Avaliação do impacto de entradas multimodais em LLMs: um estudo de caso de respostas ao POSCOMP

Florindo R. S. Carreteiro[1], **Fernando A. de Sousa**[1], **Ricardo M. Marcacini**[2], **Fábio M. F. Lobato**[1,2]

[1] Instituto de Engenharia e Geociências (IEG)
Universidade Federal do Oeste do Pará (UFOPA) – Santarém, PA – Brasil

[2]Instituto de Ciências Matemática e de Computação (ICMC)
Universidade de São Paulo (USP) – São Carlos, SP – Brasil

{florindocompsci,fernando.compsci}@gmail.com, ricardo.marcacini@usp.br,
fabio.lobato@ufopa.edu.br

Resumo. *Large Language Models (LLMs) podem apoiar alunos no estudo e docentes na criação de provas devido suportarem entradas multimodais. Porém, ainda há uma lacuna na literatura sobre a avaliação da multimodalidade dos LLMs na resolução de questões de provas. A literatura apresenta uma estrutura experimental que não abrange os possíveis tipos de entrada. Este estudo visa preencher tal lacuna, avaliando o efeito de imagens, PDFs e textos em LaTeX nas acurácias do Gemini 1.5 Pro do Google e ChatGPT-4o da OpenAI via API e interface web em questões de LFA do POSCOMP. Destarte, observou-se que o Gemini superou o ChatGPT em LFA, que questões pictóricas têm menor acurácia e que a API potencializa a vantagem do LaTeX. Em suma, os achados de pesquisa têm o potencial de impactar alunos, docentes e o setor produtivo.*

1. Introdução

Desde o lançamento do ChatGPT pela OpenAI em novembro de 2022, várias outras empresas como o Google, META e Microsoft têm investido no avanço desse tipo de ferramenta de Inteligência Artificial (IA) chamado *Large Language Model* (LLM) [Mickle 2023]. Consequentemente, novas funcionalidades vêm sendo implementadas por tais *big techs* para obter diferenciais frente às demais. Por exemplo, pode-se citar: respostas em diversos idiomas; criação de interfaces *web* intuitivas, que aqui também iremos nos referir simplesmente como *chatbots*; e suporte a entradas multimodais, isto é, outros tipos de conteúdo além de texto, como imagens, PDFs, aúdios e vídeos [Zhao et al. 2023].

A multimodalidade, por ser recente, não teve seu efeito investigado amplamente (ou completamente) no desempenho das LLMs de alguns estudos, como o de [Saldanha and Digiampietri 2024]. Esse trabalho avaliou duas LLMs perante o Exame Nacional para Ingresso na Pós-Graduação em Computação (POSCOMP) promovido pela Sociedade Brasileira de Computação (SBC) [SBC 2024], mas com entradas textuais apenas. O presente estudo, motivado por tal limitação, visa preencher essa lacuna avaliando a multimodalidade no exame com ênfase em imagens, PDFs e textos em formato LaTeX, uma linguagem de marcação para a composição tipográfica de documentos científicos.

Embora a literatura evidencie o potencial das LLMs no suporte ao estudo e na preparação de exames [Dao et al. 2023, Zhang and Tur 2024], elas possuem certas limitações, com destaque para a dificuldade em manter a consistência de suas

respostas quando solicitadas a responder mais de uma vez a uma mesma pergunta [Plevris et al. 2023]. Tal fato ainda é um questionamento em aberto quando se apresenta a pergunta em diferentes formatos (*e.g.*, LaTeX, figuras, PDF *etc.*) e por diferentes interfaces, como a *web* e a *Application Programming Interface* (API). A API é um programa intermediário entre o *backend* de uma aplicação *web* e um usuário (ou cliente) que permite ao último obter algum serviço do primeiro [Donner 2024], o que em LLMs são respostas aos *prompts* enviados, tomando como exemplo a Gemini API[1] do Google.

A fim de preencher as lacunas da literatura supraditas, este estudo objetiva avaliar os desempenhos do Gemini 1.5 Pro e do ChatGPT-4o em resolver questões de múltipla escolha a partir de entradas multimodais e modos de requisição distintos. Com base nas questões discutidas, almeja-se responder às seguintes Perguntas de Pesquisa (PP):

PP1: Apresentar questões de múltipla escolha em requisições separadas ou em requisição única pode influir a acurácia de LLMs multimodais ao resolvê-las?

PP2: Entradas utilizando a linguagem LaTeX, inclusive para representar questões com elementos pictóricos, impactam o desempenho de resolução das LLMs?

Os resultados têm o potencial de guiar o processo de engenharia de *prompt* quanto ao uso de imagens, PDFs e LaTeX para obter respostas mais precisas. Este trabalho também expande a literatura sobre o domínio de LLMs em um subcampo da Computação denominado Linguagens Formais e Autômatos (LFA) ao considerar os impactos da multimodalidade nessa competência. Em suma, são obtidos avanços metodológicos na avaliação de LLMs em contextos educacionais mediante uma cobertura experimental ampla.

O restante deste artigo encontra-se organizado como segue. A Seção 2 discute os trabalhos relacionados e os diferenciais deste. A Seção 3 descreve os materiais e métodos desta pesquisa. Em seguida, a Seção 4 apresenta as análises dos resultados obtidos. Por fim, a Seção 5 dispõe as conclusões deste estudo e projeções para trabalhos futuros.

