Engenharia de Prompt para a Geração Automatizada de Questões Assistida por LLMs: Uma Análise Comparativa

Camilla B. Quincozes[1], Diego Molinos[1], Rafael D. Araújo[1]
Silvio Quincozes[2], Gilleanes T. A. Guedes[2]

[11] FACOM – Universidade Federal de Uberlândia (UFU) – Uberlândia, Brasil
[2] Universidade Federal do Pampa (UNIPAMPA) – Brasil

{camillaquincozes, diego.molinos, rafael.araujo}@ufu.br,

{silvioquincozes, gilleanesguedes}@unipampa.edu.br

Abstract. *Education 5.0, powered by advances in Large Language Models (LLMs), fosters personalized, accessible, and student-centered learning experiences. In this context, the automated generation of multiple-choice questions (MCQs) emerges as a promising solution to support scalable and adaptive assessment. The effectiveness of this process largely depends on the quality of instructions provided to LLMs—defined through Prompt Engineering (PE). We present a comparative analysis of five PE, to MCQ generation in the domain of Requirements Engineering using four different LLMs. The evaluation combines cross-model assessments and human pedagogical validation based on linguistic and educational criteria. Results provide insights for a more critical and effective adoption of LLMs in educational contexts.*

Resumo. *A Educação 5.0, impulsionada pelos avanços em Grandes Modelos de Linguagem (LLMs), promove experiências de aprendizagem personalizadas, acessíveis e centradas no estudante. Nesse contexto, cresce o interesse por metodologias automatizadas de avaliação, como a geração de questões de múltipla escolha (MCQs). A eficácia dessas abordagens depende diretamente da qualidade das instruções fornecidas aos modelos, ou seja, da Engenharia de Prompt (PE). Este trabalho analisa comparativamente cinco técnicas de PE, aplicadas à geração de MCQs na área de Engenharia de Requisitos, utilizando quatro LLMs. Combinou-se a avaliação cruzada entre modelos e validação pedagógica humana, com base em critérios linguísticos e educacionais. Os resultados contribuem para a adoção mais crítica e eficiente dos LLMs na educação.*

1. Introdução

Com o avanço dos Grandes Modelos de Linguagem (LLMs), a Educação 5.0 propõe um novo paradigma educacional, no qual a Inteligência Artificial (IA) é integrada de maneira personalizada aos processos de ensino e aprendizagem [Maity and Deroy 2024, Parker 2024]. Um dos pilares dessa nova abordagem é a adoção de mecanismos dinâmicos para supervisionar o processo de aprendizagem, possibilitando diferentes contextos de ensino, contribuindo para a redução da sobrecarga docente [Baral et al. 2024, Xiao et al. 2024].

A adoção crescente de LLMs em ambientes de aprendizagem tem impulsionado avanços em áreas como *feedback* automatizado [Meyer et al. 2024, Koutcheme et al.

DOI: 10.5753/sbie.2025.12891

2024], correção discursiva [Jauhiainen 2024, Jia 2024] e geração de perguntas abertas ou fechadas [Tran et al. 2023, Wang et al. 2022]. Neste contexto, destaca-se o potencial de LLMs para apoiar práticas de ensino que antes exigiam alto esforço manual por parte de docentes. Entre essas práticas, a geração automática de Questões de Múltipla Escolha (MCQs, do inglês, *Multiple Choice Questions*) surge como uma abordagem estratégia, com capacidade de escalar avaliações personalizadas, rápidas e contextualizadas às necessidades dos estudantes [Li and Zhang 2024, Scaria et al. 2024a].

Mesmo diante desse cenário inovador, persistem desafios significativos relacionados à consistência, à relevância pedagógica e ao nível de complexidade das questões geradas automaticamente. Parte desses entraves decorre do uso de prompts—instruções textuais fornecidas ao modelo de linguagem para guiar sua resposta—que, quando formulados de maneira genérica ou mal estruturada, não exploram adequadamente o potencial dessas tecnologias. Portanto, a qualidade das questões produzidas depende diretamente da forma como os LLMs são instruídos, ou seja, das estratégias de Engenharia de *Prompt* utilizadas. Embora técnicas como *zero-shot*, *few-shot* e *chain-of-thought* venham sendo empregadas com frequência, ainda são escassas as análises comparativas que investigam, de forma sistemática, os efeitos dessas abordagens em contextos específicos no ensino superior [Li and Zhang 2024, Mazzullo and Bulut 2024].

Diante desse cenário, este trabalho propõe uma análise comparativa entre cinco técnicas de Engenharia de Prompt: *zero-shot*, *few-shot*, *chain-of-thought*, *exemplar-guided* e *template-based*, aplicadas à geração de MCQs utilizando quatro LLMs distintos, a saber os modelos Qwen, LLaMA, Gemma e DeepSeek. A investigação concentra-se na área de computação, especificamente no contexto de Engenharia de Requisitos. A questão de pesquisa que orienta este estudo é: qual técnica de Engenharia de *Prompt* é mais eficaz na geração de MCQs, considerando critérios de qualidade, diversidade, relevância e complexidade, no contexto da Engenharia de Requisitos? Além de responder a essa questão principal, este trabalho também busca: (a) analisar como cada critério de qualidade (fluidez, diversidade, relevância e complexidade) influencia na avaliação geral das questões produzidas; (b) examinar o desempenho relativo dos diferentes LLMs em combinação com as técnicas de *prompting* utilizadas, identificando quais configurações geram melhores resultados; (c) realizar uma análise qualitativa de exemplos de questões com alta e baixa pontuação, buscando identificar padrões recorrentes que contribuam para orientar boas práticas na geração automática de avaliações educacionais; (d) comparar as avaliações atribuídas por LLMs, por meio de avaliação cruzada, com aquelas realizadas por docentes humanos, a fim de investigar o grau de concordância entre julgamentos automatizados e pedagógicos na educação 5.0.

