunicação social	Comportamentos restritos e repetitivos
Nível 1 Exige apoio	Na ausência de apoio, a pessoa apresenta déficits na comunicação social que causam prejuízos notáveis. Há dificuldade para iniciar interações sociais e exemplos claros de respostas atípicas ou sem sucesso às tentativas de interação feitas por outros. Pode parecer ter interesse reduzido por interações sociais.	A inflexibilidade comportamental causa interferência significativa no funcionamento da pessoa em um ou mais contextos. Ela apresenta dificuldade em mudar de atividade. Problemas com organização e planejamento atrapalham sua independência.
Nível 2 Exige apoio substancial	A pessoa no Nível 2 apresenta graves déficits nas habilidades de comunicação social verbal e não verbal, resultando em prejuízos sociais aparentes mesmo com apoio. Há limitação para iniciar interações sociais e resposta reduzida ou anormal a tentativas de interação vindas de outros.	A pessoa apresenta inflexibilidade comportamental, dificuldade em lidar com mudanças e outros comportamentos restritos/repetitivos que são frequentes o suficiente para serem óbvios a um observador casual. Esses comportamentos interferem no funcionamento da pessoa em diversos contextos.
Nível 3 Exige apoio muito substancial	Pessoas possuem déficits graves de comunicação social, tanto verbal quanto não verbal, o que provoca grandes prejuízos de funcionamento. Pessoas neste nível têm limitações severas para iniciar interações sociais e respondem minimamente a tentativas de interação vindas de outras pessoas.	Há inflexibilidade comportamental, extrema dificuldade em lidar com mudanças e outros comportamentos restritos e repetitivos que interferem acentuadamente no funcionamento da pessoa em todas as áreas da vida.

Fonte: Adaptada de APA (2022).

Os desafios associados ao TEA podem persistir ao longo da vida, especialmente aqueles relacionados à atenção, à organização e ao planejamento de atividades acadêmicas (APA, 2022). Ao longo do desenvolvimento, é frequente a presença de comorbidades, como o transtorno de déficit de atenção e hiperatividade (TDAH) e transtornos de ansiedade, que podem impactar diretamente o desempenho escolar (Baker et al., 2015; APA, 2022).

Na adolescência, fase correspondente ao Ensino Médio, essas condições tendem a se intensificar em função das mudanças cognitivas, emocionais e sociais, tornando esse período particularmente desafiador para estudantes com TEA (Mutluer et al., 2022). Embora o Ensino Médio possa representar um período de desenvolvimento complexo para todos os adolescentes, as exigências sociais e acadêmicas dessa etapa configuram desafios específicos para estudantes autistas.

Apesar dessas dificuldades, esses estudantes demonstram forte motivação para aprender, profundo interesse em áreas específicas e desejo de inclusão significativa no contexto escolar (White et al., 2025). Diante das características apresentadas, observa-se que os estudantes com TEA demandam adaptações no processo educacional, o que torna importante a discussão sobre práticas pedagógicas voltadas à educação inclusiva (Madureira et al. 2022).

2.2 Processo de ensino e aprendizagem na Educação inclusiva

O processo de ensino e aprendizagem na educação inclusiva é fundamental para garantir o direito de todos a uma educação de qualidade, independentemente de suas condições, sendo a colaboração entre escola, aluno e professor fundamental para criar um ambiente motivador e inclusivo (Davidova, 2024; Tai et al., 2023; Kleinlein, 2025). Essa colaboração conjunta promove uma compreensão mais ampla sobre as habilidades e dificuldades dos estudantes, o que, como destaca Aragão (2023), permite o desenvolvimento de abordagens personalizadas voltadas ao ensino que atendam às necessidades individuais de cada aluno.

A escola deve promover uma cultura inclusiva, valorizando a diversidade e garantindo a participação ativa dos alunos por meio de recursos materiais, humanos e tecnológicos, essenciais para o desenvolvimento da aprendizagem (Aragão, 2023; Santos, 2005). Madureira et al. (2022) destacam as dificuldades de inclusão de alunos com TEA, especialmente no direito de expressão e na comunicação, que são prejudicados por comportamentos retraídos e desafios nas relações sociais (Butera et al., 2020; Tanaka et al., 2024).

Para uma escola verdadeiramente inclusiva, os autores apontam que é necessário adaptar o ensino às necessidades desses alunos, promovendo mudanças no currículo, nas avaliações e na formação docente (Azevedo et al., 2024; Fathoni et al., 2024). Sob essa perspectiva, estudos que consideram a voz dos próprios estudantes evidenciam caminhos possíveis para a superação dessas dificuldades.

Em se tratando de alunos adolescentes, observou-se maior engajamento e motivação quando a aprendizagem estava alinhada às suas preferências, bem como a importância das

relações de confiança com o professor, de oportunidades de autonomia e do reconhecimento de seus pontos fortes no processo educativo (White et al., 2025; Van de Watering et al., 2008).

Dessa forma, a ausência de suporte adequado e a adoção de abordagens tradicionais de avaliação desconsideram as especificidades desses alunos com TEA, desmotivando o aprendizado e contradizendo práticas pedagógicas inclusivas (Madureira et al. 2022). Nesse contexto, o professor ocupa uma posição central na promoção da educação inclusiva. Ele é responsável por adaptar as abordagens, métodos, práticas de ensino, identificar e superar as barreiras que possam dificultar o aprendizado dos alunos e promover um ambiente acolhedor e inclusivo que facilite a aprendizagem e a transmissão do conhecimento (Aragão, 2023; Bigcas et al., 2024).

Para isso, é importante que a escola proporcione aos docentes o devido reconhecimento, preparação, apoio, recursos e oportunidades necessários para que possam desenvolver suas habilidades e implementar um ensino efetivamente direcionado às necessidades individuais dos alunos (Jurs et al., 2023; Casagrande et al., 2025).

Segundo Jurs et al. (2023), para apoiar os docentes, é importante adotar uma abordagem educacional, que inclua planos de aprendizagem, apoio (como a colaboração de outros docentes da educação), e o registro das características individuais e preferências dos alunos. Dessa forma, torna-se possível aplicar abordagens educacionais mais estruturadas, que atendam às diversas dificuldades de aprendizagem dos estudantes, promovendo um ambiente inclusivo e adaptado às suas necessidades específicas (Missaoui et al., 2021; Fathoni et al., 2024).

Nesse sentido, tais abordagens são particularmente relevantes para estudantes com TEA, uma vez que contribuem para uma educação mais inclusiva e para a adaptação das práticas pedagógicas às suas necessidades específicas de aprendizagem (Azevedo et al., 2024; Fathoni et al., 2024).

2.3 Ensino Estruturado como Prática Inclusiva

O ensino estruturado refere-se a um conjunto de estratégias utilizadas para organizar o ambiente de aprendizagem de forma clara, previsível e acessível, favorecendo a compreensão e a participação dos estudantes (Rasmussen et al., 2021; Silva, 2022; Azevedo et al., 2024). No âmbito mais amplo da educação inclusiva, a prática inclusiva pode ser compreendida como a aplicação sistemática de estratégias pedagógicas fundamentadas em teorias que consideram as

necessidades cognitivas, comportamentais e comunicativas dos estudantes, por meio de metodologias e abordagens adaptadas (Tanaka et al., 2024).

Embora o ensino estruturado seja amplamente associado ao trabalho com alunos no TEA, sua aplicação não se limita a esse público, podendo beneficiar estudantes com diferentes perfis e necessidades educacionais (Howley et al., 2006; Suprihatin et al., 2019; Zager et al., 2012)

Essa abordagem foi originalmente desenvolvida na Carolina do Norte, no início da década de 1970, como um componente central do programa *Treatment and Education of Autistic and Related Communication Handicapped Children* (TEACCH). A estrutura do ensino estruturado é considerada adequada para uso com indivíduos de diversas idades e níveis de aprendizagem (MacDonald et al., 2017; Howley et al., 2006; Suprihatin et al., 2019; Magová et al., 2022).

A partir dessa concepção, o ensino estruturado passou a ser discutido na literatura como uma prática pedagógica inclusiva (Bastos et al., 2019), caracterizando-se por uma abordagem flexível e passível de adaptação ao plano de ensino do docente, sendo utilizada em contextos educacionais que envolvem estudantes com TEA (Tanaka et al., 2024). Estudantes com TEA tendem a demandar maior organização das atividades e rotinas durante o processo de aprendizagem, uma vez que podem apresentar dificuldades na filtragem de estímulos irrelevantes do ambiente, em decorrência de alterações sensoriais que interferem na atenção e na concentração (David et al., 2021).

Conforme Suprihatin et al., (2019), o ensino estruturado tem dois objetivos principais:

- i. Melhorar as habilidades de adaptação de cada indivíduo, o que pode ser alcançado elevando o nível de habilidades individuais, muitas vezes utilizando seus interesses específicos;
- ii. Ajustar o ambiente para acomodar as dificuldades individuais causadas pelo autismo.

Para que esses objetivos sejam possíveis de serem alcançados, do ponto de vista do processo avaliativo, tornam-se necessárias avaliações de aprendizagem planejadas e contínuas para observar o resultado e compor um programa de ensino adaptável que corresponda ao nível de habilidades de cada indivíduo (Rasmussen et al., 2021; Kleinlein, 2025).

Em termos de aplicação, o ensino estruturado envolve a organização de atividades com o mesmo objetivo de aprendizagem, utilizando instruções diretas, apoio visual e

sequenciamento de tarefas, favorecendo a compreensão e o engajamento de estudantes com autismo (Rasmussen et al., 2021; Howley et al., 2006).

Com essa finalidade, o ensino estruturado também enfatiza a organização do tempo, uma aprendizagem clara e previsível das atividades pedagógicas e da linguagem utilizada, bem como o uso de recursos de apoio que favorecem a compreensão das tarefas e a autonomia dos estudantes (Tanaka et al., 2024; Magová et al., 2022).

No que se refere aos enunciados escritos, no contexto do ensino estruturado, recomenda-se a apresentação das instruções de forma clara e sequencial, preferencialmente organizadas da esquerda para a direita ou de cima para baixo, com a exposição apenas das informações necessárias para a execução da tarefa no momento, favorecendo a compreensão e a organização do estudante (Rasmussen et al., 2021).

Além disso, recomenda-se o uso de instruções diretas e claras, apoio visual, vocabulário simplificado, destaque de informações relevantes e demonstrações práticas, de modo que as instruções sejam apresentadas de forma organizada e compreensível, favorecendo a realização das atividades de maneira mais independente, em consonância com os resultados das avaliações dos estudantes com TEA (Suprihatin et al., 2019; Howley et al., 2006).

Dessa forma, sua adoção no processo de aprendizagem de estudantes, especialmente com TEA tem se mostrado relevante, contribuindo para a redução da ansiedade, o aumento do foco nas informações essenciais e o incentivo à independência, especialmente por meio do uso de recursos visuais e da clareza na apresentação das atividades (Rasmussen et al., 2021; Tanaka et al., 2024).

2.4 Avaliação da Aprendizagem

A avaliação da aprendizagem constitui um elemento central no processo de ensino-aprendizagem, exercendo diferentes funções pedagógicas conforme seus objetivos e momentos de aplicação (Oliveira et al., 2022; Freia et al., 2023; Kodrle et al., 2024). Sua relevância torna-se ainda mais evidente no contexto da educação inclusiva, no qual se consideram as necessidades específicas dos estudantes, incluindo aqueles com TEA (APA, 2022; Azevedo et al., 2024; Fathoni et al., 2024).

No contexto educacional, as modalidades de avaliação referem-se ao papel que a avaliação desempenha no acompanhamento e na promoção da aprendizagem dos estudantes, sendo tradicionalmente classificadas em diagnóstica, formativa e somativa (Oliveira et al., 2022; Freia et al., 2023; Kodrle et al., 2024).

A avaliação diagnóstica tem como finalidade identificar as características iniciais dos estudantes em determinado momento do processo educativo, permitindo ao docente reconhecer conhecimentos prévios, dificuldades e potencialidades. Essa modalidade fornece subsídios relevantes para o planejamento das ações pedagógicas subsequentes (Souza, 2021; Izidoro et al., 2024; Oliveira et al., 2022). No entanto, quando utilizada de forma isolada, apresenta limitações, especialmente em contextos que demandam acompanhamento contínuo da aprendizagem (Kodrle et al., 2024).

Por sua vez, a avaliação somativa caracteriza-se pela síntese dos resultados de aprendizagem ao final de um período letivo, sendo frequentemente associada à atribuição de notas ou conceitos para fins de certificação do desempenho acadêmico dos estudantes (Izidoro et al., 2024).

Embora amplamente adotada, essa modalidade apresenta menor flexibilidade para ajustes pedagógicos ao longo do processo, o que pode limitar sua efetividade em contextos educacionais inclusivos que exigem adaptações contínuas (Perondi et al., 2025; Medeiros et al., 2024).

Em contraste, a avaliação formativa distingue-se por seu caráter contínuo e processual, acompanhando o desenvolvimento da aprendizagem ao longo das atividades pedagógicas. Essa abordagem busca não apenas identificar erros e acertos, mas também fornecer feedback contínuo e orientador, possibilitando intervenções pedagógicas oportunas e a adaptação das estratégias de ensino às necessidades dos estudantes (Souza, 2021; Izidoro et al., 2024; Kodrle et al., 2024). Estudos indicam que a avaliação formativa favorece a personalização do ensino e a adequação das práticas pedagógicas às características individuais dos alunos, sendo especialmente relevante no Ensino Médio, marcado pela heterogeneidade de perfis cognitivos e estilos de aprendizagem (Freia et al., 2023; Weber, 2025).

No contexto da educação inclusiva, a avaliação formativa favorece a adaptação contínua das estratégias pedagógicas às necessidades dos estudantes. Aspectos como clareza nas instruções, organização das atividades, previsibilidade das tarefas e diversificação dos formatos avaliativos contribuem para a promoção da acessibilidade e da equidade no processo de avaliação (Fathoni et al., 2024; Azevedo et al., 2024).

Embora as modalidades diagnóstica, formativa e somativa definam a finalidade da avaliação, sua operacionalização ocorre por meio de diferentes métodos avaliativos. Em outras palavras, os métodos representam as formas pelas quais as evidências da aprendizagem são coletadas, permitindo ao docente verificar conhecimentos, habilidades e competências desenvolvidas pelos estudantes. Dessa forma, uma mesma modalidade de avaliação pode

utilizar diferentes métodos, conforme os objetivos de aprendizagem, o contexto educacional e as características dos alunos (Oliveira et al., 2022; Freia et al., 2023; Kodrle et al., 2024).

Nesse contexto, destacam-se os métodos avaliativos objetivo, discursivo, oral e prático, amplamente utilizados na educação básica e profissional. Esses métodos não constituem modalidades independentes de avaliação, mas instrumentos que podem ser empregados em avaliações diagnósticas, formativas ou somativas, dependendo da finalidade pedagógica pretendida (Souza, 2021; Casagrande et al., 2025; Weber, 2025).

Por exemplo, o método de avaliação objetivo caracteriza-se pelo uso de linguagem clara e direta, exigindo respostas curtas e previamente definidas, como ocorre em questões de múltipla escolha ou verdadeiro ou falso (Engel et al., 2024; Pedaste et al., 2021). Esse formato tende a reduzir ambiguidades na interpretação das questões, podendo beneficiar estudantes que demandam maior clareza nas instruções, desde que alinhado aos objetivos de aprendizagem estabelecidos no plano de ensino (Fathoni et al., 2024; Garcia et al., 2022). A Figura 1 apresenta um exemplo desse tipo de avaliação por meio de uma questão de múltipla escolha, com alternativas previamente definidas.

Figura 1 - Exemplo de questão objetiva em uma avaliação de aprendizagem.

- A) Os pólos mundiais de alta tecnologia, chamados tecnopólos, localizando-se próximos a importantes universidades, como é o caso:
- () Região industrial dos grandes lagos nos EUA.
 - () Vale do Silício na Califórnia.
 - () Região industrial dos Urais, na Rússia.
 - () Bacia de Londres, na Inglaterra (Geografia – Escola D).

Fonte: Adaptado de Bezerra, (2008).

Um método avaliativo discursivo requer a organização das ideias e o uso adequado do vocabulário na produção escrita, podendo assumir formatos de respostas curtas ou longas (Liao et al., 2024; Bezerra, 2008). Enquanto respostas discursivas longas demandam maior elaboração textual e são avaliadas considerando aspectos como coesão e estilo de escrita, as respostas curtas tendem a focalizar principalmente o conteúdo semântico apresentado (Silva, 2023).

Métodos discursivos podem representar desafios adicionais para estudantes que apresentam dificuldades na organização textual, demandando adaptações que considerem suas características individuais. A Figura 2 ilustra um exemplo de uma questão discursiva, centrada no conteúdo da disciplina, podendo o aluno demonstrar que detém o conhecimento específico previsto (Bezerra, 2008).

Figura 2-Exemplo de questão discursiva

A) O que são economias de transição? (Geografia – Escola B)

Fonte: Adaptado de Bezerra, (2008).

Um método avaliativo oral caracteriza-se pela apresentação verbal de respostas, geralmente em situações de interação direta entre docente e estudante, com tempo delimitado, podendo ou não contar com o uso de recursos visuais de apoio. (Ozbasi, 2019; Jorge et al., 2023). Esse método pode favorecer estudantes que apresentam dificuldades na expressão escrita, desde que sejam consideradas adaptações relacionadas ao tempo, à organização das respostas e ao ambiente de aplicação (Fathoni et al., 2024; Davidova, 2024).

Nesse formato, o docente pode utilizar instrumentos como exemplos de perguntas orientadoras ou estímulos iniciais que solicitem ao estudante a explicação de conceitos, a descrição de procedimentos ou a justificativa de respostas, permitindo intervenções e aprofundamentos ao longo da interação avaliativa (Fenton, 2025; Kodrle et al., 2024).

Por fim, um método avaliativo prático envolve a realização de atividades em contextos reais ou simulados, exigindo a aplicação de conhecimentos e habilidades em situações concretas (Peña et al., 2024). Esse formato pode incluir tarefas como planejamento de atividades, execução de experimentos, organização de procedimentos e análise de resultados, mobilizando habilidades de investigação e tomada de decisão (Pedaste et al., 2021). Quando bem estruturada, a avaliação prática contribui para o engajamento dos estudantes ao aproximar o processo avaliativo de situações significativas do cotidiano acadêmico ou profissional (Weber, 2025).

Para que os diferentes formatos avaliativos sejam efetivos, é importante que apresentem critérios claros, coerentes com os objetivos de aprendizagem e acompanhados de orientações que facilitem a compreensão das tarefas propostas (Peña et al., 2024). A diversidade de formatos amplia as possibilidades de expressão do conhecimento e contribui para a redução de barreiras no processo avaliativo, favorecendo práticas mais inclusivas (Fathoni et al., 2024; Azevedo et al., 2024).

2.5 Modelos de Linguagem

Modelos de Linguagem Grandes (do inglês, *Large Language Models* – LLMs) são modelos neurais treinados em grandes volumes de dados textuais, capazes de gerar linguagem natural por meio da previsão probabilística de sequências de palavras a partir de um contexto

de entrada (Prompiengchai et al., 2024; Caseli; Nunes, 2024, p. 366). Esses modelos são, em sua maioria, fundamentados na arquitetura Transformer, a qual utiliza mecanismos de autoatenção para modelar relações contextuais entre palavras, permitindo a captura de dependências de longo alcance no texto (Caseli; Nunes, 2024, p. 382–385).

No processo de geração textual, diferentes parâmetros influenciam o comportamento dos modelos. Dentre eles, destaca-se a temperatura, responsável por controlar o nível de aleatoriedade na saída gerada. De modo geral, valores mais baixos da temperatura tendem a produzir respostas mais determinísticas, enquanto valores mais elevados favorecem maior diversidade na geração textual (Renze et al., 2024). Os LLMs (*Large Language Models* — Modelos de Linguagem de Grande Escala), como GPT e BERT, são amplamente utilizados em diferentes aplicações de processamento de linguagem natural, incluindo geração de conteúdo, classificação e análise textual (Rani et al., 2023; Khadhraoui et al., 2022).

O GPT (*Generative Pre-trained Transformer*), desenvolvido pela OpenAI, pertence à classe de modelos generativos, destacando-se por sua capacidade de produzir linguagem natural de forma coerente e contextualizada, sendo amplamente empregado na geração automatizada de textos.

Por outro lado, o BERT (*Bidirectional Encoder Representations from Transformers*) é classificado como um modelo baseado em compreensão (ou discriminativo), sendo voltado principalmente para tarefas de entendimento textual, como classificação de texto e análise de similaridade semântica (Devlin et al., 2019; Khadhraoui et al., 2022).

Modelos como o bert-base-multilingual-cased exemplificam essas capacidades, possuindo um vocabulário com aproximadamente 30.522 termos e utilizando um processo de tokenização baseado em subpalavras. Esse processo segmenta o texto em unidades menores chamadas tokens, permitindo representar palavras fora do vocabulário por meio de combinações de subunidades linguísticas frequentes (Alsulami et al., 2025; Ariyanta et al., 2025; Devlin et al., 2019). Dessa forma, o modelo gera representações vetoriais contextualizadas (*embeddings*), fundamentais para tarefas como análise de similaridade semântica textual (Khadhraoui et al., 2022).

No contexto educacional, ferramentas baseadas em modelos de linguagem, como o ChatGPT, apresentam capacidade de gerar perguntas, textos, avaliações e auxiliar na execução de tarefas educacionais, demonstrando desempenho relevante em atividades de apoio ao ensino (Leung, 2024).

No entanto, esses modelos (LLMs) dependem fortemente da estrutura dos prompts e não incorporam, de forma nativa, mecanismos pedagógicos estruturados, como a consideração

sistemática das preferências avaliativas dos estudantes. Assim, embora sejam eficazes na geração de conteúdo, tendem a operar de maneira generalista quando não integrados a abordagens que considerem o perfil do aprendiz, como em sistemas adaptativos baseados em modelagem do estudante (White, 2023; Prompiengchai et al., 2025; Missaoui et al., 2021).

2.5.1 Engenharia de Prompts

A engenharia de prompts consiste na elaboração e organização de instruções específicas, denominadas *prompts*, com o objetivo de orientar modelos de linguagem de grande escala (LLMs) na execução de diferentes tarefas (Gu et al., 2023). Essas instruções são estruturadas de modo a adaptar o comportamento do modelo às demandas do contexto de uso.

Essa prática tem se tornando cada vez mais relevante, uma vez que a forma como os *prompts* são formulados influencia diretamente a qualidade, a coerência e a adequação das respostas geradas pelos LLMs (White et al., 2023).

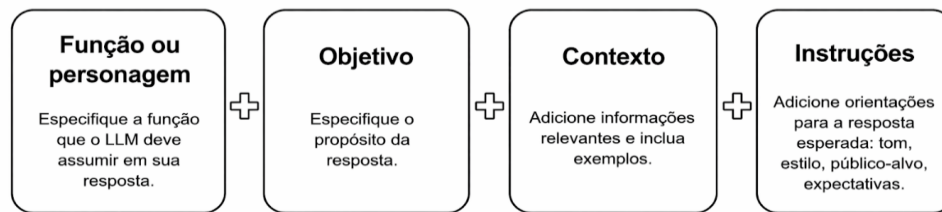
Particularmente, no contexto educacional, a engenharia de prompts tem sido aplicada para orientar modelos de linguagem na geração de perguntas diversas e contextualmente relevantes, adaptadas a diferentes objetivos de aprendizagem (Maity et al., 2024). Além disso, esses modelos podem ser aplicados em tarefas como classificação, inferência e geração de respostas, a depender da estrutura das instruções fornecidas (Caseli; Nunes, 2024, p. 373–375).

Para que essa aplicação possa ser mais eficiente, a literatura destaca a importância da estrutura dos prompts. Segundo Rodriguez-Donaire (2024), prompts estruturados são compostos por quatro elementos fundamentais (Figura 3): a atribuição de um papel ao modelo, a definição clara de um objetivo, a apresentação de um contexto bem delimitado e a formulação de instruções explícitas.

Essa organização favorece a geração de respostas mais eficazes em ambientes educacionais, contribuindo para maior profundidade, clareza e alinhamento das respostas aos propósitos de aprendizagem (Gu et al., 2023; Leung, 2024).

Nesse contexto, a atribuição de um papel ao modelo está associada à técnica conhecida como *role prompting*, também descrita na literatura como *Persona Pattern* (White et al., 2023).

Figura 3 - Elementos que compõem a estrutura de um prompt



Fonte: RODRIGUEZ-DONAIRE, (2024).

Além da técnica de *role prompting*, outras abordagens de engenharia de prompts têm sido amplamente utilizadas em diferentes contextos, sendo também aplicadas no contexto educacional (Maity et al., 2024). Entre as principais, destacam-se: (i) o *Zero-Shot Prompting*, que permite que modelos de linguagem gerem respostas a partir de instruções mínimas, utilizando apenas seu conhecimento pré-treinado, embora a qualidade das respostas possa variar conforme a complexidade da tarefa (Prompiengchai et al., 2025); (ii) o *Few-Shot Prompting*, que orienta o modelo por meio da apresentação de exemplos, contribuindo para maior qualidade e consistência em tarefas mais complexas; e (iii) o *Chain-of-Thought Prompting*, uma técnica estruturada que conduz o modelo por um processo de raciocínio passo a passo antes da geração da resposta final (Leung, 2024; Maity et al., 2024).

Nesse cenário, tais técnicas podem ser utilizadas para orientar a geração de conteúdos alinhados a objetivos de aprendizagem, apoiar a elaboração de atividades avaliativas e favorecer a adaptação às necessidades dos estudantes (Missaoui et al., 2021; White et al., 2023).