2. Trabalhos Relacionados

Diversos trabalhos têm investigado a capacidade de LLMs em resolver provas sobre vários temas, em especial as de múltipla escolha. [Dao et al. 2023], por exemplo, submeteram as edições de 2019 a 2023 do *Vietnamese High School Graduation Examination Dataset for Large Language Models* (VNHSGE) à API do ChatGPT-3.5. [Williams and Huckle 2024] avaliaram ainda mais LLMs, a saber: GPT-4 Turbo Preview (da OpenAI), Claude 3 Opus (da Anthropic), Mistral Large e Mistral 8x22B (da Mistral), Gemini Pro 1.0 e Gemini Pro 1.5 (do Google) e o Llama 3 70B (da Meta). Contudo, esses trabalhos limitaram-se a entradas exclusivamente textuais. Nossa pesquisa, embora não tão abrangente quanto às LLMs usadas, diferencia-se por utilizar também imagens, PDFs e LaTeX.

[Saldanha and Digiampietri 2024] compararam o desempenho humano no POSCOMP com o ChatGPT e o Google Bard, com e sem contexto adicional. Concluíram que eles superam os humanos no exame em média, mesmo nas abordagens sem contexto, e que, em nenhuma das edições analisadas (2016 a 2019 e 2022), as LLMs tiveram acurácia superior a dois terços das questões. Porém, os autores não usaram entradas multimodais e abordaram temas mais gerais, enquanto nós focamos em LFA, inclusive nas questões não resolvidas pelos modelos supraditos devido à incapacidade multimodal.

[1] `https://ai.google.dev/gemini-api`

[Viegas 2024] também abordou as questões do POSCOMP ao comparar os desempenhos das LLMs ChatGPT-4, Gemini 1.0 Advanced, Claude 3 Sonnet e Le Chat Mistral Large com os candidatos da prova para 2022 e 2023. Isso foi feito por meio das interfaces *web* das LLMs e de *zero-shot prompt*, considerando uma abordagem com questões em texto traduzidas para o inglês e outra abordagem com capturas de tela das questões. Os resultados demonstraram que o ChatGPT-4 apresenta melhor capacidade de lidar com imagens do que o Gemini 1.0 Advanced e que o primeiro superou as outras LLMs na maioria dos testes. O presente trabalho se diferencia por adotar um *framework* experimental mais amplo, considerando também as APIs dos modelos, utilizando PDFs e um *prompt* mais detalhado quanto ao comportamento a ser seguido pelos modelos e à saída esperada.

[Abu-Haifa et al. 2024] passaram links para imagens no Google Drive de questões do *Graduate Record Examination* (GRE) ao ChatGPT-3.5, ChatGPT-4 e ao Microsoft Copilot. [Mendonça 2024], por sua vez, aplicou questões do Brazil's 2021 National Undergraduate Exam (ENADE) por imagens a um modelo de fato multimodal da OpenAI, o ChatGPT-4 Vision. Porém, ambos os trabalhos não utilizaram APIs e PDFs, enquanto nós fizemos isso por meio de *Retrieval-Augmented Generation* (RAG), uma técnica que permite às LLMs ampliar seu conhecimento ou se especializar num domínio específico a partir de uma base de dados externa [Wu et al. 2024]. Podemos tomar como exemplo o ChatGPT, que usa RAG para trabalhar com PDFs conforme a documentação da API[2,3].

Apesar da similaridade de escopo, a maioria das pesquisas anteriores focou no uso de entradas exclusivamente textuais, exceto [Viegas 2024, Abu-Haifa et al. 2024, Mendonça 2024]. Ainda assim, estes estudos apresentam limitações por não abordarem APIs e PDFs. Nosso trabalho se diferencia pela adoção de uma cobertura experimental mais abrangente, com o uso de RAG, de *links* para imagens somente nas APIs, enquanto nos *chatbots* as imagens e PDFs foram anexados. Frente ao exposto, a revisão da literatura mostrou que ainda há lacunas quanto à investigação ampla de como entradas multimodais e interfaces distintas podem afetar a resolução de questões de múltipla escolha por LLMs.

3. Materiais e Métodos

Considerando a natureza exploratória e experimental do estudo, o presente trabalho se baseou no processo denominado *Data Science Trajectories* (DST), proposto por [Martínez-Plumed et al. 2021]. O DST expande o bem conceituado *Cross-Industry Standard Process for Data Mining* (CRISP-DM) agregando tarefas voltadas para ciência de dados como visualização da informação, uso de conhecimento de fundo e interação do humano no laço (*human-in-the-loop*) [Wu et al. 2022]. Tal conceito é imprescindível na engenharia de *prompt*, desde a definição do problema e do objetivo, definição do formato de saída e validações dos resultados, uma vez que estes são, mormente, qualitativos. Destarte, o processo de pesquisa adotado seguiu as etapas abaixo, consonantes ao DST.

1. **Compreensão do domínio:** Revisamos a literatura sobre LLMs aplicadas em provas de diferentes domínios para descobrir lacunas de conhecimento;
2. **Exploração de objetivos:** Definimos os objetivos e perguntas de pesquisa;
3. **Aquisição dos dados:** Selecionamos as questões não anuladas de LFA da maioria das provas do POSCOMP realizadas antes dos nossos experimentos;

[2] `https://platform.openai.com/docs/guides/optimizing-llm-accuracy`. Acesso em: 17 mar. 2025.
[3] `https://platform.openai.com/docs/assistants/tools/file-search`. Acesso em: 17 mar. 2025.

