Evaluation of LLMs in Answering Questions: A Case Study on POSCOMP

Carlos H. P. Silveira[1], Florindo R. S. Carreteiro[1], Fernando A. Sousa[1], Fabio M. F. Lobato[1,2]

[1] Universidade Federal do Oeste do Pará, Brasil
carlos.silveira@discente.ufopa.edu.br
[2] Universidade de São Paulo, Brasil
fabio.lobato@icmc.usp.br

Abstract. Several studies have investigated Large Language Models' (LLMs) ability to answer questions from national educational assessments. However, specialized domains such as databases remain underexplored. Furthermore, current literature lacks a comprehensive analysis of how multimodal capabilities affect accuracy and stability across multiple runs. This study aims to fill these gaps by evaluating the accuracy and stability of five modern LLMs in solving 25 database questions from POSCOMP, using API requests, with hyperparameter control and five independent runs per configuration. The models investigated were GPT-5 Nano, GPT-5.4, Grok 4.1 Fast, Gemini 3 Flash Preview, and Gemini 3.1 Pro Preview, considering both LaTeX and image inputs. Additionally, we applied non-parametric statistical tests to evaluate the significance of performance differences between these input modalities. The results show that Gemini 3.1 Pro Preview demonstrated the best overall performance, outperforming the other models in both accuracy and consistency across runs, while GPT-5 Nano and Grok 4.1 Fast exhibited significant performance degradation under image inputs. Finally, the findings reinforce the importance of API-based evaluations with multiple independent runs to analyze not only the accuracy but also the reliability of responses generated by LLMs in technical domains.

CCS Concepts: • **Computing methodologies** → **Natural language generation**.

Keywords: databases, large language models, multimodality, POSCOMP, stability

1. INTRODUÇÃO

Os *Large Language Models* (LLMs) recentes incorporam capacidades multimodais, permitindo interpretar imagens, tabelas e diagramas [Xiao et al. 2025]. Esse avanço tem ampliado o interesse pelo seu uso no cenário educacional, em especial na resolução de questões de exames nacionais [Wang et al. 2025]. No contexto brasileiro, estudos recentes investigaram o desempenho desses modelos em exames como o Exame Nacional do Ensino Médio (ENEM), o Exame Nacional de Desempenho dos Estudantes (ENADE) e o Exame Nacional para Ingresso na Pós-Graduação em Computação (POSCOMP) [Taschetto and Fileto 2024; Carvalho et al. 2025; Viegas et al. 2025]. Em particular, o POSCOMP destaca-se pelo rigor em tópicos complexos, incluindo o domínio de banco de dados, que exige a interpretação de linguagem natural, esquemas relacionais, consultas SQL e álgebra relacional.

Apesar dos avanços dos LLMs nesses exames, a literatura ainda apresenta duas limitações metodológicas: a dependência de interfaces *web*, que restringem o controle experimental, e o uso de uma única submissão por questão, o que dificulta a análise da estabilidade das respostas [Saldanha and Digiampietri 2024; Viegas et al. 2025; Carreteiro et al. 2025]. A fim de suprir as lacunas identificadas, este estudo tem como objetivo avaliar o desempenho de LLMs modernos na resolução de 25 questões de múltipla escolha de banco de dados do POSCOMP. Diferenciando-se das avaliações com interface *web*, este trabalho utiliza requisições via *Application Programming Interfaces* (APIs) para assegurar o controle dos hiperparâmetros e adota cinco execuções independentes por configuração, comparando

as entradas em LaTeX e em formato de imagem. À luz do objetivo apresentado, as seguintes Perguntas de Pesquisa (PP) foram delineadas:

PP-1: Qual é o nível de desempenho dos LLMs ao resolver questões de banco de dados do POSCOMP?

PP-2: Qual é o nível de estabilidade dos LLMs ao resolver questões de banco de dados do POSCOMP ao longo de múltiplas execuções?

PP-3: Existe diferença estatisticamente significativa no desempenho dos LLMs no que tange à multimodalidade, ao comparar entradas nos formatos LaTeX e de imagem no domínio de banco de dados?