Ao integrar análise comparativa de técnicas de Engenharia de *Prompt* entre os LLMs e docentes, esta pesquisa contribui para a compreensão crítica do uso de LLMs na geração de instrumentos de aprendizagem. Destaca-se que os resultados apresentaram razoável consistência entre as avaliações das LLMs e dos docentes. Também evidenciou-se que as técnicas de *prompt* são influenciadas pela arquitetura de cada LLM.

Este artigo está estruturado da seguinte maneira: A sessão 2 trata da fundamentação teórica, enquanto a sessão 3 discute sobre trabalhos relacionados. Já a sessão 4 discorre sobre a metodologia adotada neste estudo e a sessão 5 apresenta e discute os resultados evindenciados.

2. Fundamentação Teórica

Esta seção apresenta o arcabouço teórico desta proposta, abordando os princípios da Educação 5.0, o papel dos LLMs como ferramentas de inteligência aumentada e seu uso na geração de questões por meio de técnicas de Engenharia de *prompt*.

2.1. Educação 5.0 e o Papel das LLMs na Aprendizagem

A Educação 5.0 representa uma evolução no paradigma educacional ao integrar tecnologias digitais de forma humanizada e intencional ao processo de ensino e aprendizagem. Seu foco está em promover experiências mais ativas e personalizadas de aprendizagem, indo além da mera digitalização de conteúdos [Maity and Deroy 2024]. Neste contexto, essa abordagem incorpora tecnologias como IA para fomentar práticas educacionais mais acessíveis, inclusivas e adaptativas [Maity and Deroy 2024].

Neste cenário, os LLMs emergem como ferramentas promissoras para apoiar o trabalho docente em tarefas mediadas por linguagem natural. Entre as principais aplicações, destacam-se: produção de conteúdo, a geração de *feedback* formativo e, elaboração de avaliações automatizadas [Xiao et al. 2024, Meyer et al. 2024]. Dentro deste contexto, a geração automática de MCQ tem se mostrado relevante por viabilizar a criação rápida de instrumentos alinhados a objetivos pedagógicos diversos [Tran et al. 2023].

A elaboração de MCQs com apoio de LLMs requer estratégias instrucionais que orientem a geração de conteúdo com coerência pedagógica. Uma abordagem relevante nesse contexto consiste em instruir o modelo a planejar a resposta ideal antes da formulação da pergunta, princípio que contribui para maior controle temático e alinhamento educacional [Li and Zhang 2024]. Outra prática recorrente envolve a adaptação dos LLMs para operar com diferentes níveis da *Taxonomia de Bloom*, promovendo a geração de questões que variam desde o reconhecimento factual até a análise crítica [Scaria et al. 2024a]. Em contextos mais técnicos, como a Computação, essas abordagens têm sido aplicadas com ganhos de produtividade, especialmente em disciplinas como programação e algoritmos, embora ainda exijam validação humana para garantir o alinhamento pedagógico [Tran et al. 2023].

2.2. Geração de MCQs com LLMs e os Desafios do Prompting

Apesar dos avanços na aplicação de LLMs na geração de MCQs, existem inconstistênicas significativas na qualidade das perguntas produzidas, especialmente em aspectos como clareza, relevância e complexidade das questões [Scaria et al. 2024a,Tran et al. 2023]. Essas limitações estão diretamente relacionadas à maneira como os modelos são instruídos, ou seja, o *design* do *prompt* [Li and Zhang 2024, Maity and Deroy 2024].

Neste cenário, ganha destaque a Engenharia de *Prompt*, campo emergente dedicado à formulação de instruções textuais que orientam o comportamento dos modelos, influenciando diretamente a adequação pedagógica das respostas geradas [Baral et al. 2024, Scarlatos et al. 2024]. Diversas técnicas de *prompting* vêm sendo exploradas na literatura com o objetivo de orientar os LLMs de forma mais eficaz em tarefas educacionais específicas, como a geração de questões de múltipla escolha (MCQs).

A técnica conhecida como *zero-shot prompting* consiste em fornecer ao modelo apenas a descrição da tarefa, sem exemplos adicionais. Embora seja uma abordagem

simples e eficiente, ela tende a gerar respostas inconsistentes ou genéricas, sobretudo em domínios complexos [Maity and Deroy 2024, Baral et al. 2024].

Já o *few-shot prompting* inclui a apresentação de exemplos ilustrativos junto à tarefa. Essa estratégia melhora a consistência e a qualidade das respostas, sendo especialmente útil em tarefas que exigem inferência, estrutura ou aplicação de padrões conceituais [Li and Zhang 2024, Scaria et al. 2024a].

No caso do *chain-of-thought prompting*, busca-se induzir o modelo a explicitar as etapas do raciocínio antes de apresentar a resposta final. Essa técnica tem se mostrado eficaz em atividades que demandam análise lógica, interpretação ou resolução matemática [Koutcheme et al. 2024, Mazzullo and Bulut 2024].

O *exemplar-guided prompting* é uma variação do *few-shot prompting*, na qual os exemplos são elaborados especificamente para o domínio da tarefa. Isso proporciona maior alinhamento semântico e pedagógico, sendo particularmente vantajoso em áreas técnicas e conceituais, como a educação matemática [Baral et al. 2024].

Por fim, o *template-based prompting* utiliza estruturas fixas de enunciado ou formulários semânticos padronizados. Essa abordagem favorece a padronização e a replicabilidade na geração de questões, sendo amplamente adotada em contextos de avaliação automatizada [Wang et al. 2022].

Apesar da ampla adoção das técnicas de *prompting* na literatura, são poucos os estudos que realizam análises comparativas sistemáticas entre essas abordagens, especialmente considerando múltiplos critérios como diversidade das questões geradas, relevância pedagógica, custo computacional e escalabilidade. No trabalho de Scarlatos et al. 2024, esse desafio é parcialmente enfrentado por meio do uso de aprendizado por reforço para otimizar *prompts* com base em critérios de validade educacional. Já Estévez-Ayres et al. 2024 explora o uso de LLMs no contexto de cursos técnicos, mas sem tratar de forma aprofundada as escolhas de *prompting* como variável experimental crítica.