4. **Preparação dos dados:** As questões selecionadas foram transcritas em LaTeX e revisadas pelos autores para garantir a integridade das versões delas em PDF e em capturas de tela geradas em seguida;
5. **Modelagem:** Primeiro realizou-se um processo de engenharia de *prompt* para especificar o que as LLMs iriam receber, como deveriam se comportar e o formato de saída esperado. Em seguida, foram planejados e executados os casos de comparação entre as LLMs de acordo com o tipo de entrada, a interface de acesso e o modo de requisição, com códigos ou manualmente via interface gráfica;
6. **Avaliação:** As respostas coletadas foram comparadas com os gabaritos oficiais do POSCOMP, foram identificados os cenários em que cada LLM se sobressaiu e respondidas as perguntas de pesquisa, mediante análises gráficas e estatísticas;
7. **Liberação dos dados:** Disponibilizamos os dados dos experimentos em um repositório do GitHub[4], garantindo a facilidade de acesso e a reprodutibilidade.

3.1. *Dataset*

O *dataset* dos experimentos foi construído a partir das questões do POSCOMP e encontra-se em repositório anônimo[4] seguindo os princípios de ciência aberta [Munafò et al. 2017]. Com exceção dos anos de 2020 e 2021, em que o exame não ocorreu devido à pandemia de COVID-19, foram utilizadas questões de 2002 a 2022 cuja temática é um tópico da Computação conhecido como LFA. Esse tema foi escolhido porque as questões possuem múltiplos formatos/elementos (grafos, tabelas de transições, expressões regulares *etc.*).

Então, os autores selecionaram manualmente as questões relacionadas a LFA e, auxiliados por dois monitores da disciplina, transcreveram-nas para LaTeX, incluindo elementos pictóricos. Em seguida, os autores do estudo revisaram as transcrições, comparando-as com as questões originais a fim de garantir a confiabilidade dos dados. Ao final, o *dataset* foi composto por 55 perguntas, sendo que 9 delas (questões 1, 17, 20, 22, 34, 39, 47, 50 e 52) têm elementos pictóricos e o restante são puramente textuais. Além de LaTeX, as questões também foram representadas nos formatos PNG e PDF.

3.2. Modelos, Categorias e Abordagens de Teste

Este estudo focou em avaliar os modelos Gemini 1.5 Pro do Google e ChatGPT-4o da OpenAI, devido ao suporte de entradas multimodais por ambos e à performance superior em trabalhos relacionados. As respostas foram coletadas usando ambas as interfaces *web* (*chatbot*) e de programação (API), pois embora a primeira seja amplamente empregada na literatura [Raihan et al. 2025], ela não oferece vantagens como configurações de determinismo da saída que a última fornece [Williams and Huckle 2024]. Por isso, é válido investigar se as diferenças de desempenho decorrentes desse fator são significantes.

Para cada interface e cada tipo de entrada (imagem, LaTeX e PDF), duas abordagens foram analisadas. A primeira consistiu em realizar uma requisição por questão em uma nova sessão de *chat* junto ao *prompt* (ou seja, sem contexto de perguntas prévias), visando guiar o foco apenas ao contexto do enunciado da vez e do *prompt*. A segunda abordagem consistiu em uma única requisição com todas as questões em lote, além do *prompt*. Desse modo, totalizaram-se 12 categorias de comparação conforme a Figura 1. Em cada categoria, 2 abordagens foram testadas, uma com a LLM do Google e outra

[4] `https://anonymous.4open.science/r/lfa-llm-research`

com a LLM da OpenAI, resultando em 24 abordagens de teste. Ademais, toda questão foi submetida 3 vezes em cada abordagem visando obter estimativas mais consistentes, semelhante ao que foi feito por [Plevris et al. 2023]. Neste caso, obtém-se 55 respostas em todas as 3 tentativas, cada qual com sua métrica individual.

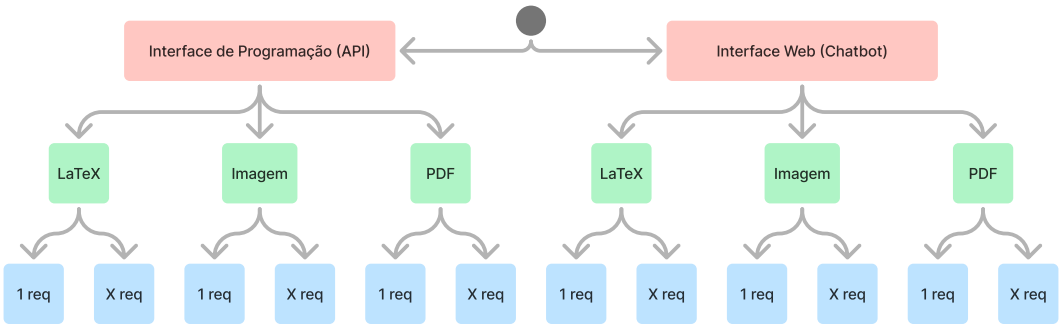

Figura 1. Categorias de comparação entre as LLMs.