Dentre as abordagens existentes, destaca-se o *zero-shot prompting*, escolhido por sua simplicidade e por não exigir exemplos prévios para a execução da tarefa. Para ilustrar sua aplicação, apresenta-se, a seguir, um exemplo adaptado da literatura (Wang et al., 2023):

Exemplo de prompt (Zero-Shot)

“Você é um professor do ensino médio especialista em Ciências. Com base no conteúdo sobre fotossíntese, elabore uma questão de múltipla escolha com quatro alternativas, indicando a resposta correta.”

Fonte: Adaptado de Wang et al. (2023).

O exemplo apresenta que a definição explícita de papel, contexto e tarefa, em um cenário zero-shot, de modo a orientar um dado modelo de linguagem na geração de respostas mais adequadas ao objetivo proposto.

Assim, a engenharia de prompts assume papel relevante na utilização de LLMs, uma vez que diferentes técnicas podem impactar significativamente o desempenho dos modelos,

reforçando a importância de sua aplicação estruturada e contextualizada (Gu et al., 2023; Rodriguez-Donaire, 2024).

2.5.2 Avaliação em LLMs

A avaliação em LLMs constitui uma tarefa complexa, uma vez que não há um método único e universalmente aceito. Os mecanismos de avaliação podem ser amplamente categorizados em métricas quantitativas, benchmarks automatizados e avaliações humanas (Awasthi et al., 2025). Embora as métricas quantitativas forneçam medidas objetivas, a avaliação humana é frequentemente considerada uma das abordagens mais confiáveis para analisar o desempenho desses modelos, especialmente em tarefas que envolvem julgamento subjetivo e interpretação contextual (Gao et al., 2025).

Nesse contexto, especialmente no âmbito das métricas quantitativas, destacam-se abordagens mais recentes, como o *BERTScore*, que utiliza *embeddings* para avaliar a similaridade semântica entre textos, superando limitações de métodos tradicionais baseados em *n-gramas* e aproximando-se, em certa medida, do julgamento humano (Hanna et al., 2021; Wu et al., 2022). Ainda assim, tais métricas não capturam plenamente aspectos subjetivos e contextuais da linguagem, o que evidencia a necessidade de métodos mais robustos que permitam avaliar, de forma sistemática, o alinhamento entre as saídas geradas pelos modelos e as expectativas humanas (Zhao, 2024; Nguyen et al., 2024).

A natureza probabilística dos LLMs pode levar à inconsistência nas respostas para uma mesma entrada, o que reforça a necessidade de métricas que considerem não apenas a qualidade, mas também a estabilidade e a confiabilidade dos resultados (Wei et al., 2024). Diante disso, garantir o alinhamento dos LLMs com valores e julgamentos humanos permanece um desafio central, sendo fundamental a adoção de estratégias de avaliação mais organizadas, combinadas à avaliação humana, para mitigar desalinhamentos e possíveis consequências indesejadas (Gao et al., 2025; Liu et al., 2025).

2.5.3 Métrica de Avaliação

A partir de representações vetoriais contextualizadas (*embeddings*), que capturam informações semânticas das palavras em contexto, podem ser aplicadas métricas de avaliação semântica, como o *BERTScore*, que mede a similaridade entre textos com base nessas representações (Zhang et al., 2020). Essa métrica realiza um alinhamento token a token entre

um texto candidato e um texto de referência (*ground truth*), entendido como a resposta padrão, padrão-ouro ou parâmetro previamente definido para fins de comparação (Caseli; Nunes, 2024).

A similaridade entre esses tokens é calculada por meio da similaridade do cosseno entre seus vetores, possibilitando capturar relações semânticas mesmo na presença de variações lexicais ou sintáticas (Yuan et al., 2021; Sun et al., 2022; Alsulami et al., 2025).

A precisão (*precision*) e a revocação (*recall*) constituem métricas clássicas de desempenho em Processamento de Linguagem Natural, sendo frequentemente combinadas pela medida F1 (Caseli; Nunes, 2024). No contexto do BERTScore, tais métricas são adaptadas para considerar similaridade semântica entre os embeddings contextualizados dos tokens comparados (Zhang et al., 2020). As métricas de precisão e recall para o BERTScore são calculadas comparando cada representação de token $\hat{x}$ do texto candidato com cada representação de token x do texto a ser o (*ground truth*), da seguinte forma (Hanna et al., 2021; Zhang et al., 2020).

$$(1) \quad P_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^\top \hat{x}_j ; (2) \quad R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^\top \hat{x}_j \text{ e}$$

$$(3) \quad F_{BERT} = \frac{2 \cdot P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}}.$$

Onde (Alsulami et al., 2025):

- **Precisão:** proporção do conteúdo da questão gerada que está semanticamente similar com a referência;
- **Revocação:** corresponde à proporção do conteúdo presente na questão de referência que é efetivamente coberto pela questão gerada.;
- **F1:** representa a média harmônica da Precisão e da Revocação, indicando uma sobreposição equilibrada.

A pontuação F1 pode ser definida como a média harmônica entre precisão e revocação. Como o BERTScore pode variar de -1 a 1, mas geralmente se situa na extremidade superior desse intervalo, essa medida tende a refletir altos níveis de similaridade semântica entre os textos comparados (Hanna et al., 2021).

Exemplo do BERTScore

Para ilustrar o funcionamento da métrica BERTScore, considere o exemplo adaptado de Alsulami (2025), no qual se avalia a resposta gerada pelo ChatGPT para uma pergunta sobre o significado da palavra "procrastinador".

Pergunta: "O que é um procrastinador?"

Texto Referência (*ground truth*): "Uma pessoa que adia fazer algo."

Texto Candidato (gerado pelo ChatGPT): "Alguém que atrasa tarefas."

Este exemplo demonstra a principal vantagem do BERTScore sobre métricas tradicionais baseadas em correspondência exata de palavras (como BLEU): ele reconhece paráfrases e variações linguísticas (ex: 'adia' como equivalente a 'atrasa'), avaliando a similaridade pelo significado, e não pela forma superficial do texto (Zhang et al., 2020; Alsulami et al., 2025; Hanna et al., 2021).

2.6 Uso de LLMs para geração da avaliação de aprendizagem

A capacidade dos LLMs de gerar linguagem natural coerente, a partir de diferentes tipos de instruções e contextos de entrada, permite sua aplicação em múltiplos domínios, incluindo o educacional (Mienye et al., 2025; Rani et al. 2023). Nesse cenário, esses modelos passam a ser utilizados como ferramentas de apoio à geração e automatização de conteúdos pedagógicos, especialmente na construção de instrumentos avaliativos (Mahamuni et al., 2024; Mienye et al., 2025).

O uso de LLMs na geração da avaliação da aprendizagem está relacionado à sua capacidade de produzir conteúdo textuais coerentes e contextualmente adequados a partir de instruções específicas, o que os torna úteis em diferentes aplicações educacionais (Mahamuni et al., 2024; Mienye et al., 2025). Nesse sentido, seu uso é particularmente voltado à automatização e ao apoio na construção de instrumentos avaliativos (Maity et al., 2024), especialmente em contextos que demandam maior flexibilidade e adaptação do processo avaliativo.

Uma das principais aplicações desses modelos no contexto educacional consiste na geração de questões avaliativas, que podem ser adaptadas a diferentes níveis de conhecimento e objetivos de aprendizagem, utilizando informações contextuais fornecidas no prompt (Owan et al., 2023; Prompiengchai et al., 2025). Essa capacidade permite maior flexibilidade na criação de avaliações, contribuindo para a diversificação das atividades propostas aos estudantes (Owan et al., 2023), o que é particularmente relevante em contextos educacionais inclusivos.

Além disso, a utilização de LLMs pode apoiar educadores e instituições no desenvolvimento de avaliações mais dinâmicas, ampliando as possibilidades de construção de instrumentos avaliativos mais adaptáveis e alinhados às necessidades dos estudantes (Kodrlé et al., 2024). Essa integração pode contribuir tanto para a melhoria da qualidade das avaliações

quanto para o aprimoramento da experiência educacional (Meissner et al., 2024; Owan et al., 2023).

Ferramentas baseadas em modelos de linguagem grandes, como ChatGPT, Gemini e Claude, exemplificam esse uso ao possibilitar a geração de questões educacionais e o apoio à construção de avaliações em diferentes formatos (Hadzhikoleva et al., 2024; Silva et al., 2024). Para fins educacionais, isso significa que os LLMs podem produzir não apenas perguntas gramaticalmente corretas, mas também conteúdos contextualizados e alinhados a objetivos pedagógicos (Maity et al., 2024; Leung, 2024).

Contudo, apesar de suas potencialidades, o uso de LLMs na avaliação da aprendizagem apresenta desafios importantes. Entre eles, destacam-se questões relacionadas à qualidade e consistência das perguntas geradas, possíveis inconsistências semânticas nas respostas e limitações na interpretação do contexto pedagógico (Sun et al., 2022; Nguyen et al., 2024). Além disso, aspectos éticos envolvendo o uso da inteligência artificial na educação também têm sido amplamente discutidos na literatura (Zhao, 2024; Owan et al., 2023).

Ainda assim, esses modelos representam uma alternativa promissora para a evolução dos processos avaliativos, especialmente no que se refere à avaliação formativa, ao possibilitar novas formas de construção e adaptação de instrumentos de avaliação (Kodrle et al., 2024). Nesse sentido, há um interesse crescente em explorar o potencial dos LLMs para apoiar práticas avaliativas mais flexíveis, adaptadas e inovadoras no contexto educacional (Kodrle et al., 2024; Prompiengchai et al., 2025).

2.7 Considerações

Este capítulo apresentou o referencial teórico sobre os principais tópicos desta dissertação. Inicialmente, foi conceituado o TEA, destacando suas características, níveis de suporte e impactos no desenvolvimento da comunicação social, comportamento e autonomia do estudante (APA, 2022). Evidenciou-se que o TEA se manifesta de forma heterogênea, o que reforça a necessidade de abordagens educacionais flexíveis e individualizadas, especialmente no contexto do Ensino Médio, onde demandas acadêmicas e sociais se intensificam.

Em seguida, foram discutidos os princípios da educação inclusiva, enfatizando o processo de ensino e aprendizagem como uma prática colaborativa entre docentes, estudantes e instituição escolar. Observou-se que abordagens pedagógicas não adaptadas podem limitar o engajamento e a aprendizagem de estudantes com TEA, principalmente quando não consideram suas preferências, formas de interação e necessidades específicas de suporte.

Posteriormente, foi apresentado o ensino estruturado como estratégia pedagógica inclusiva, caracterizado pela organização clara e previsível do ambiente, das atividades e das instruções. Essa abordagem favorece a compreensão das tarefas, reduz a ansiedade e contribui para o aumento da autonomia dos estudantes, sendo especialmente relevante para estudantes com TEA, embora aplicável a diferentes perfis de aprendizagem.

Na sequência, foram abordadas as modalidades de avaliação da aprendizagem, incluindo as avaliações diagnóstica, formativa e somativa. Destacou-se a avaliação formativa como uma abordagem processual e contínua mais flexível, capaz de orientar intervenções pedagógicas ao longo do processo de ensino. Além disso, evidenciou-se que a diversidade de instrumentos avaliativos contribui para práticas mais inclusivas ao possibilitar diferentes formas de expressão do conhecimento.

Por fim, foram apresentados os LLMs, destacando sua capacidade de geração de textos coerentes e contextualizados a partir de instruções específicas. Observou-se seu potencial de aplicação na educação, especialmente na geração de instrumentos avaliativos, embora também sejam evidenciadas limitações como inconsistências semânticas, dependência da engenharia de prompts e necessidade de validação humana podendo garantir semanticamente um alinhamento pedagógico.

Dessa forma, conclui-se que a integração entre o TEA, práticas de educação inclusiva, ensino estruturado, avaliação formativa e o uso de LLMs pode contribuir para o desenvolvimento de processos avaliativos mais adaptativos e alinhados às necessidades dos estudantes com TEA. Contudo, essa integração requer uma abordagem pedagógica organizada, na qual a tecnologia atue como suporte ao docente, preservando seu papel central no processo de avaliação da aprendizagem.

Por fim, o próximo capítulo apresenta os trabalhos relacionados, situando esta pesquisa no estado da arte e evidenciando os estudos que dialogam com os temas discutidos neste referencial teórico.

3. TRABALHOS RELACIONADOS

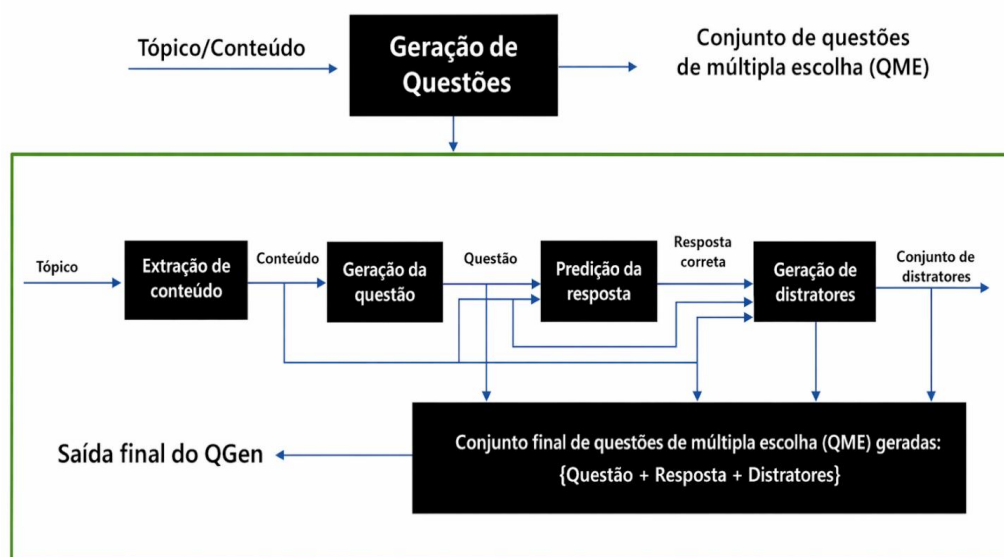
Este capítulo descreve alguns trabalhos relacionados à presente pesquisa. Os trabalhos abordam investigações acerca da geração de questões avaliativas, utilizando modelos de linguagem grandes. Considerando que as pesquisas sobre geração automática de itens avaliativos com LLMs ainda se concentram em diferentes níveis de ensino, são considerados trabalhos do Ensino Fundamental ao Ensino Superior. Particularmente, são apontados trabalhos pautados na educação inclusiva.

3.1 Bhowmick et al., (2023)

Bhowmick et al. (2023) investigam o uso de LLMs para a geração automática de instrumentos avaliativos baseados em questões. Com base em uma pesquisa inicial com docentes, os autores identificam desafios relacionados ao tempo necessário para a elaboração de questões, à limitação de recursos e à dificuldade de personalização. A partir desses achados, propõem o QGen, uma arquitetura de sistema modular voltada à geração automática de questões de múltipla escolha (QMCs) a partir de textos educacionais.

A Figura 4 apresenta a arquitetura geral do QGen, evidenciando o fluxo de processamento desde a entrada de um tópico ou conteúdo até a geração final das questões avaliativas.

Figura 4- Etapas e arquitetura do sistema QGen



Fonte: Bhowmick et al., (2023).

A arquitetura do QGen é composta por módulos independentes, porém interconectados, responsáveis pela extração de conteúdo, geração de perguntas, previsão da resposta correta, geração de distratores, entendidos como as alternativas incorretas em questões de múltipla escolha, elaboradas para serem conceitualmente próximas da resposta correta, a fim de aumentar o nível de dificuldade e evitar acertos por adivinhação e a filtragem final. Essa organização modular permite maior flexibilidade na integração de diferentes modelos de linguagem, possibilitando a combinação de abordagens distintas em cada etapa do processo.

No primeiro módulo, a entrada do sistema pode ser fornecida na forma de um tópico ou de um texto educacional. Quando um tópico é utilizado, o sistema realiza a extração automática de um trecho de conteúdo a partir de fontes externas, como repositórios educacionais, páginas da Wikipédia e sites especializados, como o History.com¹. Esse conteúdo extraído passa a ser utilizado como contexto de entrada para os módulos subsequentes, especialmente para a geração de perguntas e a previsão da resposta correta, uma vez que, conforme descrito no artigo, as questões são geradas com base em um conteúdo textual previamente definido. Por outro lado, quando o texto educacional é fornecido diretamente, essa etapa de extração é ignorada, sendo o conteúdo utilizado como entrada para o processo de geração de questões.

O segundo módulo realiza a geração automática de perguntas em linguagem natural, empregando modelos de linguagem baseados em Transformers, como o T5 e variantes do GPT, incluindo o InstructGPT. A utilização de múltiplos modelos se justifica pela abordagem modular adotada no sistema, que permite explorar diferentes técnicas e comparar seus desempenhos em tarefas específicas, além de combinar as vantagens de cada modelo para melhorar a qualidade das questões geradas, tornando o sistema mais flexível, adaptável e eficiente.

Modelos baseados no T5 são ajustados com o conjunto de dados SQuAD², composto por perguntas elaboradas a partir de artigos da Wikipédia, o que possibilita a geração de questões semanticamente coerentes e alinhadas ao conteúdo de origem. Por sua vez, o

¹ O repositório principal de conteúdo do History Channel é o seu site oficial, [History.com](https://www.history.com), que funciona como uma enciclopédia digital, abrangendo tópicos desde o Egito Antigo até a história moderna, guerras mundiais, biografia de figuras históricas e vídeos exclusivos.

² O Stanford Question Answering Dataset (SQuAD) consiste em um conjunto de dados de perguntas e respostas extraídas de artigos da Wikipédia, mantido e disponibilizado publicamente pelo Stanford NLP Group. Disponível em: <https://rajpurkar.github.io/SQuAD-explorer/>.

InstructGPT é utilizado por meio de estratégias de prompting do tipo few-shot, nas quais exemplos contendo trechos de texto e perguntas associadas são fornecidos ao modelo para orientar a geração conforme o estilo e a estrutura desejados. Essa combinação evidencia a adoção de uma abordagem híbrida, que explora diferentes técnicas de geração de linguagem natural com o objetivo de aprimorar a qualidade das questões geradas, além de tornar o sistema mais flexível e adaptável a diferentes requisitos educacionais.

No terceiro módulo, responsável pela previsão da resposta correta, os autores utilizam o modelo RoBERTa, treinado no conjunto de dados SQuAD 2.0³, com o objetivo de identificar automaticamente o trecho do texto que corresponde à resposta correta para cada pergunta gerada. O quarto módulo é dedicado à geração de distratores, componente essencial na construção de questões de múltipla escolha. Para cada par formado por pergunta e resposta correta, o sistema gera alternativas incorretas plausíveis, que apresentam similaridade semântica com a resposta correta, evitando opções trivialmente incorretas e aumentando o nível de dificuldade das questões.

Por fim, o QGen incorpora um módulo de filtragem automática, responsável por eliminar questões de baixa qualidade, como aquelas mal formuladas, inconsistentes ou não fundamentadas no conteúdo. Essa etapa assegura que apenas questões de múltipla escolha bem estruturadas e semanticamente coerentes sejam disponibilizadas ao usuário final, reforçando o caráter automatizado do sistema.

A avaliação do desempenho do QGen combina métricas quantitativas e qualitativas. No âmbito quantitativo, são utilizadas métricas baseadas em processamento de linguagem natural, como Exact Match (EM) e F1-score, que permitem comparar as respostas geradas pelo sistema com as respostas de referência do conjunto de dados SQuAD.

Adicionalmente, é realizada uma avaliação qualitativa com base em julgamento humano, com o objetivo de analisar a qualidade das questões geradas. Para essa etapa, foram avaliadas 500 questões produzidas por diferentes variantes do QGen, a partir de 100 trechos de texto da Wikipédia, sendo analisadas por um grupo de três anotadores humanos. Os anotadores foram responsáveis por avaliar as questões geradas, analisando seus diferentes componentes, pergunta, resposta correta e distratores e identificando possíveis problemas, como falhas na formulação da questão, incorreção da resposta, baixa qualidade dos distratores ou

³ combina as 100.000 perguntas do SQuAD1.1 com mais de 50.000 perguntas sem resposta, escritas de forma adversária por colaboradores remotos para parecerem semelhantes a perguntas respondíveis. <https://rajpurkar.github.io/SQuAD-explorer/>

inconsistências no conjunto da questão. Quando tais problemas não eram identificados, a questão era considerada de qualidade satisfatória.

Os resultados obtidos indicam que o QGen apresenta desempenho promissor, com aproximadamente 93,53% das questões sendo avaliadas como de alta qualidade, além de resultados consistentes nas avaliações quantitativas e qualitativas, o que evidencia seu potencial para aplicações educacionais, especialmente no que se refere à geração automatizada e personalizável de questões avaliativas.

Como forma de ilustrar o funcionamento do QGen, considere o exemplo em que o sistema recebe como entrada o tópico “Fotossíntese”. A partir desse tópico, o sistema realiza a extração automática de conteúdo de fontes externas, gerando um contexto textual que é utilizado para a geração de perguntas, previsão de respostas e criação de distratores, resultando na questão de múltipla escolha apresentada a seguir.

Exemplo completo baseado na arquitetura do QGen

Entrada (Módulo 1)

Tópico: Fotossíntese

Extração de conteúdo (Módulo 1)

Texto extraído (Parágrafo):

A fotossíntese é o processo pelo qual plantas, algas e algumas bactérias convertem energia luminosa em energia química. Esse processo ocorre nos cloroplastos e resulta na produção de glicose e oxigênio a partir de dióxido de carbono e água.

Geração de pergunta (Módulo 2)

Pergunta gerada:

Qual é o principal produto energético gerado durante o processo de fotossíntese?

Previsão da resposta correta (Módulo 3)

Resposta prevista:

Glicose

Geração de distratores (Módulo 4)

Distratores Gerados

(A) Oxigênio

(B) Dióxido de carbono

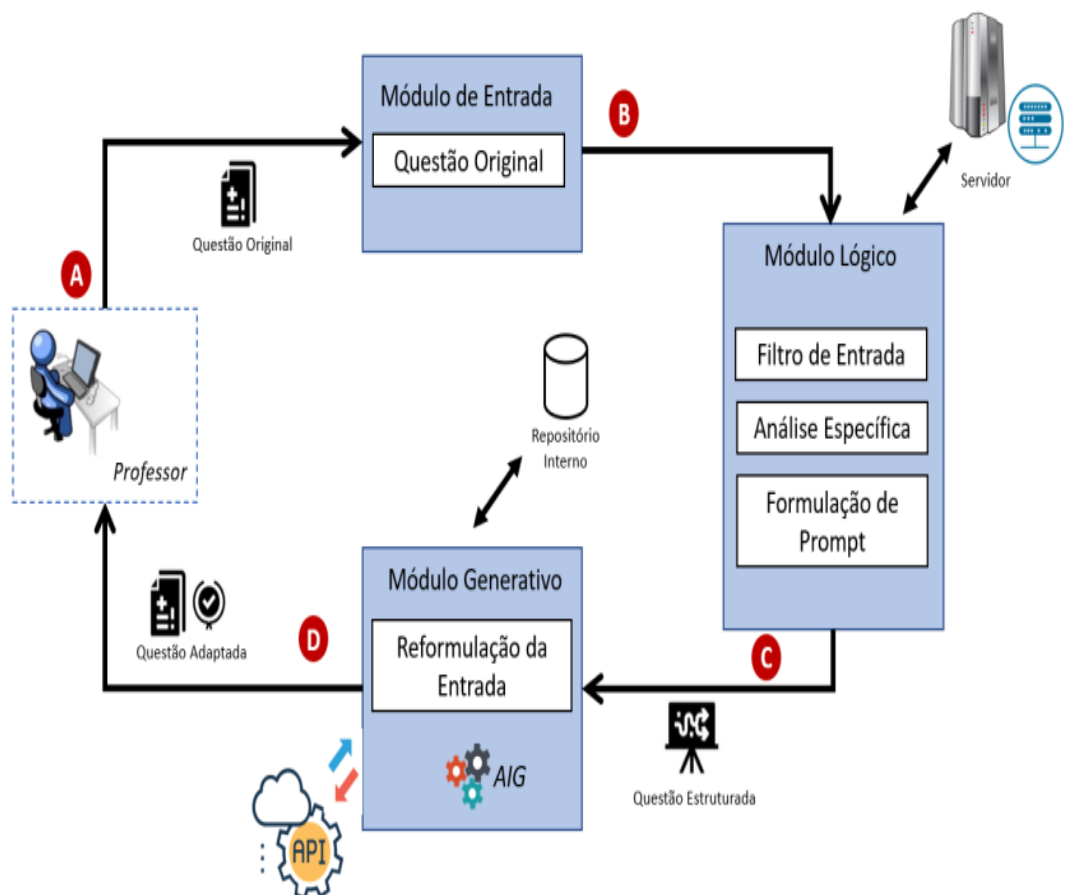
(C) Glicose ✓

(D) Água

3.2 Silva et al. (2024)

No trabalho de Silva et al. (2024), é apresentada a arquitetura Mind Bridge, desenvolvida com o objetivo de adaptar questões avaliativas para estudantes neurodivergentes por meio do uso de Inteligência Artificial Generativa. A arquitetura atua na reformulação das questões avaliativas, ajustando sua estrutura e linguagem, sem alterar os objetivos pedagógicos definidos pelo docente, com a finalidade de facilitar a compreensão por parte dos estudantes. A arquitetura é composta por três módulos principais, conforme ilustrado na Figura 5.

