UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

CAYO VINICÍUS VIEGAS

AVALIANDO A CAPACIDADE DE LLMS NA RESOLUÇÃO DE QUESTÕES DO POSCOMP

CAMPINA GRANDE - PB

2024

CAYO VINICÍUS VIEGAS

AVALIANDO A CAPACIDADE DE LLMS NA RESOLUÇÃO DE QUESTÕES DO POSCOMP

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador : Rohit Gheyi

CAMPINA GRANDE - PB

2024

CAYO VINICÍUS VIEGAS

AVALIANDO A CAPACIDADE DE LLMS NA RESOLUÇÃO DE QUESTÕES DO POSCOMP

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Rohit Gheyi

Orientador – UASC/CEEI/UFCG

Francilene Procópio Garcia

Examinador – UASC/CEEI/UFCG

Francisco Vilar Brasileiro

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 15 de MAIO de 2024.

CAMPINA GRANDE - PB

RESUMO

Avanços recentes em Modelos de Linguagem de Grande Escala (LLMs) expandiram significativamente as capacidades da inteligência artificial (IA) em tarefas de processamento de linguagem natural. No entanto, seu desempenho em domínios especializados, como a ciência da computação, permanece relativamente pouco explorado. Este estudo investiga se os LLMs podem igualar ou superar o desempenho humano no POSCOMP, um exame brasileiro prestigiado usado para admissões de pós-graduação em ciência da computação. Quatro LLMs—ChatGPT-4, Gemini 1.0 Advanced, Claude 3 Sonnet e Le Chat Mistral Large—foram avaliados nos exames POSCOMP de 2022 e 2023. A avaliação consistiu em duas análises: uma envolvendo interpretação de imagens e outra somente de texto, para determinar a proficiência dos modelos em lidar com questões complexas típicas do exame. Os resultados indicaram que os LLMs tiveram um desempenho significativamente melhor nas questões baseadas em texto, com a interpretação de imagens representando um grande desafio. Por exemplo, na avaliação baseada em imagens, o ChatGPT-4 respondeu corretamente 40 de 70 perguntas, enquanto o Gemini 1.0 Advanced conseguiu apenas 11 respostas corretas. Na avaliação baseada em texto de 2022, o ChatGPT-4 liderou com 57 respostas corretas, seguido por Gemini 1.0 Advanced (49), Le Chat Mistral (48) e Claude 3 Sonnet (44). O exame de 2023 mostrou tendências semelhantes.

EVALUATING THE ABILITY OF LLMS TO SOLVE POSCOMP QUESTIONS

ABSTRACT

Recent advancements in Large Language Models (LLMs) have significantly expanded the capabilities of artificial intelligence in natural language processing tasks. However, their performance in specialized domains like computer science remains relatively underexplored. This study investigates whether LLMs can match or surpass human performance on the POSCOMP, a prestigious Brazilian examination used for graduate admissions in computer science. Four LLMs—ChatGPT-4, Gemini 1.0 Advanced, Claude 3 Sonnet, and Le Chat Mistral Large—were evaluated on the 2022 and 2023 POSCOMP exams. The evaluation consisted of two assessments: one involving image interpretation and another text-only format, to determine the models' proficiency in handling complex questions typical of the exam. Results indicated that LLMs performed significantly better on text-based questions, with image interpretation posing a major challenge. For instance, in the image-based assessment, ChatGPT-4 answered 40 out of 70 questions correctly, while Gemini 1.0 Advanced managed only 11 correct answers. In the text-based assessment of 2022, ChatGPT-4 led with 57 correct answers, followed by Gemini 1.0 Advanced (49), Le Chat Mistral (48), and Claude 3 Sonnet (44). The 2023 exam showed similar trends.