Para usar as APIs das LLMs, foram feitos códigos Python (v3.10.12) em *colab notebooks*. Devido à atualização constante dos modelos e visando a replicabilidade dos testes com API, foi preciso fixar nos códigos o modelo escolhido: *gemini-1.5-pro-001*, do Google, ou *gpt-4o-2024-05-13*, da OpenAI. Ademais, visando garantir um formato de resposta consistente, definiu-se o parâmetro de formato de resposta para JSON e o parâmetro "*temperature*" como zero, pois esse valor torna a saída determinística e contribui para que nossos testes sejam replicáveis [Williams and Huckle 2024]. Quanto mais alto esse parâmetro, maior a probabilidade de obter respostas variadas[5,6], o que não é desejado já que pretendemos avaliar se as respostas se repetem em mais de uma solicitação.

Nos testes com *chatbots*, embora eles não permitissem a definição explícita da versão do modelo *pro* usado em termos temporais (as opções de escolha eram apenas a versão básica e a versão *pro*), as execuções foram feitas antes que as interfaces *web* apontassem para as versões subsequentes dos modelos *gemini-1.5-pro-001* e *gpt-4o-2024-05-13* (respectivamente *gemini-1.5-pro-002* e *gpt-4o-2024-08-06*)[7,8]. Assim, nosso estudo teve um cuidado a mais comparado aos trabalhos correlatos que usaram interfaces *web*, pois verificamos se os modelos escolhidos graficamente continuaram apontando para os mesmos modelos-base durante o período de teste com a referida interface.

3.3. Engenharia de *Prompt*

Na literatura podem ser encontrados padrões bem conhecidos de engenharia de *prompt*, como *zero-shot*, *few-shots* e *Chain-of-Thoughts* (CoT) [Brown et al. 2020, Wei et al. 2022, Kojima et al. 2022]. Como coletaríamos respostas por APIs e muitas requisições seriam feitas por essa interface (cerca de 1.980), entradas e saídas com muitos *tokens* como em *few-shots* e em CoT poderiam aumentar consideravelmente os custos e resultar em saídas com padrões não determinísticos, dificultando a coleta e a análise

[5] https://ai.google.dev/gemini-api/docs/models/generative-models?hl=pt-br#model-parameters. Acesso em: 10 jan. 2025.

[6] https://platform.openai.com/docs/api-reference/chat/create#chat-create-temperature. Acesso em: 10 jan. 2025.

[7] https://ai.google.dev/gemini-api/docs/changelog?hl=pt-br#09-24-24. Acesso em: 9 jan. 2025.

[8] https://community.openai.com/t/reminder-gpt-4o-default-model-will-be-updated-to-latest-version-on-october-2nd-2024/962350. Acesso em: 9 jan. 2025.

quantitativa das respostas. Esses dois tipos de *prompt* também requerem maior esforço de engenharia [Kojima et al. 2022], o que para nós poderia ser agravado considerando a variabilidade dos enunciados do *dataset* e, consequentemente, a especificidade com que as instruções teriam que ser adaptadas. Por essas razões, optou-se pelo uso de *zero-shot prompt* como em [Viegas 2024] e [Saldanha and Digiampietri 2024], mas com melhorias que possibilitam um equilíbrio entre clareza da instrução e determinismo na resposta.

Sendo assim, no *prompt*, primeiro foi solicitado que o modelo agisse como uma pessoa respondendo a uma prova de múltipla escolha sobre o tema de Linguagens Formais e Autômatos. Em seguida, solicitou-se que cada resposta estivesse no formato "{Questão: Alternativa}", também denominado JSON, deixando explícito que a chave "Questão" deveria ser um número inteiro e que "Alternativa" deveria ser uma, e somente uma, das seguintes opções: "A", "B", "C", "D" ou "E". Também foi definido que o modelo não deveria retornar o conteúdo da alternativa, apenas sua letra representante. Dessa forma, disponibilizamos o *prompt* nos *colab notebooks* de teste do repositório anônimo[4].

3.4. Limitações e Tratamentos

Na execução da abordagem {API+IMG+1Req+GPT}, descobriu-se que a API da OpenAI no *Usage Tier 1* - categoria que define os limites de uso da API e que evolui conforme a recarga de créditos - limitava-se a 30.000 tokens em uma única requisição[9]. Embora todas as questões juntas em LaTeX não ultrapassassem esse limite, em imagens ele foi ultrapassado. Por essa razão, admitiu-se que aquela abordagem em específico seria executada em 2 requisições: uma com as 27 primeiras imagens e outra com as 28 últimas.

Ademais, para processar PDFs com a API da OpenAI, é necessário o uso de um objeto *Assistant* com ferramenta *File Search*[3] de RAG. Este, porém, mostrou dificuldades em manter um formato de resposta consistente com o solicitado na abordagem {API+PDF+1Req+GPT} apenas com o *prompt* original. Por isso, foi usado o seguinte *assistant prompt* adicional: *"Responda às 55 questões de LFA do PDF informando o número da questão e a letra da resposta em formato JSON"*. Ainda assim, apenas a primeira tentativa retornou um dicionário com todas as questões e suas respectivas respostas. Nas tentativas subsequentes, mesmo que idênticas à primeira em termos de *prompts* e parâmetros de requisição, o *gpt-4o* passou a requisitar que o usuário fornecesse a questão para a qual ele desejava a resposta ao invés de retornar o dicionário referido. Como o documento com todas as questões já estava incorporado, isso evidencia o comportamento não determinístico de LLMs mesmo quando *temperature = 0*, relatado por [Williams and Huckle 2024].

