



Análise do Potencial Cognitivo e Pedagógico de Modelos de Linguagem: Uma

Avaliação sobre IA Generativa no Processo de Ensino-Aprendizagem

Hellen Guterres França, Universidade Federal do Maranhão,
hellen.guterres@nca.ufma.br, <https://orcid.org/0009-0002-7128-6678>

Anderson Lopes Silva, Universidade Federal do Maranhão,
anderson.silva@nca.ufma.br, <https://orcid.org/0009-0003-7721-0700>

Lisandra de Melo Lima, Universidade Federal do Maranhão,
lisandra@nca.ufma.br, <https://orcid.org/0009-0001-6262-7218>

Matheus Lopes Raposo Lima, Universidade Federal do Maranhão,
matheus@nca.ufma.br, <https://orcid.org/0009-0006-6358-3540>

Lisle Faray de Paiva, Essen University Hospital,
lisle.faraydepaiva@uk-essen.de, <https://orcid.org/0000-0002-4594-5628>

Anselmo Cardoso de Paiva, Universidade Federal do Maranhão,
paiva@nca.ufma.br, <https://orcid.org/0000-0003-4921-0626>

Jan Egger, Essen University Hospital,
jan.egger@uk-essen.de, <https://orcid.org/0000-0002-5225-1982>

Resumo: Os *Large Language Models* (LLMs) têm ampliado a aplicação de educação mediada por Inteligência Artificial (IA), especialmente em tarefas que envolvem interpretação e raciocínio. Este trabalho analisa o potencial pedagógico e cognitivo de atividades interdisciplinares dos modelos Claude Sonnet 4.5, DeepSeek-V3, Gemma 2 9B-IT, GPT-5 e Llama 4 Scout, a partir da geração de respostas a questões extraídas do Exame Nacional do Ensino Médio (ENEM). A avaliação foi desenvolvida utilizando uma metodologia que contempla a verificação da contaminação de dados, a padronização das interações e a avaliação do desempenho por meio de métricas sintáticas e semânticas. Os resultados indicam habilidades significativas em interpretação textual, porém com limitações em abstração matemática e em inferência lógica. Este estudo propõe uma abordagem avaliativa, associada à reflexão sobre a assistência de IA na educação.

Palavras-chave: Modelos de Linguagem de Grande Escala; Inteligência Artificial Generativa; Educação Mediada por IA; Avaliação de Potencial Cognitivo.

Analysis of the Cognitive and Pedagogical Potential of Language Models: An Evaluation of Generative AI for the Teaching-Learning Process

Abstract: Large Language Models (LLMs) have expanded the possibilities of Artificial Intelligence (AI)-mediated education, particularly in tasks that involve interpretation and reasoning. This study analyzes the pedagogical and cognitive potential of models Claude Sonnet 4.5, DeepSeek-V3, Gemma 2 9B-IT, GPT-5, and Llama 4 Scout in interdisciplinary activities, based on their ability to generate answers to questions drawn from Brazil's National High School Exam (ENEM). The methodology includes data contamination checks, interaction standardization, and performance evaluation through syntactic and semantic metrics. The results indicate strong interpretive abilities but limitations in mathematical abstraction and logical inference. This study proposes an evaluative approach that reflects on the role of AI assistance in education.

Keywords: Large Language Models; Generative Artificial Intelligence; AI-Mediated Education; Cognitive Potential Evaluation.

1. Introdução

Nos últimos anos, os Modelos de Linguagem de Grande Escala (LLMs - Large Language Models) têm impactado profundamente a forma como interagimos com o



conhecimento, abrindo novas possibilidades para a educação mediada por inteligência artificial (IA). Esses modelos de linguagem, com bilhões de parâmetros e treinados com textos de múltiplas origens, demonstram grande capacidade de compreender, gerar e explicar conteúdos em linguagem natural (Chang *et al.*, 2024). No contexto educacional, isso se traduz em oportunidades de apoio personalizado à aprendizagem, na geração de materiais didáticos e na avaliação automática de competências cognitivas (Fabre e Bertagnolli, 2025).

As tecnologias educacionais baseadas em IA generativa têm mostrado benefícios, como o aumento da motivação dos estudantes, a melhoria da retenção de conteúdo e a promoção do pensamento crítico e colaborativo (Paradedá; Torres e Takahashi, 2024). Contudo, a utilização de LLMs como mediadores do processo educacional deve ser orientada por princípios pedagógicos e éticos, assegurando que essas tecnologias complementem, e não substituam, o papel do educador (Peláez-Sánchez; Velarde-Camaqui e Glasserman-Morales, 2024). É importante compreender até que ponto os modelos de linguagem reproduzem ou compreendem os processos cognitivos humanos envolvidos na resolução de problemas educacionais reais. Também é importante considerar que, além de medir acertos e erros, precisamos avaliar como os LLMs interpretam textos, integram conhecimentos de diferentes áreas e produzem raciocínios coerentes, que se constituem em capacidades centrais para o aprendizado humano (Sachete *et al.*, 2025; Mammides e Papadopoulos, 2024).