O estudo de Schorcht et al. 2024 investiga diferentes estilos de *prompting* aplicados à resolução de problemas matemáticos, com ênfase na qualidade pedagógica das respostas. Embora traga contribuições relevantes sobre o impacto desses estilos, não aborda a geração de questões nem realiza avaliação automatizada. Já Amyeen 2023 propõe uma arquitetura baseada em *transformers* para geração e avaliação automática de questões educacionais, mas limita-se à utilização de um único modelo de linguagem.

Em Maity et al. 2023, os autores exploram técnicas como *zero-shot* e *few-shot* na geração de questões, com avaliação conduzida por especialistas humanos. Apesar dos resultados promissores, não há comparação entre os LLMs, o que também ocorre em Li and Zhang 2024, que examina o impacto da variação de exemplos em *prompts*, mas sem diversificação de modelos ou análise automatizada.

Estudos como Hang et al. 2024 e Tran et al. 2023 concentram-se na comparação entre LLMs. O primeiro apresenta o sistema *MCQGen* e propõe critérios objetivos de avaliação automatizada através de uma abordagem que inspira diretamente a metodologia deste trabalho; contudo, ambos não exploram diferentes técnicas de engenharia de *prompt* e carecem de avaliação pedagógica estruturada por humanos.

Os trabalhos de Scaria et al. 2024a e Scaria et al. 2024b se destacam por englo-

barem múltiplas técnicas de *prompting* e validação humana, mas não realizam avaliação cruzada automatizada entre LLMs nem apresentam análise comparativa sistemática entre modelos. De forma complementar, Scarlatos et al. 2024 propõe o uso de aprendizado por reforço para melhorar a qualidade das saídas geradas por LLMs, enquanto Estévez-Ayres et al. 2024 avalia esses modelos em cursos técnicos, com foco no uso de ferramentas de IA para *feedback*. Por fim, de Amorim da Silva et al. 2025 se diferencia pela aplicação da teoria de resposta ao item (IRT) para avaliar MCQs geradas por LLMs, mas não explora múltiplas técnicas de engenharia de *prompt*.

De modo geral, a literatura atual trata o problema de forma fragmentada, ora enfatizando os diferentes estilos de *prompting*, ora focando na qualidade das respostas ou nos modelos utilizados. Em contraste, este trabalho propõe uma abordagem mais integrada e sistemática, comparando múltiplas técnicas de engenharia de *prompt* em quatro LLMs distintos. A avaliação combina métricas automatizadas e análise pedagógica conduzida por especialistas humanos, promovendo uma perspectiva mais ampla e criteriosa. Com isso, busca-se fornecer evidências mais sólidas para o uso crítico, reflexivo e alinhado aos princípios da Educação 5.0.

A Tabela 1 apresenta uma comparação entre esta proposta e os principais trabalhos da literatura, considerando os seguintes aspectos: uso de múltiplas técnicas de engenharia de *prompt*, comparação entre diferentes LLMs, presença de avaliação automatizada por LLM, participação de avaliadores humanos e foco específico em questões de MCQs.

Tabela 1. Comparação entre o presente estudo e trabalhos relacionados

Referência	PE	LLMs	Avaliação por LLMs	Avaliação por Humanos	MCQs
Este trabalho	✓	✓	✓	✓	✓
[Schorcht et al. 2024]	✓	✗	✗	✓	✗
[Amyeen 2023]	✓	✗	✓	✗	✓
[Maity et al. 2023]	✓	✗	✗	✓	✓
[Hang et al. 2024]	✗	✓	✓	✗	✓
[Tran et al. 2023]	✗	✓	✗	✓	✓
[Scaria et al. 2024a]	✓	✓	✗	✓	✓
[Scaria et al. 2024b]	✓	✓	✗	✓	✓
[Li and Zhang 2024]	✓	✗	✗	✓	✓
[Scarlatos et al. 2024]	✗	✗	✓	✓	✗
[Estévez-Ayres et al. 2024]	✗	✓	✓	✓	✗
[de Amorim da Silva et al. 2025]	✗	✓	✓	✓	✓

Nota: PE = compara técnicas de Engenharia de *Prompt*; LLMs = compara diferentes modelos de linguagem; Aval. LLM = avaliação automática realizada por LLMs; Aval. Hum. = avaliação realizada por humanos; MCQs = foco na geração de questões de múltipla escolha.

3. Metodologia

Este estudo propõe uma análise comparativa do desempenho de cinco técnicas de PE, são elas: *zero-shot, few-shot, chain-of-thought, exemplar-guided* e *template-based*, aplicadas a quatro LLMs de código aberto: Qwen [Team 2024b], LLaMA [AI 2024], Gemma [Team 2024a] e DeepSeek [DeepSeek-AI 2025]. O objetivo central é investigar como diferentes estratégias de *prompting* influenciam a qualidade, a diversidade e a complexidade das MCQs geradas por LLMs.

3.1. Geração das Questões de Múltipla Escolha

Foram geradas MCQs a partir de modelos representativos das LLMs selecionadas para este estudo, combinadas com as técnicas de engenharia de *prompt*. Em particular, os respectivos modelos para cada LLM são: Qwen (*qwen-qwq-32b*), LLaMA (*LLaMA-3.1-8b-instant*), Gemma (gemma2-9b-it) e DeepSeek (*deepseek-r1-distill-LLaMA-70b*). Já as cinco técnicas de engenharia de prompt selecionadas são *zero-shot*, *few-shot*, *chain-of-thought*, *exemplar-guided* e *template-based*. Cada LLM foi responsável por gerar cinco questões para cada técnica, resultando em 25 questões por modelo e um total de 100 questões. A geração foi realizada via Interface de Programação de Aplicações (API, do inglês, *Application Programming Interface*) da Groq[1], com automação implementada em `Python 3.10`, utilizando bibliotecas como `pandas` e `dotenv`.