Figura 5- Módulos da arquitetura da plataforma Mind Bridge



Fonte: Adaptado Silva et al., (2024).

No Módulo de Entrada, o professor fornece a questão original em formato textual ou visual, sendo este último passível de conversão para texto para posterior processamento.

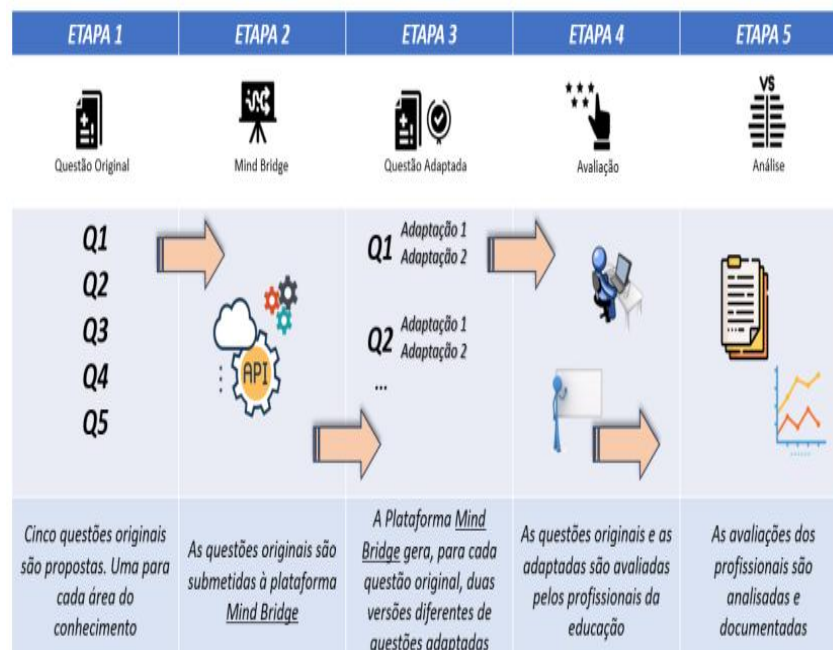
No Módulo Lógico, as entradas são processadas por um servidor que aplica filtros para garantir a integridade dos dados. Inicialmente, um filtro de entrada utiliza recursos de Inteligência Artificial para detectar e classificar conteúdos inválidos, como spam. Em seguida, um filtro de análise específica identifica as necessidades do estudante e determina como a questão deve ser adaptada. Por fim, a entrada passa por um filtro de formulação de prompt, no qual um prompt padronizado é selecionado e a estrutura do texto é ajustada conforme as necessidades do aluno.

No Módulo Generativo, a entrada estruturada é processada por um LLM, responsável por reformular a questão com base nas análises realizadas anteriormente, adaptando-a às necessidades específicas do estudante. Esse processo conta com o suporte de um repositório interno, que auxilia na continuidade do processamento. A versão adaptada da questão é, então, encaminhada ao professor para revisão e possíveis ajustes antes de sua aplicação.

O processo metodológico adotado pelos autores foi estruturado em cinco etapas, conforme ilustrado na Figura 6. Inicialmente, foram elaboradas cinco questões avaliativas de diferentes áreas do conhecimento, que foram submetidas à arquitetura para análise. Em seguida, o sistema gerou duas versões adaptadas para cada questão original.

Posteriormente, profissionais da educação avaliaram tanto as versões originais quanto as adaptadas, fornecendo feedback sobre as modificações realizadas. Por fim, as avaliações e os comentários dos participantes foram sistematizados para análise.

Figura 6- O processo metodológico da plataforma



Fonte: Adaptado de Silva et al., 2024

Na Etapa 5, um questionário para avaliação foi compartilhado com sete profissionais da educação, incluindo professores de diferentes níveis de ensino, docentes do Atendimento Educacional Especializado (AEE) e um pedagogo. O objetivo da avaliação foi verificar a eficácia da plataforma na adaptação de questões, bem como sua utilidade percebida pelos professores e seu potencial de apoio a estudantes neurodivergentes.

A proposta foi avaliada de modo qualitativo, no qual os participantes analisaram tanto as versões originais quanto as versões adaptadas das questões. Os resultados indicaram predominância de respostas positivas, evidenciando que as adaptações contribuíram para tornar as questões mais claras, compreensíveis e adequadas, mantendo o significado original.

Ademais, o pedagogo participante destacou positivamente as adaptações realizadas pela plataforma, como o uso de grifo e de cores na organização visual das perguntas, reconhecendo esses recursos como facilitadores da compreensão.

Os autores destacam o potencial da plataforma como apoio à adaptação das questões avaliativas, tornando-as mais acessíveis e compreensíveis. No entanto, a avaliação foi realizada apenas com profissionais da educação, não havendo avaliação direta com estudantes, o que indica a necessidade de investigações futuras em contextos reais de aplicação.

Como trabalhos futuros, os autores apontam a ampliação das formas de entrada de dados, como áudio e imagem, bem como a implementação de um sistema de feedback contínuo e a aplicação da proposta em diferentes níveis de ensino.

3.3 Meissner et al., (2024)

O estudo de Meissner et al. (2024) explora como LLMs, aplicados à produção de materiais instrucionais no ensino superior, podem contribuir para a geração automática de avaliações de aprendizagem com itens de resposta aberta. Isso se aplica tanto para avaliações formativas quanto somativas, em conteúdos matemáticos típicos de cursos do ensino superior.

Para isso, é apresentada uma ferramenta baseada em LLM, denominada ItemForge, na qual a geração dos itens matemáticos é orientada por instruções (prompts) que especificam conceitos matemáticos e níveis cognitivos, conforme a Taxonomia de Bloom revisada, que organiza os objetivos educacionais em uma estrutura bidimensional composta por conhecimento e processos cognitivos (Krathwohl, 2002).

A ferramenta foi desenvolvida especificamente para a geração automática de itens de avaliação eletrônica na área de matemática, destinados ao uso em sistemas digitais de ensino.

A abordagem é alinhada aos níveis e à hierarquia dos objetivos cognitivos de aprendizagem da Taxonomia de Bloom, os quais são compostos pela dimensão do processo cognitivo, que delineia seis processos (lembrar, entender, aplicar, analisar, avaliar e criar) necessários para que os alunos abordem um item ou problema, e pela dimensão do conhecimento, composta por quatro categorias (factual, conceitual, processual e metacognitivo).

Esse processo é utilizado para orientar cada item da avaliação antes de ser gerado, considerando os conteúdos matemáticos específicos empregados na geração dos itens avaliativos.

Na Figura 7, é apresentado um exemplo de prompt que orienta a geração de itens por competências, entendidos como itens avaliativos concebidos para medir o desempenho do estudante em competências específicas, descritas pela combinação entre o tipo de conhecimento e a dimensão do processo cognitivo requerida para a resolução da tarefa.

Nesse prompt, são especificadas as dimensões da Taxonomia de Bloom revisada, bem como os conceitos matemáticos da disciplina.

As variáveis de entrada responsáveis por receber as dimensões taxonômicas e os conteúdos são representadas por $\{ . . . \}$, sendo utilizadas para parametrizar dinamicamente a geração dos itens avaliativos.

Figura 7 - Prompt para geração dos itens matemáticos

Crie um exercício de matemática sobre o conceito matemático {concept}, representando a dimensão do processo {pd} e a dimensão do conhecimento {kd}.

**INSTRUÇÕES**

1. Os itens devem conter notações matemáticas e fórmulas em formato LaTeX!
2. Os itens devem ser formulados em alemão!
3. Os itens devem conter números aleatórios e não apenas números enumerados (1, 2, 3, 4, 5, ...)!
4. Os itens não devem fornecer quaisquer dicas, ajuda, exemplos ou explicações!
5. Pense passo a passo sobre como gerar este item corretamente.

**CONTEXTO**

- Os itens gerados são destinados a estudantes do primeiro semestre de Bacharelado em Ciência da Computação, que devem ser apoiados e desafiados individualmente, de modo que atendam às exigências do curso de matemática.
- Os estudantes recebem os itens gerados por meio de um sistema automático de recomendação e podem respondê-los em formato textual.
- Os itens são selecionados de forma personalizada para os estudantes, de acordo com seu nível de competência (representado por uma dimensão de conhecimento e uma dimensão de processo) e os conceitos atualmente relevantes.
- Portanto, é extremamente importante que os itens correspondam apenas ao conceito matemático especificado, à dimensão do processo e à dimensão do conhecimento!

Fonte :Adaptado Meissner et al. (2024)

Nesse prompt, o conceito matemático, a dimensão do processo cognitivo e a dimensão do conhecimento, conforme a Taxonomia de Bloom, são definidos antes da geração do item. Além disso, são estabelecidas instruções, como o uso de notação matemática, a geração dos itens em língua alemã, a ausência de dicas ou explicações nos itens e o contexto de aplicação desses itens para alunos de cursos superiores da área de Computação.

A qualidade dos itens gerados, considerando critérios como adequação ao conceito matemático, completude da tarefa, correção da tarefa, completude da solução apresentada e

correção da solução apresentada, bem como sua adequação ao nível cognitivo e ao conteúdo, foi analisada por meio de um estudo de pequena escala. Essa avaliação foi conduzida por três especialistas em Matemática, que analisaram um total de 240 itens gerados automaticamente, examinando sua eficácia e relevância como adequação ao conceito matemático, completude e correção da tarefa, bem como completude e correção da solução apresentada.

Os resultados indicaram que a ferramenta computacional baseada em LLMs, destinada à geração automática de itens avaliativos orientados por competências, é capaz de produzir itens de alta qualidade, alinhados aos conceitos matemáticos e aos níveis cognitivos pretendidos.

Contudo, observou-se que algumas soluções apresentaram imprecisões ou incompletudes nos itens, apontando a necessidade de aperfeiçoamentos adicionais no processo de geração e, para trabalhos futuros, a realização de questões fechadas de múltipla escolha.

3.4 Wang et al., (2025)

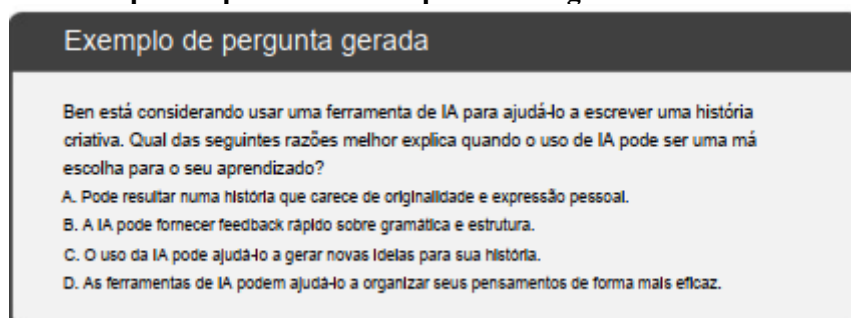
Wang et al. (2025) propõem uma abordagem baseada em LLM para a geração automática de questões de múltipla escolha (QMC) voltadas à alfabetização em Inteligência Artificial (IA) na educação básica.

Os autores destacam a carência de materiais avaliativos escaláveis e confiáveis para o ensino de IA nesse contexto. A proposta baseia-se em um fluxo de trabalho multiagente, no qual diferentes agentes, orientados por prompts, desempenham funções como a geração das questões, a realização de críticas linguísticas e a verificação de falhas na construção dos itens avaliativos.

As questões geradas são alinhadas aos objetivos de aprendizagem da alfabetização em IA, aos níveis escolares e aos níveis da Taxonomia de Bloom (Krathwohl, 2002).

Como forma de exemplificação, a Figura 8 ilustra uma questão de múltipla escolha gerada pelo sistema, correspondente ao nível de alfabetização.

Figura 8- Exemplo da questão de múltipla escolha gerada



Fonte: Wang et al., (2025)

Dessa forma, a questão foi elaborada por meio de um fluxo iterativo baseado em um sistema multiagente, no qual diferentes agentes atuam na geração, revisão linguística, verificação de falhas e validação final dos itens, com o objetivo de assegurar sua qualidade pedagógica.

Os itens são refinados até atenderem a critérios de qualidade que abrangem aspectos linguísticos, como clareza, compreensibilidade e correção gramatical e pedagógicos como alinhamento aos objetivos de aprendizagem, adequação ao nível escolar e ao nível cognitivo da Taxonomia de Bloom, além da ausência de falhas na elaboração das alternativas.

Foram geradas 40 questões destinadas a estudantes do 7º ao 9º ano, avaliadas por três especialistas por meio de uma rubrica composta por dez critérios. Os resultados indicam elevados níveis de concordância entre avaliadores (97%), alta qualidade linguística (93%) e forte alinhamento aos objetivos de aprendizagem, evidenciando o potencial da abordagem para a geração de avaliações escaláveis e pedagogicamente consistentes.

No entanto, os resultados também sugerem que preferências individuais e estilos de ensino podem influenciar a disposição dos docentes para utilizar determinadas questões geradas pelo sistema. Ademais, os autores reconhecem limitações no estudo, destacando a necessidade de investigações futuras em contextos reais de sala de aula, uma vez que a proposta não foi testada diretamente nesse ambiente. Como trabalhos futuros, indicam ainda a ampliação da adaptabilidade do sistema a diferentes níveis de escolaridade, considerando de forma mais abrangente os níveis da Taxonomia de Bloom.

3.5 Diyab et al., (2025)

Diyab et al., (2025) apresentam o AI Assess, um sistema de avaliação educacional baseado no ChatGPT, desenvolvido para cursos de Engenharia de Software e Ciência da Computação. A proposta integra engenharia de prompts e é composta por quatro funções principais: correção automática, identificação de lacunas de aprendizagem, fornecimento de feedback e geração de perguntas.

A função de correção automática refere-se à capacidade do sistema de atribuir notas às respostas dos estudantes, especialmente em questões discursivas, considerando o contexto do curso, seus tópicos e objetivos de aprendizagem. As respostas são avaliadas e recebem uma pontuação em escala percentual (0 a 100), podendo incluir justificativas para a nota atribuída.

Para questões objetivas, como múltipla escolha e verdadeiro/falso, a correção é realizada por meio da comparação direta com o gabarito fornecido.

No que se refere à geração de perguntas, o sistema é capaz de produzir automaticamente diferentes tipos de questões, tais como múltiplas escolhas, verdadeiro/falso, preenchimento de lacunas, resposta curta e questões dissertativas. A Figura 9 apresenta o template do prompt utilizado nesse processo, o qual retorna uma lista de perguntas com base nas informações de entrada fornecidas

Figura 9 - Prompt de geração de perguntas.

Solicito que você crie uma avaliação para alunos de graduação da disciplina [TÍTULO DA DISCIPLINA]. A avaliação deve incluir [NÚMERO DE QUESTÕES DE MÚLTIPLA ESCOLHA] questões de múltipla escolha com lacunas para preencher, [NÚMERO DE QUESTÕES DE VERDADEIRO OU FALSO] questões de verdadeiro ou falso, [NÚMERO DE QUESTÕES DE RESPOSTA CURTA] questões de resposta curta e [NÚMERO DE QUESTÕES DE REDAÇÃO] questões dissertativas. A avaliação deve abordar o tópico intitulado [TÍTULO DO TÓPICO], que possui os seguintes conceitos-chave: [CONCEITOS-CHAVE]. As questões de múltipla escolha devem conter apenas uma resposta correta, e as questões de resposta curta devem apresentar algum grau de subjetividade. Segue abaixo uma visão geral do tópico: [VISÃO GERAL DO TÓPICO]

O prompt é baseado em técnica zero-shot, contudo, de acordo com os autores, esse mesmo prompt também pode ser utilizado no modo few-shot, por meio da inclusão de exemplos, nos quais os docentes dos cursos podem fornecer perguntas que o ChatGPT pode usar para orientar a geração das novas questões.

Os autores demonstraram que o uso de prompts bem projetados permite gerar questões e exercícios personalizados, bem como avaliar automaticamente as respostas dos estudantes e oferecer feedback imediato. Os testes com respostas de até 25 estudantes indicam que a inclusão de exemplos humanos no processo melhora a consistência da correção automática, reforçando a importância da supervisão humana.

O AI Assess também se mostrou eficaz na identificação de tópicos em que a maioria dos alunos apresenta dificuldades. Além disso, o sistema contribuiu para a redução da carga de trabalho docente e para a possível melhoria da aprendizagem dos estudantes. O artigo evidencia que, embora sistemas baseados em LLMs, como o AI Assess, apresentem avanços relevantes na automação da avaliação educacional, sua adoção deve ser realizada de forma cautelosa.

Os resultados indicam ganhos em eficiência, personalização e feedback imediato, mas também revelam desafios importantes, como a presença de vieses, limitações na precisão e riscos de alucinação. Nesse sentido, os autores defendem a integração obrigatória da supervisão humana, tanto para validação das avaliações quanto para garantir a confiabilidade no processo avaliativo.

Como linhas futuras de investigação, os autores indicam a integração dos componentes do AI Assess com outras tecnologias educacionais, como os ambientes virtuais de aprendizagem (LMS), de modo a promover uma integração no processo de ensino e aprendizagem. Além disso, destacam a ampliação do sistema para suportar avaliações multimodais, incluindo vídeos, produções criativas e outras formas não textuais de expressão, ampliando as possibilidades de aplicação em diferentes contextos educacionais.

3.6 Síntese dos trabalhos relacionados

Os trabalhos descritos estão resumidamente comparados no Quadro 1.

Quadro 1 - Síntese dos Trabalhos relacionados

Trabalho	Ferramenta apresentada	Público-alvo	Apoio ao docente	Prática inclusiva	Adaptação com base em preferências dos alunos	Utiliza o conteúdo do plano de ensino da disciplina	Geração de questões avaliativas
Bhowmick et al., (2023)	QGen	Estudantes do fundamental	Sim	Não	Não	Sim	Sim
Silva et al. (2024)	Mind Bridge	Estudantes com TEA	Sim	Sim	Não	Não	Sim
Meissner et al., (2024)	itemForge	Estudantes do Ensino Superior	Sim	Não	Não	Não	Sim
Wang et al., (2025)	Sistema Multiagente	Estudantes do Ensino Infantil	Sim	Não	Não	Não	Sim
Diyab et al., (2025)	AI Assess	Estudantes do Ensino Superior	Sim	Não	Não	Sim	Sim
Este trabalho	MeuAvalia	Estudantes do ensino médio com ou sem o TEA	Sim	Sim	Sim	Sim	Sim

Trabalhos Relacionados

Conforme mostra o Quadro 1, os trabalhos analisados apresentam contribuições relevantes no uso de modelos de linguagem para apoiar a geração automática de questões avaliativas. Entretanto, ainda são limitadas as propostas que contemplam avaliações formativas adaptadas de maneira integrada, ou seja, que articulem, em um único fluxo, a geração de questões, o alinhamento aos objetivos de aprendizagem dos conteúdos, a adaptação conforme o perfil do estudante e a incorporação de princípios pedagógicos, como uma prática inclusiva.

Nesse contexto, tanto o estudo de Bhowmick et al. (2023) quanto a presente pesquisa buscam reduzir o tempo despendido pelo docente na elaboração de avaliações e utilizam modelos de linguagem para geração automática de questões. Embora o estudo não aborde explicitamente a inclusão, propõe a personalização como forma de facilitar a aprendizagem. No entanto, utiliza conteúdos de fontes externas selecionadas pelo docente, enquanto esta pesquisa baseia-se nos tópicos do plano de ensino, com foco no Ensino Médio. Além disso, não especifica claramente o nível de ensino ao qual as questões se destinam.

No que se refere às práticas inclusivas, o trabalho de Silva et al. (2024) apresenta uma aproximação com esta pesquisa ao empregar linguagem simples na adaptação de questões para estudantes neurodivergentes. Embora os autores não façam referência explícita a uma prática inclusiva específica voltada ao TEA, a simplificação da linguagem pode contribuir para a acessibilidade e compreensão dos enunciados. Tanto esta pesquisa quanto Silva et al. (2024) buscam tornar o processo avaliativo mais inclusivo por meio de ajustes na linguagem e na estrutura das questões. Diferentemente, a presente pesquisa incorpora princípios do Ensino Estruturado e considera as preferências avaliativas dos alunos. Em ambos os casos, o professor atua como mediador, revisando as questões antes de sua aplicação.

O estudo de Meissner et al. (2024) propõe a geração automática de itens discursivos em Matemática para o ensino superior, orientados pela Taxonomia de Bloom e por conteúdo definidos pelos docentes, sem abordar a inclusão. Em contraste, a presente pesquisa amplia essa abordagem ao incorporar, além dos tópicos do plano de ensino, as preferências avaliativas dos alunos e práticas inclusivas, utilizando diferentes tipos de questões, e não apenas questões discursivas.

Em relação ao trabalho de Wang et al. (2025), destaca-se a utilização de LLMs para gerar apenas questões de múltipla escolha com base em objetivos de aprendizagem e níveis da Taxonomia de Bloom no contexto da alfabetização. Por sua vez, a presente pesquisa não se limita a esse tipo de questão. Além disso, embora Wang et al. (2025) não utilizem diretamente

as preferências e estilos de aprendizagem dos alunos como parâmetros de geração, os autores indicam que tais fatores podem influenciar a formulação das questões.

Em contraste, a presente pesquisa destaca que para alguns contextos educacionais a abordagem precisa de um refinamento mais apropriado para gerar questões de múltipla escolha pedagogicamente de alta qualidade.

Por fim, o sistema AI Assess, proposto por Diyab et al. (2025), destaca-se por oferecer múltiplas funcionalidades de apoio ao docente, incluindo acompanhamento do desempenho dos estudantes e geração de avaliações. O sistema utiliza tópicos de conteúdo como base e contempla diferentes tipos de questões, sendo aplicado ao ensino superior. Este estudo aproxima-se da presente pesquisa ao demonstrar a viabilidade da geração de avaliações com questões abertas e de múltipla escolha, embora não aborde aspectos de inclusão ou preferências avaliativas dos alunos.

Diante dessas lacunas, a presente pesquisa diferencia-se ao propor uma abordagem voltada à geração de modelos de avaliações formativas adaptadas, integrando o plano de ensino do docente, as preferências avaliativas dos alunos e a prática inclusiva do Ensino Estruturado, com foco no Ensino Médio e em adolescentes com ou sem Transtorno do Espectro Autista (TEA). A abordagem proposta é detalhada no próximo capítulo.

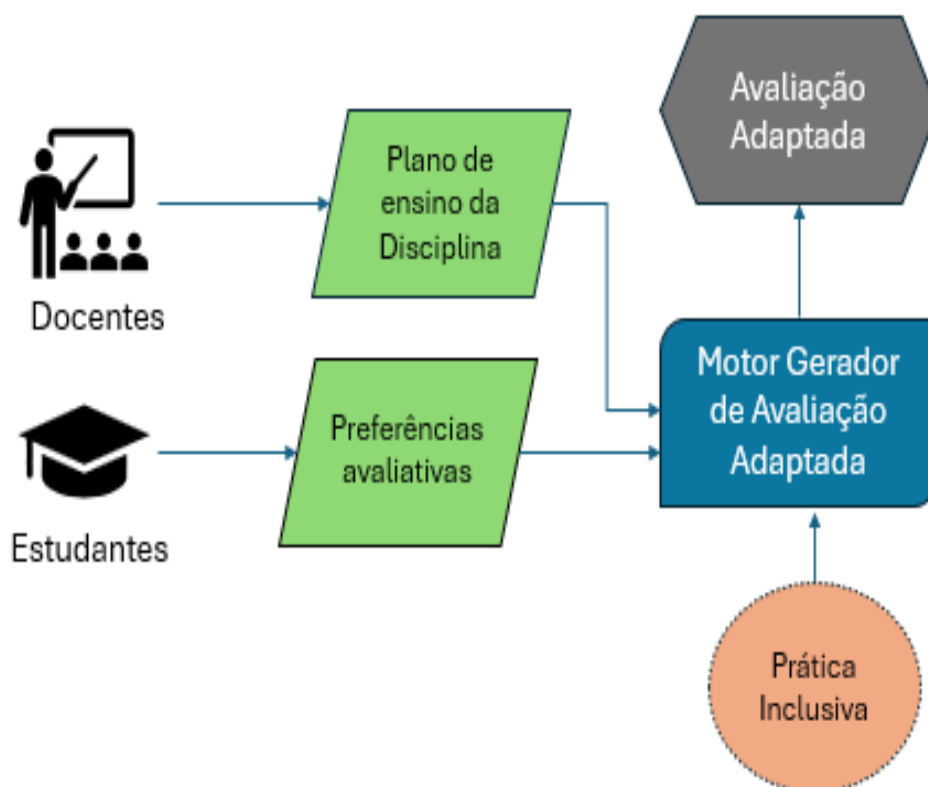
4. Abordagem MeuAvalia

Este capítulo apresenta a abordagem **MeuAvalia**, contribuição central deste trabalho de mestrado. O capítulo está organizado da seguinte forma: a Seção 4.1 apresenta a visão geral da abordagem, descrevendo seus elementos e como eles se articulam; a Seção 4.2 propõe o motor de geração de avaliações; a Seção 4.3 apresenta um exemplo de uso da abordagem; e, por fim, a Seção 4.4 apresenta considerações sobre o capítulo.

4.1 Visão geral da abordagem MeuAvalia

A Figura 10 apresenta o fluxograma da abordagem **MeuAvalia**, elaborado utilizando a notação de fluxograma. Esse diagrama representa o fluxo do processo de geração automatizada de modelos de avaliação adaptadas, considerando três elementos centrais da abordagem: a prática inclusiva, o plano de ensino da disciplina e as preferências avaliativas dos estudantes.

Figura 10 – Visão geral da abordagem MeuAvalia



Fonte: Elaborado pelo autor (2025).