Nesse cenário, o Exame Nacional do Ensino Médio (ENEM) constitui um importante ponto de partida para investigar o potencial educacional dos LLMs (Almeida *et al.*, 2024). Aplicado anualmente a milhões de estudantes brasileiros, o ENEM é reconhecido por avaliar não apenas a memorização de conteúdos, mas também habilidades de leitura crítica, raciocínio lógico, interpretação e resolução de problemas interdisciplinares (Nunes *et al.*, 2023). Por essa razão, o exame pode ser compreendido como um reflexo das competências cognitivas e linguísticas esperadas ao final do ensino médio, tornando-se um ambiente ideal para avaliar o quanto os LLMs são capazes de reproduzir processos de raciocínio alinhados às demandas cognitivas da educação brasileira (Almeida *et al.*, 2023).

Este trabalho propõe avaliar o desempenho de diferentes modelos de linguagem, abertos e proprietários, na resolução de questões do ENEM 2024, considerando o ENEM como um ambiente de referência para a análise das capacidades cognitivas e pedagógicas dessas tecnologias. Mais do que um comparativo técnico, o estudo busca compreender o papel que os LLMs podem desempenhar como ferramentas de apoio ao ensino e à avaliação, bem como suas limitações e riscos no contexto educacional. A partir dessa análise, pretende-se responder a questões como: Até que ponto os LLMs são capazes de raciocinar sobre conteúdos interdisciplinares em língua portuguesa? Quais áreas do conhecimento os modelos demonstram maior ou menor domínio? Como os resultados obtidos podem orientar o uso pedagógico responsável de modelos de linguagem?

2. Trabalhos Relacionados

Os avanços recentes dos LLMs têm impulsionado novas abordagens para a educação assistida por IA, especialmente na personalização do aprendizado e na avaliação automatizada. Estudos indicam que esses modelos têm potencial para redefinir práticas pedagógicas, ao oferecer suporte individualizado aos estudantes e ao reduzir a carga de trabalho docente (Vogel; Ramos e Franzoni, 2025; Silva e Aranha, 2025; Silva; Pedreira e Alves, 2025). Contudo, a compreensão da capacidade real dos LLMs ainda representa um desafio, sobretudo em contextos educacionais e linguísticos específicos.



No cenário internacional, benchmarks como o *Measuring Massive Multitask Language Understanding* (MMLU) (Hendrycks *et al.*, 2020) são utilizados para avaliar as respostas dos LLMs em diversas áreas do conhecimento. Essas avaliações fornecem uma visão da abrangência conceitual dos modelos, mas não capturam as nuances culturais, linguísticas e cognitivas presentes em idiomas distintos do inglês e em contextos socioculturais variados. Assim, permanece uma lacuna na literatura quanto à avaliação dos LLMs sob uma perspectiva educacional contextualizada, especialmente no que se refere à formação crítica e interdisciplinar exigida pelo sistema de ensino brasileiro.

No contexto nacional, estudos como os de Nunes *et al.* (2023), Pires *et al.* (2023) e Almeida *et al.* (2024) iniciaram o uso de provas do ENEM como instrumento de avaliação de LLMs. Esses trabalhos mostraram que modelos como GPT-4 e Sabiá conseguem reproduzir parte do raciocínio interdisciplinar exigido pelo exame, mas apresentam limitações em questões que demandam interpretação crítica e integração semântica entre diferentes áreas do conhecimento.

Nesse contexto, (Vogel; Ramos e Franzoni, 2025) destaca a importância de adotar uma perspectiva ética e crítica no uso de IA na educação, reconhecendo desafios como o viés algorítmico, a desigualdade de acesso e a necessidade de capacitação docente. Ressalta-se que os LLMs devem ser utilizados como instrumentos complementares ao ensino, promovendo a autonomia intelectual e o pensamento crítico, sem substituir o papel mediador do educador.

Silva e Pedreira (2025) propõem uma ferramenta que utiliza LLMs e dados do Exame Nacional de Desempenho dos Estudantes (ENADE) para correlacionar disciplinas, questões e resultados dos estudantes avaliados, visando apoiar a revisão curricular e a análise institucional. De modo semelhante, (Silva e Aranha, 2025) investiga o uso de LLMs na educação em programação, com ênfase no feedback formativo e na engenharia de prompts. Em ambos os estudos, destaca-se o papel dos modelos de linguagem como mediadores do processo de aprendizagem, capazes de fornecer diagnósticos pedagógicos e de oferecer suporte adaptativo aos alunos.

Apesar desses avanços, a literatura ainda carece de avaliações que considerem o potencial cognitivo dos LLMs com base em referenciais avaliativos nacionais e multidisciplinares, como o ENEM. Este trabalho busca preencher essa lacuna ao propor uma análise que integra o caráter empírico dos benchmarks internacionais a uma perspectiva educacional, visando mensurar não apenas o desempenho técnico, mas também a adequação pedagógica e linguística dos modelos. Dessa forma, a pesquisa contribui para a investigação de tecnologias educacionais voltadas à integração ética e pedagógica dos LLMs no contexto brasileiro.