3.2. Etapas Avaliativas

Na avaliação cruzada entre LLMs, cada modelo atuou como avaliador das 75 MCQs geradas pelos outros três modelos (3 modelos × 5 técnicas × 5 questões = 75), evitando possíveis vieses decorrentes de autoavaliação. Cada questão foi avaliada em 5 critérios (conforme definido na Seção 3.3), resultando em 1.500 avaliações automatizadas (4 avaliadores × 75 questões × 5 critérios). Para tanto, foram realizadas requisições às APIs das LLMs. Em etapa posterior, todos os arquivos foram lidos e unificados em um único `dataframe` por meio da biblioteca `pandas` e da linguagem de programação *python*. Cada uma das 700 linhas da base consolidada representa uma avaliação aplicada a um critério específico de uma questão.

A etapa de avaliação humana fundamentou-se na estratégia de triangulação proposta por [Denzin 1978], que defende o uso de múltiplos observadores como forma de aumentar a confiabilidade e a validade das interpretações científicas, além de minimizar os vieses individuais e promover uma análise mais robusta, ao integrar diferentes pontos de vista. Portanto, as 100 questões de múltipla escolha geradas pelos quatro modelos foram analisadas de forma independente por quatro docentes com experiência em Computação e Engenharia de Software, atuantes nos cursos de Sistemas de Informação, Ciência da Computação e Engenharia de Software. A avaliação seguiu rigorosamente os mesmos critérios e escalas definidos na Seção 3.3, permitindo a comparação direta com os julgamentos automatizados realizados pelas LLMs.

Para viabilizar o processo, foi desenvolvida uma aplicação web personalizada, composta por um formulário em `HTML`, backend em `Flask` e lógica interativa em `JavaScript`. As questões foram apresentadas sequencialmente, uma por vez, para que os avaliadores atribuíssem as pontuações conforme os critérios estabelecidos. As respostas foram registradas automaticamente em arquivos no formato `JSON`, contendo os identificadores do modelo de LLM, da técnica de engenharia de prompt, da questão avaliada e as respectivas pontuações atribuídas.

3.3. Critérios de Avaliação

Ambas as etapas avaliativas, apresentadas na Seção 3.2, utilizaram a mesma métrica, composta por cinco critérios qualitativos propostos por Hang et al. 2024: fluidez gramatical,

[1] `https://console.groq.com/`

capacidade de resposta, diversidade, complexidade e relevância. Cada critério foi avaliado em escala *Likert* de 1 a 5, conforme descrito na Tabela 2, que apresenta a rubrica utilizada para interpretação dos escores.

Tabela 2. Critérios de avaliação adotados para avaliação docente e por LLMs.

Critério	Descrição e Exemplo
Fluidez gramatical	Linguagem clara, coesa e correta. *Ex: sem erros gramaticais; frases bem estruturadas.*
Capacidade de resposta	Apenas uma alternativa deve ser claramente correta, sem ambiguidade. *Ex: evita empates entre duas opções igualmente plausíveis.*
Diversidade	Variedade na formulação das perguntas, evitando repetições. *Ex: não repetir sempre o mesmo tipo de estrutura ou foco.*
Complexidade	A questão deve exigir algum nível de raciocínio ou interpretação. *Ex: evita perguntas óbvias ou cuja resposta salta aos olhos.*
Relevância	Conexão direta com o conteúdo de Engenharia de Requisitos. *Ex: evita perguntas genéricas ou fora do escopo temático.*

O critério fluidez gramatical considera a clareza, coesão e correção do enunciado, assegurando que as frases estejam bem estruturadas e livres de erros. A capacidade de resposta refere-se à clareza da formulação da pergunta, avaliando se ela permite identificar uma única alternativa correta, sem ambiguidade entre as opções. O critério diversidade diz respeito à variação temática, estrutural e estilística das questões, evitando repetições excessivas ou padronizações mecânicas. Já a complexidade analisa o nível de raciocínio exigido para responder corretamente, sendo inspirado na Taxonomia de Bloom e considerando desde conhecimentos factuais até habilidades analíticas. Por fim, a relevância avalia o grau de alinhamento entre a questão e o conteúdo programático da disciplina, neste caso, Engenharia de Requisitos, rejeitando formulações genéricas ou fora do escopo.

4. Resultados e Discussões

Nesta seção, apresentam-se os principais resultados da avaliação das técnicas de PE utilizadas para geração das MCQs. A análise considerou cinco critérios qualitativos: fluidez, resposta, diversidade, complexidade e relevância. Os dados foram avaliados por docentes especialistas e, paralelamente, por meio de uma avaliação cruzada entre os próprios LLMs. A seguir, discutem-se quatro abordagens complementares para identificar a técnica de prompt mais eficaz.

4.1. Desempenho Geral das Técnicas de Engenharia de Prompt

Inicialmente, calculou-se a média geral das notas atribuídas a cada técnica, considerando todos os critérios combinados. Os resultados, apresentados na Figura 1, indicam que a técnica *zero-shot* obteve o melhor desempenho médio global, destacando-se consistentemente na fluidez (4,57), resposta (4,81) e relevância (4,42). Essa técnica demonstrou equilíbrio entre clareza, precisão e adequação sem exigir exemplos adicionais no prompt, configurando-se como uma estratégia eficiente e de baixo custo de implementação.

A técnica *few-shot* também apresentou bom desempenho geral, com destaque para a resposta (4,58) e fluidez (4,37). Por sua vez, a técnica *template-based* obteve médias equilibradas, sendo relevância (4,33) um de seus pontos altos, reforçando sua capacidade de gerar questões semanticamente adequadas. Em contraste, a técnica *chain-of-thought*

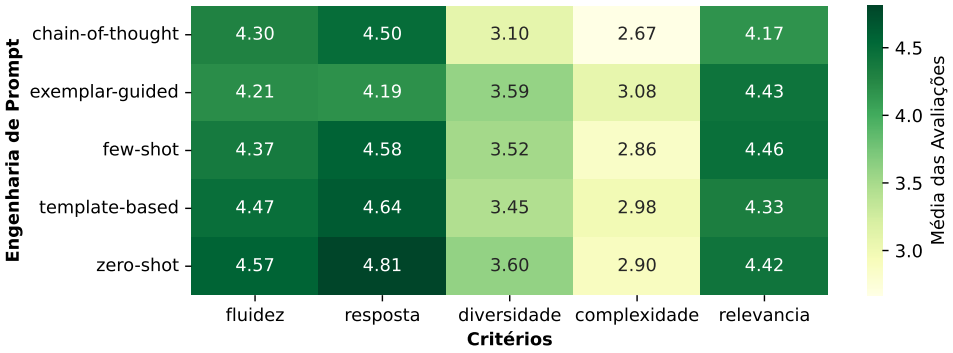

Figura 1. Média das avaliações

registrou os menores índices médios entre todas as abordagens, notadamente em complexidade (2,67) e diversidade (3,10), evidenciando sua limitação no contexto específico da geração de questões de múltipla escolha com propósito educacional.