Em síntese, as principais contribuições deste artigo são: (i) a expansão da literatura sobre o desempenho de LLMs no subcampo de banco de dados, ao analisar a acurácia e o impacto multimodal em 25 questões das edições do POSCOMP de 2016 até 2025; (ii) a mensuração da estabilidade das respostas dos modelos *GPT-5 Nano*, *GPT-5.4*, *Grok 4.1 Fast*, *Gemini 3 Flash Preview* e *Gemini 3.1 Pro Preview* por meio de múltiplas execuções independentes.

O restante deste artigo está estruturado como segue. Na Seção 2 são discutidos os trabalhos relacionados e o posicionamento desta pesquisa. Na Seção 3 estão descritos os materiais e métodos. Na Seção 4 são apresentados os resultados, suas análises e as respostas às perguntas de pesquisa. Por fim, a Seção 5 apresenta as considerações finais e as perspectivas para pesquisas futuras.

2. TRABALHOS RELACIONADOS

No POSCOMP, [Viegas et al. 2025] avaliaram oito modelos em 206 questões das edições de 2022 a 2024, utilizando textos traduzidos para o inglês, capturas de tela e arquivos em *Portable Document Format* (PDF). Os testes foram realizados via interface *web*, e os melhores resultados foram obtidos por modelos mais recentes, embora questões com dependência visual tenham se mostrado mais desafiadoras. Apesar disso, o estudo não utilizou API nem investigou a estabilidade das respostas em múltiplas execuções.

No ENADE, [Carvalho et al. 2025] investigaram modelos multimodais em 25 questões de ciência da computação, utilizando imagens e 10 submissões independentes por questão. Os resultados mostram uma variação relevante entre acurácia e estabilidade, evidenciando a importância de múltiplas execuções. No entanto, o estudo não utilizou API nem explorou o formato LaTeX.

O trabalho de [Carreteiro et al. 2025] avaliou o *GPT-4o* e o *Gemini-1.5-pro* em 55 questões de linguagens formais e autômatos de edições do POSCOMP de 2002 a 2022, comparando LaTeX, imagens e PDF, tanto via API quanto via interface *web*. Os autores observaram desempenho superior no formato LaTeX, especialmente com o uso de API. No entanto, a estabilidade das respostas dos modelos em múltiplas execuções e sob métricas específicas não foi avaliada em profundidade.

Embora a literatura recente evidencie o potencial dos LLMs no contexto educacional, nota-se que grande parte dos estudos realiza apenas uma única submissão via interface *web* ou não avalia a estabilidade das respostas. Para preencher essas lacunas, o presente trabalho analisa o desempenho e a estabilidade de LLMs na resolução de questões de banco de dados do POSCOMP, adotando um rigor metodológico alinhado às abordagens de [Carreteiro et al. 2025] e de [Carvalho et al. 2025].

3. MATERIAIS E MÉTODOS

Este estudo adotou a metodologia *Data Science Trajectories* (DST), proposta por [Martínez-Plumed et al. 2021]. A Figura 1 apresenta o fluxograma metodológico utilizado que foi dividido em cinco etapas: (i) Entendimento do Domínio, focado na revisão da literatura; (ii) Entendimento e Preparação dos Dados, que consistiu na seleção de questões do POSCOMP e na transcrição para LaTeX e imagem; (iii) Modelagem, que compreende a engenharia de *prompt* e o *design* experimental; (iv) Avaliação, baseada na análise de desempenho e estabilidade; e (v) Liberação dos Dados, com a disponibilização em repositório. Essas etapas são detalhadas a seguir.