3. Metodologia

A metodologia adotada neste trabalho foi estruturada para avaliar o desempenho de diferentes LLMs em tarefas que exigem raciocínio cognitivo e interpretação de textos em língua portuguesa. O processo metodológico baseia-se no estudo de (Myrzakhan; Bsharat e Shen, 2024), que é dividido em quatro etapas principais: seleção e preparação do corpus, interação com os modelos de linguagem, coleta e processamento das respostas e, por fim, análise quantitativa e qualitativa dos resultados. Essa abordagem permite avaliar o desempenho dos modelos quanto ao conhecimento técnico e também examinar sua capacidade de compreender contextos educacionais e de reproduzir formas de raciocínio próximas às competências humanas. Para a avaliação dos modelos de linguagem, este estudo adotou a técnica de *prompt zero-shot*, na qual a instrução recebida pelo modelo consiste apenas na tarefa de gerar respostas, sem exemplos. Assim, é



possível analisar a capacidade de generalização e de interpretação em distintas áreas de conhecimento.

3.1. Base de Dados

A base de dados utilizada neste estudo é composta pelas questões do ENEM 2024. O exame é composto por 180 questões interdisciplinares de múltipla escolha, distribuídas em quatro grandes áreas do conhecimento: Linguagens, Códigos e suas Tecnologias; Ciências Humanas e suas Tecnologias; Ciências da Natureza e suas Tecnologias; e Matemática e suas Tecnologias. Cada questão avalia competências cognitivas previstas na Base Nacional Comum Curricular, tais como leitura crítica, interpretação de textos, integração de informações e resolução de problemas. Por isso, o ENEM constitui um ambiente de avaliação adequado para investigar o potencial cognitivo e pedagógico dos modelos de linguagem.

A preparação do corpus seguiu o procedimento proposto em (Pires *et al.*, 2023b). Inicialmente, realizamos a limpeza e filtragem do texto para garantir a consistência das entradas nos modelos. Removemos todas as questões com imagens, gráficos, tabelas ou outros elementos visuais que exigissem interpretação visual, pois o objetivo deste trabalho é avaliar o desempenho linguístico e cognitivo, e não a capacidade de interpretação visual dos LLMs. Após a filtragem, restaram 124 questões compostas apenas por texto, abrangendo os quatro domínios do exame.

Em seguida, realizou-se a estruturação padronizada das questões, em um formato que preserva o enunciado original e as cinco alternativas (A–E), indicando explicitamente a resposta correta. Essa padronização visa garantir interações consistentes com os modelos e evitar ruídos semânticos que possam comprometer a interpretação das instruções. As questões foram revisadas manualmente para manter a coerência linguística e remover caracteres incorretos decorrentes da extração digital.

3.2. Seleção dos Modelos de Linguagem

Foram selecionados cinco LLMs com diferentes arquiteturas, tamanhos e origens de treinamento, a fim de possibilitar uma análise comparativa mais ampla do desempenho de sistemas com distintas abordagens de modelagem e alinhamento instrucional. Essa diversidade visa observar como fatores como o tamanho do modelo e o foco de treinamento influenciam a capacidade de raciocínio e interpretação em língua portuguesa. Sendo assim, os modelos escolhidos foram os seguintes:

Claude Sonnet 4.5: desenvolvido pela Anthropic, com foco em segurança, coerência textual e manutenção de contexto prolongado (Anthropic, 2025).

DeepSeek-V3: modelo de código aberto voltado ao desempenho e à eficiência computacional (DeepSeek-AI *et al.*, 2025).

Gemma 2 9B-IT: desenvolvido pelo Google DeepMind, e instruções ajustadas para uso interativo (Team, 2024).

GPT-5: sucessor dos modelos da OpenAI, amplamente conhecido por seu alto desempenho em benchmarks de compreensão linguística (OpenAI, 2025).

Llama 4 Scout: modelo aberto desenvolvido pela Meta, com boa capacidade de raciocínio contextual e de adaptação ao português. (Meta AI, 2025)

Os modelos selecionados apresentam diversidade arquitetural e representatividade entre soluções abertas e proprietárias, proporcionando uma visão equilibrada do estado atual das tecnologias de linguagem natural. Devido às limitações de acesso às APIs, os modelos Claude Sonnet 4.5, DeepSeek-V3 e GPT-5 foram avaliados exclusivamente por meio de suas interfaces de chat, enquanto os demais foram acessados por meio de ambientes de execução controlados.



3.3. Contamination Check

Uma das principais preocupações metodológicas em estudos que avaliam LLMs é a possibilidade de contaminação dos dados, isto é, o modelo já ter sido exposto, durante seu pré-treinamento, a parte ou a totalidade do conjunto de dados que será utilizado na avaliação. Esse acesso prévio a esses textos no treinamento do LLM compromete a validade do experimento, pois o modelo poderia simplesmente reproduzir respostas memorizadas, em vez de demonstrar raciocínio ou compreensão genuína do problema proposto. Considerando que alguns exames nacionais são amplamente divulgados e utilizados em materiais didáticos e em plataformas digitais, essa verificação torna-se relevante para garantir a integridade dos resultados.