4.2. Técnicas de Engenharia de Prompt *vs* Critérios de Avaliação

Com o objetivo de comparar o desempenho das técnicas de engenharia de prompt segundo os critérios estabelecidos, foram gerados gráficos de barras individuais por métrica: fluidez, resposta, diversidade, complexidade e relevância. Cada gráfico apresenta as notas médias atribuídas por docentes e por modelos LLMs, permitindo uma análise detalhada e contrastiva entre as perspectivas humana e automatizada.

A Figura 2 apresenta as médias obtidas nas avaliações de cada técnica de PE para o critério fluidez. Observa-se que todas as técnicas obtiveram avaliações altas em fluidez, com destaque para *zero-shot* e *template-based*, que superaram a média de 4,7 nas avaliações das LLMs. Os docentes também atribuíram boas notas a essas técnicas, indicando consenso quanto à clareza textual das questões geradas. No critério de resposta, mostrado na Figura 3, as médias continuam elevadas em ambas as fontes avaliativas. A técnica *zero-shot* novamente se sobressai, sendo reconhecida por docentes e LLMs pela coerência e adequação das alternativas em relação ao enunciado.

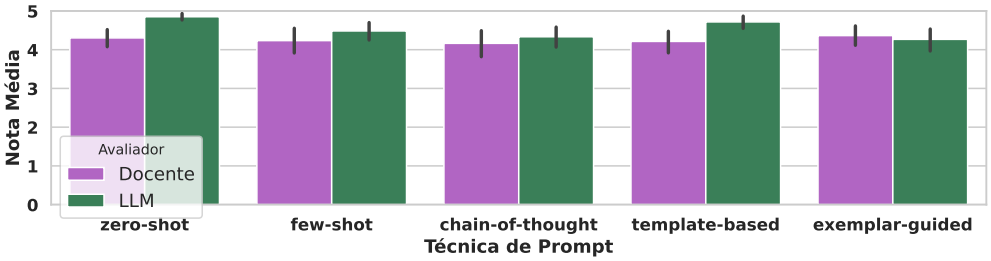

Figura 2. Média das avaliações para o critério fluidez.

As divergências mais expressivas entre os grupos de avaliadores surgem nos critérios de complexidade e diversidade. A Figura 4 contém o resultado das médias obtidas por cada técnica de PE para o critério complexidade. Nota-se uma tendência geral de avaliações mais baixas, com docentes atribuindo médias inferiores a 2,8 em várias

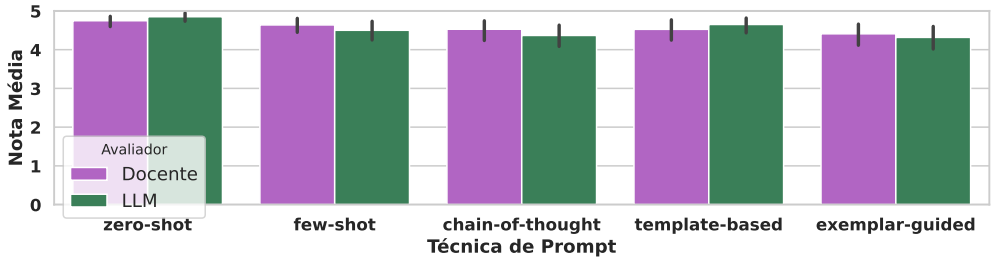

Figura 3. Média das avaliações para o critério resposta.

técnicas. A exceção é *exemplar-guided*, que se destaca com a maior média docente nesse critério, sugerindo maior capacidade de gerar questões desafiadoras.

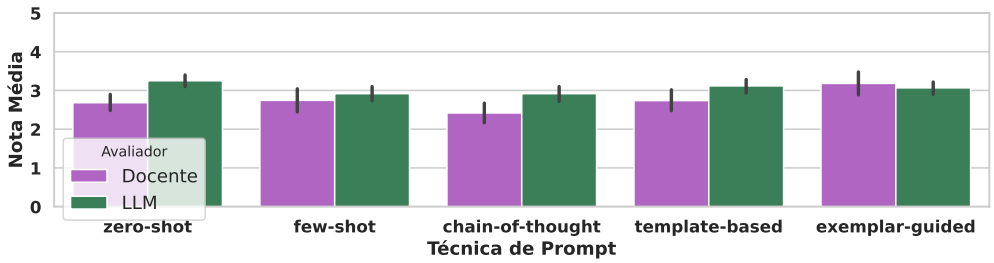

Figura 4. Média das avaliações para o critério complexidade.

No caso do critério diversidade, cujos resultados são mostrados na Figura 5, enquanto as LLMs conferem notas mais homogêneas e altas em diversidade, os docentes são mais críticos, especialmente em *chain-of-thought* e *template-based*, indicando limitações na variação das alternativas geradas por essas técnicas.

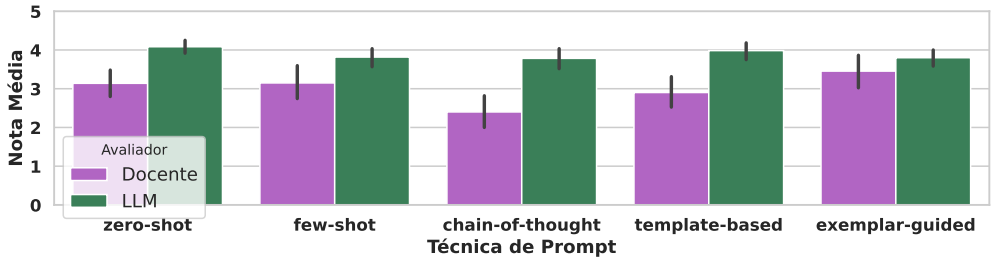

Figura 5. Média das avaliações para o critério diversidade.