Avaliando a Capacidade de LLMs na Resolução de Questões do POSCOMP

Cayo Vinicíus Viegas
Universidade Federal de Campina Grande
Campina Grande, Paraíba
cayo.viegas@ccc.ufcg.edu.br

Rohit Gheyi
Universidade Federal de Campina Grande
Campina Grande, Paraíba
rohit@dsc.ufcg.edu.br

ABSTRACT

Recent advancements in Large Language Models (LLMs) have significantly expanded the capabilities of artificial intelligence in natural language processing tasks. However, their performance in specialized domains like computer science remains relatively underexplored. This study investigates whether LLMs can match or surpass human performance on the POSCOMP, a prestigious Brazilian examination used for graduate admissions in computer science. Four LLMs—ChatGPT-4, Gemini 1.0 Advanced, Claude 3 Sonnet, and Le Chat Mistral Large—were evaluated on the 2022 and 2023 POSCOMP exams. The evaluation consisted of two assessments: one involving image interpretation and another text-only format, to determine the models' proficiency in handling complex questions typical of the exam. Results indicated that LLMs performed significantly better on text-based questions, with image interpretation posing a major challenge. For instance, in the image-based assessment, ChatGPT-4 answered 40 out of 70 questions correctly, while Gemini 1.0 Advanced managed only 11 correct answers. In the text-based assessment of 2022, ChatGPT-4 led with 57 correct answers, followed by Gemini 1.0 Advanced (49), Le Chat Mistral (48), and Claude 3 Sonnet (44). The 2023 exam showed similar trends.

RESUMO

Avanços recentes em Modelos de Linguagem de Grande Escala (LLMs) expandiram significativamente as capacidades da inteligência artificial (IA) em tarefas de processamento de linguagem natural. No entanto, seu desempenho em domínios especializados, como a ciência da computação, permanece relativamente pouco explorado. Este estudo investiga se os LLMs podem igualar ou superar o desempenho humano no POSCOMP, um exame brasileiro prestigiado usado para admissões de pós-graduação em ciência da computação. Quatro LLMs—ChatGPT-4, Gemini 1.0 Advanced, Claude 3 Sonnet e Le Chat Mistral Large—foram avaliados nos exames POSCOMP de 2022 e 2023. A avaliação consistiu em duas análises: uma envolvendo interpretação de imagens e outra somente de texto, para determinar a proficiência dos modelos em lidar com questões complexas típicas do exame. Os resultados indicaram que os LLMs tiveram um desempenho significativamente melhor nas questões baseadas em texto, com a interpretação de imagens representando um grande desafio. Por exemplo, na avaliação baseada em imagens, o ChatGPT-4 respondeu corretamente 40 de 70 perguntas, enquanto o Gemini 1.0 Advanced conseguiu apenas 11 respostas corretas. Na avaliação baseada em texto de 2022, o ChatGPT-4 liderou com 57 respostas corretas, seguido por Gemini 1.0 Advanced (49), Le Chat Mistral (48) e Claude 3 Sonnet (44). O exame de 2023 mostrou tendências semelhantes.

Keywords

LLMs, POSCOMP, NLP.

1. INTRODUÇÃO

Avanços recentes em *Large Language Models* (LLMs) [8, 4] têm ampliado as fronteiras do que a inteligência artificial pode realizar, especialmente em tarefas de processamento de linguagem natural. Esses desenvolvimentos têm suscitado questões sobre a capacidade dos LLMs de compreender e resolver problemas em domínios especializados, como a ciência da computação.

Apesar de suas habilidades notáveis em processamento de linguagem natural, o desempenho dos LLMs em áreas especializadas, como a parte de fundamentos da ciência da computação, permanece relativamente inexplorado. A questão central é se os LLMs podem competir ou mesmo igualar os seres humanos em avaliações rigorosas, como o Exame Nacional para Ingresso na Pós-Graduação em Computação (POSCOMP) no Brasil.

O POSCOMP é uma avaliação conceituada, projetada para testar os conhecimentos de futuros estudantes de pós-graduação em ciência da computação e é utilizada como critério de entrada em várias universidades pelo Brasil. Sua utilização como campo de testes para LLMs permite uma comparação direta entre a capacidade de inteligência artificial e os padrões humanos, oferecendo insights sobre as potencialidades e limitações dos modelos de Inteligência Artificial (IA).