Devido ao incidente incomum mencionado, uma possível solução seria modificar o *prompt* original para tentar obter (sem nenhuma garantia) respostas idênticas à Tentativa 1 da referida abordagem. Porém verificamos a inviabilidade de tal abordagem, pois violaria um dos métodos pré-estabelecidos da pesquisa (o *prompt*) e implicaria em refazer os testes das 23 abordagens bem-sucedidas para manter a padronização. Logo, a solução adotada foi tratar os dados ausentes considerando os desempenhos das tentativas 2 e 3 da abordagem em questão iguais ao da Tentativa 1 (imputação de dados por última consulta).

Também houve limitações nas abordagens {WEB+IMG+1Req+GPT} e {WEB+IMG+1Req+Gemini}, pois o ChatGPT permitia no máximo 10 imagens anexa-

[9] `https://platform.openai.com/docs/guides/rate-limits/usage-tiers?context=tier-one`. Acesso em: 10 jan. 2025.

das por requisição e o Gemini no máximo uma, além de exigir o envio de um *prompt* junto a ela. Então, para simular o envio em lote das 55 questões, optou-se por aproveitar uma mesma sessão de *chat*, mas com várias requisições, até que todas as imagens fossem usadas. O ChatGPT exigiu o *prompt* uma só vez na sessão e o Gemini o fez 55 vezes.

3.5. Avaliação das Respostas

O processo de avaliação das respostas foi qualitativo-quantitativo. Qualitativamente, conferiu-se se elas seguiam o formato JSON solicitado e se havia falhas como as da subseção anterior. Isso foi importante para que as alternativas respondidas pudessem ser extraídas com expressões regulares (*Regex*) ao usar APIs e tornar o processo mais automatizado, embora nos *chatbots* as respostas fossem extraídas e armazenadas manualmente. Quantitativamente, realizou-se a análise das taxas de acerto ($accuracy$), sendo a acurácia resultante em uma abordagem a média das acurácias de suas 3 tentativas.

Visando a robustez experimental, testes estatísticos foram realizados baseados nas 24 acurácias resultantes, dividindo-as em grupos com base nos fatores: modo de requisição (uma requisição ou múltiplas), interface de LLM (API ou *web*) e tipo de entrada (imagem, LaTeX e PDF). Esses grupos geraram 16 amostras, com um *boxplot* associado a cada uma conforme a Figura 4, importantes para responder às perguntas de pesquisa com base em testes de hipóteses sobre as médias. Para tanto, verificou-se o requisito de normalidade nas 16 amostras com o teste Shapiro-Wilk[10], a igualdade de variâncias com o teste Bartlett[11] e as médias das amostras foram comparadas com testes T-Student[12].

4. Resultados e Discussões

Partindo dos materiais, métodos e objetivos estipulados, foram avaliadas 55 questões do POSCOMP em 24 combinações entre interface, formato de entrada, modo de requisição e LLMs. A seguir apresentamos os resultados para cada cenário, buscando discutí-los à luz da literatura correlata. Ao todo foram coletadas 3.960 respostas das LLMs e obtidas 72 acurácias parciais - 3 por abordagem. Com base nesses dados, as 24 abordagens foram ordenadas pelas respectivas acurácias médias conforme o *ranking* da Figura 2A.

Nota-se que as LLMs não apresentaram diferenciação significativa entre conhecimentos de computação, uma vez que em LFA não acertaram dois terços das questões, consoante ao desempenho geral do estudo de [Saldanha and Digiampietri 2024]. Vale notar que o Gemini teve melhor desempenho em 8 de 12 categorias (Figura 2B), com diferença média de aproximadamente 6,67% em desempenho. Isso nos permite sugerir que o Gemini é a opção preferível para a resolução de problemas de LFA. Este achado tem o potencial de guiar estudantes no seu processo formativo, aumentando a segurança no uso de LLMs no suporte ao processo de ensino-aprendizagem.

Por sua vez, a Figura 3 permite analisar separadamente questões com e sem elementos pictóricos. Como esperado, questões puramente textuais são de mais fáceis resoluções para as LLMs, com desempenho inclusive superior à barreira de dois terços supradita. É possível verificar também que o desempenho baseado em questões puramente textuais possui uma curva mais suave em comparação com o comportamento de

[10] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html
[11] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bartlett.html
[12] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html