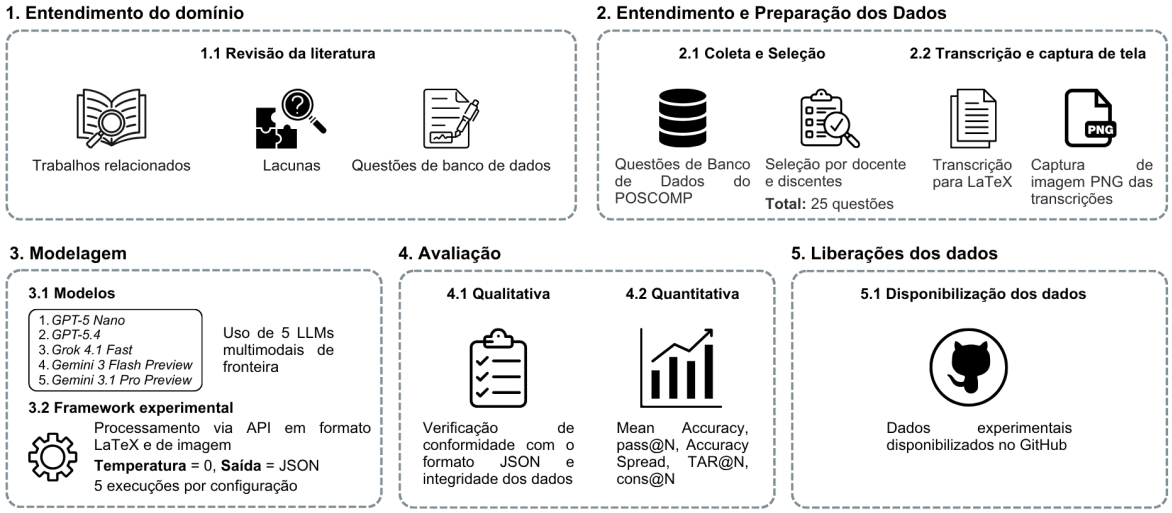

Figura. 1. Fluxograma do *pipeline* experimental do estudo.

3.1 Entendimento do Domínio

Essa etapa consistiu em uma revisão da literatura sobre o uso de LLMs em exames educacionais nacionais. Esse levantamento indicou que, embora esses modelos apresentem bons resultados em provas gerais, ainda há lacunas na comparação entre formatos de entrada, na falta de controle sobre hiperparâmetros em interfaces *web* e na análise da estabilidade das respostas em múltiplas execuções. Diante disso, o estudo concentrou-se no domínio de banco de dados do POSCOMP. Para mitigar limitações da literatura, definiu-se um protocolo via API com controle de hiperparâmetros [Liang et al. 2022] e múltiplas execuções independentes para cada configuração. Essa estratégia permitiu comparar entradas em formato LaTeX e em imagem, considerando a acurácia e a estabilidade. Essa comparação é essencial no domínio de banco de dados, pois ele exige a interpretação de linguagem natural e de estruturas complexas, como operadores de álgebra relacional, esquemas relacionais e consultas SQL.

3.2 Entendimento e Preparação dos Dados

Foram analisadas questões das edições do POSCOMP publicadas entre 2016 e 2025, com exceção de 2020 e 2021 em razão da suspensão do exame durante a pandemia de COVID-19. Nas edições em que houve dois tipos de prova, foram selecionadas as do tipo 1. A seleção das questões foi realizada por um docente e por dois discentes de pós-graduação em computação. O critério de inclusão considerou a abrangência e a representatividade de tópicos de banco de dados, como álgebra relacional, SQL, normalização e modelagem entidade-relacionamento.

Ao final, foi consolidado um *dataset* com 25 questões. A preparação dos dados envolveu duas etapas: i) transcrição para LaTeX, em que as questões foram transcritas para preservar a estrutura lógica e reduzir ambiguidades comuns em conversões automáticas de *Optical Character Recognition* (OCR) [Almeida and Caminha 2024]; e ii) representação em formato de imagem, para avaliar a capacidade multimodal dos LLMs, as questões transcritas foram renderizadas e capturadas em formato *Portable Network Graphics* (PNG) de alta resolução. Por fim, todas as questões foram revisadas manualmente com base nas provas originais, garantindo fidelidade às versões oficiais.

3.3 Modelagem

Foram avaliados os cinco LLMs: *GPT-5 Nano*, *GPT-5.4*, *Grok 4.1 Fast*, *Gemini 3 Flash Preview* e *Gemini 3.1 Pro Preview*, selecionados por sua relevância na literatura recente e por sua capacidade multimodal. As inferências foram realizadas via API, com saída em *JavaScript Object Notation*

(JSON). Com o objetivo de reduzir a variabilidade das respostas e favorecer a reprodutibilidade, o único hiperparâmetro modificado foi a *temperature*, fixada em zero [Romero et al. 2025], enquanto os demais parâmetros, como *thinking* e *seed*, foram mantidos nas configurações padrão em todas as requisições. Os experimentos foram realizados no *Google Colab*.