Com o intuito de mitigar esse risco, foi realizado um procedimento de *Contamination Check* (Paiva *et al.*, 2025). A verificação consistiu em selecionar aleatoriamente dez questões do conjunto de dados e suas respectivas alternativas, instruindo os modelos, por meio de um *prompt* específico, como exemplificado na Tabela 1), a gerar novos enunciados com base nas opções apresentadas. O objetivo dessa abordagem foi avaliar se os modelos seriam capazes de reproduzir, de forma idêntica ou muito similar, o texto original das questões, o que indicaria a presença de memorização ou de acesso prévio ao conteúdo durante o treinamento.

Tabela 1. *Prompts* utilizados para APIs e interface de chat.

Tipo de Tarefa	Prompt para APIs	Prompt para Chats
Contamination Check	Gere uma pergunta relevante e bem formulada com base na seguinte resposta: Resposta: <code>{resposta['answer']}</code> Gere uma pergunta coerente com essa resposta. Responda apenas com a pergunta gerada, sem explicações adicionais.	Gere uma pergunta relevante e bem formulada com base nas respostas do seguinte arquivo. Gere uma pergunta coerente com essa resposta. Responda apenas com a pergunta gerada, sem explicações adicionais. Retorne um arquivo <code>jsonl</code> com colunas <code>id</code> e <code>resposta</code> .
Gerar Respostas	Responda à pergunta abaixo de acordo com as opções. A pergunta vem de um exame de vestibular. Questão: <code>{questao['id']}</code> Pergunta: <code>{questao['question']}</code> Opcoes: <code>{"", " .join(questao['alternatives'])}</code> Qual é a resposta correta? Responda apenas com a alternativa correta, sem explicações adicionais.	Responda às perguntas do seguinte arquivo de acordo com as opções. As perguntas vêm de um exame de vestibular. Qual é a resposta correta? Responda apenas com o texto da alternativa correta, sem explicações adicionais. Retorne um arquivo em formato <code>jsonl</code> com colunas <code>id</code> e <code>answer</code> .

3.4. Avaliação

Após a realização do *Contamination Check*, inicia-se o processo de avaliação do desempenho dos modelos. Para isso, foi utilizado um *prompt* padronizado, apresentado na Tabela 1), contendo o enunciado de cada questão e suas cinco alternativas, no formato de instrução *zero-shot*. Esse formato foi escolhido para observar o comportamento natural dos modelos, sem exemplos prévios de resposta, a fim de analisar sua capacidade de interpretação, inferência e tomada de decisão com base apenas no texto fornecido. Cada modelo deveria selecionar a alternativa que considerasse correta, sendo a resposta ideal aquela que apresentasse apenas a letra correspondente à opção escolhida. Em alguns casos, no entanto, os modelos incluíram explicações adicionais ou justificativas, o que foi



considerado no cálculo das métricas e na análise qualitativa.

A escolha das métricas de avaliação buscou equilibrar a análise técnica e pedagógica dos resultados. Foram empregadas três métricas: *ROUGE* (Lin, 2004), *BERTScore* (Zhang *et al.*, 2020) e Acurácia (Li *et al.*, 2024), responsáveis por mensurar, respectivamente, a similaridade sintática, a proximidade semântica e a exatidão das respostas. A combinação de métricas propostas visa permitir a avaliação da corretude dos modelos (acertam a alternativa correta) e também da qualidade de suas respostas quanto à coerência textual e proximidade conceitual ao conteúdo de referência.

As métricas *ROUGE*, amplamente utilizadas na avaliação de geração de texto, medem o grau de sobreposição entre partes do texto produzido e o texto de referência, verificando o quanto suas expressões e estruturas coincidem. A variante *ROUGE-N* avalia a similaridade com base em sequências curtas de palavras (n-gramas) compartilhadas entre as duas respostas, enquanto a *ROUGE-1* considera a maior sequência comum de palavras preservada na mesma ordem. Já o *BERTScore* avalia a correspondência semântica entre os textos, comparando cada *token* da resposta gerada ao texto de referência por meio de características contextuais derivadas de modelos pré-treinados.

A métrica de Acurácia foi utilizada como indicador de desempenho objetivo, medindo o percentual de respostas em que a saída do modelo coincide exatamente com a alternativa correta. Essa métrica reflete a capacidade do modelo de selecionar a opção correta entre múltiplas alternativas, aproximando-se da lógica dos exames tradicionais. Essas três métricas combinadas fornecem uma visão mais abrangente do comportamento dos modelos, avaliando tanto a precisão literal quanto a coerência linguística e semântica de suas respostas.

4. Resultados e Discussão

Os resultados da avaliação permitem uma análise do potencial de uso dos LLMs como instrumentos de apoio à aprendizagem e à avaliação educacional. As respostas geradas pelos modelos foram examinadas sob uma perspectiva técnica e pedagógica, com o objetivo de compreender não apenas os índices de desempenho, mas também a natureza do raciocínio exibido em cada tipo de questão. A seguir, são apresentados os resultados, tanto quantitativos quanto qualitativos, acompanhados de uma discussão sobre suas implicações para o uso ético e efetivo da IA generativa na educação.