Este estudo avaliou o desempenho de quatro modelos LLM, ChatGPT-4 [22], Gemini 1.0 Advanced [23], Claude 3 Sonnet [24] e Le Chat Mistral Large [25] nos exames POSCOMP de 2022 e 2023. Foram conduzidas duas avaliações, uma envolvendo questões que requeriam a interpretação de imagens e outra com versões baseadas apenas em texto, para determinar a capacidade dos LLMs de navegar e resolver as questões complexas apresentadas no exame.

Os resultados indicaram que os LLMs tiveram um desempenho significativamente melhor em questões baseadas apenas em texto, destacando a dificuldade de interpretação de imagens como um obstáculo para esses modelos nas suas versões mais atuais. Na primeira avaliação, onde a interpretação de imagens era necessária, o ChatGPT-4 acertou 40 de 70 questões, enquanto o

Gemini 1.0 Advanced acertou apenas 11. Na segunda avaliação, na prova de 2022, o ChatGPT-4 liderou com 57 acertos, seguido pelo Gemini 1.0 Advanced (49), Le Chat Mistral (48) e Claude 3 Sonnet (44). A prova de 2023 teve resultados semelhantes: ChatGPT-4 (59), Gemini 1.0 Advanced (53), Le Chat Mistral (50) e Claude 3 Sonnet (45). Todos os dados dessa pesquisa estão disponíveis online [26].

Esse artigo está estruturado da seguinte forma. A Seção 2 descreve sobre o POSCOMP. As Seções 3 e 4 apresentam avaliações dos usos de LLMs para responder questões do POSCOMP usando *prompts* contendo imagens, e apenas texto, respectivamente. A Seção 5 descreve as ameaças à validade dos estudos. Por fim, as Seções 6 e 7 apresentam os trabalhos relacionados e as conclusões, respectivamente.

2. POSCOMP

O Exame Nacional para Ingresso na Pós-Graduação em Computação (POSCOMP) [21] é uma prova no Brasil, elaborada para avaliar as competências fundamentais de candidatos que buscam cursar programas de pós-graduação em Ciência da Computação e áreas relacionadas. Realizado pela primeira vez em 2000 e organizado pela Sociedade Brasileira de Computação (SBC) desde 2002, o POSCOMP desempenha um papel crucial na agilização do processo de seleção para a maioria dos programas de pós-graduação em Ciência da Computação em todo o país. O exame é realizado anualmente, sendo a edição de 2023 a vigésima organizada pela SBC (não houve exame em 2020 e 2021 devido à pandemia de COVID-19).

A prova em si avalia três áreas fundamentais: Matemática (Figura 1), Fundamentos da Computação (Figura 2) e Tecnologia de Computação (Figura 3). Os candidatos devem responder a 70 questões de múltipla escolha cuidadosamente elaboradas, que se alinham aos currículos padrão dos principais programas de graduação em Ciência da Computação do Brasil. A prova tem duração de 4 horas. Na edição de 2023, se inscreveram 761 candidatos, sendo que 617 efetivamente realizaram a prova, que foi aplicada 100% online. A edição anterior, de 2022, foi realizada presencialmente.

QUESTÃO 12 - Determine a distância aproximada entre o ponto J(3, 1) e a reta s : 6x - 2y + 11 = 0.

A)1,3
B)2,6
C)4,3
D)12,1
E)18,5

Figura 1: Questão 12 do POSCOMP 2023 (Tema geral: Matemática).

QUESTÃO 32 - Um grafo não direcionado no qual todos os pares de vértices são adjacentes, isto é, possui arestas ligando todos os vértices entre si, é um grafo:

A) Desconexo.
B) Completo.
C) Ponderado.
D) Livre.
E) Hipergrafo.

Figura 2: Questão 32 do POSCOMP 2022 (Tema geral: Fundamentos da Computação).

QUESTÃO 65 - Uma rede conectada à Internet possui a máscara de sub-rede 255.255.255.128. Qual o número máximo de computadores que a rede suporta?

A) 126
B) 128
C) 254
D) 255.255.255.128
E) 256

Figura 3: Questão 65 do POSCOMP 2023 (Tema geral: Tecnologia de Computação).