Cada modelo foi avaliado via API em duas modalidades de entrada, LaTeX e imagem. Em cada configuração, foi feita uma única requisição contendo as 25 questões do *dataset*. Esse procedimento foi repetido cinco vezes, de forma independente, totalizando cinco execuções por configuração, seguindo uma estratégia semelhante à adotada por [Carvalho et al. 2025] para lidar com o não-determinismo dos LLMs. Adotou-se um *prompt* padronizado, elaborado com a abordagem *zero-shot*, para instruir os modelos a responder às questões de múltipla escolha de forma estruturada. Foram definidas restrições explícitas quanto ao formato da saída, exigindo que cada modelo retornasse um objeto JSON válido contendo apenas o número da questão e a respectiva alternativa selecionada. O *prompt* completo utilizado nos experimentos está disponível no repositório deste trabalho.

3.4 Avaliação

A avaliação das respostas foi conduzida em duas etapas. Na etapa qualitativa, focada na conformidade estrutural das saídas, verificou-se que todas as respostas retornadas pelas APIs eram objetos JSON válidos. Na etapa quantitativa, foram adotadas cinco métricas consolidadas na literatura [Chen et al. 2021; OpenAI 2024; Atil et al. 2025], com foco especial na metodologia utilizada por [Carvalho et al. 2025] ao avaliar a acurácia e a estabilidade de LLMs em questões do ENADE, detalhadas a seguir:

—***Mean Accuracy*:** proporção média de respostas corretas;

—***pass@N*:** probabilidade de obter ao menos uma resposta correta em N execuções;

—***Accuracy Spread*:** diferença entre a maior e a menor acurácia média observadas;

—***Total Agreement Rate (TAR@N)*:** avalia se todas as N respostas a uma mesma questão são idênticas entre si, independentemente de estarem corretas ou incorretas;

—***Consensus Correctness (cons@N)*:** determina se a resposta de maior frequência entre as N execuções coincide com o gabarito oficial.

Para a análise da PP-3, calculou-se a acurácia média de cada questão para cada modelo, a partir das cinco execuções independentes de cada configuração e de cada formato de entrada. Em seguida, as acurácias foram pareadas por questão para comparar os dois formatos em cada modelo. Inicialmente, avaliou-se a normalidade da distribuição dessas diferenças por meio do teste de Shapiro-Wilk. Diante da rejeição da hipótese de normalidade, definiu-se a aplicação do teste não paramétrico de Wilcoxon Signed-Rank para a comparação pareada, com a correção de Holm-Bonferroni para o controle de múltiplas comparações.

3.5 Liberação dos dados

Visando garantir a transparência e a reprodutibilidade desta pesquisa, em consonância com os princípios de ciência aberta [Munafò et al. 2017], o *prompt* utilizado, as provas do POSCOMP selecionadas, seus respectivos gabaritos, as respostas coletadas nos experimentos com as LLMs e os gráficos gerados encontram-se disponíveis em um repositório[1].

4. RESULTADOS E DISCUSSÕES

As Figuras 2 e 3 apresentam os resultados das cinco métricas adotadas para ambos os formatos de entrada. Outrossim, a Tabela I sintetiza a análise de significância estatística da PP-3. A seguir, apresentam-se as discussões para cada pergunta de pesquisa.