4.1. Análise dos Resultados

Primeiramente, os resultados do *Contamination Check* confirmaram a adequação dos modelos selecionados para a etapa de avaliação. A Tabela 2 apresenta as pontuações obtidas pelos modelos nas métricas *BERTScore F1*, *ROUGE-1*, *ROUGE-2* e *ROUGE-L*, aplicadas às perguntas originais e às questões geradas artificialmente. Esses indicadores medem o grau de semelhança textual e semântica entre dois textos para avaliar se uma resposta ou um enunciado se aproxima de um texto de referência. Valores mais altos nessas métricas indicam maior similaridade entre os textos comparados, enquanto valores mais baixos sugerem diferenças significativas tanto na estrutura linguística quanto no conteúdo semântico.

No contexto deste estudo, esses resultados indicam que as perguntas geradas pelos modelos não apresentaram sobreposição textual ou semântica relevante às questões originais. Assim, a combinação de métricas em contexto sintático (*ROUGE*) e semântico (*BERTScore*) fornece uma dupla confirmação: por um lado, de que não há coincidências estruturais entre os textos; por outro, de que não há equivalência de significado entre as questões geradas e as originais. Dessa forma, esses resultados nas métricas, abaixo de



0,65 para o *BERTScore* e próximos de zero nas variantes *ROUGE*, indicam que os LLMs não reproduziram frases idênticas ou muito similares às do exame, o que evidencia a ausência de memorização prévia e, portanto, a validade do corpus para fins avaliativos.

Tabela 2. *Contamination Check* dos modelos usando as métricas *BERTScore F1*, *ROUGE-1*, *ROUGE-2* e *ROUGE-L*.

Modelo	<i>BERTScore F1</i>	<i>ROUGE-1</i>	<i>ROUGE-2</i>	<i>ROUGE-L</i>
Claude Sonnet 4.5	0,59	0,05	0,01	0,04
DeepSeek-V3	0,60	0,06	0,00	0,05
Gemma 2	0,60	0,07	0,01	0,06
GPT-5	0,61	0,07	0,00	0,06
Llama 4 Scout	0,60	0,08	0,02	0,06

Em seguida, realizou-se a avaliação dos modelos na tarefa proposta. Esta avaliação teve como objetivo analisar o desempenho geral dos modelos quanto à precisão das respostas e à qualidade textual das saídas geradas. As métricas empregadas permitem observar diferentes dimensões de desempenho. Valores mais elevados de Acurácia indicam que o modelo foi capaz de selecionar corretamente a alternativa correta, refletindo competência em raciocínio objetivo. Por outro lado, valores elevados de *BERTScore* e *ROUGE*, mesmo em casos de respostas incorretas, sugerem que o modelo produziu explicações semanticamente próximas ao conteúdo esperado, o que é relevante do ponto de vista pedagógico, pois demonstra compreensão parcial e capacidade de formulação coerente.

A Tabela 3 apresenta os valores da avaliação para cada modelo, organizados por área de conhecimento e com o total geral. Nestes resultados, podemos observar tendências de comportamento que ressaltam os modelos que apresentam resultados consistentes em domínios distintos e os que apresentam especialização em determinadas áreas.

Tabela 3. Resultados das métricas para as respostas geradas pelos modelos em porcentagem (%).

Área	Métrica	Claude Sonnet 4.5	DeepSeek-V3	Gemma 2	GPT-5	Llama 4 Scout
Linguagens	Acurácia	0,95	0,92	0,81	0,43	0,86
	<i>BERTScore F1</i>	0,99	0,98	0,95	0,86	0,96
	<i>ROUGE-1</i>	0,97	0,92	0,87	0,53	0,87
	<i>ROUGE-2</i>	0,97	0,92	0,84	0,46	0,86
	<i>ROUGE-L</i>	0,97	0,92	0,86	0,52	0,87
Ciências Humanas	Acurácia	1,00	0,97	0,78	0,28	0,95
	<i>BERTScore F1</i>	1,00	1,00	0,98	0,86	0,99
	<i>ROUGE-1</i>	1,00	1,00	0,96	0,41	0,96
	<i>ROUGE-2</i>	1,00	1,00	0,95	0,31	0,95
	<i>ROUGE-L</i>	1,00	1,00	0,96	0,41	0,96
Ciências da Natureza	Acurácia	0,94	0,83	0,83	0,27	0,66
	<i>BERTScore F1</i>	0,99	0,99	0,99	0,84	0,94
	<i>ROUGE-1</i>	0,98	0,91	0,98	0,45	0,79
	<i>ROUGE-2</i>	0,86	0,74	0,86	0,29	0,66
	<i>ROUGE-L</i>	0,98	0,91	0,98	0,45	0,79
Ciências Exatas	Acurácia	0,81	0,66	0,88	0,18	0,29
	<i>BERTScore F1</i>	0,98	0,95	0,99	0,86	0,90
	<i>ROUGE-1</i>	0,84	0,73	0,96	0,33	0,52
	<i>ROUGE-2</i>	0,62	0,47	0,63	0,21	0,34
	<i>ROUGE-L</i>	0,84	0,73	0,96	0,31	0,51
Total	Acurácia	0,93	0,87	0,82	0,30	0,74
	<i>BERTScore F1</i>	0,99	0,98	0,97	0,86	0,95
	<i>ROUGE-1</i>	0,95	0,90	0,93	0,43	0,81
	<i>ROUGE-2</i>	0,89	0,82	0,83	0,33	0,75
	<i>ROUGE-L</i>	0,95	0,90	0,93	0,43	0,81