Com o compromisso de transparência, o POSCOMP fornece aos participantes resultados individuais detalhados, incluindo respostas corretas/incorretas, bem como médias gerais e desvio padrão. Além disso, as provas e gabaritos oficiais são publicados online [21].

Desde 2006, o alcance do POSCOMP se estende além do Brasil por meio de uma parceria estratégica com a Sociedade Peruana de Computação. Essa colaboração permite que o exame seja aplicado também no Peru, ampliando ainda mais as oportunidades para aspirantes a pós-graduandos em Ciência da Computação.

3. AVALIAÇÃO 1: IMAGENS

A seguir apresentamos a avaliação considerando *prompts* usando apenas enunciados das questões em imagens.

3.1 GQM

A seguir estruturamos a nossa avaliação usando GQM [10]. O objetivo deste estudo é avaliar a eficácia dos *Large Language Models*, especificamente — ChatGPT e Gemini — na resolução de questões do exame POSCOMP 2023, com o propósito de identificar os domínios e tipos de perguntas que esses modelos

mostram proficiência e onde podem enfrentar desafios no contexto de *prompts* com o enunciado das questões em imagens. Para alcançar este objetivo, responderemos as seguintes questões de pesquisa (QP):

- **QP_1.** Até que ponto o ChatGPT-4 é capaz de resolver questões do POSCOMP?
- **QP_2.** Até que ponto o Gemini 1.0 Advanced é capaz de resolver questões do POSCOMP?

Nesta avaliação, cada modelo receberá uma captura de tela da pergunta juntamente com a consulta "Qual é a resposta?". As respostas corretas e incorretas serão contadas de acordo com o gabarito oficial do POSCOMP, fornecendo uma medida da precisão e eficácia dos modelos.

3.2 Planejamento

Para esta avaliação, as versões dos modelos usadas foram ChatGPT-4 e Gemini 1.0 Advanced e a prova utilizada foi o POSCOMP 2023. Para o *prompt*, foi tirada uma captura de tela, usando o aplicativo Captura e Esboço do Windows 10, da questão contida no PDF da prova. Nessa captura de tela, estão contidos o enunciado, as alternativas e imagens e/ou tabelas, caso exista na questão. Junto da captura de tela, o texto "Qual é a resposta?" também é enviado em cada *prompt*. A avaliação foi conduzida entre os dias 29 de fevereiro e 15 de março de 2024. A avaliação considerou *zero-shot prompt*, onde nenhum exemplo é apresentado para o LLM [9, 5].

3.3 Resultados

O resultado da avaliação evidencia um domínio do ChatGPT-4, com 40 acertos de 70 questões, superando o Gemini 1.0 Advanced, que obteve 11 acertos de 70 questões. Esse padrão é evidente em todos os temas gerais, como apresentado na Tabela 1.

Tema Geral	ChatGPT	Gemini
Matemática	11/20	2/20
Fundamentos da Computação	16/30	7/30
Tecnologia de Computação	13/20	2/20
TOTAL	40/70	11/70

Tabela 1: Acertos POSCOMP 2023 - *prompt* com imagem + "Qual é a resposta?".

3.4 Discussão

As seções a seguir discutem os resultados da avaliação.

3.4.1 Corretude

Entre todas as questões, ChatGPT-4 e Gemini apontaram a mesma alternativa, de forma correta, para 7 questões específicas (20, 22, 23, 26, 45, 52 e 53) dentre as 70 questões da prova. Em 26 questões, nenhum dos modelos conseguiu responder corretamente de acordo com o gabarito oficial. Essas perguntas abrangiam uma variedade de tópicos, indicando lacunas potenciais ou desafios nos dados de treinamento ou algoritmos dos modelos que poderiam ser abordados para melhorar seu desempenho.

A avaliação também revelou uma discrepância no desempenho individual dos dois modelos. O ChatGPT-4 respondeu corretamente 33 questões sozinho, mostrando sua robustez em uma variedade mais ampla de tópicos em comparação com o Gemini, que respondeu corretamente apenas quatro questões que o ChatGPT-4 não respondeu corretamente. Isso sugere que o ChatGPT-4 pode ter uma compreensão mais ampla ou melhores capacidades de generalização no contexto deste conjunto específico de exames, além de um mecanismo de interpretação de imagem mais avançado.