[1] `https://github.com/carloshsilveira/Kdmile-2026`

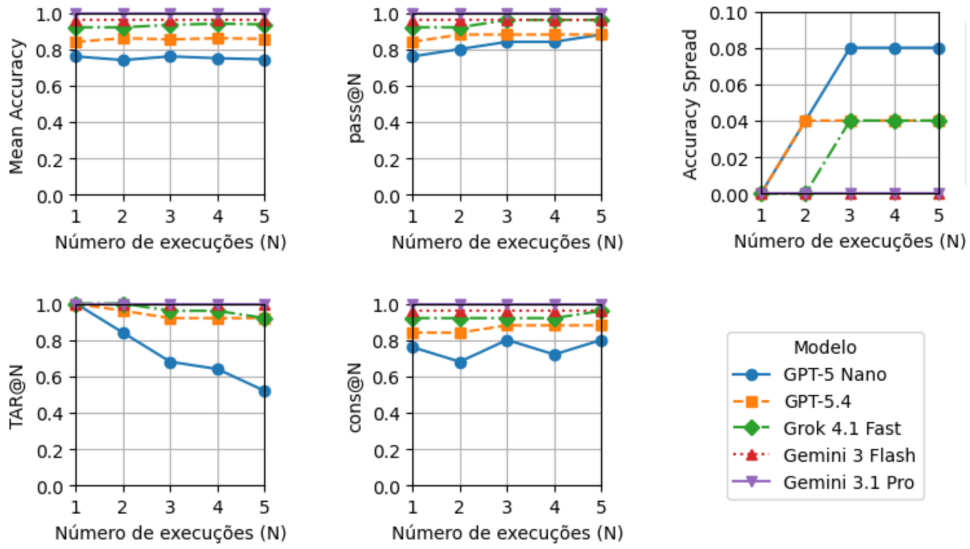

Figura. 2. Desempenho médio dos cinco modelos em função do número de execuções (N) para questões no formato LaTeX, considerando cinco métricas: *Mean Accuracy* (superior esquerdo), *pass@N* (superior centro), *Accuracy Spread* (superior direito), *TAR@N* (inferior esquerdo) e *cons@N* (inferior centro).

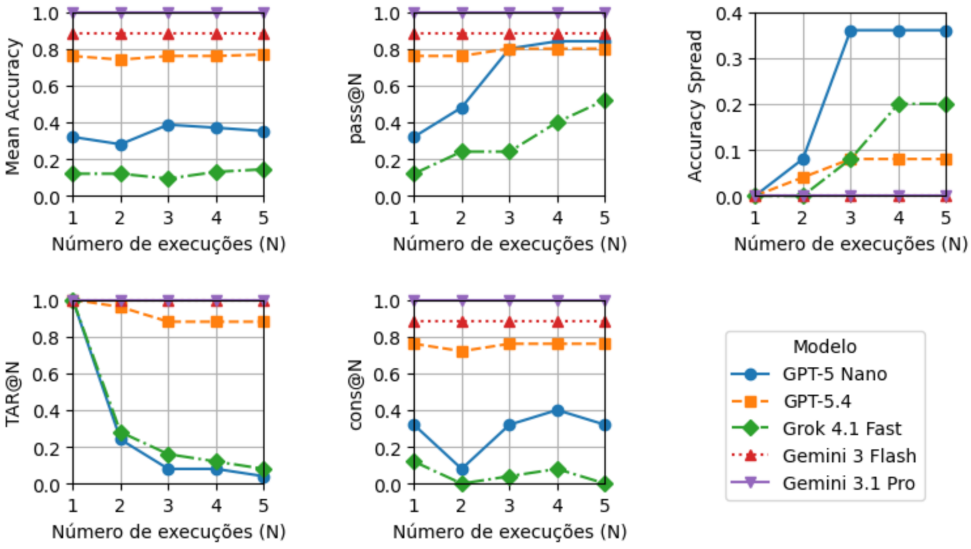

Figura. 3. Desempenho médio dos cinco modelos em função do número de execuções (N) para questões no formato de imagem, considerando cinco métricas: *Mean Accuracy* (superior esquerdo), *pass@N* (superior centro), *Accuracy Spread* (superior direito), *TAR@N* (inferior esquerdo) e *cons@N* (inferior centro).

4.1 PP-1: Acurácia

Na métrica *Mean Accuracy*, no formato LaTeX, o *Gemini 3.1 Pro Preview* manteve desempenho máximo em todas as execuções. Por sua vez, o *Gemini 3 Flash Preview* e o *Grok 4.1 Fast* também apresentaram valores elevados. No formato de imagem, o *Gemini 3.1 Pro Preview* preservou a pontuação máxima, enquanto o *Gemini 3 Flash Preview* e o *GPT-5.4* mantiveram bom desempenho. No entanto, o *GPT-5 Nano* e o *Grok 4.1 Fast* apresentaram quedas acentuadas, evidenciando maior sensibilidade ao processamento de conteúdo visual nas questões. A métrica *pass@N* mostrou que

o aumento do número de tentativas eleva a probabilidade de ao menos um acerto, sobretudo para modelos menos consistentes. Em LaTeX, todos os modelos convergiram para valores elevados.