A análise dos resultados revela disparidades de desempenho entre os LLMs avaliados. O Claude Sonnet 4.5 da Anthropic obteve as melhores pontuações gerais, destacando-se em quase todos os domínios, especialmente em Ciências Humanas, onde atingiu desempenho perfeito em todas as métricas. Suas médias foram de 0,94 em acurácia, 0,99 em BERTScore F1, 0,95 em ROUGE-1, 0,89 em ROUGE-2 e 0,95 em ROUGE-L. O Claude só não liderou em Ciências Exatas. O DeepSeek-V3 apresentou desempenho semelhante ao do Claude em Ciências Humanas, com acurácia de 0,97 e pontuação máxima nas demais métricas. Suas médias gerais foram 0,87 em acurácia, 0,98 em BERTScore F1, 0,90 em ROUGE-1, 0,82 em ROUGE-2 e 0,90 em ROUGE-L, ficando em segundo lugar entre os modelos avaliados.

O modelo Gemma apresentou bons resultados em Ciências Exatas, superando DeepSeek e Claude em acurácia, com 0,89. Em Ciências da Natureza, igualou-se ao DeepSeek em acurácia, mas obteve pontuações superiores nas demais métricas. Suas médias foram 0,82 em acurácia, 0,97 em BERTScore F1, 0,93 em ROUGE-1, 0,83 em ROUGE-2 e 0,93 em ROUGE-L, evidenciando um desempenho promissor, especialmente em Ciências Exatas e Ciências da Natureza.

O Llama obteve desempenho inferior ao dos modelos anteriores, mas apresentou resultados promissores em Ciências Humanas, com acurácia de 0,95. No entanto, teve desempenho baixo em Ciências da Natureza e Ciências Exatas, o que reduziu suas médias gerais. O GPT-5 da OpenAI registrou o pior desempenho, com médias de 0,31 em acurácia, 0,86 em BERTScore F1, 0,43 em ROUGE-1, 0,33 em ROUGE-2 e 0,43 em ROUGE-L, indicando alinhamento textual inferior. Apesar de pontuações semelhantes em todas as áreas, destacou-se em Linguagens. É importante notar que o GPT-5 manteve precisão lexical, fornecendo respostas semanticamente corretas e detalhadas, mas selecionando alternativas incorretas.

Esta avaliação abrangente revela uma hierarquia clara de desempenho entre os modelos de linguagem da última geração. O Claude Sonnet 4.5 é o modelo dominante na avaliação, alcançando pontuações métricas mais altas em todos os domínios, destacando-se particularmente em Ciências Humanas e demonstrando uma integração robusta do conhecimento. O DeepSeek-V3 ficou em segundo lugar no geral, com pontuação quase perfeita no campo das Ciências Humanas. Além disso, o modelo Gemma apresentou bons resultados, superando o Claude e o DeepSeek em Ciências Exatas e mantendo bons resultados nos demais domínios. O Llama ainda demonstrou pontos fortes de especialização em Ciências Humanas, porém, o GPT-5 apresentou pontuações consistentemente baixas, sugerindo limitações fundamentais de alinhamento textual. Estes resultados destacam que a especialização por domínio apresenta variação significativa, mesmo entre modelos amplamente utilizados.

A avaliação evidencia uma hierarquia clara entre os modelos de linguagem de última geração. O Claude Sonnet 4.5 lidera, com as maiores pontuações em todos os domínios, especialmente em Ciências Humanas. O DeepSeek-V3 ocupa a segunda posição, com o melhor desempenho em Ciências Humanas. O Gemma superou os modelos Claude e DeepSeek em Ciências Exatas e manteve resultados consistentes nos demais domínios. O Llama apresentou bons resultados em Ciências Humanas, enquanto o GPT-5 apresentou desempenho consistentemente baixo, o que indica limitações de alinhamento textual. Os resultados mostram que a especialização por domínio varia significativamente, mesmo entre modelos amplamente utilizados. Feita por Claude Sonnet 4.5 no campo de Ciências Humanas, pois sugere que os modelos podem se especializar em determinado(s) domínio(s) de conhecimento. Outro exemplo é a diferença de desempenho entre os 2 modelos com melhores resultados (Claude e DeepSeek) na área de Ciências



Exatas e o modelo Gemma, que os ultrapassou em quase todas as métricas.

Observa-se que os modelos tendem a enfrentar dificuldades em Ciências da Natureza e Ciências Exatas, especialmente ao combinar raciocínio interpretativo com cálculos matemáticos. A tabela 4 ilustra o desempenho dos modelos em uma questão específica, evidenciando tais limitações. Os modelos apresentaram dificuldades significativas em problemas que exigem essa combinação de habilidades.