A abordagem proposta está centrada no apoio ao docente de uma turma do ensino médio, no contexto de uma disciplina específica. Então, o docente é responsável pela elaboração do plano de ensino da disciplina que contém os tópicos (conteúdos) alvo de aprendizagem. O docente é o ator que inicia o processo de geração de avaliação.

Por outro lado, os alunos são os atores responsáveis por definir suas preferências avaliativas. Essas preferências dizem respeito às formas pelas quais os estudantes se sentem mais confortáveis para serem avaliados e estão alinhadas com o princípio da diversidade e da educação inclusiva.

Como princípio pedagógico da abordagem, destaca-se o uso de prática inclusiva de aprendizagem, neste caso, o Ensino Estruturado.

Essa prática é utilizada por favorecer a organização, clareza e previsibilidade das atividades educacionais, especialmente com respeito às avaliações da aprendizagem, de modo que elas se tornem mais acessíveis, especialmente para estudantes com TEA. Assim, a prática inclusiva atua como um elemento orientador no processo de adaptação das avaliações.

As informações provenientes do plano de ensino, das preferências avaliativas e das diretrizes de prática inclusiva são processadas pelo Motor gerador de Avaliação Adaptada, para geração, amparada em um LLM, de avaliações adaptadas ao contexto da turma. Esse componente constitui o núcleo da abordagem e estará detalhadamente descrito na Seção 4.2.

Como saída da abordagem, o docente obtém as avaliações. A Avaliação de Aprendizagem Adaptada (AAA), doravante denominada Avaliação Adaptada (AA), resulta da integração dos elementos considerados na abordagem. Esse formato de avaliação é organizado de acordo com os tópicos do plano da disciplina referentes ao conteúdo previsto para aquela avaliação e as preferências diversas dos alunos quanto ao tipo de avaliação (discursiva, objetiva, oral ou prática), permitindo ao docente ter em mãos opções de avaliações mais formativas, acessíveis e alinhadas às necessidades dos estudantes.

Como premissa da abordagem, a avaliação formativa contribui para um processo avaliativo mais inclusivo, ao favorecer o acompanhamento contínuo da aprendizagem e a adaptação das estratégias avaliativas às características dos alunos. No contexto da avaliação formativa, consideram-se, neste trabalho, quatro tipos de métodos possíveis de avaliação, a saber (Neiva et al., 2021): objetivo, discursivo, oral ou prático.

Adicionalmente, nesta pesquisa, entende-se por preferências avaliativas as escolhas dos alunos em relação às características dos métodos avaliativos (oral, prático, dissertativo ou objetivo), refletindo como se sentem mais confortáveis e confiantes para demonstrar seu

aprendizado dentro do formato selecionado. Tais preferências podem estar associadas a um ou mais métodos avaliativos.

Para fins de organização, as preferências consideradas neste trabalho foram classificadas em duas categorias: (i) preferências de formato avaliativo, relacionadas ao tipo de avaliação adotado (objetiva, oral, prática ou discursiva), e (ii) preferências instrucionais, que correspondem às condições, adaptações e orientações de aplicação da avaliação para o docente, influenciando a forma como o estudante interage com a atividade avaliativa.

Essas preferências incluem aspectos como tempo disponível para realização, possibilidade de trabalho em dupla, antecipação de instruções e outras estratégias de apoio que podem favorecer a participação e o desempenho dos estudantes.

As preferências avaliativas e as instrucionais consideradas neste estudo foram definidas com base na literatura relacionada à avaliação inclusiva, aos estilos de aprendizagem e às preferências avaliativas dos estudantes (Kleinlein, 2025; Tai et al., 2023; Weber, 2025). Além disso, foram complementadas pelos dados obtidos por meio de questionários aplicados a estudantes do Ensino Médio das escolas estaduais Liceu Paraibano e Horácio de Almeida (Apêndice A), permitindo identificar características e preferências relevantes para o contexto investigado.

Para isso, inicialmente, os estudantes participaram de uma apresentação em que foram explicados os conceitos de métodos avaliativos (objetivo, discursivo, oral e prático) e de preferências avaliativas.

Em seguida, responderam a um questionário (Apêndice A), no qual indicaram as características que consideravam mais adequadas ou confortáveis para cada método avaliativo. As respostas obtidas foram analisadas e utilizadas para a definição das preferências de formato avaliativo e das preferências instrucionais adotadas nesta pesquisa.

O resultado dessa análise encontra-se sintetizado nos Quadros 2 e 3. O Quadro 2 apresenta as preferências avaliativas associadas aos métodos avaliativos, contemplando características relacionadas ao formato e à estrutura das atividades de avaliação, como organização das questões, tipo de resposta esperada, forma de apresentação dos conteúdos e estratégias de contextualização.

Essas preferências representam elementos que podem tornar as avaliações mais adequadas às características e necessidades dos estudantes.

Por sua vez, o Quadro 3 apresenta as preferências instrucionais associadas aos métodos avaliativos, evidenciando as condições de aplicação da avaliação. Neste trabalho, as preferências instrucionais dizem respeito a aspectos relacionados ao tempo disponível, à

preparação prévia e às orientações fornecidas aos estudantes. Essa distinção permite representar tanto características inerentes ao formato da avaliação quanto elementos que podem favorecer sua acessibilidade, compreensão e participação durante o processo avaliativo.

Mais detalhes sobre o questionário aplicado são apresentados no Apêndice A.

Quadro 2 - Preferências Avaliativas associadas aos Métodos Avaliativos

Preferência Avaliativa	Objetiva	Oral	Prática	Discursiva
Organizar as questões da mais fácil para a mais difícil	✓	✓	✓	✓
Destacar palavras importantes em maiúsculas	✓		✓	✓
Usar enunciados curtos e diretos	✓			✓
Basear as questões em situações-problema	✓	✓		
Explicar termos técnicos para facilitar a compreensão	✓			
Seguir exemplo demonstrado pelo professor			✓	
Receber instruções para uso dos materiais			✓	
Resolver situações práticas			✓	
Poder revisar material de apoio durante a prática			✓	
Explicar conceito por meio de exemplos		✓		
Explicar pontos principais do conteúdo		✓		
Resumo oral de até 2 minutos (pitch)		✓		
Enviar áudio ou vídeo em vez de apresentação ao vivo		✓		
Permitir respostas curtas e objetivas				✓
Permitir respostas longas e explicativas				✓
Relacionar questões ao cotidiano ou notícias				✓
Iniciar questão com verbo de comando (Descreva, Compare, etc.)				✓

Quadro 3- Preferências Instrucionais associadas aos Métodos Avaliativos.

Preferência Instrucional	Objetiva	Oral	Prática	Discursiva
Conceder tempo extra para realizar a avaliação	✓	✓	✓	✓
Possibilitar responder à avaliação em dupla	✓	✓	✓	✓
Saber previamente se a resposta deve ser curta ou explicativa		✓		
Ser informado sobre o tempo disponível para responder		✓		
Ter tempo extra para se preparar antes de responder		✓		
Receber perguntas possíveis antecipadamente para estudo		✓		

Observa-se que as preferências instrucionais não alteram o formato da avaliação em si, mas definem condições e estratégias de aplicação que podem contribuir para maior acessibilidade, compreensão das tarefas e participação dos estudantes durante o processo avaliativo.

4.2 Motor gerador de Avaliação Adaptada

O Motor gerador de Avaliação Adaptada constitui o núcleo central da abordagem MeuAvalia, sendo responsável por processar os elementos de entrada e gerar uma Avaliação Adaptada (AA).

O formato esperado para uma AA é definido a seguir:

1. Identificação da Avaliação

Disciplina:

Série

Nome;

2. Método Avaliativo Selecionado

Tipo de avaliação (objetiva, discursiva, prática ou oral)

3. Estrutura das Questões

Linguagem clara e Objetiva (prática inclusiva)

Incorporação das preferências avaliativas.

Conforme formato esperado, Motor é desenvolvido considerando as seguintes entradas (Algoritmo_1).

1. Tópicos de conteúdo extraídos do plano de ensino pelo docente, que definem o foco da avaliação em questão;
2. Preferências avaliativas dos estudantes, que refletem as formas mais adequadas de avaliação para cada aluno;
3. Critérios do docente, como a quantidade de questões a serem geradas, a seleção da turma e o nome da avaliação.

O processo de geração envolve a correspondência entre os elementos de entrada, associando os tópicos de conteúdo aos métodos avaliativos e às preferências dos estudantes. A partir dessa associação, o motor instancia um template de prompt, que é utilizado para gerar as questões. Como saída, o motor produz os modelos AA, como mostra o Algoritmo_1.

Figura 11 – Algoritmo 1: Lógica do Motor Gerador de Avaliação Adaptada

Algoritmo 1 – Geração de Avaliação Adaptada

Input: plano_de_ensino, preferencias_avaliativas, criterios_do_docente

Output: modelo_avaliacao

// Etapa 1: Verificação das entradas

01: If (plano_de_ensino.tem_topicos_definidos() AND
criterios_do_docente.estao_definidos()) then

02: topicos_de_conteudo := plano_de_ensino.obter_topicos();

03: Else

04: Retornar erro: "Plano de ensino ou critérios do docente não cadastrados.";

05: End If;

// Etapa 2: Validação da existência de tópicos

06: If (topicos_de_conteudo.empty()) then

07: Retornar erro: "É necessário selecionar ao menos 1 tópico de conteúdo.";

08: End If;

// Etapa 3: Seleção do método avaliativo

09: If (metodo_selecionado_manual) then

10: metodo_selecionado := selecionar_metodo_manual();

11: Else

12: metodo_selecionado := selecionar_metodo_automatico(preferencias_avaliativas);

13: End If;

// Etapa 4: Validação do método selecionado

14: If (metodo_selecionado.empty()) then

15: Retornar erro: "Erro na geração da avaliação sem preferências.";

16: End If;

// Etapa 5: Filtragem das preferências conforme o método selecionado

17: preferencias_filtradas := filtrar_preferencias(
metodo_selecionado,

preferencias_avaliativas

```

);
// Etapa 6: Cálculo do peso agregado por preferência
18: Para cada preferencia em preferencias_filtradas faça
19: peso_total[preferencia] := quantidade_votos(preferencia) *
   peso_docente(preferencia);
20: Fim Para

21: preferencias_ordenadas := ordenar_por_relevancia(
   quantidade_votos,
   peso_total
);

22: preferencias_finais := selecionar_preferencias_predominantes(
   preferencias_ordenadas
); // Etapa 9: Construção do prompt
23: prompt := gerar_prompt(
   topicos_de_conteudo,
   metodo_selecionado,
   preferencias_finais,
   criterios_do_docente,
   pratica_inclusiva
);
// Etapa 10: Geração da avaliação
24: questoes := gerar_questoes(prompt);
25: questoes_ajustadas := aplicar_criterios(
   questoes,
   criterios_do_docente
);
// Etapa 11: Saída final
26: modelo_avaliacao := montar_modelo(questoes_ajustadas);
27: Retornar modelo_avaliacao;

End

```

Inicialmente, nas linhas 01 a 05, o motor verifica se o plano de ensino contém tópicos previamente cadastrados e se os critérios definidos pelo docente foram informados. Caso alguma dessas informações esteja ausente, o processo é interrompido e uma mensagem de erro é retornada. Estando as entradas corretas, os tópicos de conteúdo são obtidos a partir do plano de ensino oferecido pelo docente.

Em seguida, nas linhas 06 a 08, realiza-se a validação da existência de tópicos de conteúdo. Caso nenhum tópico tenha sido selecionado, o motor retorna erro, indicando a necessidade de informar ao menos um conteúdo para a geração da avaliação.

Nas linhas 09 a 16, ocorre a seleção e validação do método avaliativo. Se houver definição manual por parte do docente, o método informado é utilizado. Caso contrário, o motor

realiza a seleção automática com base nas preferências avaliativas registradas pelos estudantes. Na linha 17, as preferências avaliativas são filtradas conforme o método selecionado, de modo que apenas preferências compatíveis com a modalidade escolhida sejam consideradas no processo.

Nas linhas 18 a 20, o motor realiza o cálculo dos pesos associados às preferências avaliativas. Inicialmente, cada preferência é contabilizada de acordo com a quantidade de estudantes que a selecionaram, permitindo identificar aquelas com maior representatividade na turma. Além da frequência de votos dos estudantes, a abordagem permite associar pesos às preferências avaliativas. Esses pesos representam a relevância pedagógica atribuída a cada preferência, considerando aspectos como os objetivos de aprendizagem da disciplina e as características do conteúdo trabalhado. Dessa forma, o mecanismo de pesos possibilita incorporar o julgamento pedagógico do docente ao processo de seleção das preferências, sem desconsiderar as escolhas realizadas pelos estudantes.

O docente pode atribuir um peso a cada preferência em uma escala de 1 a 10. Quanto maior o valor atribuído, maior será a prioridade daquela preferência em situações que exijam um critério adicional de seleção.

Para identificar as preferências predominantes da turma, o motor considera inicialmente a frequência de ocorrência de cada preferência avaliativa entre os estudantes. As preferências são contabilizadas e ordenadas de forma decrescente conforme o número de votos recebidos. Quando não há empate, a seleção é realizada exclusivamente com base nessa frequência.

Entretanto, quando duas ou mais preferências apresentam a mesma quantidade de votos, a abordagem utiliza os pesos atribuídos pelo docente como critério de desempate. Nesses casos, é calculado um peso total para cada preferência por meio da fórmula do *PesoTotal*.

Nesses casos, é calculado um peso total para cada preferência por meio da fórmula seguinte:

$$PesoTotal = QuantidadeDeVotos \times PesoDocente$$

Em que:

- **QuantidadeDeVotos** corresponde ao número de estudantes que selecionaram a preferência;
- **Peso Docente** representa o valor atribuído pelo professor;
- **PesoTotal** é utilizado para estabelecer a prioridade entre preferências que apresentem a mesma frequência de ocorrência.

Após esse cálculo, as preferências são ordenadas segundo sua relevância e aquelas consideradas predominantes são selecionadas para compor o processo de geração da Avaliação Adaptada (AA).

O Quadro 4 apresenta um exemplo ilustrativo dessa estratégia.

Quadro 4- Exemplo de preferências com pesos

Preferência avaliativa	Votos	Peso	Peso Total
Organizar da mais fácil para a mais difícil	8	5	40
Basear as questões em situações-problema	6	3	18
Destacar palavras em MAIÚSCULAS	6	4	24
Usar enunciados curtos e diretos	6	2	12
Explicar termos técnicos	5	1	5

Para o exemplo apresentado, a preferência “Organizar da mais fácil para a mais difícil” ocupa a primeira posição por possuir a maior quantidade de votos entre os estudantes. Em seguida, observa-se um empate entre as preferências “Basear as questões em situações-problema”, “Destacar palavras em MAIÚSCULAS” e “Usar enunciados curtos e diretos”, todas com seis votos. Nessa situação, o sistema aplica o critério de desempate baseado no peso total de cada preferência.

Como “Destacar palavras em MAIÚSCULAS” apresenta peso total igual a 24, ela é posicionada à frente das demais preferências empatadas. Em seguida, “Basear as questões em situações-problema”, com peso total igual a 18, ocupa a posição seguinte, enquanto “Usar enunciados curtos e diretos”, com peso total igual a 12, é posicionada posteriormente.

Considerando que a abordagem seleciona apenas as três preferências predominantes para compor o prompt de geração da avaliação, o ranking final é formado por: (1) “Organizar da mais fácil para a mais difícil”, (2) “Destacar palavras em MAIÚSCULAS” e (3) “Basear as questões em situações-problema”. Nesse caso, a preferência “Usar enunciados curtos e diretos”, apesar de possuir a mesma quantidade de votos que as demais preferências empatadas, não é selecionada em razão de apresentar menor peso total.

Esse exemplo evidencia que os pesos não alteram a ordenação quando as quantidades de votos são diferentes, mas desempenham um papel determinante em situações de empate, influenciando diretamente quais preferências serão incluídas no conjunto final utilizado pelo motor de geração da Avaliação Adaptada (AA).

Nas linhas 21 e 22, as preferências avaliativas são organizadas em ordem de relevância e as preferências predominantes são selecionadas para compor a geração da avaliação. Na linha 23, é construído o template de prompt que será encaminhado ao motor gerador. Esse prompt reúne os tópicos de conteúdo, o método avaliativo selecionado, as preferências finais, os critérios definidos pelo docente e os princípios de prática inclusiva. Em seguida, nas linhas 24 e 25, o motor gera as questões e aplica os ajustes necessários conforme os critérios estabelecidos pelo docente.

O prompt apresentado corresponde ao método avaliativo objetivo. Para os demais métodos avaliativos (discursivo, oral e prático), a abordagem utiliza templates específicos contendo instruções compatíveis com cada modalidade de avaliação. Esse prompt reúne os tópicos de conteúdo, o método avaliativo selecionado, as preferências finais, os critérios definidos pelo docente e os princípios de prática inclusiva. Sobre a prática inclusiva seus princípios já estão definidos no template do prompt de forma fixa. Com base nessas informações, o motor gera as questões da avaliação e aplica os ajustes necessários conforme os parâmetros estabelecidos pelo professor. Por fim, nas linhas 26 e 27, o modelo final de avaliação é retornado como saída do algoritmo. Como exemplo, considere o seguinte prompt do Motor, como mostra na Etapa 9 do Algoritmo_1.

Quadro 5 – Exemplo do prompt utilizado pelo motor gerador de avaliações adaptadas

Você é um(a) professor(a) do ensino técnico de nível médio.
Elabore uma avaliação de aprendizagem do tipo {{METODO_AVALIATIVO}} para a turma {{TURMA_NOME}}.

A avaliação deve contemplar os seguintes tópicos:
{{TOPICOS}}

As preferências avaliativas já foram processadas pelo sistema e devem ser incorporadas:
{{PREFERENCIAS}}

Critérios do docente:
{{CRITERIOS}}

Prática Inclusiva:

- Linguagem clara e objetiva
- Estrutura organizada e previsível
- Adequação ao método avaliativo selecionado

Instruções:

- Cada questão deve conter 5 alternativas (a–e)
- Apenas uma alternativa correta
- Não incluir gabarito
- Distribuir as questões entre os tópicos

Saída:

Apenas as questões numeradas.

O prompt apresentado constitui o núcleo do motor gerador de avaliações da abordagem proposta, sendo responsável por orientar o modelo LLM na criação das questões avaliativas definidas pelo docente e às preferências dos estudantes.

4.3 Exemplo de uso

Para tornar mais claro o entendimento da proposta de abordagem, apresenta-se um exemplo ilustrativo de seu uso. Para tal, suponha uma disciplina de Física, ofertada no 1º ano do Ensino Médio, na qual são considerados os seguintes elementos de entrada da abordagem para a geração de uma (AA):

Informações iniciais: Plano de ensino da disciplina, Preferências avaliativas;

Núcleo central: Motor gerador de Avaliação Adaptada;

Saída: Modelo de Avaliação Adaptada;

Elemento orientador: Prática Inclusiva;

Atores envolvidos Professor(a) e Alunos.

Plano de Ensino

- Disciplina: Física
- Série: 1º ano
- Tópicos de conteúdos alvo da avaliação: velocidade média dentro de grandezas e medidas / razão e proporção.

Preferências Avaliativas da Turma

- Organizar as questões da mais fácil para a mais difícil;
- Destacar palavras importantes em maiúsculas;

Critérios Definidos pelo Docente

- Quantidade de questões: duas

Seleção do método avaliativo

- O método selecionado foi o objetivo, definido automaticamente com base na maior incidência de preferências da turma.

Para exemplificar a maior incidência por turma, considere as seguintes preferências informadas pelos estudantes:

- Pedro – Objetiva e Oral
- Liana – Objetiva e Prática
- Italo – Discursiva

Após a contagem das preferências, obteve-se o seguinte resultado:

- Objetiva: 2 indicações
- Oral: 1 indicação
- Prática: 1 indicação
- Discursiva: 1 indicação

A partir desses dados, o motor gera automaticamente uma AA. Para fins comparativos, apresenta-se a seguir uma avaliação tradicional elaborada pelo docente do Liceu Paraibano

(Figura 12) e, em seguida, a avaliação gerada pela abordagem proposta (Figura 13), ambas com o mesmo conteúdo programático.

Figura 12 - Fragmento de uma avaliação tradicional.

Nome Completo: _____

Série/Turma: _____

QUESTÃO 1. **LICEU PARAIBANO**

Você decidiu fazer uma viagem de avião para uma cidade distante. O voo cobre 1200 km em 4 horas. Qual foi a velocidade média do avião?

- a) 200 km/h
- b) 250 km/h
- c) 300 km/h
- d) 350 km/h
- e) 400 km/h

QUESTÃO 2. **LICEU PARAIBANO**

Você decidiu fazer um passeio de bicicleta na cidade. Você percorreu 45 km em 1 hora e 30 minutos. Qual foi sua velocidade média?

- a) 25 km/h
- b) 30 km/h
- c) 35 km/h
- d) 40 km/h
- e) 45 km/h

Figura 13 - Fragmento da AA

- 1) Um carro percorre uma DISTÂNCIA de 150 km em 2 horas. Qual é a VELOCIDADE MÉDIA do carro em km/h?
 - a) 50 km/h
 - b) 60 km/h
 - c) 70 km/h
 - d) 75 km/h
 - e) 80 km/h
- 2) Em uma receita, a proporção de água para açúcar é de 3:2. Se você utilizar 6 litros de água, quantos litros de açúcar serão necessários para manter a mesma PROPORÇÃO?
 - a) 2 litros
 - b) 3 litros
 - c) 4 litros
 - d) 5 litros
 - e) 6 litros

Observa-se que, embora ambas as avaliações sejam estruturadas no método objetivo, há variação significativa quanto ao tipo de questão apresentado. A primeira avaliação, de caráter mais tradicional, demanda a aplicação direta da fórmula de velocidade média em ambas as questões, evidenciando um padrão homogêneo. Já a segunda avaliação contempla diferentes conteúdos, incluindo velocidade média e razão e proporção, além de apresentar elementos adicionais no enunciado, como o uso de palavras-chave em letras maiúsculas, o que pode favorecer a compreensão por determinados perfis de alunos.

Essa variação demonstra que a abordagem proposta é capaz de gerar questões que atendem a diferentes necessidades e perfis de aprendizagem, contribuindo para uma possível adaptação do processo avaliativo.

4.4 Considerações do capítulo

A abordagem MeuAvalia foi concebida como uma ferramenta de apoio ao docente na elaboração de Avaliações Adaptadas (AA), não tendo como finalidade substituir seu papel pedagógico no processo avaliativo. Assim, o professor permanece como agente central e decisório quanto à utilização, escolha, modificação ou descarte das avaliações geradas pelo sistema.

A efetividade da abordagem depende diretamente da qualidade e consistência das informações submetidas, especialmente do plano de ensino, dos tópicos-alvo da avaliação e das preferências avaliativas dos estudantes.

Em se tratando das preferências avaliativas, nesta pesquisa, ultrapassam a ideia de simples escolha de formatos de avaliação, sendo compreendidas como elementos relacionados ao perfil do estudante e ao contexto escolar (Ozbasi, 2019). Quando articuladas aos objetivos de aprendizagem, podem favorecer maior engajamento, motivação e participação discente, além de contribuir para práticas avaliativas mais inovadoras e alinhadas às necessidades da turma (Woods et al., 2024; Lozano Rodríguez et al., 2020).

No âmbito da educação inclusiva, sua consideração possibilita avaliações mais flexíveis, personalizadas e adequadas às especificidades dos estudantes, reforçando o caráter formativo e inclusivo da avaliação (Rosa et al., 2025).

5.Implementação e Experimentos

O presente capítulo descreve a ferramenta que implementa a abordagem MeuAvalia e a avaliação experimental realizada para verificar o alinhamento semântico da geração automática de avaliações de aprendizagem por meio da ferramenta. Para isso, este capítulo está organizado da seguinte forma: a Seção 5.1 introduz a visão geral da ferramenta computacional MeuAvalia; a Seção 5.2 apresenta a implementação da geração das avaliações de aprendizagem na ferramenta; e a Seção 5.3 aborda a avaliação realizada e a Seção 5.4 apresenta o experimento e os resultados obtidos.

5.1 Ferramenta MeuAvalia

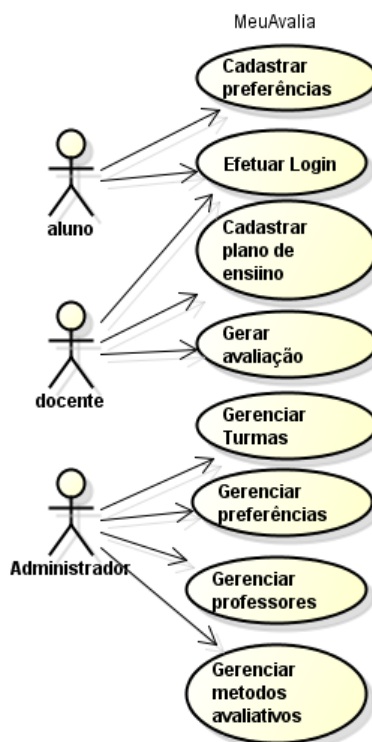
Com o objetivo de tornar todo o processo da abordagem MeuAvalia acessível, compreensível e transparente aos usuários (docentes e discentes), desenvolveu-se uma ferramenta computacional de apoio à geração de modelos de avaliação de aprendizagem.

Seguindo o contexto de desenvolvimento da ferramenta MeuAvalia, os requisitos funcionais foram eliciados por meio de observação direta e da aplicação de questionários iniciais junto a docentes e discentes de duas escolas estaduais participantes do projeto de extensão que originou a ferramenta (Apêndice A).