Por fim, no critério de relevância, mostrado na Figura 6, é possível notar uma avaliação bem alta feita por LLMs para todas as técnicas de PE, o que não se repete na avaliação feita por docentes. Esse descompasso evidencia que, embora os modelos apresentem coerência em suas autoavaliações, os critérios pedagógicos aplicados por especialistas humanos ainda se mostram mais rigorosos e sensíveis ao contexto educacional, reforçando a necessidade de uma supervisão crítica no uso dessas ferramentas em práticas avaliativas. Dentre as técnicas de PE, a *zero-shot* foi aquela melhor avaliada por docentes para o critério relevância, seguida pela técnica *exemplar-guided*.

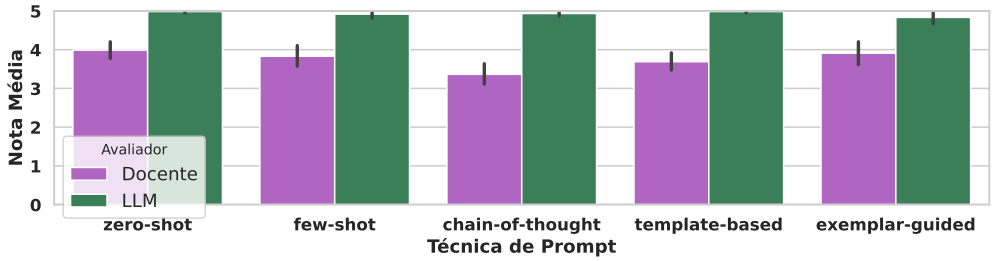

Figura 6. Média das avaliações para o critério relevância.

Em conjunto, os dados indicam que a técnica *zero-shot* é consistentemente bem avaliada pelas LLMs, enquanto os docentes demonstram preferência por abordagens como *exemplar-guided* e *few-shot*, especialmente nos critérios de maior subjetividade interpretativa. Esses resultados reforçam a importância de considerar diferentes perspectivas na construção de estratégias de geração de questões educacionais assistidas por IA.

4.3. Técnicas de Engenharia de Prompt *vs* LLMs

Para identificar a técnica de engenharia de prompt com melhor desempenho para cada LLM, foi gerado um mapa de calor único (Figura 7), o qual apresenta a média das notas atribuídas às cinco técnicas avaliadas (*zero-shot*, *few-shot*, *chain-of-thought*, *exemplar-guided* e *template-based*) em cada um dos cinco critérios: fluidez, resposta, diversidade, complexidade e relevância. Cada subfigura corresponde a uma LLM específica.

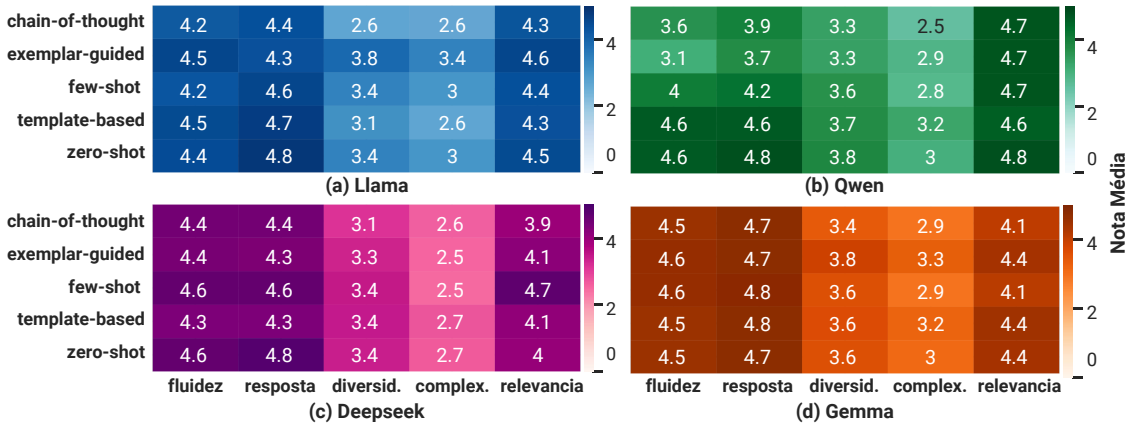

Figura 7. Média das avaliações para o critério relevância.

Na Figura 7 (a), o modelo LLaMA apresenta médias entre 3,1 e 4,6, com desempenho relativamente bom em fluidez, resposta e relevância, mas com quedas visíveis nos critérios de complexidade e relevância, especialmente para algumas técnicas de PE como *template-based* e *chain-of-thought*. A técnica *exemplar-guided* obteve questões mais complexas, diversas e relevantes para o LLaMA, mas obteve o menor índice de resposta—critério em que o *zero shot* se destacou para este modelo, seguido do *template-based* e do *few-shot*. Técnicas como *exemplar-guided* e *template-based* demonstraram ser mais consistentes em termos de fluidez, sendo ligeiramente superiores ao *zero-shot*.

Já na Figura 7 (b), referente ao modelo Qwen, observa-se uma alta relevância, em geral, com notas de 4,6 a 4,8, onde o *zero-shot* demonstrou atingir a maior nota. O

zero-shot também obteve desempenho superior nos quesitos resposta, diversidade, fluidez e relevância. Já a técnica *template-based* além de empatar com o *zero-shot* no quesito fluidez, obteve o maior índice de complexidade. Em contraste, a técnica *chain-of-thought* obteve desempenho consideravelmente inferior no quesito complexidade.