Nesse exemplo, nenhum dos cinco modelos conseguiu chegar à resposta correta, revelando uma limitação em lidar com problemas que envolvem pensamento estratégico e compreensão de sistemas temporais. Os modelos tiveram especial dificuldade em entender a natureza contraintuitiva do problema, em que, para reduzir o custo total, é preciso aumentar o valor das parcelas que serão gratuitas. Esse comportamento indica que os modelos possuem capacidade de realizar cálculos matemáticos diretos, mas ainda enfrentam dificuldades em problemas que exigem raciocínio lógico estratégico, especialmente quando a solução ideal vai contra a intuição inicial.

Tabela 4. Comparativo de Respostas dos Modelos - Questão 177 (Ciências Exatas)

Item	Descrição
Questão	Uma imobiliária iniciou uma campanha de divulgação para promover a venda de apartamentos que podem ser pagos em 100 parcelas mensais. O valor da primeira delas é fixado no momento da compra, com o pagamento dessa primeira parcela. A partir da segunda parcela, o valor é determinado pela aplicação de um acréscimo percentual fixo ao valor da parcela anterior. Como atrativo, a imobiliária fará o pagamento de todas as parcelas correspondentes ao mês de aniversário do comprador. Um cliente, que faz aniversário no mês de maio, decidiu comprar um desses apartamentos por meio do financiamento oferecido pela imobiliária, e pretende escolher o mês mais adequado para realizar essa compra, de modo que o valor total dos pagamentos seja o menor possível. Qual é o mês que esse cliente deverá escolher para realizar a compra do apartamento?
Alternativas	A) Fevereiro B) Abril C) Maio D) Junho E) Agosto
Alternativa correta	A. Fevereiro
Claude Sonnet 4.5	D. Junho
DeepSeek-V3	D. Junho
Gemma 2	C. Maio
GPT-5	D. Junho
Llama 4 Scout	C. Maio

4.2. Implicações e Considerações no Contexto Educacional

A análise dos resultados obtidos permite refletir sobre o papel dos modelos de linguagem como ferramentas de apoio à aprendizagem e à prática docente. O desempenho dos modelos em tarefas que envolvem interpretação de texto e raciocínio interdisciplinar evidencia o potencial dessas tecnologias para auxiliar estudantes na compreensão de conteúdos e na resolução de problemas complexos. As limitações observadas, especialmente em questões que exigem abstração matemática, indicam que esses sistemas ainda necessitam de mediação pedagógica.

Do ponto de vista educacional, os resultados sugerem que os LLMs podem contribuir como recursos de apoio ao estudo autônomo, oferecendo explicações detalhadas, reformulando conceitos e simulando situações avaliativas. Em ambientes formais de ensino, podem ser incorporados como ferramentas para o desenvolvimento



do pensamento crítico, desde que seu uso seja mediado por professores que orientem a análise e a verificação das respostas. Essa mediação é importante para que os estudantes aprendam a interpretar criticamente as informações geradas pela IA, reconhecendo tanto seu potencial explicativo quanto seus limites cognitivos e éticos. Assim, a utilização de LLMs em contextos educacionais deve ser vista como uma oportunidade para diversificar estratégias pedagógicas e estimular a autonomia intelectual.

Os LLMs podem ser utilizados para fomentar o diálogo e a reflexão, apoiando o processo de ensino-aprendizagem, mantendo o protagonismo do estudante e a função do professor. Os resultados apresentados sugerem que os LLMs podem se tornar aliados na educação, desde que empregados com responsabilidade pedagógica e consciência crítica. O equilíbrio entre a inovação tecnológica e os princípios educacionais deve orientar sua integração às práticas de ensino, assegurando que essas ferramentas ampliem o acesso ao conhecimento e fortaleçam o desenvolvimento cognitivo dos estudantes.

5. Conclusão

Este trabalho apresentou resultados de uma avaliação de modelos de linguagem no desempenho em tarefas que exigem raciocínio interdisciplinar e compreensão textual. Os resultados sugerem que os modelos têm boa capacidade de interpretação e coerência linguística, mas enfrentam limitações em situações que exigem abstração matemática e inferência lógica. Isto apresenta-os como ferramentas técnicas e recursos educacionais, cujo impacto depende da mediação docente e da intencionalidade pedagógica de seu uso no ambiente educacional.

O trabalho apresenta uma contribuição ao uso de IA generativa na educação brasileira, demonstrando que a análise de LLMs em contextos avaliativos pode oferecer subsídios para o alinhamento de estratégias formativas mais personalizadas e críticas. Como perspectivas futuras, propõe-se expandir as investigações para outras tarefas que envolvam conhecimento interdisciplinar, como a produção textual e a resolução de problemas abertos. Além disso, também se busca o desenvolvimento de estudos empíricos que envolvam diretamente docentes e estudantes em situações de uso dessas tecnologias, de modo a compreender seu impacto real no processo de ensino-aprendizagem e na formação de competências cognitivas e éticas.