O objetivo foi compreender as necessidades, dificuldades e percepções dos usuários em relação ao processo avaliativo, com ênfase nas preferências avaliativas e nos desafios enfrentados em contextos inclusivos.

O Apêndice A apresenta os instrumentos de coleta utilizados, sendo composto por: (A.1) questionário aplicado aos docentes e (A.2) questionário aplicado aos discentes. A Figura 14 apresenta o diagrama de casos de uso da ferramenta MeuAvalia, modelado por meio da notação UML. Seu objetivo é representar as funcionalidades disponibilizadas ao ator Professor e sua interação com o sistema.

Figura 14- Diagrama de Casos de Uso do MeuAvalia



Fonte: Elaborado pela autora (2026)

Três atores compõem o cenário de uso da ferramenta. O ator Administrador tem a atribuição de gerenciar e manter a infraestrutura da ferramenta.

Esse perfil é responsável pela administração dos dados e o funcionamento contínuo da ferramenta. Por meio desse perfil, é possível realizar o cadastro, atualização e remoção de alunos e professores, bem como gerenciar a vinculação dos alunos às suas respectivas turmas. Compete ainda ao administrador o gerenciamento dos métodos avaliativos disponíveis na ferramenta e das preferências avaliativas registradas, assegurando a integridade e consistência dos dados armazenados. Além disso, o administrador é responsável pelo gerenciamento das permissões de acesso ao sistema, garantindo a disponibilidade, segurança e funcionamento adequado dos serviços oferecidos pela ferramenta.

O ator Docente representa o núcleo operacional e pedagógico da ferramenta, uma vez que é o usuário responsável por solicitar a geração das AA. Cabe ao docente:

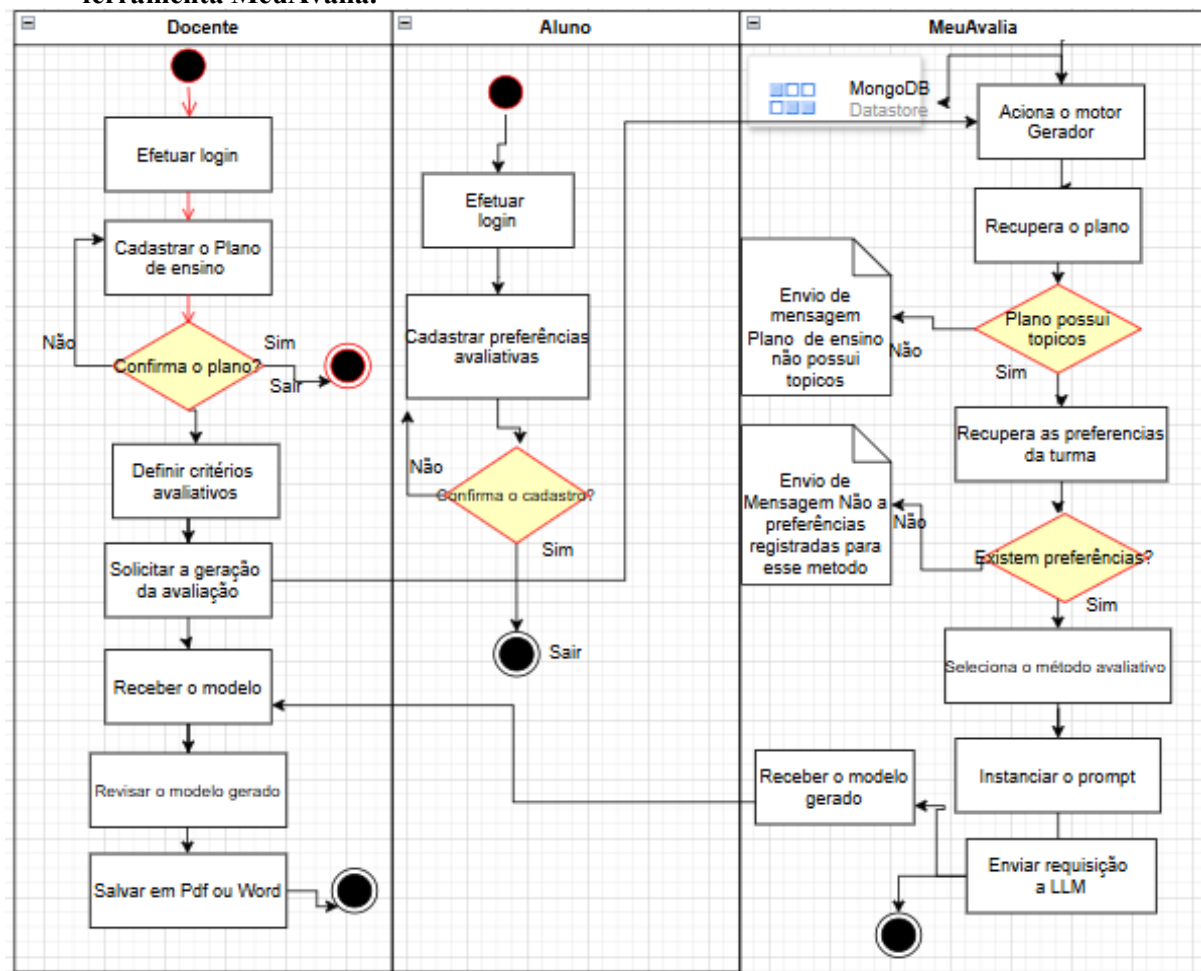
- Cadastrar e editar plano de ensino de disciplina;
- Indicar, entre os conteúdos do plano, aqueles que serão alvo da geração de AA;
- Configurar os critérios da avaliação (quantidade de questões, turma e título);
- Solicitar a geração da AA;

- Visualizar, salvar e revisar as avaliações geradas.

No que se refere ao ator Aluno, este realiza o login na ferramenta, seleciona os métodos avaliativos de sua preferência e define as preferências associadas a esse método, as quais são armazenadas na ferramenta e consideradas no momento de geração da AA vinculadas à turma do aluno.

A Figura 15 apresenta o diagrama de atividades relacionado às ações desempenhadas pelo docente e pelo aluno, bem como às respostas geradas pelo MeuAvalia. O diagrama segue a notação UML e descreve as principais etapas do funcionamento geral do sistema.

Figura 15 – Diagrama de atividades representando o funcionamento geral da ferramenta MeuAvalia.



Fonte: Elaborado pela autora (2026).

O diagrama está organizado em três faixas, correspondentes ao Docente, ao Aluno e à ferramenta MeuAvalia, permitindo identificar as responsabilidades de cada participante durante o processo de geração da avaliação adaptada. Além disso, os losangos representam os pontos de decisão do fluxo, nos quais são verificadas condições necessárias para a continuidade do

processo, como a existência de tópicos de conteúdo no plano de ensino e de preferências avaliativas cadastradas para a turma.

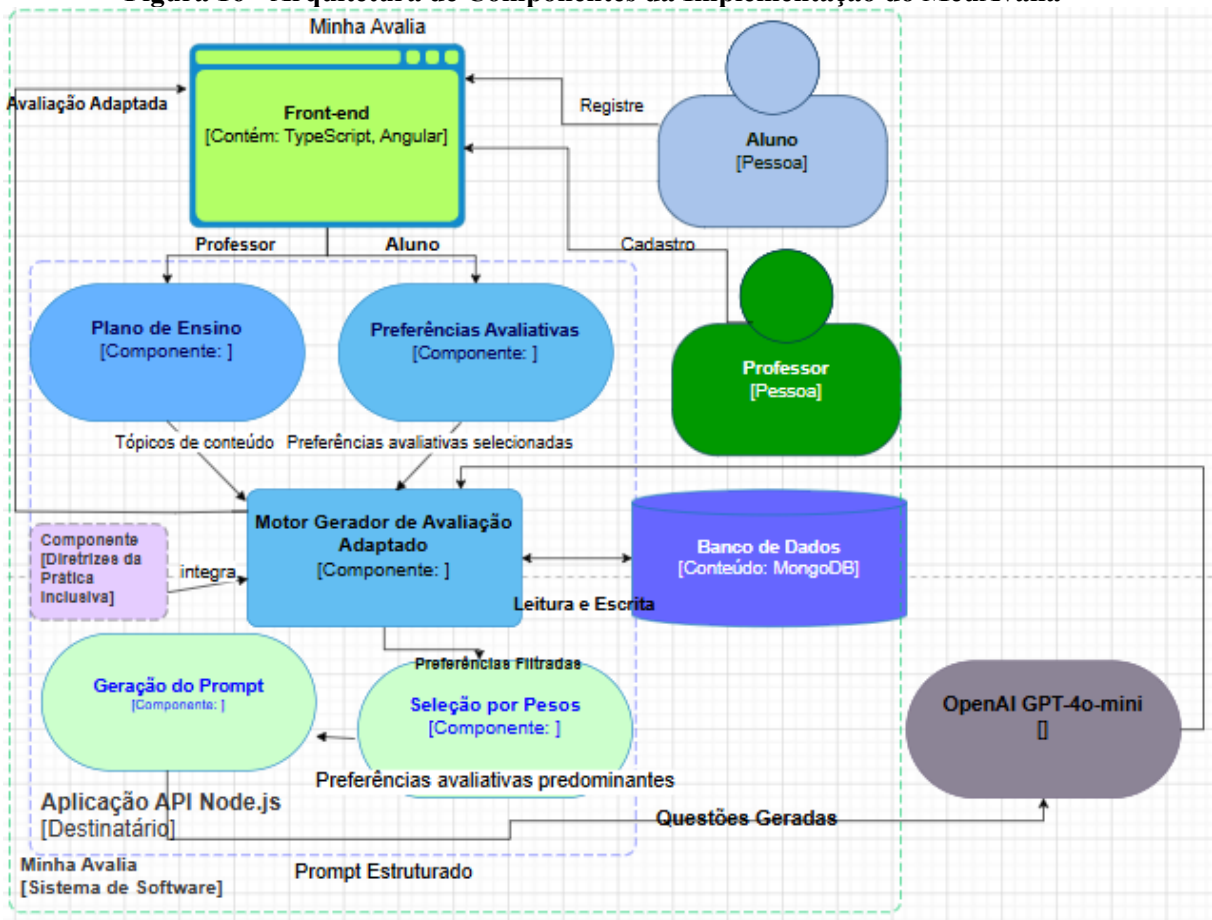
O fluxo inicia-se com a autenticação dos usuários. Na raia do Docente, são realizadas as atividades de cadastro e confirmação do plano de ensino, definição dos critérios avaliativos, como a quantidade de questões, e solicitação da geração da avaliação. Paralelamente, na raia do Aluno, ocorre o cadastro das preferências avaliativas, que são armazenadas e posteriormente utilizadas na geração da avaliação.

Na raia correspondente à ferramenta MeuAvalia, o Motor Gerador de Avaliação Adaptada é acionado após a solicitação do docente. A ferramenta recupera as informações do plano de ensino e as preferências avaliativas da turma, realizando as validações necessárias. Caso sejam identificadas inconsistências, como a ausência de tópicos de conteúdo ou de preferências avaliativas, uma mensagem é apresentada ao usuário e o fluxo retorna à etapa correspondente para correção.

Após a validação das informações, a ferramenta seleciona o método avaliativo com base nas preferências cadastradas, instancia o prompt e encaminha a requisição ao modelo de linguagem para a geração da avaliação adaptada.

Por fim, a avaliação gerada é disponibilizada ao docente, que pode revisá-la e exportá-la nos formatos PDF ou Word, concluindo o processo. Após a apresentação das funcionalidades disponibilizadas aos usuários e do fluxo geral de funcionamento da ferramenta, apresenta-se, na Figura 16, a arquitetura de software da implementação do MeuAvalia.

Figura 16 - Arquitetura de Componentes da Implementação do MeuAvalia



Fonte: Autoria própria (2026).

A Figura 16 apresenta a arquitetura da ferramenta, elaborada utilizando a notação C4 Model (Context, Containers, Components and Code). Para isso, foi empregado o diagrama no nível de Componentes, adaptado para representar os principais componentes da solução, os atores Professor e Aluno, bem como o fluxo de informações entre os módulos da abordagem.

A arquitetura é composta pelo Front-End, pelo Banco de Dados, pelo Motor Gerador de Avaliação Adaptada, pelo componente de Seleção por Pesos, pelo componente de Construção do Prompt e pelo modelo de linguagem OpenAI GPT-4o-mini. O Motor Gerador de Avaliação Adaptada coordena o processo de geração das avaliações, integrando as informações do Plano de Ensino e das Preferências Avaliativas. Inicialmente, o componente de Seleção por Pesos identifica as preferências avaliativas mais adequadas.

Em seguida, essas preferências, juntamente com as diretrizes da prática inclusiva fundamentadas no Ensino Estruturado e as informações do Motor, são utilizadas pelo componente para a Construção do Prompt para gerar um prompt estruturado. Esse prompt é enviado ao modelo OpenAI GPT-4o-mini, responsável pela geração da avaliação adaptada. Por

fim, a avaliação gerada é retornada ao Motor Gerador de Avaliação Adaptada, que a disponibiliza ao Front-End para que o professor possa visualizá-la, revisá-la e utilizá-la.

A Figura 17 apresenta a interface destinada ao aluno, na qual é possível registrar suas preferências avaliativas.

Nessa, o estudante pode indicar os métodos de avaliação com os quais possui maior afinidade, entre as opções de avaliação objetiva, discursiva, prática ou oral. Além da seleção do tipo de avaliação, a interface permite ao aluno especificar características desejáveis no formato das questões, como a organização do nível de dificuldade (do mais fácil ao mais difícil), o uso de linguagem mais clara e direta, a explicitação de termos técnicos, a contextualização por meio de situações-problema e a possibilidade de adaptações no processo avaliativo, como tempo adicional ou realização em dupla. Essas preferências permitem que o processo avaliativo seja ajustado às necessidades dos estudantes, contribuindo para a personalização das avaliações.

Dessa forma, o sistema não apenas identifica o tipo de avaliação preferido, mas também considera elementos que podem influenciar a forma como os enunciados são apresentados e compreendidos pelos alunos.

A Figura 18 apresenta a interface principal utilizada pelo docente para a geração das avaliações. Por meio dela, o professor pode selecionar a turma, definir os critérios da avaliação e solicitar a geração automática do modelo AA.

Figura 17 - Registro das preferências avaliativas na visão do aluno.

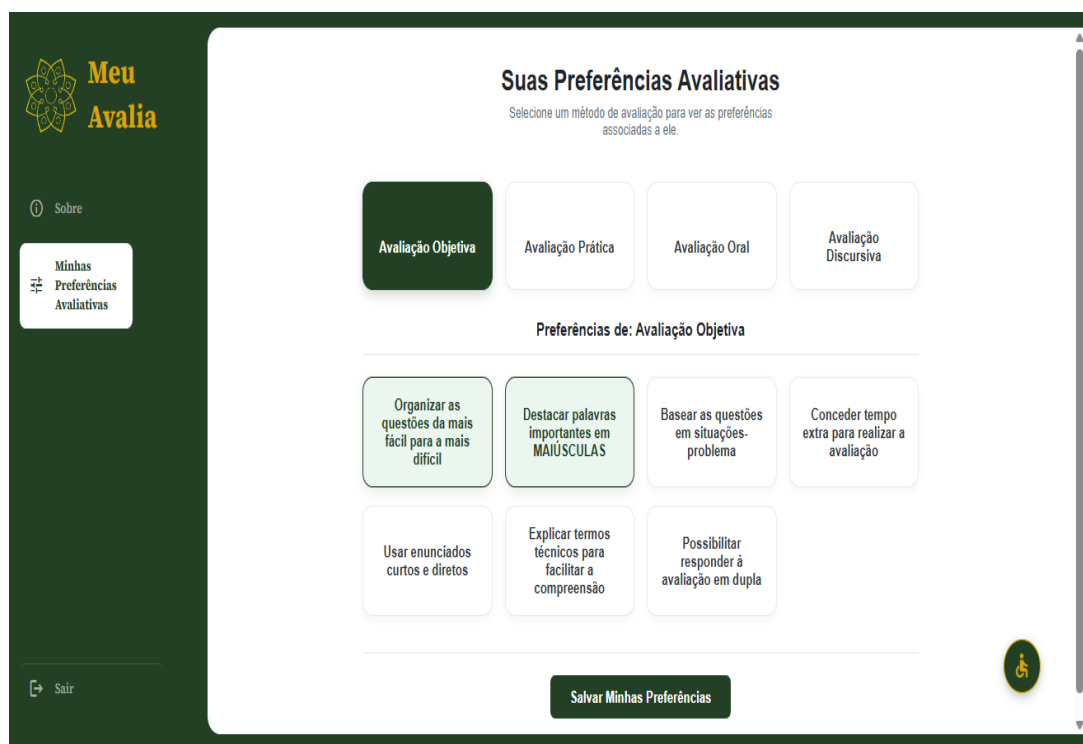


Figura 18 - Geração da avaliação na visão do docente.

Meu Avalia

Sobre

Ver Planos de Ensino

Gerar Avaliação

Sair

Gerar Nova Avaliação

Cancelar Gerar Avaliação

1. Configuração da Avaliação

Turma: 1 ano

Título da Avaliação: 1 Bim

Quantidade de Questões: 04

2. Tópicos (Conteúdo)

Selecione os tópicos do plano de ensino que esta avaliação irá cobrir:

- ☐ Linguagem Sql e comandos
- ☐ USO E MANUSEIO DE MATERIAIS DE LABORATÓRIOS DE FÍSICA
- ☐ SISTEMA INTERNACIONAL DE UNIDADE DE MEDIDAS
- ☐ CINEMÁTICA, A PRÁTICA É SOBRE MOVIMENTO UNIFORME

Métodos Preferidos da Turma

- 1 aluno(s) Preferem: Avaliação Oral
- 1 aluno(s) Preferem: Avaliação Prática

Preferências Específicas

- 1 aluno(s) Preferem: Organizar as questões da mais fácil para a mais difícil

Fonte: MeuAvalia (2026).

5.2 Implementação do Motor gerador de Avaliação Adaptada

Conforme descrita no Capítulo 4, a abordagem MeuAvalia está centrada em um componente principal que é o Motor gerador de Avaliação Adaptada. Para isso, o Motor leva em consideração três elementos: (i) Lista de tópicos de conteúdo selecionados pelo docente para compor a avaliação; (ii) Lista de preferências avaliativas dos alunos da turma e (iii) princípios do Ensino Estruturado como prática inclusiva. O Motor utiliza um modelo de linguagem para gerar questões avaliativas alinhadas aos tópicos do plano de ensino alvo da avaliação e às preferências avaliativas registradas pelos alunos. Isso é realizado a partir da instanciamento de um prompt com as informações necessárias.

5.2.1 Modelo de Linguagem e Estratégia de Prompting

Para a implementação do Motor gerador de Avaliação Adaptada foi utilizado um LLM, especificamente o GPT-4o-mini, disponibilizado pela OpenAI, responsável pela geração automática das questões avaliativas a partir das instruções fornecidas pela ferramenta. Alguns experimentos preliminares foram realizados para avaliar e escolher o LLM mais adequado a essa tarefa. Os experimentos estão descritos no Anexo A.

Foram investigados os seguintes modelos: Llama 3.1 (8B), Claude 3 Haiku, DeepSeek R1T2 Chimera, DeepSeek V3.2, DeepSeek V3.1 Nex N1 e OpenAI GPT-4o Mini.

Para isso, inicialmente, foram realizados testes utilizando diferentes métodos avaliativos, com base na estratégia de prompting zero-shot.

Esses testes permitiram analisar o desempenho dos modelos na geração de diferentes formatos de questões avaliativas, como questões objetivas e discursivas, considerando as instruções fornecidas.

O Apêndice B apresenta o prompt inicial utilizado no modelo Llama 3.1 (8B), empregado no experimento de geração de avaliações objetivas. Para os demais métodos avaliativos (discursivo, prático e oral), foram utilizados prompts específicos, ajustados de acordo com as características de cada tipo de avaliação.

Durante esta etapa inicial, foram realizadas execuções experimentais com diferentes combinações de preferências avaliativas e tópicos de conteúdo, a fim de analisar o comportamento dos modelos na geração de questões. Observou-se que a inclusão de um número elevado de preferências no prompt comprometia a consistência das saídas, levando à definição de um limite de até três preferências por geração, visando maior aderência às instruções.

Para a segunda etapa dos testes, o template de prompt já se encontrava mais estruturado e organizado, sendo então aplicado a diferentes modelos de linguagem mencionados inicialmente, mantendo-se constantes os parâmetros de entrada, com o objetivo de possibilitar uma análise comparativa entre os resultados obtidos.

Os testes consideraram aspectos como coerência textual, aderência aos tópicos e incorporação das preferências avaliativas.

Os resultados indicaram variações no desempenho entre os modelos, sendo o GPT-4o Mini o que apresentou melhor desempenho na organização das questões, diversidade dos conteúdos e alinhamento com o contexto educacional. Com base nesses resultados, esse modelo foi selecionado para a implementação na ferramenta MeuAvalia.

Detalhes adicionais sobre os experimentos e exemplos das questões geradas encontram-se no Apêndice/Anexo.

A Figura 19 apresenta o template final de prompt adotado na abordagem MeuAvalia, concebido como um modelo parametrizável. Esse template permite instanciar diferentes prompts a partir de uma mesma estrutura base, conforme o método avaliativo selecionado (objetivo, discursivo, prático ou oral).

Dessa forma, o sistema não utiliza um prompt único e estático, mas sim um mecanismo dinâmico de geração de prompts, promovendo flexibilidade, reutilização e consistência pedagógica na geração das avaliações.

Figura 19 - Estrutura do template de prompt parametrizado utilizado no MeuAvalia.

‘Você é um(a) professor(a) do ensino técnico de nível médio, com experiência em educação básica. Elabore uma avaliação adaptada considerando o conjunto de preferências avaliativas selecionadas pelos alunos. As preferências avaliativas devem obrigatoriamente ser aplicadas em todas as questões, são elas: {{PREFERENCIAS}}. A avaliação será do tipo {{METODO_AVALIATIVO}} e deve incluir {{NUMERO_QUESTOES}} questões. A avaliação deve contemplar o conteúdo da turma {{TURMA_NOME}} e os tópicos: {{TOPICOS}}. {{INSTRUcoes_FORMATO}}

Instruções de geração:

- Faça todo o planejamento e raciocínio internamente e não os inclua na resposta.
- Cada questão não deve incluir o gabarito.
- Cada questão deve estar vinculada a exatamente um dos tópicos da lista {{TOPICOS}}, garantindo distribuição equilibrada.
- Use linguagem simples, direta e objetiva, adequada ao ensino técnico de nível médio.
- Organize todas as questões em lista numerada.

FORMATO DE SAÍDA (obrigatório):

- Responda APENAS com as QUESTÕES.
- Não inclua título, cabeçalho, explicações ou comentários.
- Use lista numerada:

{{FORMATO_SAIDA}}

Fonte: elaborado pelo autor (2025).

O template de prompt inclui blocos com parâmetros responsáveis por definir instruções específicas de acordo com o método avaliativo selecionado a partir das preferências dos estudantes da turma. Entre esses blocos, destacam-se as variáveis INSTRUcoes_FORMATO e FORMATO_SAIDA, que determinam, respectivamente, as regras de construção das questões e a estrutura esperada da resposta gerada pelo modelo.

Para cada método avaliativo, essas variáveis são instanciadas de forma distinta. No caso da avaliação objetiva, por exemplo, as configurações adotadas são apresentadas na Figura 20.

Figura 20- Instanciiação do prompt para avaliação objetiva.

```
INSTRUcoes_FORMATO:
- Cada questão deve conter exatamente cinco alternativas (a–e), com apenas
uma correta.
FORMATO_SAIDA:
1) <enunciado>
a) ...
b) ...
c) ...
d) ...
e) ...
```

Para os métodos avaliativos que envolvem respostas abertas, como os tipos discursivo, prático e oral, o template de prompt adota uma estrutura comum, baseada na formulação de enunciados sem alternativas, conforme apresentado na Figura 21.

Figura 21 - Instanciiação do prompt para avaliações dos tipos discursiva, prática e oral

```
{{INSTRUcoes_FORMATO}}:
- As questões devem exigir resposta aberta compatível com o tipo
{{METODO_AVALIATIVO}}.

{{FORMATO_SAIDA}}:
1) <enunciado>

2) <enunciado>

... até {{NÚMERO_QUESTOES}}
```

Fonte: Elaborado pelo autor (2025).

5.3 Exemplo de uma avaliação gerada pela ferramenta.

Para demonstrar o uso da ferramenta e do Motor gerador de Avaliação Adaptada, apresenta-se nesta seção um exemplo de geração automática de avaliação. Para tal, considera-se o método avaliativo do tipo objetivo, com a geração de cinco questões de múltipla escolha.

O cenário de entrada foi definido com base em um tópico previamente cadastrado na ferramenta, referente à disciplina de Física do primeiro ano do Ensino Médio: Sistema Internacional de Unidades (SI). Foram consideradas preferências avaliativas dos estudantes, a saber: linguagem simples e direta, enunciados objetivos e organização estruturada das questões.

A partir desses parâmetros, foi gerada automaticamente uma Avaliação Adaptada do tipo objetiva, conforme apresentada na Figura 22. Observa-se que todas as questões estão alinhadas ao tópico selecionado pelo docente, explorando diferentes aspectos desse conteúdo ao longo da avaliação.

Adicionalmente, verifica-se a incorporação das preferências avaliativas, evidenciada pelo uso de enunciados curtos, linguagem clara e estrutura padronizada das alternativas.