A Figura 7 (c), correspondente ao DeepSeek, revela um padrão mais homogêneo, com médias equilibradas entre as técnicas, embora com leve vantagem para *few-shot* e *zero-shot* em critérios como fluidez e resposta. O *few-shot* também se destacou de forma significativa em termos de relevância. O Deepseek apresentou as questões com menor complexidade, sobretudo quando combinado com as técnicas *exemplar-guided* e *few-shot*.

Por fim, a Figura 7 (d) mostra o desempenho do modelo Gemma, que apresenta altos e consistentes índices em fluidez e resposta para todas as técnicas, com destaque para *few-shot* e *template-based*, sobretudo nos critérios de complexidade e relevância. O principal resultado observado para o Gemma se reflete no desempenho da técnica *exemplar-guided*, a qual obteve os mais altos índices para quase todos os critérios, exceto pelo critério de resposta—onde ficou ligeiramente inferior.

Esses resultados evidenciam que o desempenho das técnicas de prompt não é uniforme entre os modelos: enquanto o *zero-shot* se destacou no Qwen e DeepSeek, técnicas como *exemplar-guided* e *template-based* apresentaram maior efetividade no Gemma e LLaMA. Isso indica que a interação entre a arquitetura da LLM e a estrutura do *3* influencia diretamente na qualidade das questões geradas.

4.4. Avaliação Geral Docente

A Figura 8 ilustra as médias das avaliações realizadas exclusivamente pelos docentes para cada técnica de PE, discriminadas por critério. Observa-se que a técnica *exemplar-guided* obteve os melhores resultados em critérios como fluidez (4,36) e diversidade (3,45). Em contraste, a técnica *zero-shot* manteve-se robusta em resposta (4,75) e relevância (3,99), sendo a mais bem avaliada para o critério resposta pelos docentes. A técnica *chain-of-thought* obteve a avaliação mais baixa pelos docentes, apresentando as menores médias em todos os critérios avaliados, incluindo fluidez (4,16), diversidade (2,40) e complexidade (2,42), o que sugere sua inadequação ao formato de múltipla escolha com foco educacional na percepção pedagógica.

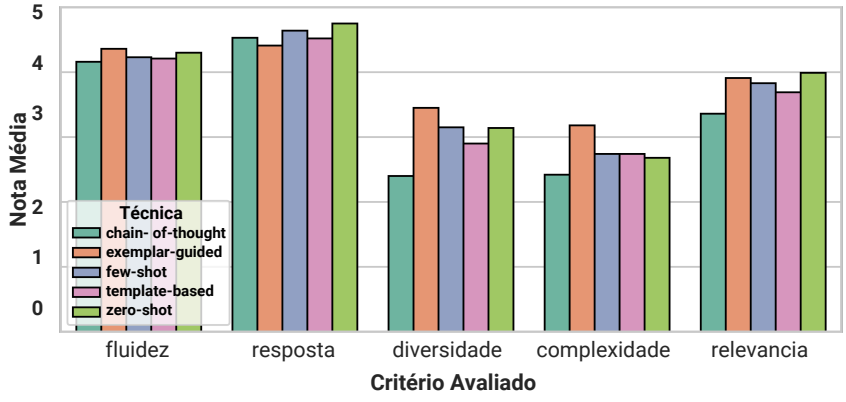

Figura 8. Média das avaliações docentes por critério avaliativo.

A análise visual do gráfico apresentado na Figura 9 reforça a percepção de que

exemplar-guided foi a que obteve as melhores avaliações sob o olhar humano, considerando todos os critérios. Por outro lado, o gráfico confirma que a *chain-of-thought* obteve a menor avaliação na média de todos os critérios.

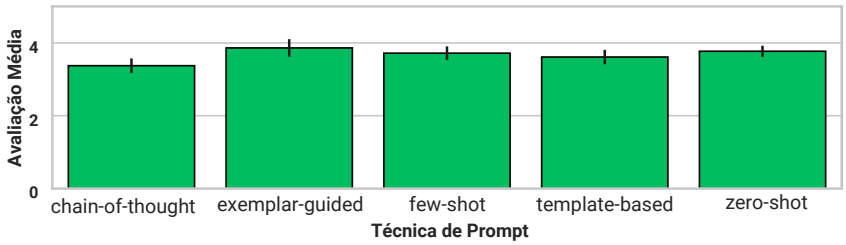

Figura 9. Média das avaliações docentes por técnica de prompt.

5. Conclusão e Trabalhos Futuros

Este estudo investigou comparativamente o desempenho de cinco técnicas de PE aplicadas à geração automatizada de questões de múltipla escolha MCQs, por quatro modelos LLMs, sob duas perspectivas avaliativas: agentes humanos (docentes especialistas) e avaliações cruzadas entre modelos. Os resultados evidenciam que a técnica *zero-shot* apresentou o melhor desempenho médio geral, destacando-se especialmente em critérios como fluidez e coerência da resposta. Por outro lado, a técnica *chain-of-thought*, embora amplamente explorada na literatura, mostrou desempenho inferior na maioria dos critérios, indicando limitações em sua aplicação direta ao contexto educacional de avaliação sob as condições em que foi experimentada neste trabalho.

A análise por modelo também demonstrou que a eficácia das técnicas de prompt varia de acordo com a arquitetura da LLM, o que sugere a importância de calibragem contextualizada. Além disso, observou-se relativa consistência entre os julgamentos dos docentes e das LLMs, com variações pontuais que apontam para diferenças na percepção de qualidade entre avaliadores humanos e modelos.

Como limitações do estudo, destaca-se a utilização de versões gratuitas das LLMs, que restringem ajustes finos nos parâmetros de geração (e.g., *temperature*, *top-p*). Destaca-se o impacto do elevado consumo de tokens em técnicas mais complexas, como *chain-of-thought*, que podem comprometer a completude das questões. Embora exista uma forte indicação de que o desempenho observado nessa técnica decorra da limitação de contexto, especialmente em *prompts* mais longos, não se pode descartar a influência de fatores como a adequação semântica da técnica ao formato de MCQs. Ademais, não foi realizada análise de possíveis vieses linguísticos, culturais ou de gênero nas questões geradas, o que pode afetar a equidade das avaliações. Por fim, o estudo concentrou-se exclusivamente no domínio da Engenharia de Requisitos, sendo recomendável que futuras pesquisas ampliem a investigação para outros contextos educacionais.