No formato de imagem, esse ganho foi visível nos modelos *GPT-5 Nano* e *Grok 4.1 Fast*, cujos crescimentos em *pass@N* contrastam com baixos valores de *Mean Accuracy*, sugerindo que os acertos ocorrem de forma esporádica e instável. Dessa forma, a análise dessas métricas indica que o *Gemini 3.1 Pro Preview* foi o modelo com o melhor desempenho em ambos os formatos, seguido pelo *Gemini 3 Flash Preview*. O *GPT-5.4* apresentou desempenho intermediário, enquanto o *GPT-5 Nano* e o *Grok 4.1 Fast* foram os mais afetados pela conversão para formato de imagem. Esse comportamento reforça os achados de [Carreteiro et al. 2025], segundo os quais o formato LaTeX tende a ser melhor para resolução de questões.

4.2 PP-2: Estabilidade

A métrica *Accuracy Spread* mostrou que, em LaTeX, todos os modelos apresentaram baixa variação entre execuções, com destaque para o *Gemini 3.1 Pro Preview* e o *Gemini 3 Flash Preview*, ambos com valor zero ao longo do experimento. Já no formato de imagem, a instabilidade tornou-se mais evidente nos modelos *GPT-5 Nano* e *Grok 4.1 Fast*, enquanto o *GPT-5.4* permaneceu estável e os modelos *Gemini 3.1 Pro Preview* e *Gemini 3 Flash Preview* mantiveram o *Accuracy Spread* nulo. Já na métrica *TAR@N*, que avalia concordância total entre as cinco respostas, o padrão foi semelhante. Em LaTeX, o *Gemini 3.1 Pro Preview* e o *Gemini 3 Flash Preview* mantiveram estabilidade absoluta. Os modelos *GPT-5.4* e *Grok 4.1 Fast* apresentaram valores elevados, mas o *GPT-5 Nano* exibiu uma queda progressiva à medida que o número de tentativas aumentava.

Para a entrada de imagem, a queda de *TAR@N* foi acentuada para o *GPT-5 Nano* e o *Grok 4.1 Fast*, aproximando-se de zero, ao passo que o *GPT-5.4* permaneceu próximo de 0,90 e os modelos *Gemini 3.1 Pro Preview* e *Gemini 3 Flash Preview* apresentaram concordância total. Para a métrica *cons@N*, em LaTeX, os modelos *Gemini 3.1 Pro Preview* e *Gemini 3 Flash Preview* permaneceram no topo, seguidos pelo *Grok 4.1 Fast* e pelo *GPT-5.4*, enquanto o *GPT-5 Nano* apresentou maior oscilação. Para imagens, o *Gemini 3.1 Pro Preview* manteve desempenho máximo, já o *Gemini 3 Flash Preview* e o *GPT-5.4* continuaram acima de 0,80, mas o *GPT-5 Nano* e o *Grok 4.1 Fast* apresentaram forte degradação, indicando dificuldade em manter um consenso correto quando expostos a entradas visuais. Considerando a estabilidade, as três métricas convergem para a mesma conclusão: o *Gemini 3.1 Pro Preview* e o *Gemini 3 Flash Preview* foram os modelos mais consistentes nos dois formatos; o *GPT-5.4* apresentou estabilidade intermediária, mas ainda robusta; e os modelos *GPT-5 Nano* e *Grok 4.1 Fast* revelaram alta volatilidade no formato de imagem. Esse padrão está alinhado a [Carvalho et al. 2025], que também observaram maior consistência dos modelos *Gemini 3.1 Pro Preview* e *Gemini 3 Flash Preview* em avaliações com múltiplas execuções.

4.3 PP-3: Análise de significância estatística

A avaliação da normalidade por meio do teste de Shapiro-Wilk indicou que a distribuição das diferenças não segue uma distribuição normal na maioria dos modelos: *Gemini 3 Flash Preview* ($p < 0{,}000001$), *GPT-5 Nano* ($p = 0{,}016731$), *GPT-5.4* ($p < 0{,}000001$) e *Grok 4.1 Fast* ($p = 0{,}000079$). Para o modelo *Gemini 3.1 Pro Preview*, as diferenças mantiveram-se constantes. Em seguida, foi aplicado o teste de Wilcoxon Signed-Rank com a correção de Holm-Bonferroni, os resultados experimentais obtidos estão sintetizados na Tabela I.