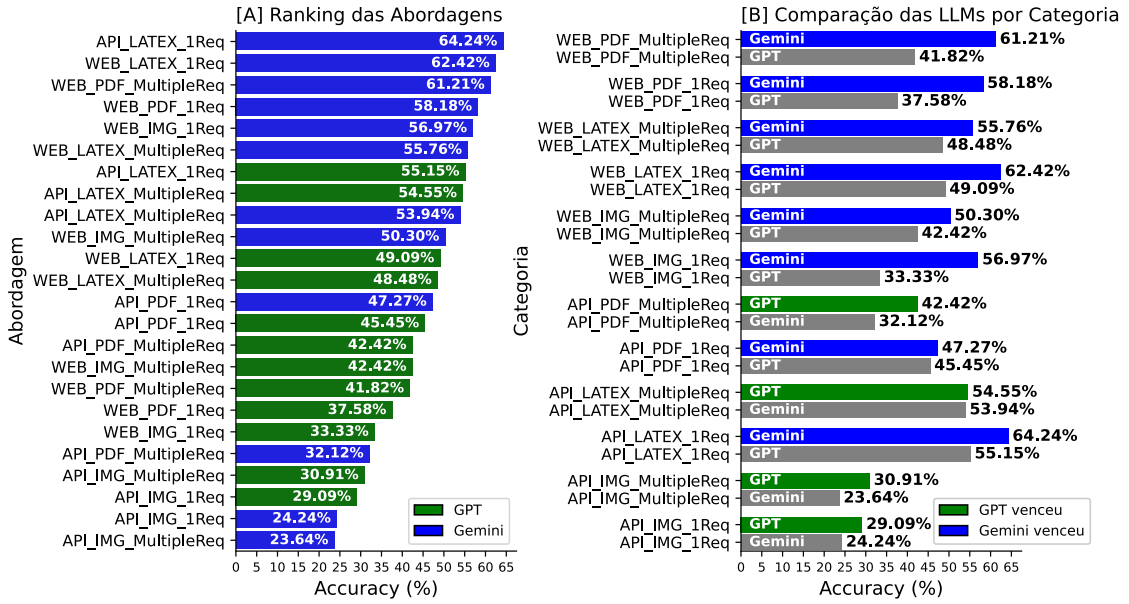

Figura 2. *Ranking* das abordagens e comparação das LLMs por categoria.

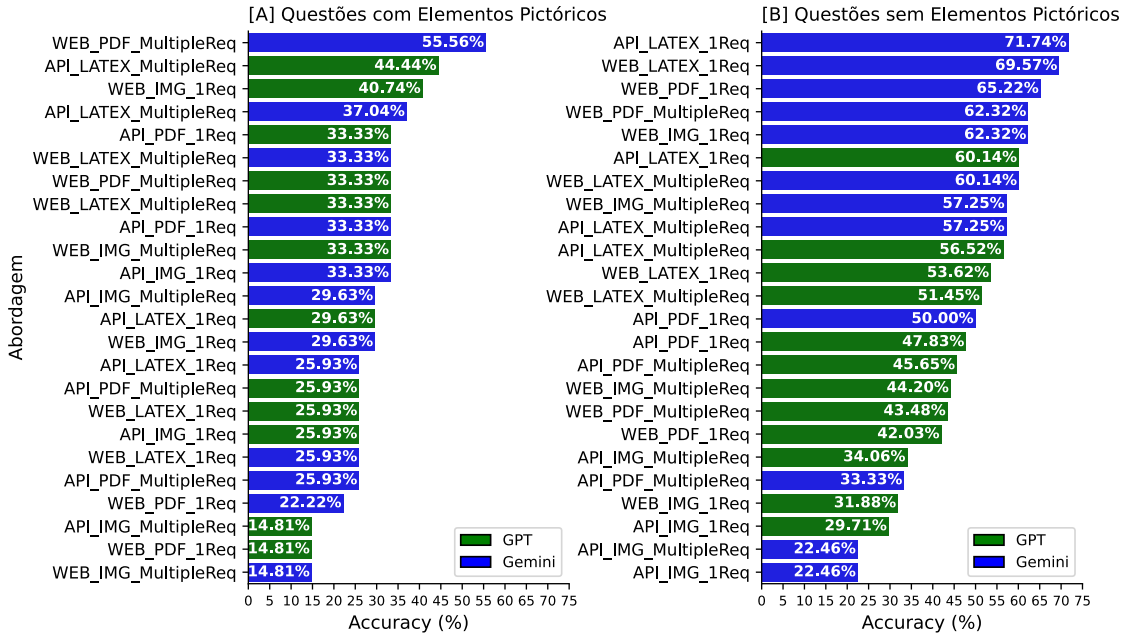

Figura 3. *Ranking* das abordagens - subamostras.

escada das questões pictóricas. Este achado de pesquisa permite-nos sugerir que questões pictóricas são excelente alternativa em aplicações que visam evitar potenciais fraudes em provas pelo uso de LLMs. Assim, o referido achado tem o potencial de guiar docentes de LFA e bancas de concursos na construção de testes mais efetivos e seguros.

Considerando agora as PPs e a Figura 4, notam-se poucas diferenças entre as variações dos fatores Modo de Requisição e Interface de LLM (*boxplots* vermelhos, com 12 acurácias cada). Porém, em um mesmo grupo de *boxplots*, nota-se uma aparente liderança das abordagens com LaTeX. Então, para obter conclusões mais embasadas, recorremos aos testes de hipóteses do Quadro 1, admitindo nível de significância $\alpha = 5\%$.

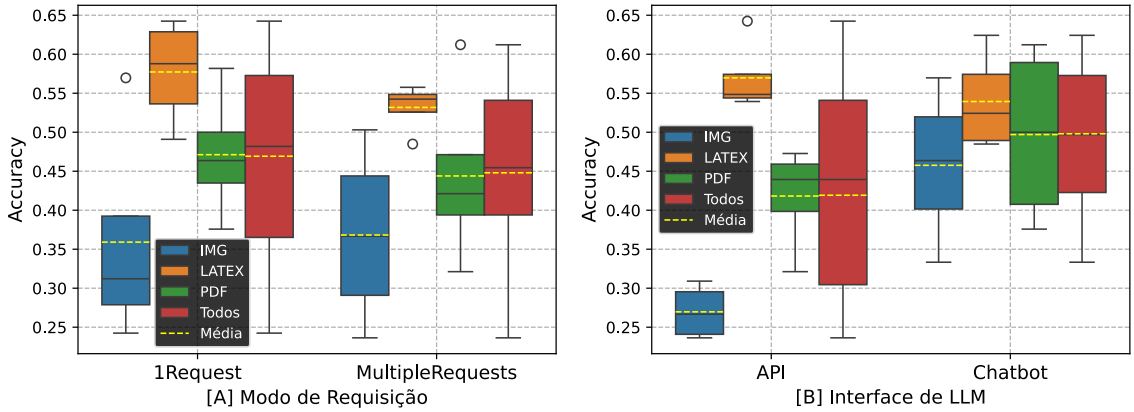

Figura 4. *Boxplots* das acurácias em função dos fatores Modo de Requisição, Interface de LLM e Tipo de Entrada.