Referências

- Almeida, T. S.; Abonizio, H.; Nogueira, R.; Pires, R. **Sabiá-2: A New Generation of Portuguese Large Language Models**. 2024. Disponível em: <https://arxiv.org/abs/2403.09887>.
- Almeida, T. S.; Laitz, T.; Bonás, G. K.; Nogueira, R. Bluex: A benchmark based on brazilian leading universities entrance exams. In: Springer. **Brazilian Conference on Intelligent Systems**. [S.l.], 2023. p. 337–347.
- Anthropic. **Claude 4 Announcement or Technical Overview**. [S.l.]: Anthropic, 2025. <https://www.anthropic.com/news/claude-4>. Accessed on: May, 29, 2025.
- Chang, Y. *et al.* A survey on evaluation of large language models. **ACM transactions on intelligent systems and technology**, ACM New York, NY, v. 15, n. 3, p. 1–45, 2024.
- DeepSeek-AI *et al.* **DeepSeek-V3 Technical Report**. 2025. Disponível em: <https://arxiv.org/abs/2412.19437>.
- Fabre, M. C. J. M.; Bertagnolli, S. de C. Contribuições da inteligência artificial generativa para o design de experiências de aprendizagem: uma revisão sistemática de literatura. **RENOTE**, v. 23, n. 1, p. 298–309, 2025.



- Hendrycks, D. *et al.* Measuring massive multitask language understanding. **arXiv preprint arXiv:2009.03300**, 2020.
- Li, H. *et al.* Llms-as-judges: a comprehensive survey on llm-based evaluation methods. **arXiv preprint arXiv:2412.05579**, 2024.
- Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In: **Text Summarization Branches Out**. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 74–81. Disponível em: <https://aclanthology.org/W04-1013/>.
- Mammides, C.; Papadopoulos, H. The role of large language models in interdisciplinary research: Opportunities, challenges and ways forward. **Methods in Ecology and Evolution**, Wiley Online Library, v. 15, n. 10, p. 1774–1776, 2024.
- Meta AI. **Llama-4-Scout-17B-16E**. 2025. <https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E>. Acessado em 19 de outubro de 2025. Disponível em: <https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E>.
- Myrzakhan, A.; Bsharat, S. M.; Shen, Z. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena. **arXiv preprint arXiv:2406.07545**, 2024.
- Nunes, D.; Primi, R.; Pires, R.; Lotufo, R.; Nogueira, R. Evaluating gpt-3.5 and gpt-4 models on brazilian university admission exams. **arXiv preprint arXiv:2303.17003**, 2023.
- OpenAI. **Introducing GPT-5**. 2025. <https://openai.com/pt-BR/index/introducing-gpt-5/>. Acessado em 19 de outubro de 2025. Disponível em: <https://openai.com/pt-BR/index/introducing-gpt-5/>.
- Paiva, L. Faray de; Luijten, G.; Puladi, B.; Egger, J. How does deepseek-r1 perform on usmle? **medRxiv**, Cold Spring Harbor Laboratory Press, 2025. Disponível em: <https://www.medrxiv.org/content/early/2025/02/10/2025.02.06.25321749>.
- Paradedá, R. B.; Torres, D. T. B.; Takahashi, A. Generative ai in education: A study of undergraduate students' expectations of chatgpt. **RENOTE**, v. 22, n. 3, p. 174–185, 2024.
- Peláez-Sánchez, I. C.; Velarde-Camaqui, D.; Glasserman-Morales, L. D. The impact of large language models on higher education: exploring the connection between ai and education 4.0. In: Frontiers Media SA. **Frontiers in Education**. [S.l.], 2024. v. 9, p. 1392091.
- Pires, R.; Almeida, T. S.; Abonizio, H.; Nogueira, R. Evaluating gpt-4's vision capabilities on brazilian university admission exams. **arXiv preprint arXiv:2311.14169**, 2023.
- Pires, R.; Almeida, T. S.; Abonizio, H.; Nogueira, R. **Evaluating GPT-4's Vision Capabilities on Brazilian University Admission Exams**. 2023. Disponível em: <https://arxiv.org/abs/2311.14169>.
- Sachete, A. dos S.; Rossi, F. D.; Gomes, R. S. *et al.* Você se torna responsável por aquilo que gera: análise semântica e temporal de modelos de linguagem. **RENOTE**, v. 23, n. 1, p. 346–356, 2025.
- Silva, F. G.; Aranha, E. H. Llms na educação em programação: Estratégias para avaliação e feedback formativo. In: SBC. **Simpósio Brasileiro de Educação em Computação (EDUCOMP)**. [S.l.], 2025. p. 81–87.
- Silva, J. A.; Pedreira, K. S.; Alves, T. O. M. Avalia: Ferramenta baseada em inteligência artificial e dados do enade para análise curricular do curso de sistemas de informação da faculdade metodista granbery. In: SBC. **Simpósio Brasileiro de Sistemas de Informação (SBSI)**. [S.l.], 2025. p. 122–131.
- Team, G. Gemma. Kaggle, 2024. Disponível em: <https://www.kaggle.com/m/3301>.



Vogel, D.; Ramos, A. M.; Franzoni, A. M. B. Transformando a educação com large language models (llms): benefícios, limitações e perspectivas. **Caderno Pedagógico**, v. 22, n. 4, p. e13846–e13846, 2025.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; Artzi, Y. **BERTScore: Evaluating Text Generation with BERT**. 2020. Disponível em: <https://arxiv.org/abs/1904.09675>.