Tabela I. Resultados do Teste de Wilcoxon Signed-Rank com Correção de Holm-Bonferroni

Modelo	**W**	**p-valor (Holm-Bonferroni)**	**Significância** ($p < 0{,}05$)
Gemini 3.1 Pro Preview	-	-	Estabilidade Absoluta
Gemini 3 Flash Preview	0	0,314598	Não Significativo
GPT-5.4	4,5	0,314598	Não Significativo
GPT-5 Nano	19,5	0,001491	Significativo
Grok 4.1 Fast	0	0,000056	Significativo

Os resultados na Tabela I indicaram diferença estatisticamente significativa entre os formatos apenas para o *Grok 4.1 Fast* ($p = 0,000056$) e o *GPT-5 Nano* ($p = 0,001491$), evidenciando que, nesses modelos, a entrada em formato de imagem afeta negativamente a acurácia. Para o *Gemini 3 Flash Preview* e o *GPT-5.4* ($p = 0,314598$ em ambos os casos), não houve evidência estatística de diferença entre LaTeX e imagem. O *Gemini 3.1 Pro Preview*, por sua vez, apresentou o mesmo desempenho em ambos os formatos, confirmando sua estabilidade absoluta. Além disso, o *boxplot* correspondente, disponível no repositório do projeto, reforça essa interpretação ao evidenciar uma separação visual mais acentuada entre os formatos nos modelos *Grok 4.1 Fast* e *GPT-5 Nano*.

5. CONCLUSÕES

Este trabalho avaliou o desempenho de cinco LLMs na resolução de 25 questões de banco de dados do POSCOMP, considerando entradas em LaTeX e em imagem, com cinco execuções independentes por configuração e métricas de desempenho voltadas à análise de acurácia e estabilidade, a saber: *Mean Accuracy*, *pass@N*, *Accuracy Spread*, *TAR@N* e *cons@N*. O *Gemini 3.1 Pro Preview* apresentou o melhor desempenho geral, combinando acurácia máxima e estabilidade total em ambos os formatos. O *Gemini 3 Flash Preview* também obteve resultados elevados e consistentes, enquanto o *GPT-5.4* apresentou desempenho intermediário. Em contraste, os modelos mais leves *GPT-5 Nano* e *Grok 4.1 Fast* apresentaram forte degradação no formato de imagem, resultado confirmado pela análise estatística. Portanto, os resultados indicam que tanto a maturidade da arquitetura multimodal quanto o porte e a categoria do modelo atuam como fatores determinantes para assegurar o desempenho e a estabilidade na resolução de questões do domínio de banco de dados.

Ademais, o presente estudo oferece contribuições metodológicas para o avanço da literatura ao estabelecer um protocolo de avaliação rigoroso e reprodutível. Ao diferenciar-se das abordagens convencionais baseadas em submissões únicas via interfaces *web*, esta pesquisa preenche uma lacuna ao introduzir o controle de hiperparâmetros via API combinado a múltiplas execuções independentes. Essa estratégia permitiu avaliar não apenas a acurácia, mas também a estabilidade de LLMs modernos no domínio de banco de dados do POSCOMP. Esses achados têm implicações práticas para o desenvolvimento de tecnologias educacionais e de sistemas de avaliação automatizada baseados em IA, sugerindo que, no domínio analisado, o formato LaTeX pode ser mais vantajoso em alguns modelos, especialmente *GPT-5 Nano* e *Grok 4.1 Fast*, enquanto outros modelos apresentaram desempenho semelhante nas duas modalidades avaliadas.

Como limitação, as provas do POSCOMP (2016–2025) são de acesso público, o que possibilita *data leakage* nos dados de treinamento dos modelos avaliados. Ademais, destaca-se que o estudo foi conduzido com um *dataset* de 25 questões. Ainda que esse quantitativo de questões seja compatível com o adotado por [Carvalho et al. 2025], a ampliação do número de questões e a inclusão de diferentes edições podem contribuir para a generalização dos achados e para a robustez das análises. Além disso, os custos do uso de modelos modernos via API com entradas multimodais exigiram um equilíbrio entre profundidade analítica e viabilidade econômica do experimento, o que levou à adoção de cinco execuções independentes por configuração. Como trabalhos futuros, pretende-se: (i) ampliar o *dataset*; (ii) incluir outras subáreas da computação; (iii) adicionar LLMs *open-weight*; e (iv) analisar qualitativamente em quais tópicos ocorreram os erros das questões.