Esses aspectos demonstram a aplicação dos princípios da abordagem MeuAvalia, especialmente no que se refere à adaptação das preferências e ao alinhamento com o conteúdo selecionado.

Figura 22 – Exemplo de Avaliação Adaptada

1. O que é QUÍMICA ORGÂNICA?
 - a) Estudo de compostos inorgânicos.
 - b) Estudo de compostos que contêm CARBONO.
 - c) Estudo de reações de ácidos.
 - d) Estudo de gases nobres.
 - e) Estudo de metais pesados.
2. Qual é a ESTRUTURA MOLECULAR do etanol?
 - a) CH_3COOH
 - b) $\text{C}_2\text{H}_5\text{OH}$
 - c) C_3H_8
 - d) C_2H_4
 - e) CH_4
3. Qual FUNÇÃO ORGÂNICA é representada pela fórmula $\text{C}_n\text{H}_{2n+2}$?
 - a) Álcoois
 - b) Ésteres
 - c) Hidrocarbonetos saturados
 - d) Ácidos carboxílicos
 - e) Aminas

Fonte: MeuAvalia (2025).

5.4 Avaliação experimental

A avaliação da abordagem MeuAvalia foi conduzida por meio de um experimento de caráter comparativo e controlado, com o objetivo de analisar as avaliações geradas, considerando a similaridade semântica dos resultados em relação a avaliações de referência (*Ground Truth*).

Para isso, foi adotada uma métrica de similaridade textual, utilizando como base um conjunto de avaliações reais (*Ground Truth*). Esse conjunto de dados foi composto por questões elaboradas por docentes da Educação Básica, das escolas Liceu Paraibano e Horácio de Almeida. Cenários simulados considerando escopos de ensino no contexto das escolas foram também aplicados.

A investigação experimental busca responder à segunda questão de pesquisa definida no Capítulo 1 desta dissertação:

(QP2) Como avaliar o alinhamento semântico das avaliações geradas em relação às avaliações de referência?

A métrica de similaridade textual adotada foi a BERTScore (Seção 2.5.3), utilizada para mensurar a preservação do conteúdo e a fidelidade semântica das questões produzidas automaticamente. A BERTScore é implementada por meio do uso das pontuações máximas de similaridade de cada token para calcular BERTRecall, BERTPrecision e FBERT, ou seja, revocação, precisão e F1, respectivamente.

A comparação foi realizada exclusivamente sobre os enunciados das questões, desconsiderando as alternativas, a fim de evitar interferências estruturais na avaliação. Adicionalmente, foram considerados aspectos como a extensão dos enunciados e a formulação textual deles, uma vez que essas características podem influenciar os valores de similaridade semântica.

Para responder à questão QP2, o alinhamento das avaliações foi mensurado por meio da métrica BERTScore, considerando-se principalmente o valor de FBERT como indicador global de similaridade semântica entre os enunciados das questões gerados e os enunciados de referência.

5.4.1 Ambiente Experimental

O experimento foi realizado utilizando o modelo pré-treinado bert-base-multilingual-cased, disponibilizado pela biblioteca Transformers da plataforma Hugging Face. Esse modelo foi empregado no cálculo da métrica BERTScore, sendo responsável por gerar representações vetoriais dos textos e possibilitar a mensuração da similaridade semântica entre as questões geradas e as questões de referência. A escolha desse modelo justifica-se por sua capacidade multilíngue, permitindo a comparação semântica de avaliações em português, inglês e espanhol (Alsulami et al., 2025; Ariyanta et al., 2025).

5.4.2 Conjunto de dados

O conjunto de dados usado como *ground truth* é composto por 27 avaliações de aprendizagem elaboradas por docentes de diferentes disciplinas do Ensino Médio das escolas Liceu Paraibano e Horácio de Almeida. No total, foram analisadas aproximadamente 173 questões, distribuídas entre os métodos avaliativos objetiva (12), discursivo (5), prático (5) e

oral (5). Ressalta-se que o conjunto de dados não apresenta distribuição uniforme entre as disciplinas. Em termos de distribuição, observa-se o quantitativo seguinte: História (5), Química (4), Física (4), Língua Portuguesa (4), Matemática (4), Sociologia (2), área técnica (2), Geografia (1) e Inglês (1), conforme quantitativo apresentado na Tabela 2.

Tabela 2- Caracterização quantitativa do conjunto de dados de avaliações.

Tipo de avaliação	Quantidade
Objetiva	12
Oral	5
Prática	5
Discursiva	5
Total	27

No total, o conjunto de dados reúne aproximadamente 173 questões, com variação entre 2 e 10 questões por avaliação, evidenciando a diversidade entre as avaliações analisadas.

As avaliações passaram por uma etapa de pré-processamento, que incluiu a remoção de elementos de formatação (como cabeçalhos e quebras de linha) e a padronização da numeração das questões. Em seguida, os textos foram extraídos dos arquivos PDF por meio da biblioteca pdfplumber e organizados em listas contendo os enunciados das questões, utilizados como unidades de análise no processo comparativo.

Para garantir a consistência na comparação par a par entre as questões de referência e as geradas, foi adotada a correspondência por posição dentro das listas de questões. Ou seja, o primeiro índice da questão gerada foi comparado com o primeiro índice da questão de referência.

Além disso, foi realizada a correspondência entre o número de questões nas avaliações analisadas, permitindo a comparação direta entre os enunciados das questões das avaliações de referência e as avaliações geradas.

5.4.2 Resultados e Discussão

A Tabela 3 apresenta um exemplo comparativo entre questões de uma avaliação de História do 3º ano elaborada por um docente e as questões geradas automaticamente pela abordagem MeuAvalia. Observa-se que, embora ambas abordem o mesmo conteúdo temático,

há diferenças na estrutura e no nível de detalhamento dos enunciados, sendo a questão docente mais contextualizada e interpretativa, enquanto a questão gerada apresenta formulação mais direta e objetiva.

Tabela 3 - Comparação entre questões da avaliação original e da avaliação gerada.

Índice	Questão de Referência	Questão Gerada
0	A Revolta da Vacina (1904) mostrou claramente o aspecto defensivo, desorganizado e fragmentado da ação popular. Não se negava o Estado, não se reivindicava participação nas decisões políticas; defendiam-se valores e direitos considerados acima da intervenção do Estado. CARVALHO, J. M. Os bestializados: o Rio de Janeiro Ouro e a República que não foi. São Paulo: Cia. das Letras, 1987 adaptado A mobilização analisada representou um alerta, na medida em que a ação popular questionava:	O que foi a REVOLTA DA VACINA?
1	Rodrigo havia sido indicado pela oposição para fiscal duma das mesas eleitorais. Pôs o revólver na cintura, uma caixa de balas no bolso e encaminhou-se para seu posto. A chamada 3. o texto I dos eleitores começou às sete da manhã. Canudos não se rendeu. Exemplo único em toda Plantados junto da porta, as capangas do Trin- a história, resistiu até o esgotamento completo. òade ofereciam cédulas com o nome dos candi- (Adaptado)	Qual foi uma característica da POLÍTICA OLIGÁRQUICA no Brasil?

Com base na comparação entre as questões de referência e as questões geradas, foram calculados os valores de precisão, revocação e F1-score em nível de questão, cujos resultados completos estão apresentados no Apêndice C. Essa análise permite observar o comportamento da similaridade semântica ao longo das avaliações.

Para mensurar a similaridade semântica entre as questões geradas e as questões de referência, foi utilizada a métrica BERTScore, considerando as medidas de precisão, recall e F1-score (Seção 2.5.3).

A comparação foi realizada de forma pareada, em que cada questão gerada foi comparada à sua respectiva questão de referência. As métricas foram calculadas individualmente para cada par de questões, considerando exclusivamente os enunciados, e posteriormente agregadas por tipo de método avaliativo. Para cada método, foram obtidas estatísticas descritivas, incluindo valores mínimo, primeiro quartil (Q1), mediana, terceiro quartil (Q3), máximo, média e desvio padrão, conforme mostra a Tabela 4.

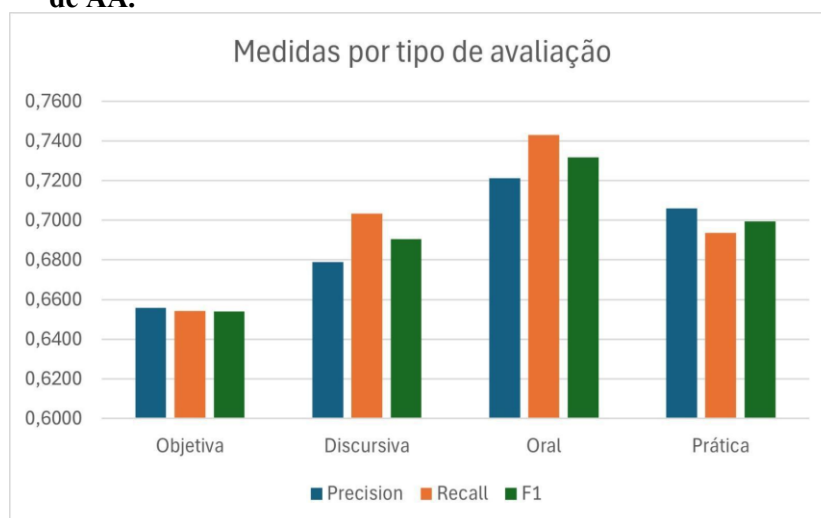
Tabela 4- Estatísticas descritivas conforme tipo de método avaliativo.

Método	Métrica	Min	Q1	Mediana	Q3	Max	Média	Desvio
Discursiva	F1	0.6553	0.6747	0.6943	0.7004	0.7283	0.6905	0.0256
	Precisão	0.6517	0.6725	0.6812	0.6839	0.7047	0.6789	0.0175
	Revocação	0.6589	0.6742	0.6976	0.7151	0.7793	0.7035	0.0432
Objetiva	F1	0.5732	0.6377	0.6540	0.6678	0.7572	0.6540	0.0492
	Precisão	0.5695	0.6125	0.6558	0.6800	0.7766	0.6558	0.0570
	Revocação	0.5681	0.6157	0.6543	0.6856	0.7387	0.6543	0.0553
Oral	F1	0.6893	0.7079	0.7421	0.7559	0.7601	0.7318	0.0306
	Precisão	0.6805	0.6920	0.7288	0.7456	0.7580	0.7212	0.0332
	Revocação	0.6964	0.7263	0.7495	0.7608	0.7789	0.7429	0.0300
Prática	F1	0.6696	0.6756	0.6884	0.7194	0.7495	0.6995	0.0322
	Precisão	0.6840	0.6950	0.7031	0.7169	0.7324	0.7061	0.0178
	Revocação	0.6476	0.6584	0.6800	0.7221	0.7674	0.6937	0.0473

Fonte: autoria própria (2026).

A partir dos dados apresentados, torna-se relevante observar na (Figura 23) as médias de precisão, revocação e F1-score por tipo de avaliação adaptada (AA), permitindo visualizar comparativamente o desempenho dos diferentes métodos avaliativos analisados

Figura 23 - Médias de precisão, revocação e F1-score por tipo de AA.



Fonte: autoria própria (2026).

A Figura 23 apresenta uma visualização comparativa do desempenho dos diferentes métodos avaliativos, evidenciando que as avaliações orais obtiveram os melhores resultados, seguidas pelas práticas, discursivas e, por fim, pelas objetivas.

Conforme observado na Tabela 4, o método oral apresentou os maiores valores médios de precisão, revocação e F1-score, além de menor dispersão dos resultados, indicando maior consistência entre as avaliações analisadas.

Esse comportamento pode ser explicado pela proximidade estrutural entre as questões elaboradas pelos docentes e aquelas geradas pela abordagem, especialmente em avaliações abertas que utilizam verbos de comando semelhantes. Em contraste, o método objetivo apresentou os menores valores médios de F1-score e revocação, sugerindo maior dificuldade na preservação do conteúdo semântico das questões de referência.

Esse resultado pode estar relacionado ao fato de que os enunciados produzidos pelos docentes tendem a ser mais extensos e contextualizados, enquanto as questões geradas pela ferramenta apresentam formulações mais diretas e simplificadas.

As avaliações discursivas e práticas apresentaram desempenho intermediário. No caso das discursivas, observou-se elevada revocação média (0,7035) e baixa variabilidade dos resultados, indicando boa preservação do conteúdo semântico. Já as avaliações práticas apresentaram resultados consistentes, porém com maior dispersão, sugerindo diferenças mais acentuadas entre as avaliações analisadas.

A análise dos quartis reforça esses achados, evidenciando concentração da maior parte dos valores de F1-score no intervalo entre 0,64 e 0,76, a depender do método avaliativo.

Para complementar a análise quantitativa, a Figura 24 apresenta uma Avaliação Adaptada (AA) oral de Português gerada pela abordagem MeuAvalia, na qual se observam verbos de comando como “comente” e “dê”, semanticamente próximos aos utilizados na avaliação original do docente, o que pode contribuir para maior similaridade entre os enunciados comparados.

Figura 24: Exemplo da AA oral de Português gerada pelo MeuAvalia.

- 1.O que caracteriza a ruptura do Modernismo Brasileiro em relação ao passado literário? Cite pelo menos duas características principais.
- 2.Como a inovação se manifesta nas obras de autores modernistas? Dê um exemplo de um autor e uma obra que exemplifique essa inovação.
- 3.Quais foram as principais influências que levaram os escritores modernistas a buscar novas formas de expressão? Comente sobre pelo menos duas dessas influências.

Para fins de análise dos resultados, foram definidos três níveis de desempenho do F1-score: baixa similaridade semântica ($F1 < 0,64$), similaridade média ($0,64 \leq F1 < 0,75$) e alta similaridade ($F1 \geq 0,75$). Essa classificação foi estabelecida com base na distribuição dos

valores observados na Tabela 4. A análise dos quartis indica uma concentração dos resultados entre aproximadamente 0,64 e 0,75, sugerindo que a maior parte das avaliações geradas apresentou similaridade semântica intermediária em relação às avaliações de referência.

Adicionalmente, a análise qualitativa dos exemplos gerados revelou que algumas questões apresentaram enunciados mais concisos quando comparadas às avaliações elaboradas pelos docentes, aspecto que pode ter influenciado os valores de similaridade semântica obtidos.

Como ilustração, a Figura 25 apresenta um exemplo de avaliação objetiva de Português do 3º ano gerada pela abordagem MeuAvalia, na qual se observam enunciados de menor extensão textual e com pouco contexto sobre o Modernismo.

Figura 25 – Exemplo de avaliação objetiva de Português do 3º ano gerada pela abordagem MeuAvalia.

1. O Modernismo brasileiro teve como marco inicial:
 - a) A publicação de Os Sertões
 - b) A Semana de Arte Moderna de 1922
 - c) O Manifesto Pau-Brasil
 - d) A Semana da Literatura Romântica
2. Assinale uma característica do Modernismo:
 - a) Linguagem rebuscada e formal
 - b) Idealização amorosa
 - c) Liberdade estética e experimentalismo
 - d) Valorização do passado clássico
3. Um dos principais autores da primeira fase do Modernismo é:
 - a) Olavo Bilac
 - b) Machado de Assis
 - c) Mário de Andrade
 - d) Castro Alves

Fonte: MeuAvalia(2025).

Essa diferença pode influenciar os valores da similaridade semântica. Esses resultados indicam que, embora existam variações na formulação textual dos enunciados, as avaliações geradas automaticamente pela ferramenta MeuAvalia preservam, em parte, o conteúdo semântico presente nas avaliações elaboradas pelos docentes.

A predominância de valores classificados como similaridade média pode ser explicada pelo fato de que textos mais curtos tendem a apresentar maior sensibilidade a variações lexicais em métricas baseadas em alinhamento de tokens, como o BERTScore, podendo resultar em valores de similaridade inferiores ao esperado mesmo quando os enunciados apresentam significado semanticamente equivalente (Zhang et al., 2020).

Para complementar a análise quantitativa, foi realizada uma análise qualitativa comparativa das questões. A Tabela 5 apresenta exemplos de questões extraídas das avaliações elaboradas pelos docentes e as correspondentes geradas automaticamente pela ferramenta

MeuAvalia. Esses exemplos permitem observar diferenças na formulação dos enunciados, ao mesmo tempo em que evidenciam a preservação do conteúdo semântico das questões.

Tabela 5- Exemplos de questões das avaliações docentes e das avaliações geradas pelo MeuAvalia.

Disciplina	Tópico	Tipo de avaliação	Questão do Docente	Questão do MeuAvalia
História	Revoltas sociais	Objetiva	1.A Revolta da Vacina (1904) mostrou claramente o aspecto defensivo, desorganizado, fragmentado da ação popular. Não se negava o Estado, não se reivindicava participação nas decisões políticas; defendiam-se valores e direitos considerados acima da intervenção do Estado	1.O que foi a REVOLTA DA VACINA?
Física	Movimento Uniforme	Prática	1.Analisar o movimento uniforme e compreender a relação entre espaço, tempo e velocidade. a) A velocidade permaneceu constante? b) O gráfico espaço $\times$ tempo é uma reta? c) Que tipo de movimento foi observado	1.Um carro se desloca em linha reta com uma velocidade constante de 60 km/h. Calcule a distância percorrida pelo carro em 2 horas. Utilize o material de apoio para revisar conceitos sobre movimento uniforme.
Português	Linguagem e Crítica social no modernismo.	Discursiva	1. Explique como a linguagem modernista rompe com os padrões formais da literatura tradicional.	1.Explique como os autores modernistas USARAM a LINGUAGEM para criticar a SOCIEDADE da época. Dê exemplos de obras.

Para complementar a análise quantitativa, foi realizada uma análise qualitativa comparativa das questões elaboradas pelos docentes e das questões geradas pela abordagem MeuAvalia. As amostras foram organizadas considerando a disciplina e o método avaliativo, permitindo observar como a ferramenta preserva os conteúdos abordados e quais modificações são introduzidas na formulação dos enunciados.

No caso de História, o tema abordado refere-se à Revolta da Vacina. A questão elaborada pelo docente apresenta um texto contextualizador extenso, exigindo do estudante

interpretação de informações históricas para responder ao problema proposto. Em contrapartida, a questão gerada pelo MeuAvalia apresenta uma formulação mais direta, centrada na identificação do conceito principal relacionado ao conteúdo.

Observa-se que o tema central foi preservado, mantendo o alinhamento com o conteúdo previsto. Entretanto, a questão gerada reduz o nível de contextualização e interpretação exigido do estudante, resultando em uma abordagem mais objetiva. Essa simplificação pode estar associada à proposta da ferramenta de produzir avaliações adaptadas e acessíveis a diferentes perfis de aprendizagem.

Na disciplina de Física, ambas as questões abordam o conteúdo de Movimento Uniforme. A questão elaborada pelo docente solicita a análise de características do movimento por meio de perguntas interpretativas relacionadas à velocidade, ao gráfico espaço-tempo e ao tipo de movimento observado. Já a questão gerada pelo MeuAvalia apresenta uma situação-problema envolvendo o cálculo da distância percorrida por um veículo com velocidade constante.

Embora as estratégias avaliativas sejam diferentes, ambas as questões exploram o mesmo conteúdo curricular. Enquanto a avaliação docente enfatiza a compreensão conceitual, a questão gerada privilegia a aplicação prática dos conceitos, evidenciando a preservação do conteúdo com alteração da forma de abordagem.

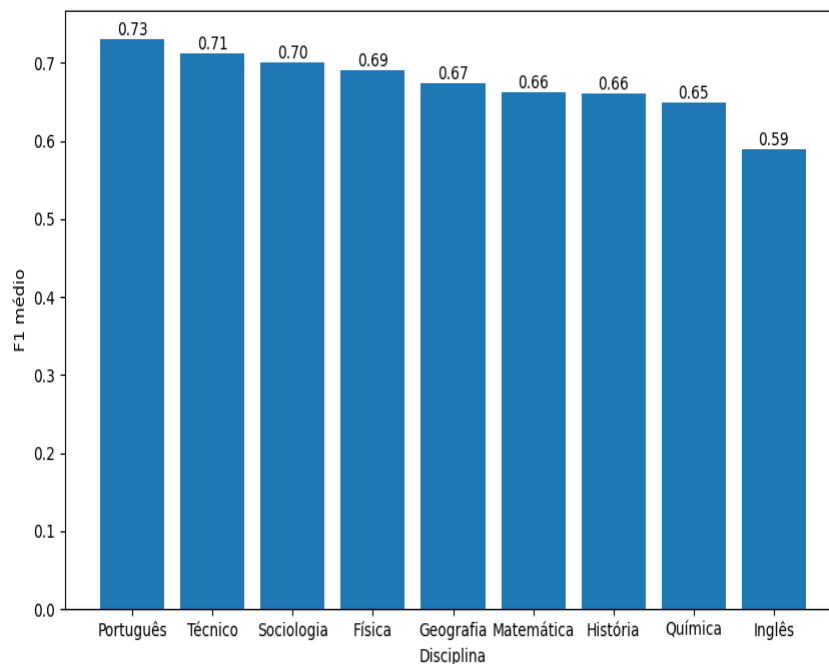
Em português, as questões abordam aspectos relacionados ao Modernismo e ao uso da linguagem como instrumento de crítica social. A questão elaborada pelo docente solicita que o estudante explique como a linguagem modernista rompe com padrões formais da literatura tradicional. A versão gerada pelo MeuAvalia mantém o verbo de comando “explique” e amplia a solicitação ao requerer exemplos de obras modernistas.

Nesse caso, observa-se elevada proximidade entre as duas formulações, uma vez que ambas preservam o tema, o objetivo pedagógico e a natureza discursiva da avaliação. A principal diferença está na ampliação do escopo da resposta esperada, sem comprometer o alinhamento com o conteúdo originalmente proposto.

De forma geral, as amostras analisadas demonstram que o MeuAvalia preserva os temas centrais e os objetivos de aprendizagem presentes nas avaliações elaboradas pelos docentes. As principais diferenças observadas estão relacionadas ao nível de contextualização dos enunciados e à estratégia utilizada para explorar os conteúdos, aspectos que podem influenciar os valores de similaridade semântica identificados na análise quantitativa.

A Figura 26 apresenta a média do F1 por disciplina, permitindo visualizar essas variações de forma comparativa.

Figura 26- Média F1 por Disciplina



Fonte: Autoria Própria (2026).

Destacam-se, por exemplo, as avaliações das disciplinas de Português e Técnico, que apresentaram os maiores valores médios de F1-score, aproximadamente 0,73 e 0,71, respectivamente. Resultados semelhantes foram observados em Sociologia (0,70) e física (0,69), indicando que a abordagem foi capaz de manter níveis consistentes de similaridade semântica em diferentes áreas do conhecimento. De modo geral, a maior parte das disciplinas apresentou valores de F1-score entre 0,65 e 0,73, sugerindo um alinhamento satisfatório entre as avaliações geradas pelo MeuAvalia e as avaliações de referência elaboradas pelos docentes.

Por outro lado, algumas disciplinas apresentaram valores inferiores aos demais, com destaque para inglês (0,59) e, em menor grau, Química (0,65). Uma possível explicação para esse comportamento está relacionada às diferenças estruturais entre os enunciados comparados. Em alguns casos, as avaliações de referência elaboradas pelos docentes continham descrições mais detalhadas, textos mais extensos ou elementos específicos da área de conhecimento, características que podem influenciar os valores de similaridade semântica obtidos.

Apesar dessas variações, os resultados demonstram que a ferramenta tem a possibilidade de manter alinhamento semântico na maior parte dos cenários de disciplinas analisados, mesmo diante de diferenças na forma de apresentação das questões. Isso indica que a geração das avaliações adaptadas (AAs) não compromete o alinhamento com os conteúdos definidos pelos docentes.

Além disso, evidencia-se que a similaridade semântica não depende exclusivamente do alinhamento do conteúdo na geração, mas também das características das questões utilizadas como referência na comparação. Embora a ferramenta produza avaliações mais objetivas e simplificadas, em decorrência da incorporação da prática inclusiva voltada a estudantes com TEA, sua aplicação ainda depende do contexto individual do aluno e de seu nível de aprendizagem. Esse aspecto é relevante pois a ferramenta pode atuar como apoio ao docente na elaboração de avaliações mais adequadas às necessidades dos estudantes (Azevedo et al., 2024; Tanaka et al., 2024).

Nesse contexto, a ferramenta deve ser compreendida como um recurso voltado à avaliação formativa de caráter contínuo, favorecendo o acompanhamento progressivo da aprendizagem, e não como substituta da atuação docente (Furtak, 2009; Medeiros et al., 2024; Mahamuni et al., 2024).

Por fim, destaca-se que a adaptação às preferências avaliativas ocorre em uma etapa anterior ao cálculo da similaridade semântica, sendo utilizada na definição do tipo de avaliação gerada, não interferindo diretamente na etapa de mensuração do alinhamento semântico entre as questões.

5.5 Considerações sobre a QP1

Em relação à **QP1 - Como apoiar o docente na elaboração automatizada de avaliações de aprendizagem adaptativas, considerando as preferências avaliativas dos estudantes**, observa-se que a abordagem, por meio da integração entre o plano de ensino, as preferências avaliativas dos estudantes e princípios de prática inclusiva, estrutura uma abordagem de apoio ao processo de elaboração de avaliações. A ferramenta implementada configura-se como resposta à QP1, indicando que o docente pode ser assistido por ela sem perder sua autonomia nas decisões. Essa integração permite que o docente concentre sua atuação na definição dos conteúdos e objetivos de aprendizagem, enquanto a ferramenta assume a geração das avaliações adaptadas aos diferentes perfis avaliativos. Com isso, reduz-se a necessidade de construção manual completa das avaliações, promovendo um deslocamento do esforço docente para decisões de caráter pedagógico.