Como trabalhos futuros, propõe-se ampliar a análise com técnicas adicionais de PE e modelos de maior capacidade, bem como investigar mecanismos automáticos de avaliação com explicabilidade para apoiar docentes na curadoria de questões geradas por IA. Este trabalho representa um passo inicial na sistematização do uso pedagógico de LLMs para avaliação formativa, com potencial de impacto na personalização do ensino e na redução da carga docente.

Referências

AI, M. (2024). Introducing Meta Llama 3.1: The Latest Generation of Meta's Open-Source Large Language Models. Blog Post. Accessed: 11 June 2025. Specific model version: Llama 3.1 8B Instant.

Amyeen, R. (2023). Prompt-Engineering and Transformer-based Question Generation and Evaluation. *arXiv.org*.

Baral, S., Worden, E., Lim, W.-C., Luo, Z., Santorelli, C., Gurung, A., and Heffernan, N. (2024). Automated feedback in math education: A comparative analysis of llms for open-ended responses. *arXiv*. Pré-publicação.

de Amorim da Silva, A., Moreira, J. P., Silva, G. P. G., Cintra, J. P., and Ciferri, R. R. (2025). Anaquest: Uma ferramenta para geração e validação de questões de múltipla escolha com llms e teoria de resposta ao item. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED)*, Brasil. Springer. No prelo.

DeepSeek-AI (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. Technical Report arXiv:2501.12948v1, DeepSeek AI.

Denzin, N. K. (1978). *The Research Act: A Theoretical Introduction to Sociological Methods*. McGraw-Hill, New York, 2nd edition.

Estévez-Ayres, I., Callejo, P., Hombrados-Herrera, M. Á., Alario-Hoyos, C., and Delgado Kloos, C. (2024). Evaluation of LLM Tools for Feedback Generation in a Course on Concurrent Programming. *International Journal of Artificial Intelligence in Education*.

Hang, C. N., Wei Tan, C., and Yu, P.-D. (2024). Mcqgen: A Large Language Model-Driven MCQ Generator for Personalized Learning. *IEEE Access*, 12:102261–102273.

Jauhiainen, Nome do Primeiro Autor e Garagorry, N. d. S. A. (2024). Avaliação de respostas discursivas de estudantes com llms: uso do framework rag. *arXiv*. Pré-publicação.

Jia, N. d. P. A. e. o. (2024). Avaliando a fidelidade do feedback gerado por llms. In *Anais do EDM 2024 – Educational Data Mining*.

Koutcheme, C., Dainese, N., Sarsa, S., Hellas, A., Leinonen, J., and Denny, P. (2024). Open source language models can provide feedback: Evaluating llms' ability to help students using gpt-4-as-a-judge. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education*, pages 52–58.

Li, K. and Zhang, Y. (2024). Planejamento primeiro, pergunta depois: Um método guiado por llms para geração controlável de questões. In *Anais da ACL 2024 – Findings of the Association for Computational Linguistics*, pages 4715–4729, Bangkok, Tailândia.

Maity, S. and Deroy, A. (2024). O futuro da aprendizagem na era da ia generativa: geração e avaliação automatizada de questões com grandes modelos de linguagem. *arXiv*. Pré-publicação.

Maity, S., Deroy, A., and Sarkar, S. (2023). Harnessing the Power of Prompt-based Techniques for Generating School-Level Questions using Large Language Models. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 30–39. ACM.

Mazzullo, N. d. P. A. and Bulut, N. d. S. A. (2024). Automated feedback generation for open-ended questions. In *NeurIPS 2024*.

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., and Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: Ai-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6:100199.

Parker, N. d. P. A. e. o. (2024). Uma abordagem com grandes modelos de linguagem para análise de feedback em pesquisas educacionais. *SpringerLink*.

Scaria, N., Chenna, S. D., and Subramani, D. (2024a). Geração automatizada de questões educacionais nos diferentes níveis da taxonomia de bloom utilizando llms. *arXiv*. Pré-publicação.

Scaria, N., Chenna, S. D., and Subramani, D. (2024b). Quão bons são os llms modernos na geração de perguntas relevantes e de alta qualidade nos diferentes níveis da taxonomia de bloom para o currículo de ciências sociais do ensino médio na Índia? In *Anais do 19º Workshop sobre Uso Inovador de PLN na Educação (BEA 2024)*, Cidade do México, México.

Scarlatos, A., Smith, D., Woodhead, S., and Lan, A. (2024). *Improving the Validity of Automatically Generated Feedback via Reinforcement Learning*, pages 280–294. Springer Nature Switzerland.

Schorcht, S., Buchholtz, N., and Baumanns, L. (2024). Prompt the problem – investigating the mathematics educational quality of AI-supported problem solving by comparing prompt techniques. *Frontiers in Education*, 9.

Team, G. (2024a). Gemma 2: Improving Open Language Models at a Practical Size. Technical Report arXiv:2408.00118v1, Google DeepMind.

Team, Q. (2024b). Qwen-qwq-32b Model Card and Documentation. Technical report, Alibaba Cloud.

Tran, A., Angelikas, K., Rama, E., Okechukwu, C., Smith, D. H., and MacNeil, S. (2023). Generating Multiple Choice Questions for Computing Courses Using Large Language Models. In *2023 IEEE Frontiers in Education Conference (FIE)*, pages 1–8. IEEE.

Wang, Z., Valdez, J., Basu Mallick, D., and Baraniuk, R. G. (2022). Towards human-like educational question generation with large language models. In *Lecture Notes in Computer Science*, pages 153–166.

Xiao, C., Ma, W., Song, Q., Xu, S. X., Zhang, K., Wang, Y., and Fu, Q. (2024). Human-AI Collaborative Essay Scoring: A Dual-Process Framework with LLMs. *arXiv*.