Trouxemos apenas os *T-Tests* mais significativos (*p-value* $< 10\%$), embora tenhamos feito 20 testes relativos às médias dos 16 *boxplots*, como: *as médias do fator 1Request são significantemente maiores do que as correspondentes em MultipleRequests?* (T1 a T4); *as médias do fator Chatbot são significantemente maiores do que as correspondentes em API?* (T11 a T14); *os boxplots de LaTeX têm médias significantemente maiores do que as médias dos boxplots de PDF e de imagem do mesmo grupo?* (T5, T6, T8, T9, T15, T16, T18 e T19); e *os boxplots de imagem têm médias significantemente menores do que as médias dos boxplots de PDF do mesmo grupo?* (T7, T10, T17 e T20).

Quadro 1. Testes T com hipóteses alternativas mais significantes ($p < 10\%$).

Test	Sample 1	Sample 2	Bartlett statistic	Bartlett p-value	T-Test equal_var param	T-Test alternative param	T-Test statistic	T-Test p-value
T5	1Req, LaTeX	1Req, PDF	9.95e-02	7.52e-01	True	greater	1.93e+00	5.08e-02
T6	1Req, LaTeX	1Req, IMG	1.28e+00	2.58e-01	True	greater	2.71e+00	**1.76e-02**
T9	MultipleReq, LaTeX	MultipleReq, IMG	3.50e+00	6.12e-02	True	greater	2.66e+00	**1.87e-02**
T11	Chatbot, IMG	API, IMG	2.41e+00	1.21e-01	True	greater	3.48e+00	**6.60e-03**
T14	Chatbot, All	API, All	1.37e+00	2.42e-01	True	greater	1.64e+00	5.73e-02
T15	API, LaTeX	API, PDF	2.72e-01	6.02e-01	True	greater	3.63e+00	**5.46e-03**
T16	API, LaTeX	API, IMG	2.39e-01	6.25e-01	True	greater	9.92e+00	**3.00e-05**
T17	API, IMG	API, PDF	9.74e-01	3.24e-01	True	less	-3.88e+00	**4.10e-03**

Em relação à PP1, podemos inferir que sua resposta é negativa, pois foram rejeitadas as hipóteses alternativas dos testes T1 a T4 de que as médias das amostras de *1Request* seriam significantemente maiores do que as correspondentes em *MultipleRequest* (*T-Test p-value* $> 5\%$). Esse achado tem potencial de impacto na forma como empresas que entregam serviços baseados em LLMs o fazem. Uma vez que usar uma só requisição tem eficácia semelhante ao uso de várias requisições em separado, tais *startups* poderiam fazer uso dessa estratégia como uma forma de otimizar seus custos, pois na requisição única é possível evitar redundâncias, consequentemente, consumindo menos créditos da API.

Quanto à PP2, é possível afirmar que LaTeX se sobressai aos outros formatos mais significativamente pela API do que pela interface *web* (T15 e T16). Isso reforça às *startups* que prestam serviços baseados em LLMs de *big techs* que elas precisam adequar seus formatos de entrada para ter melhores resultados. LaTeX também se sobressaiu significantemente perante imagens com *1Request* e *MultipleRequests* (T6 e T9, respectivamente),

uma vez que *T-Test p-value* $< 5\%$, corroborando [Viegas 2024] quanto à inferioridade das imagens. Nos demais casos, LaTeX não foi significativamente superior. Tais achados fazem necessário destacar o escopo estrito (LFA) deste estudo. Para uma visão mais abrangente, é preciso investigar outras áreas com elementos pictóricos em suas questões.

5. Conclusões

Apesar dos avanços no estado da arte sobre LLMs no processo de ensino-aprendizagem, foram identificadas lacunas na literatura quanto à avaliação de desempenho de LLMs no tema de LFA e em função de diferentes tipos de entrada. Este artigo busca preencher tais lacunas, mediante a avaliação das acurácias dos modelos *gemini-1.5-pro-001* e *gpt-4o-2024-05-13* num *dataset* de 55 questões do POSCOMP sobre LFA em três formatos distintos (LaTeX, imagens e PDF), usando duas interfaces de LLM (*web* e API) e por dois modos de requisição (requisição única e múltiplas), num total de 24 abordagens de teste.

As principais contribuições desta pesquisa são as seguintes. Os resultados mostram que perguntas em formato LaTeX costumam ter melhores resultados em comparação com imagens e PDFs, em geral mais significativos usando APIs. Isso é útil especialmente para estudantes interessados no uso de LLMs como ferramentas de estudo em LFA. Também notou-se que questões pictóricas apresentam desempenho inferior às puramente textuais, o que pode guiar docentes e bancas de concursos na elaboração de testes à prova de fraudes por uso de LLM. Em síntese, os achados têm o potencial de guiar o processo de engenharia de *prompt* quanto à multimodalidade, aos tipos de interface e de requisição, trazendo impactos a alunos, docentes e ao mercado de ferramentas de ensino com base em LLMs, sugerindo estratégias que entregam melhor desempenho com o menor custo.