Agradecimentos e Uso de IA generativa

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) – DT-303031/2023-9 e da Financiadora de Estudos e Projetos (FINEP) (ProAmazonia – 2373/24).

Declara-se que os modelos de IA generativa *Gemini 3* e a ferramenta *Grammarly Pro* foram utilizados exclusivamente para a revisão gramatical, enquanto o *NotebookLM* foi utilizado para auxiliar na

seleção de artigos relevantes para compor as referências da pesquisa. A autoria e a responsabilidade integral pelo conteúdo final, incluindo a verificação de plágio e as correções, são da primeira pessoa autora.

REFERÊNCIAS

Almeida, F. and Caminha, C. Evaluation of entry-level open-source large language models for information extraction from digitized documents. In *Anais do XII Symposium on Knowledge Discovery, Mining and Learning*. Belém, Brasil, pp. 25–32, 2024.

Atil, B., Aykent, S., Chittams, A., Fu, L., Passonneau, R. J., Radcliffe, E., Rajagopal, G. R., Sloan, A., Tudrej, T., Ture, F., Wu, Z., Xu, L., and Baldwin, B. Non-determinism of "deterministic"llm settings. `https://arxiv.org/abs/2408.04667`, 2025.

Carreteiro, F. R. S., Sousa, F. A. d., Marcacini, R. M., and Lobato, F. M. F. Avaliação do impacto de entradas multimodais em LLMs: um estudo de caso de respostas ao POSCOMP. In *Workshop sobre Educação em Computação (WEI)*. Maceió, Brasil, pp. 678–689, 2025.

Carvalho, L., Junior, C. C., and Mendonça, N. Evaluating the accuracy and stability of frontier llms on enade computer science questions. In *Anais da XXXV Brazilian Conference on Intelligent Systems*. Fortaleza, Brasil, pp. 260–274, 2025.

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. Evaluating large language models trained on code. `https://arxiv.org/abs/2107.03374`, 2021.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., Wang, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N., Khattab, O., Henderson, P., Huang, Q., Chi, R., Xie, S. M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., and Koreeda, Y. Holistic evaluation of language models. `https://arxiv.org/abs/2211.09110`, 2022.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., and Flach, P. Crisp-dm twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering* 33 (8): 3048–3061, 2021.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., and Ioannidis, J. P. A. A manifesto for reproducible science. *Nature Human Behaviour* 1 (1): 0021, 2017.

OpenAI. Gpt-4o system card. `https://openai.com/pt-BR/index/gpt-4o-system-card/`, 2024.

Romero, V., Assis, G., Carvalho, J., Mann, P., and Paes, A. Opportunities vs. risks: Exploring automatic annotation of financial polarity biases via large language models. In *Anais do XIII Symposium on Knowledge Discovery, Mining and Learning*. Fortaleza, Brasil, pp. 1–8, 2025.

Saldanha, M. S. and Digiampietri, L. A. Chatgpt and bard performance on the poscomp exam. In *Proceedings of the 20th Brazilian Symposium on Information Systems*. Juiz de Fora, Brazil, 2024.

Taschetto, L. and Fileto, R. Using retrieval-augmented generation to improve performance of large language models on the brazilian university admission exam. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*. Florianópolis, Brazil, pp. 799–805, 2024.

Viegas, C., Gheyi, R., and Ribeiro, M. Assessing the Capability of LLMs in Solving POSCOMP Questions. *Journal of the Brazilian Computer Society* 31 (1): 990–1003, 2025.

Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P. S., and Wen, Q. Large language models for education: A survey and outlook. *IEEE Signal Processing Magazine* 42 (6): 51–63, 2025.

Xiao, H., Zhou, F., Liu, X., Liu, T., Li, Z., Liu, X., and Huang, X. A comprehensive survey of large language models and multimodal large language models in medicine. *Information Fusion* vol. 117, pp. 102888, 2025.