Além disso, a abordagem incorpora informações relacionadas às preferências avaliativas dos estudantes, que são utilizadas para a geração das avaliações adaptadas. Isso fornece ao docente uma visão dessas preferências, auxiliando no planejamento de avaliações mais alinhadas ao perfil da turma.

5.6 Considerações sobre a QP2

Em relação à QP2, investigada por meio da análise do alinhamento semântico das avaliações geradas utilizando a métrica BERTScore, os resultados indicaram uma similaridade semântica média ($F1 \approx 0,69$), sugerindo um alinhamento entre as avaliações geradas e as avaliações de referência, ainda que com simplificações nos enunciados. Os resultados experimentais demonstram que as avaliações produzidas pela abordagem mantêm alinhamento semântico moderado com as avaliações elaboradas por docentes, embora existam variações relacionadas ao método avaliativo e à disciplina analisada. Esses resultados sugerem que a abordagem é capaz de preservar os principais conteúdos pedagógicos definidos no plano de ensino durante o processo de geração automática das avaliações.

6. Considerações Finais

Este trabalho apresentou a abordagem MeuAvalia, desenvolvida para apoiar docentes na geração automatizada de avaliações de aprendizagem adaptativas, considerando o plano de ensino, as preferências avaliativas dos estudantes e princípios de práticas inclusivas, com foco no contexto do ensino médio.

Os resultados obtidos indicam que a abordagem é capaz de gerar avaliações alinhadas aos conteúdos pedagógicos definidos pelo docente, incorporando adaptações relacionadas às preferências avaliativas dos estudantes. A avaliação realizada por meio da métrica BERTScore apresentou um F1 médio de aproximadamente 0,69, sugerindo alinhamento semântico moderado entre as avaliações geradas automaticamente e as avaliações elaboradas por docentes. Como principal contribuição, a abordagem propõe um mecanismo de apoio ao processo avaliativo que combina personalização, avaliação formativa e práticas inclusivas em uma única solução computacional. Dessa forma, a ferramenta atua como suporte à tomada de decisão docente, sem substituir sua autonomia pedagógica.

6.1 Trabalhos Futuros

Como trabalhos futuros, pretende-se realizar estudos com docentes em contextos educacionais reais para avaliar a utilidade, a usabilidade e a aceitação da abordagem em situações autênticas de uso. Esses estudos poderão fornecer evidências sobre a aplicabilidade da abordagem no contexto escolar e identificar possíveis oportunidades de aprimoramento.

Além disso, pretende-se investigar o impacto da utilização de diferentes modelos de linguagem na geração das avaliações, incluindo análises comparativas entre Large Language Models (LLMs) e Small Language Models (SLMs). Estudos recentes indicam que modelos

menores podem apresentar desempenho semelhante ao de modelos maiores em tarefas de geração de questões educacionais, ao mesmo tempo em que demandam menor custo computacional e menor infraestrutura para implantação. Tais características tornam os SLMs alternativas promissoras para aplicações educacionais que exigem eficiência operacional, controle institucional dos dados e viabilidade de implementação em ambientes escolares (Fawzi et al., 2024).

6.2 Cuidados éticos

As avaliações de aprendizagem utilizadas nesta pesquisa foram obtidas com autorização dos docentes participantes de duas escolas técnicas estaduais. Os dados coletados foram utilizados exclusivamente para fins acadêmicos e científicos, preservando-se a identidade dos participantes e das instituições envolvidas. As informações referentes às preferências avaliativas dos estudantes foram tratadas de forma agregada e anonimizada, sem identificação pessoal dos participantes, respeitando princípios de confidencialidade e uso ético dos dados.

6.3 Declaração sobre uso de Inteligência Artificial

Neste trabalho, ferramentas de Inteligência Artificial Generativa foram utilizadas em duas finalidades distintas. A primeira refere-se à própria abordagem proposta MeuAvalia, na qual o modelo GPT-4o-mini foi empregado para a geração automatizada das avaliações adaptativas a partir de prompts definidos pelos autores. A segunda corresponde ao apoio à redação do manuscrito, sendo utilizada para correções ortográficas, ajustes gramaticais, sugestões de coesão textual e revisão linguística de trechos do texto.

A utilização da IA ocorreu exclusivamente como ferramenta de apoio computacional e editorial. Todas as decisões relacionadas à concepção da pesquisa, definição da metodologia, análise dos resultados, interpretação dos dados, elaboração das conclusões e revisão final do conteúdo permaneceram sob responsabilidade dos autores. As ferramentas de IA não foram empregadas como autoras do documento e seu uso não exime os autores da responsabilidade integral pelo conteúdo apresentado.

REFERÊNCIAS BIBLIOGRÁFICAS

ADLIT, Marlon F.; ADLIT, Marlene F. Exploring the Efficacy of Differentiated Assessment in UCSP: Perceptions of Senior High School Teachers and Students on Enhancing Learning Outcomes. **Asian Pendidikan**, v. 3, n. 2, p. 29-43, 2023.

AGUIAR-AGUIAR, Giselvis; MAINEGRA-FERNÁNDEZ, Débora; GARCÍA-REYES, Olivia. Enseñar comprensión de textos a escolares con trastornos del espectro autista: Secretos desde la experiencia Teaching Reading Comprehension to School Children With Autism Spectrum Disorders: Secrets From Experience Ensinar a compreensão de textos para crianças em idade escolar com Distúrbios do Espectro do Autismo: segredos da experiência. **Revista Electrónica Educare**, [S. l.], v. 24, n. 2, p. 1–16, 2020.

ALMEIDA, Agnólia Pereira et al. AVALIAÇÃO FORMATIVA: A AVALIAÇÃO FORMATIVA COMO FERRAMENTA DE AVALIAÇÃO DA ESCOLA: USO CONSCIENTE DAS PLATAFORMAS DIGITAIS DE APRENDIZAGEM. **Revista Ilustração**, v. 4, n. 5, p. 127-134, 2023.

ALSULAMI, Mashael M. Evaluating ChatGPT's Semantic Alignment with Community Answers: A Topic-Aware Analysis Using BERTScore and BERTopic. 2025.

AMERICAN PSYCHIATRIC ASSOCIATION. *Manual diagnóstico e estatístico de transtornos mentais: DSM-5* [recurso eletrônico]. Tradução: Maria Inês Corrêa Nascimento et al.; revisão técnica: Aristides Volpato Cordioli et al. 5. ed. Porto Alegre: Artmed, 2014.

ARIYANTA, Nadindra Dwi et al. Assessing the Semantic Alignment in Multilingual Student-Teacher Concept Maps Using mBERT. **MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer**, v. 25, n. 1, p. 113-126, 2025.

AZEVEDO, L. F. A.; DA SILVA, M. C.; MINETTO, V. A.; SILVA, J. P.; ANDRADE, H. R. I.; DA SILVA, M. H. F. R.; LIMA, F. F. R. R.; FREIRES, K. C. P.; MAIA, L. E. de O.; DE SOUZA, M.; PERIN, T. A.; MEDA, M. P. de O.; TEODORO, F. C. A.; DOS SANTOS, R. M. Inclusive education for students with Autism Spectrum Disorder: pedagogical strategies, challenges and perspectives. **CONTRIBUCIONES A LAS CIENCIAS SOCIALES**, [S. l.], v. 17, n. 2, p. e5204, 2024.

AWASTHI, Raghav et al. Human evaluation of large language models in healthcare: gaps, challenges, and the need for standardization. **npj Health Systems**, v. 2, n. 1, p. 40, 2025.

BEZERRA, Maria Auxiliadora. Questões discursivas para avaliação escolar. **Acta Scientiarum. Language and Culture**, v. 30, n. 2, p. 149-157, 2008.

BUTERA, Christiana et al. Impact of sensory processing on school performance outcomes in high functioning individuals with autism spectrum disorder. **Mind, Brain, and Education**, v. 14, n. 3, p. 243-254, 2020.

CASELI, H. M.; NUNES, M. G. V. (org.). *Processamento de linguagem natural: conceitos, técnicas e aplicações em português*. São Carlos: BPLN, 2024.

CARRANZA, Leyda Maria Cedeño et al. Planificación Curricular: Elemento Fundamental en el Proceso Enseñanza-Aprendizaje en la Educación Ecuatoriana. **Ciencia Latina Revista Científica Multidisciplinar**, v. 8, n. 2, p. 6657-6667, 2024.

CASAGRANDE, Juliane Beal. MÉTODOS DE AVALIAÇÃO NO ENSINO MÉDIO: ANÁLISE DAS PRÁTICAS APLICADAS NA EDUCAÇÃO BÁSICA. **Revista Ibero-Americana de Humanidades, Ciências e Educação**, v. 11, n. 9, p. 574-585, 2025. Citado na página 12.

DAVIDOVA, Sayyora. The role of assessment in inclusive education. **News of the NUUz**, v. 1, n. 1.1, p. 68-71, 2024.

DAVID R. KRATHWOHL. A Revision of Bloom's Taxonomy: An Overview. *Elementary Education Online*, [S. l.], v. 8, n. 3, p. 1–7, 2023.

DEEPSEEK. The temperature parameter. 2025. Disponível em: https://api-docs.deepseek.com/quick_start/parameter_settings. Acesso em: 21 maio 2025.

DEVLIN, Jacob et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In: **Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)**. 2019. p. 4171-4186.

ENGEL NOVITA RAMADANI; DINA FITRIA HANDAYANI. Instrumento Penilaian Hasil Pembelajaran Kognitif Pada Tes Objektif . **JURNAL PENDIDIKAN DAN ILMU SOSIAL (JUPENDIS)** , [S. l.] , v. 4, pág. 86–96, 2024.

FALCÃO, Filipe et al. A suggestive approach for assessing item quality, usability and validity of Automatic Item Generation. **Advances in Health Sciences Education**, v. 28, n. 5, p. 1441-1465, 2023.

FATHONI, Arifuddin; SUTIMIN, Leo Agung; SUMMARYANTI, Sri. Optimizing Assessment Techniques for Students with Diverse Special Needs in Schools. In: **Proceeding of International Conference on Special Education in South East Asia Region**. 2024. p. 86-91.

FAWZI, Fares; BALAN, Sarang; CUKUROVA, Mutlu; YILMAZ, Emine; BULATHWELA, Sahan. **Towards Human-like Educational Question Generation with Small Language Models**. In: *International Conference on Artificial Intelligence in Education*, 25., 2024. *Proceedings [...]*. Cham: Springer, 2024. p. 295-303

FENTON, Andrea. Reconsidering the use of oral exams and assessments: An old way to move into a new future. **Educational Researcher**, v. 54, n. 7, p. 430-436, 2025.

FREIA, A. C. B.; GOMUNDANHE, A. M. Avaliação da aprendizagem e inclusão de alunos no espaço escolar em Moçambique: uma análise com base em documentos normativos. **Horizontes**, [S. l.], v. 41, n. 1, p. e023009, 2023.

FREIRE, Paulo. *Pedagogia da Autonomia: Saberes necessários à prática educativa*. São Paulo: Paz e Terra, 1996.

FURTAK, Erin Marie. Avaliação formativa . Corwin: EUA, 2009.

GARCIA, Fabrício Wickey da Silva; OLIVEIRA, Sandro Ronaldo Bezerra. Aplicação de um Plano de Ensino para Disciplina de Algoritmos com Metodologias Ativas: Um Relato de Estudo de Caso Piloto. In: **SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (SBIE)**,

33. , 2022, Manaus. **Anais [...]**. Porto Alegre: Sociedade Brasileira de Computação, 2022 . p. 301-310.

GAO, Mingqi et al. Re-evaluating automatic LLM system ranking for alignment with human preference. In: **Findings of the Association for Computational Linguistics: NAACL 2025**. 2025. p. 4605-4629.

GU, Jindong et al. A systematic survey of prompt engineering on vision-language foundation models. **arXiv preprint arXiv:2307.12980**, 2023.

GHAFAR, Zanyar Nathir et al. Avaliando habilidades em língua inglesa: O papel das avaliações formativas e centradas no aluno na aprendizagem de línguas. *Journal of Language and Pragmatics Studies* , v. 4, n. 2, p. 94-106, 2025.

HADZHIKOLEVA, Stanka et al. Automated test creation using large language models: A practical application. **Applied Sciences**, v. 14, n. 19, p. 9125, 2024.

HANNA, Michael; BOJAR, Ondřej. A fine-grained analysis of BERTScore. In: **Proceedings of the Sixth Conference on Machine Translation**. 2021. p. 507-517.

HOWLEY, Marie; PREECE, David. Structured teaching for individuals with visual impairments. *British Journal of Visual Impairment*, v. 21, n. 2, p. 78-83, 2003.

HOWLEY, Marie et al. Structured Teaching for Pupils with Autistic Spectrum Disorders: Meaningful or Meaningless?. *REACH: Journal of Inclusive Education in Ireland*, v. 19, n. 2, p. 94-101, 2006.

IZIDORO, Paula Elisie Madoglio; LUCCAS, Simone. AVALIAÇÃO EDUCACIONAL: ANÁLISE DA PERCEPÇÃO DOS PROFESSORES EM FORMAÇÃO INICIAL DO CURSO DE LETRAS. **REPPE-Revista de Produtos Educacionais e Pesquisas em Ensino**, v. 8, n. 2, p. 2895-2916, 2024.

JOONBAKHSH, Alireza et al. Evidence-Driven Decision Support for AI Model Selection in Research Software Engineering. **arXiv preprint arXiv:2512.11984**, 2025.

JORGE, Noémia; CARREIRA, Inês. Afinal, o que é uma apresentação oral?. **Palavras**, n. 60-61, p. 33-44, 2023.

KHADHRAOUI, Mayara et al. Survey of BERT-base models for scientific text classification: COVID-19 case study. **Applied Sciences**, v. 12, n. 6, p. 2891, 2022.

KLEINLEIN, Eva. Conceptualising inclusive assessment. A critical literature review at the nexus of education, inclusion, and assessment. **European Journal of Special Needs Education**, p. 1-20, 2025.

KNIPP, Franz; WINIWARTER, Werner. AI-AFACT: Designing AI-Assisted Formative Assessment of Coding Tasks in Web Development Education. In: **Proceedings of the 17th International Conference on Computer Supported Education (CSEDU 2025)**. 2025.

KODRLE, SVETLANA & LOPATINA, NATALIA. (2024). The essence of formative assessment as a didactic concept of higher education. *Pedagogy. Theory & Practice*. 9. 169-176. 10.30853/ped20240022.

LEUNG, Chi Hong. Promoting Optimal Learning with ChatGPT: A Comprehensive Exploration of Prompt Engineering in Education. **Asian Journal of Contemporary Education**, v. 8, n. 2, p. 104-114, 2024.

LIAO, Ray JT; OHTA, Renka; LEE, Kwangmin. The relationship between written discourse features and integrated listening-to-write scores for adolescent English language learners. **Language Testing**, v. 41, n. 1, p. 35-59, 2024.

LIU, Yang; CHU, Chenhui. Do LLMs Align Human Values Regarding Social Biases? Judging and Explaining Social Biases with LLMs. **arXiv preprint arXiv:2509.13869**, 2025.

LIMA, L. S. de .; GAUQUELIN, E. M. G. de O. .; OLIVEIRA, E. J. G. de .; SILVA , R. M. da . Os Instrumentos Avaliativos Utilizados pelos Professores de Jovens e Adultos (EJA) . **RACE - Revista de Administração do Cesmac**, [S. l.], v. 9, p. 140–151, 2021

LOZANO RODRÍGUEZ, Armando; PORTILHO, Evelise Maria Labatut; VIEIRA, Alboni Marisa Dudeque Pianovski. Apresentação. **Revista Diálogo Educacional**, [S. l.], v. 20, n. 64, 2020. DOI: 10.7213/1981-416X.20.064.ED01.

MACDONALD, Elizabeth. The Use of Structured Teaching Strategies to Support Students on the Autism Spectrum to Stay on Task in Mainstream Classrooms. 2017.

MADUREIRA, Nila Luciana Vilhena et al. Práticas pedagógicas para alunos com TEA: estado da arte em dissertações brasileiras dos últimos dez anos. Nova revista amazônica, v. 10, n. 2, p. 23-37, 2022.

MADUREIRA, Jamille Silva; SCHNEIDER, Henrique Nou. DESIGN SCIENCE RESEARCH APLICADA A UMA PESQUISA EM EDUCAÇÃO: UM RELATO DE EXPERIÊNCIA. **Simpósio Internacional de Educação e Comunicação-SIMEDUC**, n. 12, 2025.

MAGOVÁ, Martina; HOCHMUTHOVÁ, Miroslava. Využitie prvkov štruktúrovaného učenia vo vzdelávaní žiakov s poruchou autistického spektra v kombinácii s mentálnym postihnutím The Use of Structured Teaching Elements in the Education of Pupils with Autism Spectrum. **Studia Scientifica Facultatis Paedagogicae**, p. 92, 2022.

MAHAMUNI, Ashwini Jagannath et al. Enhancing educational assessment with artificial intelligence: Challenges and opportunities. In: **2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)**. IEEE, 2024. p. 1-5.

MAITY, Subhankar; DEROY, Aniket. 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.

MARCHISIO, Marina; MARGARIA, Tiziana; SACCHET, Matteo. Automatic formative assessment in computer science: Guidance to model-driven design. In: **2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)**. IEEE, 2020. p. 201-206.

MEDEIROS, Geisiéli Aparecida Carvalho Marin et al. AVALIAÇÃO EDUCACIONAL: NOVAS ABORDAGENS E ESTRATÉGIAS PARA UMA AVALIAÇÃO FORMATIVA. **ARACÊ**, v. 6, n. 2, p. 1522-1538, 2024.

MEISSNER, Roy et al. LLM-generated competence-based e-assessment items for higher education mathematics: methodology and evaluation. In: **Frontiers in Education**. Frontiers Media SA, 2024. p. 1427502.

MIENYE, Ibomoiye Domor et al. Large language models: an overview of foundational architectures, recent trends, and a new taxonomy. **Discover Applied Sciences**, v. 7, n. 9, p. 1027, 2025.

MISSAOUI, Siwar; MAALEL, Ahmed. Student's profile modeling in an adaptive gamified learning environment. **Education and Information Technologies**, v. 26, n. 5, p. 6367-6381, 2021.

MITHI, Joseph et al. Generative Artificial Intelligence and Formative Assessment: Perspectives from Higher Education in South Africa. In: **Proceedings of the International Conference on Education Research**. Academic Conferences and publishing limited (2024).

MOUNDRIDOU, Maria; MATZAKOS, Nikolaos; DOUKAKIS, Spyridon. Generative AI tools as educators' assistants: Designing and implementing inquiry-based lesson plans. **Computers and Education: Artificial Intelligence**, v. 7, p. 100277, 2024.

MONTESORI, Maria. **The montessori method**. Transaction publishers, 2013.

MREMA, K. J. Exploring Formative Assessment Strategies for Pupils with Learning Disabilities in Inclusive Schools in Temeke Municipality, Tanzania. *Asian Journal of Education and Social Studies*, [S. l.], v. 50, n. 11, p. 268–284, 2024.

NEIVA, G. de J. S.; MELO, Ângelo G. de. School evaluation before and during the pandemic: how students see this moment. *Research, Society and Development*, [S. l.], v. 10, n. 16, p. e562101624062, 2021.

NGUYEN, Bang et al. Reference-based metrics disprove themselves in question generation. **arXiv preprint arXiv:2403.12242**, 2024.

OLIVEIRA, Ricardo Gavioli; MOTA, Amôna Almeida; DE SOUSA, Jayne Araújo. Avaliação educacional-uma breve análise das modalidades: diagnóstica, formativa e somativa. *Cadernos da Pedagogia*, v. 16, n. 34, 2022.

OWAN, Valentine Joseph et al. Exploring the potential of artificial intelligence tools in educational measurement and assessment. **Eurasia journal of mathematics, science and technology education**, v. 19, n. 8, p. em2307, 2023.

OZBASI, Durmus. Using rank-order judgments scaling to determine students' evaluation preferences. **Eurasian Journal of Educational Research**, n. 82, p. 63-80, 2019.

PEDASTE, Margus; BAUCAL, Aleksandar; REISENBURK, Elle. Towards a science inquiry test in primary education: development of items and scales. **International Journal of STEM Education**, v. 8, n. 1, p. 19, 2021.

PEFFERS, K. et al. "Design Science Research Methodology for Information Systems Research". *Journal of Management Information Systems*, vol. 24, n. 3, 2007

PEÑA, Driannet; FERNÁNDEZ VÉLEZ, Yumy; ACOSTA CARRAZANA, Sacher. Toward achieving authentic assessment of learning in the practical physiology classroom. **Salud, Ciencia y Tecnología**, [S. l.], v. 4, p. .608, 2024.

PERONDI, L.; CATOSSO , M. C. E.; BARBA, C. H. de; BRITO, L. F. de; SOUSA, A. A. de; SOARES, H. dos S. Transformando a avaliação escolar: do exame classificatório à aprendizagem. **OBSERVATÓRIO DE LA ECONOMÍA LATINOAMERICANA**, [S. l.], v. 23, n. 4, p. e9439, 2025.

PROMPIENGCHAI, Sapolnach; NARREDDY, Charith; JOORDENS, Steve. A Practical Guide for Supporting Formative Assessment and Feedback Using Generative AI. **arXiv preprint arXiv:2505.23405**, 2025.

PROMPT ENGINEERING GUIDE. 2022. Disponível em: <https://www.promptingguide.ai/pt>. Acesso em: 15 jul. 2025.

RANI, Geeta; SINGH, Jagjit; KHANNA, Abhishek. Comparative analysis of generative AI models. In: **2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)**. IEEE, 2023. p. 760-765.

RASMUSSEN, Fernanda de Souza Machado; SILVA, Rosemeire da Costa; NEIX, Carine da Silva Vieira. O ensino e a atividade estruturada para a aprendizagem de pessoas com transtorno do espectro autista. **Construção psicopedagógica**, v. 30, n. 31, p. 101-112, 2021.

RENZE, Matthew. The effect of sampling temperature on problem solving in large language models. In: **Findings of the association for computational linguistics: EMNLP 2024**. 2024. p. 7346-7356.

RODRIGUEZ-DONAIRE, Silvia. Influence of Prompts Structure on the Perception and Enhancement of Learning through LLMs in Online Educational Contexts. **Massive Open Online Courses-Learning Frontiers and Novel Innovations, 2024**, 2024.

ROSA, Rafael Gomes; BARCELLOS, Nathan Santos. Avaliação da aprendizagem de língua adicional de estudantes com transtorno do espectro autista e deficiência intelectual. **Trabalhos em Linguística Aplicada**, v. 64, p. e025035, 2025.

ROSE, David H.; MEYER, Anne. **Teaching every student in the digital age: Universal design for learning**. Association for Supervision and Curriculum Development, 1703 N. Beauregard St., Alexandria, VA 22311-1714 (Product no. 101042: \$22.95 ASCD members; \$26.95 nonmembers), 2002.

SANTOS GALVÃO, Nadielli Maria; MADUREIRA, Jamille Silva; SCHNEIDER, Henrique Nou. Design Science Research para o desenvolvimento de artefatos educacionais. **Boletim de Conjuntura (BOCA)**, v. 18, n. 54, p. 194-220, 2024.

SILVA, Adinairde Neves. Prática pedagógica: Desafios de Transformar a Teoria na Práxis Inclusiva. **HUMANIDADES E TECNOLOGIA (FINOM)**, v. 40, n. 1, p. 398-410, 2023.

SILVA, Alécio Charles. Correção automática de questões discursivas de resposta curta: uma abordagem baseada em extração de informação. 2023.

SILVA, Carolina Gouvêia. 40. Inclusão Escolar de Alunos com TEA–Transtorno do Espectro Autista: O Ensino Estruturado Como Possibilidade. p.299,2022

SILVA, Júlio César; FERREIRA, Hiran Nonato M. Adaptação de Questões Avaliativas para Estudantes Neurodivergentes: Uma Arquitetura Computacional Fundamentada em Inteligência Artificial Generativa. In: **Workshop de Informática na Educação Inclusiva (WIEI)**. SBC, 2024. p. 36-43.

SILVA, Teresinha Letícia da; VIDOTTO, Kajiana Nuernberg Sartor; TAROUCO, Liane Margarida Rockenbach. Investigando as percepções de estudantes e professores do ensino médio e técnico sobre o uso do ChatGPT em suas atividades escolares. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (SBIE), 35. , 2024, Rio de Janeiro/RJ. **Anais [...]**. Porto Alegre: Sociedade Brasileira de Computação, 2024 . p. 1851-1864

SOUZA, F. C. G. de. A IMPORTÂNCIA DA DIVERSIDADE DOS INSTRUMENTOS AVALIATIVOS. *Revista Científica FESA*, [S. l.], v. 1, n. 3, p. 36–46, 2021

SUN, Tianxiang et al. BERTScore is unfair: On social bias in language model-based metrics for text generation. In: **Proceedings of the 2022 conference on empirical methods in natural language processing**. 2022. p. 3726-3739.

SUPRIHATIN, Suprihatin; TARJIAH, Indina. Evaluating the outcome of structured teaching intervention for children with autism. In: **5th International Conference on Education and Technology (ICET 2019)**. Atlantis Press, 2019. p. 286-289.

SCHOPLER, Eric; MESIBOV, Gary B.; HEARSEY, Kathy. Structured teaching in the TEACCH system. In: **Learning and cognition in autism**. Boston, MA: Springer US, 1995. p. 243-268.

TAI, Joanna Hong-Meng et al. Designing assessment for inclusion: An exploration of diverse students' assessment experiences. **Assessment & Evaluation in Higher Education**, v. 48, n. 3, p. 403-417, 2023.