Algumas ameaças à validade desta investigação podem ser consideradas. Primeiro, visando uma comparação justa entre as APIs das empresas Google e OpenAI, o parâmetro *seed* da API da segunda, teoricamente relevante para garantir respostas determinísticas, não foi utilizado devido a API da primeira não suportá-lo durante a condução dos testes, até o conhecimento dos autores. Por isso, apenas *temperature = 0* foi explorado para a finalidade supradita, consonante a [Williams and Huckle 2024]. Embora a reprodução dos experimentos possa ter pequenas variações nas respostas como consequência, nossa abordagem minimiza essas chances por termos repetido 3 vezes a coleta de cada resposta. A segunda limitação é a coleta parcial das respostas da abordagem {API+PDF+1Req+GPT}, sugerindo que outro trabalho poderia começar resolvendo esse problema e só então estender o *prompt* e demais parâmetros às demais abordagens.

Por isso, trabalhos futuros envolverão a aplicação de outras técnicas de engenharia de *prompt* para alcançar o referido objetivo e de outras LLMs promissoras com técnicas de treinamento diferentes das usadas pelo Gemini e ChatGPT, como o DeepSeek[13] e seu aprendizado por reforço [Bi et al. 2024]. Como a maioria das LLMs do mercado são treinadas com mais ênfase na língua inglesa, embora [Mendonça 2024] tenha relatado poucas diferenças em comparação à lingua portuguesa para os modelos da OpenAI, também almeja-se avaliar questões com elementos pictóricos em inglês ou usar nosso mesmo *dataset* com uma LLM especializada em português, como um modelo da Maritaca AI[14].

[13] `https://github.com/deepseek-ai/DeepSeek-R1/blob/main/DeepSeek_R1.pdf`. Acesso em 4 fev. 2025.
[14] `https://www.maritaca.ai/`

Agradecimentos

Este trabalho foi apoiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) - DT-303031/2023-9, PDS - 101057/2024-5, PQ - 2023/10100-4. Agradecemos também aos monitores da disciplina de Linguagens Formais e Autômatos, que ao longo dos semestres letivos, selecionaram as questões e as passaram para o LaTeX. Por fim, agradecemos às pessoas revisoras pelas valorosas contribuições ao estudo.

Referências

Abu-Haifa, M., Etawi, B., Alkhatatbeh, H., and Ababneh, A. (2024). Comparative analysis of chatgpt, gpt-4, and microsoft copilot chatbots for gre test. *International Journal of Learning, Teaching and Educational Research*, 23:327–347.

Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., et al. (2024). Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Dao, X.-Q., Le, N.-B., Phan, X.-D., and Ngo, B.-B. (2023). Can chatgpt pass the vietnamese national high school graduation examination? *arXiv preprint arXiv:2306.09170*.

Donner, C. G. G. (2024). Misinformation detection methods using large language models and evaluation of application programming interfaces. Master's thesis, University of Oklahoma.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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

Mendonça, N. C. (2024). Evaluating chatgpt-4 vision on brazil's national undergraduate computer science exam. *ACM Trans. Comput. Educ.*, 24(3).

Mickle, T. (2023). Big tech rebounds and preps for transformative a.i. investments. `https://link.gale.com/apps/doc/A759698393/AONE?u=anon~4972424c&sid=googleScholar&xid=2696c45b`. Acesso em: 8 ago 2024.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., and Ioannidis, J. (2017). A manifesto for reproducible science. *Nature human behaviour*, 1(1):1–9.

Plevris, V., Papazafeiropoulos, G., and Rios, A. J. (2023). Chatbots put to the test in math and logic problems: A preliminary comparison and assessment of chatgpt-3.5, chatgpt-4, and google bard.

Raihan, N., Siddiq, M. L., Santos, J. C., and Zampieri, M. (2025). Large language models in computer science education: A systematic literature review. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, pages 938–944.

Saldanha, M. S. and Digiampietri, L. A. (2024). Chatgpt and bard performance on the poscomp exam. In *Proceedings of the 20th Brazilian Symposium on Information Systems*, SBSI '24, New York, NY, USA. Association for Computing Machinery.

SBC, S. B. d. C. (2024). Exame Nacional para Ingresso na Pós-Graduação em Computação (POSCOMP). https://www.sbc.org.br/educacao/poscomp. Acesso em: 16 jul. 2024.

Viegas, C. V. (2024). Avaliando a capacidade de llms na resolução de questões do poscomp. *Repositório Institucional da UFCG*. http://dspace.sti.ufcg.edu.br:8080/xmlui/handle/riufcg/38035.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Williams, S. and Huckle, J. (2024). Easy problems that llms get wrong. https://arxiv.org/abs/2405.19616.

Wu, S., Xiong, Y., Cui, Y., Wu, H., Chen, C., Yuan, Y., Huang, L., Liu, X., Kuo, T.-W., Guan, N., et al. (2024). Retrieval-augmented generation for natural language processing: A survey. *arXiv preprint arXiv:2407.13193*.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., and He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.

Zhang, P. and Tur, G. (2024). A systematic review of chatgpt use in k-12 education. *European Journal of Education*, 59(2):e12599.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.