TANAKA, Allan Hideick; TANAKA, Gabrielly C. Espinosa Belinello; BELINELLO, Rogéria Cristina Espinosa. Educação Inclusiva e Comportamento no Transtorno do Espectro Autista: Desafios e Estratégias de Intervenção para a Promoção do Desenvolvimento Escolar. **RCMOS-Revista Científica Multidisciplinar O Saber**, v. 1, n. 2, 2024.

TOMASINE, Joseph S. The multimodal organization of feedback and its documentation: multiactivity during formal, formative reading assessment. **Frontiers in Communication**, v. 9, p. 1390843, 2024.

VAN DE WATERING, Gerard et al. Students' assessment preferences, perceptions of assessment and their relationships to study results. **Higher Education**, v. 56, n. 6, p. 645-658, 2008.

WANG, Jiayi; XIAO, Ruiwei; TSENG, Ying-Jui. Generating ai literacy mcqs: A multi-agent llm approach. In: **Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 2**. 2025. p. 1651-1652.

WANG, Zhiqiang; PANG, Yiran; LIN, Yanbin. Large language models are zero-shot text classifiers. **arXiv preprint arXiv:2312.01044**, 2023.

WEI, Hui et al. Systematic evaluation of llm-as-a-judge in llm alignment tasks: Explainable metrics and diverse prompt templates. **arXiv preprint arXiv:2408.13006**, 2024.

WEBER, Viktoriya. Flexible assessment as an authentic learning strategy: addressing diversity and enhancing student engagement in higher education. **Discover Education**, 2025.

WHITE, J. A prompt pattern catalog to enhance prompt engineering with ChatGPT. **arXiv preprint arXiv:2302.11382**, 2023.

WHITE, Jia et al. ‘We think differently, we learn differently, but at the end of the day we’re not that different’: Strengths-based approaches in high school from the perspectives of autistic adolescents. **Autism**, p. 13623613251348557, 2025.

WOODS, Kevin et al. Student views on the assessment medium for General Certificates of Secondary Education in England: insights from the 2020 examination cancellations. **Assessment in Education: Principles, Policy & Practice**, v. 31, n. 2, p. 135-153, 2024.

WU, Zhanglin et al. Kg-bertscore: Incorporating knowledge graph into bertscore for reference-free machine translation evaluation. In: **Proceedings of the 11th International Joint Conference on Knowledge Graphs**. 2022. p. 121-125.

YANG, Lan; PANG, Feifan; SIN, Kuen-Fung. Assessing the psychometric properties of the practice and product inventory of supporting students with ASD (PPI-SSA): a concise assessment tool for teachers in inclusive classrooms. **Sustainability**, v. 15, n. 19, p. 14576, 2023. Citado na página 12.

YARLAGADDA (2025) IA na Educação: Aprendizagem Personalizada e Sistemas Tutores Inteligentes, **European Journal of Computer Science and Information Technology** , 13 (32), 15-27

ZAGER, Dianne Berkell; WEHMEYER, Michael L.; SIMPSON, Richard L. (Ed.). **Educating students with autism spectrum disorders: Research-based principles and practices**. Routledge, 2012.

ZHANG, Tianyi et al. Bertscore: Evaluating text generation with contextualized embeddings. **Proceedings of ICLR**, 2020.

ZHAO, Chunyi. Avaliação assistida por IA no ensino superior: uma revisão sistemática. **Journal of Educational Technology and Innovation** , v. 6, n. 4, 2024.

APÊNDICES

APÊNDICE A – Questionários utilizados na elicitação de requisitos

A.1 Questionário aplicado aos docentes

Este questionário tem como objetivo coletar a percepção de docentes das Escolas Liceu Paraibano e Escola Escritor Horácio de Almeida sobre a aplicação de avaliações adaptadas na educação básica, considerando três elementos centrais: as preferências avaliativas dos estudantes, os tópicos (conteúdos) definidos nos planos de ensino das disciplinas para uma dada avaliação e uma prática inclusiva voltada a apoiar alunos, especialmente aqueles com Transtorno do Espectro Autista (TEA).

As informações obtidas contribuirão para aprimorar a abordagem e a implementação da ferramenta associada chamada de “MeuAvalia”. Agradecemos sua colaboração!

1) Na sua prática docente, com que frequência você considera as preferências avaliativas dos alunos ao planejar e aplicar as avaliações? (1 - Nunca são consideradas

2- Raramente são consideradas

3- Às vezes são consideradas

4- Frequentemente são consideradas

5- Sempre são consideradas)

1	2	3	4	5
☐	☐	☐	☐	☐

2) Em sua experiência docente, até que ponto uma prática inclusiva voltada para estudantes com Transtorno do Espectro Autista (TEA) pode contribuir para o processo de ensino, quando incorporada às avaliações de aprendizagem realizadas em sala de aula? (1 – Nenhuma contribuição

2 – Pouca contribuição

3 – Contribuição moderada

4 – Contribuição significativa

5 – Contribuição muito significativa)

1	2	3	4	5
☐	☐	☐	☐	☐

3) Uma prática inclusiva adotada em uma avaliação pode contribuir para aumentar a motivação e o engajamento dos alunos com Transtorno do Espectro Autista (TEA) e também dos estudantes sem o transtorno? (1 – Discordo totalmente *
2 – Discordo parcialmente
3 – Nem concordo nem discordo
4 – Concordo parcialmente
5 – Concordo totalmente)

1	2	3	4	5
☐	☐	☐	☐	☐

4) Você, como docente, sente que possui autonomia para adotar práticas inclusivas no processo de avaliação em sua disciplina? (1 – Discordo totalmente *
2 – Discordo parcialmente
3 – Nem concordo nem discordo
4 – Concordo parcialmente
5 – Concordo totalmente)

1	2	3	4	5
☐	☐	☐	☐	☐

5) Considera que oferecer aos alunos a possibilidade de escolher entre diferentes métodos avaliativos (oral, discursivo, objetivo e prático) contribui para tornar o processo avaliativo mais justo e inclusivo? (1 – Discordo totalmente
2 – Discordo parcialmente
3 – Nem concordo nem discordo
4 – Concordo parcialmente
5 – Concordo totalmente) *

1	2	3	4	5
☐	☐	☐	☐	☐

6) Em sua percepção, quais métodos avaliativos os alunos em sua disciplina demonstram maior conforto ou preferências ao serem avaliados? *

Texto de resposta longa
.....

7) Quais desafios podem surgir na aplicação da ferramenta educacional **MeuAvalia**, no contexto escolar? *

Texto de resposta longa
.....

8) Como a integração entre as preferências avaliativas dos alunos, o conteúdo da disciplina e a prática inclusiva propostas pelo **MeuAvalia** pode contribuir para o aprimoramento do processo de ensino e aprendizagem nas avaliações ? *

A.2 Questionário aplicado aos discentes

Queremos conhecer sua opinião sobre as preferências avaliativas e o que torna cada tipo de avaliação (discursiva, oral, prática ou objetiva) mais claro, justo e acessível para você. Por favor, responda com sinceridade!

1. Avaliação Discursiva - Como você prefere que as avaliações discursivas sejam aplicadas? *
(Marque todas que vocês gostariam ou apenas que seja de sua preferência principal)

- ☐ Quero que as questões usem linguagem simples, clara e direta
- ☐ Criar questões em que mais de uma resposta é possível e que permitam criatividade.
- ☐ Verbos de comando claros no início das questões (ex: Descreva , Compare)
- ☐ Permita questões com respostas curtas .
- ☐ Palavras-chave destacadas em Maiúsculas

Você tem alguma sugestão de preferência para avaliação discursiva? Se sim, indique algumas

Texto de resposta longa
.....

2. Avaliação Oral Como você prefere que as avaliações orais sejam aplicadas? (Marque *
todas que vocês gostariam ou apenas que seja de sua preferência principal)

- ☐ Saber antes se a resposta deve ser curta ou longa
- ☐ Indicar claramente que a resposta deve ser oral.
- ☐ Possibilidade de apresentar vídeo ou podcast em vez de falar ao vivo

3. Avaliação Prática Como você prefere que as avaliações práticas sejam aplicadas? (*
Marque todas que vocês gostariam ou apenas que seja de sua preferência principal)

- ☐ Realizar a avaliação por um exemplo fornecido pelo professor .
- ☐ Utilizar telas, aplicativos ou ferramentas gráficas para responder
- ☐ Simular situações reais do ambiente escolar ou de outro contexto.
- ☐ Fornecer um roteiro ou guia para criar um vídeo explicativo sobre a questão.
- ☐ Trabalhar em dupla

Você tem alguma sugestão de preferência para avaliação prática? Se sim, indique algumas

Texto de resposta longa

4. Avaliação Objetiva - Como você prefere que as avaliações objetivas sejam aplicadas? (*
Marque todas que vocês gostariam ou apenas que seja de sua preferência principal)

- ☐ Basear as questões em situações problema.
- ☐ Organizar as questões da mais fácil para a mais difícil.
- ☐ Palavras-chave em MAIÚSCULAS para destacar conceitos importantes.
- ☐ As questões devem usar uma linguagem simples e fácil de entender.
- ☐ Antes de cada questão, escrever um exemplo simples relacionado ao tópico.

APÊNDICE B – Experimentos Para Escolha Do LLM

B.1 Estratégia inicial de prompting

Com o objetivo de analisar o comportamento do modelo de linguagem na geração de avaliações, foram realizados testes iniciais utilizando o prompting zero-shot. Nessa abordagem, o modelo recebe apenas instruções diretas no prompt, sem exemplos prévios, sendo responsável por gerar as questões avaliativas com base exclusivamente nas informações fornecidas.

Além disso, o prompt elaborado incorpora a técnica de role prompting, ao definir explicitamente o papel do modelo como um(a) professor(a) do ensino técnico de nível médio. Essa estratégia contribui para orientar o estilo da resposta, o nível de complexidade das questões e a adequação da linguagem ao contexto educacional pretendido. A Figura 19 apresenta o prompt inicial utilizado nos experimentos, contendo as instruções definidas para a geração de avaliações no formato objetivo.

Prompt inicial da geração da avaliação Objetiva

```
ollama run llama3.1:8b "Você é um(a) professor(a) do ensino técnico de nível médio, com experiência em educação básica. Elabore uma avaliação de aprendizagem considerando o conjunto de preferências avaliativas selecionadas pelos alunos. As preferências avaliativas devem obrigatoriamente ser aplicadas em todas as questões; são elas :Destacar palavras importantes em maiúsculas ;Possibilitar responder à avaliação em dupla;Conceder tempo extra para fazer a avaliação. A avaliação de aprendizagem será do tipo discursiva e deve incluir o número de questões 5. A avaliação de aprendizagem deve contemplar o conteúdo da disciplina Linguagem de Programação I e a lista de tópicos Introdução à Lógica de Programação, Variáveis e Tipos de Dados, Estruturas Condicionais (if/else) de acordo com a série 1ª série. Instruções de geração: - Faça todo o planejamento e raciocínio internamente e não os inclua na resposta. Cada questão não deve incluir o gabarito. - Cada questão deve estar vinculada a exatamente um dos tópicos da lista Introdução à Lógica de Programação; Variáveis e Tipos de Dados , Estruturas Condicionais (if/else) de forma que todos os tópicos sejam contemplados e a distribuição da questão entre eles seja a mais equilibrada possível. - Use linguagem simples ,direta e objetiva, adequada ao ensino técnico de nível médio.- Organize todas as questões em lista numerada"
```

B.2 Configuração dos experimentos

Os experimentos foram conduzidos utilizando o modelo Llama 3.1 (8B), executado localmente por meio da plataforma Ollama. As entradas utilizadas consideraram:

- Três tópicos de conteúdo da disciplina de Programação, referentes ao primeiro ano do Ensino Médio;

- Diferentes combinações de preferências avaliativas dos estudantes;
- Variação na quantidade de preferências incluídas no prompt.

Cada método avaliativo foi executado múltiplas vezes, com o objetivo de observar o comportamento do modelo diante de diferentes configurações de entrada.

APÊNDICE C – Resultados da similaridade semântica por algumas questões

Este apêndice apresenta os resultados detalhados da similaridade semântica calculada por meio do BERTScore, considerando as métricas de precisão (precision), revocação (recall) e F1 em nível de questão. Os resultados estão organizados por disciplina e tipo de avaliação, permitindo uma observação mais detalhada no comportamento da similaridade semântica entre as questões das avaliações elaboradas pelos docentes e aquelas geradas automaticamente pela ferramenta MeuAvalia.

C.1 Avaliação de História – Oral (3º ano)

Análise dos resultados por questão – Avaliação de História (3º ano, modalidade oral)

Questão	Precisão	Revocação	F1
Q1	0.6803	0.6797	0.6800
Q2	0.6758	0.6873	0.6815
Q3	0.6701	0.6833	0.6767
Q4	0.6880	0.6682	0.6780

Médias:

Precisão: 0.6807

Revocação: 0.7196

F1: 0,6992

C.2 Avaliação de Química – Discursiva

Resultados por questão

Questão	Precisão	Revocação	F1
Q1	0.6219	0.6693	0.6447
Q2	0.6091	0.6600	0.6335
Q3	0.5583	0.6702	0.6091
Q4	0.6711	0.7019	0.6862
Q5	0.6185	0.6412	0.6297
Q6	0.5803	0.5829	0.5816
Q7	0.6332	0.6821	0.6567
Q8	0.5994	0.5986	0.5990
Q9	0.6389	0.6321	0.6355
Q10	0.6371	0.6434	0.6402

Média

Precisão: 0,6168

Revocação:0,6482

F1:0,6316

A comparação entre essas duas avaliações, considerando as questões de cada uma, mostra que, na avaliação de História, a precisão e a revocação apresentam valores bem próximos. No entanto, das quatro questões analisadas, observa-se que em duas (Q2 e Q4) a revocação apresentou valores mais baixos, indicando menor cobertura do conteúdo das questões elaboradas pelos docentes.

Esse resultado pode estar relacionado ao fato de se tratar de uma avaliação oral, na qual há menor presença do conceito ou da ideia que ajudem a embasar melhor o conteúdo da

questão. Além disso, o enunciado mais curto das questões geradas pode não contemplar totalmente o conteúdo presente nas questões de referência.

Já na avaliação de Química discursiva, das dez questões analisadas, seis apresentaram valores de revocação superiores aos de precisão. Isso indica que, embora as questões geradas consigam cobrir parcialmente o conteúdo das questões dos docentes, há certa variação na precisão do alinhamento semântico entre elas. Esse comportamento pode estar relacionado à natureza da disciplina, uma vez que as avaliações de Química elaboradas pelos docentes tendem a apresentar mais fórmulas e maior detalhamento técnico, enquanto as questões geradas pelo MeuAvalia são, em geral, mais teóricas, curtas e com menor presença de fórmulas, o que pode impactar diretamente os valores de similaridade.

ANEXO A -Avaliação gerada pelos Modelos

As avaliações apresentadas neste anexo foram geradas a partir do mesmo cenário experimental, mantendo constantes o plano de ensino que são os tópicos de conteúdo, as preferências avaliativas e os parâmetros de geração.

A.1 – Avaliação gerada pelo modelo DeepSeek R1T2 Chimera

Avaliação de Aprendizagem – Linguagem de Programação I

1ª Série do Ensino Técnico de Nível Médio

1. **João precisa armazenar sua idade (17 anos) em um programa. Qual tipo de dado é mais adequado para essa situação?**
 - a) float
 - b) string
 - c) int
 - d) boolean
 - e) char
2. **Em um fluxograma, qual símbolo é utilizado para representar o início e o fim de um algoritmo?**
 - a) Retângulo
 - b) Losango
 - c) Oval
 - d) Paralelogramo
 - e) Triângulo
3. **Maria deseja calcular a temperatura média diária de uma cidade, que pode ter valores como 22.5°C. Qual tipo de dado ela deve usar para armazenar esse número?**
 - a) boolean
 - b) int
 - c) char
 - d) string
 - e) float
4. **Um programa verifica se um aluno foi aprovado com a condição: se a nota final for maior ou igual a 6, exibe "Aprovado". Se o aluno tirar nota 7, o que será exibido?**
 - a) Nada
 - b) "Reprovado"
 - c) "Aprovado"
 - d) "Nota inválida"
 - e) 7
5. **Pedro precisa escrever um código que verifique se um número é par. Qual condicional utiliza o operador correto para essa verificação?**
 - a) `if (numero / 2 == 0)`

b) `if (numero * 2 == 0)`

A.2 – Avaliação gerada pelo modelo DeepSeek V3.1 Nex N1

AValiação DE APRENDIZAGEM – LINGUAGEM DE PROGRAMAÇÃO I

1ª SÉRIE – ENSINO TÉCNICO DE NÍVEL MÉDIO

Instruções

- Leia cada questão com atenção.
 - Escolha a **única alternativa correta** para cada questão.
 - Marque sua resposta na folha de respostas.
-

1. **João precisa escrever um algoritmo para calcular a média de duas notas de um aluno. Qual das opções abaixo representa o conjunto correto de passos para resolver este problema?**
 - a) Digitar as notas no computador, esperar o resultado, imprimir na tela.
 - b) Pedir ao usuário que insira as duas notas, somar as duas notas, dividir o resultado por 2, mostrar o resultado na tela.
 - c) Criar uma variável, mostrar uma mensagem, fechar o programa.
 - d) Somar os números, criar uma tela bonita, salvar no computador.
 - e) Escrever o código, compilar, executar sem testar.
2. **Maria está desenvolvendo um programa para armazenar a idade de uma pessoa. Qual tipo de dado ela deve usar para declarar a variável que guardará essa informação?**
 - a) texto (string)
 - b) número inteiro (int)
 - c) número decimal (float)
 - d) verdadeiro/falso (boolean)
 - e) caractere (char)
3. **Carlos criou uma variável chamada “salario” e atribuiu o valor 2500.50. Que tipo de dado está sendo armazenado nessa variável?**
 - a) int

- b) string
 - c) boolean
 - d) float
 - e) char
4. Um programa precisa verificar se um aluno foi aprovado ou reprovado. A média para aprovação é 7.0. Qual estrutura condicional deve ser utilizada para tomar essa decisão?
- a) loop (for)
 - b) função (function)
 - c) condicional (if/else)
 - d) array (vetor)
 - e) classe (class)
5. Ana desenvolveu um sistema que calcula o desconto de um produto. Se o valor da compra for maior que R\$ 100,00, deve aplicar 10% de desconto. Qual seria o trecho de código correto usando if/else para resolver este problema?
- a) `if (valor > 100) { desconto = valor * 0.10; }`
 - b) `if (valor < 100) { desconto = valor * 0.10; }`
 - c) `if (valor >= 100) { desconto = valor * 10; }`
 - d) `if (valor > 100) { desconto = valor + 0.10; }`
 - e) `if (valor == 100) { desconto = valor * 0.10; }`

A.3 – Avaliação gerada pelo modelo DeepSeek V3.2

Avaliação de Aprendizagem – Linguagem de Programação I – 1ª Série

1. João precisa armazenar em um programa a idade de uma pessoa. Qual dos seguintes itens representa corretamente a criação de uma variável para guardar esse valor?
- a) `idade = "18"`
 - b) `idade = 18`
 - c) `idade = 18 anos`
 - d) `var idade: texto = 18`
 - e) `idade = verdadeiro`

Maria está escrevendo um algoritmo para calcular a média de duas notas. Ela escreveu:

```
nota1 = 8
nota2 = 7
media = nota1 + nota2 / 2
```

0. O resultado que apareceu para a média foi 11.5, mas ela esperava 7.5. O que precisa ser corrigido para o cálculo ficar certo?
- a) Trocar o sinal de + por –
 - b) Usar parênteses: $\text{media} = (\text{nota1} + \text{nota2}) / 2$
 - c) Mudar o 2 para 1
 - d) Inverter as notas: $\text{media} = \text{nota2} + \text{nota1} / 2$
 - e) Trocar os valores das notas
0. Um programa precisa verificar se um aluno foi aprovado. A regra é: média 7 ou acima significa "Aprovado", caso contrário, "Reprovado". Qual estrutura condicional usa a lógica correta?
- a) se ($\text{media} > 7$) então "Aprovado" senão "Reprovado"
 - b) se ($\text{media} == 7$) então "Aprovado" senão "Reprovado"
 - c) se ($\text{media} >= 7$) então "Aprovado" senão "Reprovado"
 - d) se ($\text{media} <= 7$) então "Aprovado" senão "Reprovado"
 - e) se ($\text{media} != 7$) então "Aprovado" senão "Reprovado"
0. Em um sistema de cadastro, é necessário armazenar se um usuário é maior de idade. Qual tipo de dado é mais adequado para guardar apenas "sim" ou "não"?
- a) Número inteiro
 - b) Número decimal
 - c) Texto
 - d) Lógico (verdadeiro/falso)
 - e) Caractere único
0. Uma lanchonete automatizada deve decidir se entrega um brinde. O brinde é dado apenas se o pedido custar mais de R\$ 30,00 E o cliente tiver cartão fidelidade. As variáveis são: valorPedido e temCartaoFidelidade (verdadeiro/falso). Qual condicional implementa essa regra corretamente?
- a) se ($\text{valorPedido} > 30$) então "Brinde"
 - b) se ($\text{temCartaoFidelidade} == \text{verdadeiro}$) então "Brinde"
 - c) se ($\text{valorPedido} > 30$) e ($\text{temCartaoFidelidade} == \text{verdadeiro}$) então "Brinde"
 - d) se ($\text{valorPedido} > 30$) ou ($\text{temCartaoFidelidade} == \text{verdadeiro}$) então "Brinde"
 - e) se ($\text{valorPedido} < 30$) e ($\text{temCartaoFidelidade} == \text{falso}$) então "Brinde"

A.4 – Avaliação gerada pelo modelo Claude 3 Haiku

Avaliação de Aprendizagem – Linguagem de Programação I

1. Qual das seguintes afirmações sobre variáveis em programação é verdadeira?
- a) Variáveis podem armazenar apenas números inteiros.
 - b) Variáveis devem sempre ter nomes longos e complexos.
 - c) Variáveis são espaços na memória do computador que armazenam valores.

- d) Variáveis não podem ser modificadas durante a execução do programa.
 - e) Variáveis são usadas apenas em programas de computador avançados.
-

2. Considere a seguinte situação-problema:

Uma loja de eletrônicos precisa calcular o valor total de uma compra, incluindo o desconto aplicado.

O cliente comprou um produto com valor de R\$ 1.500,00 e recebeu um desconto de 10%.

Qual das seguintes estruturas condicionais pode ser usada para resolver esse problema?

- a) Apenas a estrutura **if**.
 - b) Apenas a estrutura **else**.
 - c) Apenas a estrutura **if-else**.
 - d) Apenas a estrutura **switch**.
 - e) Tanto a estrutura **if-else** quanto a estrutura **switch**.
-

3. Qual das seguintes afirmações sobre lógica de programação é correta?

- a) A lógica de programação é uma habilidade que não pode ser aprendida.
 - b) A lógica de programação envolve apenas a escrita de código de computador.
 - c) A lógica de programação é o processo de pensar de forma estruturada para resolver problemas.
 - d) A lógica de programação é uma disciplina apenas para estudantes de Ciência da Computação.
 - e) A lógica de programação é uma habilidade útil apenas para programadores profissionais.
-

4. Considere a seguinte situação-problema:

Um programa precisa verificar se um número é par ou ímpar.

Se o número for par, o programa deve exibir a mensagem "O número é par".

Caso contrário, deve exibir "O número é ímpar".

Qual das seguintes estruturas condicionais pode ser usada para resolver esse problema?

- a) Apenas a estrutura **if**.
- b) Apenas a estrutura **else**.
- c) Apenas a estrutura **if-else**.
- d) Apenas a estrutura **switch**.
- e) Tanto a estrutura **if-else** quanto a estrutura **switch**.

A.5 – Avaliação gerada pelo modelo OpenAI 4o Mini

Avaliação de Aprendizagem – Linguagem de Programação I

Nome do aluno: _____

Data: _____

Instruções

Leia cada questão com atenção e marque a alternativa que você considera correta.

1. (Introdução à Lógica de Programação)

O que é um algoritmo?

- a) Um tipo de programação em rede
 - b) Uma sequência de passos para resolver um problema
 - c) Um programa que não funciona
 - d) Um erro de programação
 - e) Um tipo de dado
-

2. (Variáveis e Tipos de Dados)

Em uma linguagem de programação, uma variável é:

- a) Um valor fixo que não pode ser mudado
 - b) Um espaço na memória que armazena dados
 - c) Um tipo de erro comum
 - d) Um comando que encerra o programa
 - e) Um tipo de função
-

3. (Estruturas Condicionais – if/else)

O que acontece se a condição de um comando **if** for verdadeira?

- a) O programa termina
- b) O bloco de código dentro do **if** é executado
- c) O bloco de código dentro do **else** é executado
- d) O programa ignora o **if**
- e) O código é duplicado

4. (Introdução à Lógica de Programação)

Qual é o objetivo principal de um diagrama de fluxo?

- a) Mostrar a aparência do programa
- b) Representar graficamente a lógica de um algoritmo
- c) Criar variáveis automaticamente
- d) Armazenar dados em um banco de dados
- e) Executar o programa mais rapidamente

5. (Estruturas Condicionais – if/else)

Em um programa, se quisermos verificar se um número é maior que 10, qual seria a estrutura correta?

- a) `if (número < 10) { ... }`
- b) `if (número == 10) { ... }`
- c) `if (número > 10) { ... }`
- d) `if (número <= 10) { ... }`
- e) `if (número != 10) { ... }`