No POSCOMP 2023, houve uma questão que foi anulada — questão 17 relacionada à Matemática — embora tenha sido tentada por ambos os modelos. Ambos selecionaram a mesma opção incorreta, destacando um possível mal-entendido comum ou interpretação errônea no conteúdo ou contexto matemático fornecido.

Os dados também mostram que o ChatGPT-4 não apontou para nenhuma alternativa em 3 questões, enquanto o Gemini não conseguiu sugerir uma alternativa em 27 questões. Esse alto número para o Gemini foi consequência de uma série de alucinações a partir da segunda metade do exame que afetou significativamente o desempenho do modelo (fato que será explicado em maiores detalhes na Seção 3.4.4).

No geral, esses resultados demonstram as capacidades variadas e limitações dos modelos de linguagem atuais e sugerem a necessidade de refinamento contínuo e melhorias direcionadas para abordar fraquezas específicas.

3.4.2 Explicação

Os dados apresentados na Tabela 2 mostram o número de questões do POSCOMP usando os seguintes critérios, baseados em suas respostas:

- **Explica sobre o tema:** significa que a resposta do LLM contém alguma explicação sobre o tema da questão.
- **Explica sobre a alternativa apontada:** significa que a resposta do LLM contém explicação sobre a alternativa que foi apontada como correta pelo modelo.
- **Explica sobre cada alternativa:** significa que a resposta do LLM contém explicação sobre as outras alternativas que não foram apontadas como correta (por exemplo, porque elas são incorretas).
- **Apenas a alternativa:** significa que o LLM escreveu apenas a alternativa correta em sua resposta, sem explicar adicional ou o porquê.

Explica:	ChatGPT	Gemini
sobre o tema	32/70	40/70
sobre a alternativa apontada	13/70	31/70
sobre cada alternativa	20/70	23/70
apenas a alternativa	21/70	0/70

Tabela 2: Respostas dos modelos ao POSCOMP 2023 de acordo com o nível de explicação (*prompt* com imagem).

O ChatGPT-4 demonstra uma abordagem variada em suas respostas. O LLM explica o tópico em quase metade dos casos, sugerindo um foco moderado em fornecer um entendimento contextual. No entanto, ele se mostra preciso ao acertar todas as questões nas quais forneceu apenas a resposta correta, sem explicações adicionais. Isso pode indicar um foco em respostas diretas em vez de explicações detalhadas em determinados cenários, especialmente em questões de Matemática.

Por outro lado, o Gemini consistentemente oferece explicações mais abrangentes. O LLM se destaca particularmente na discussão do tópico e das alternativas, nunca se limitando a apenas declarar a resposta correta sem explicação. Essa profundidade consistente nas respostas destaca a robustez do Gemini no manejo de conteúdo educacional detalhado, visando melhorar a compreensão em vez de apenas a identificação correta.

Apesar dos pontos fortes do Gemini, nessa avaliação, a partir da metade da prova, o modelo começou a alucinar (evento que será explicado em maiores detalhes na seção 3.4.4), gerando respostas verborrágicas sobre o mesmo tema por várias questões seguidas.

3.4.3 Temas

Em uma análise detalhada das taxas de acerto do ChatGPT-4 e do Gemini ao responder questões do POSCOMP nos seus diversos temas, padrões distintos e variações aparecem. A edição de 2023 do POSCOMP conta apenas com os Temas Gerais. Para identificarmos as questões em Temas Específicos, tivemos como base os Temas Específicos que estavam presentes na edição de 2022 do exame e utilizamos o ChatGPT-4 para fazer a categorização. Após as sugestões do ChatGPT-4, os autores revisaram as categorias atribuídas a cada questão. Concentrando-se nos Temas Gerais, o ChatGPT-4 consistentemente supera o Gemini em todas as categorias. Em Matemática, o ChatGPT-4 alcançou uma taxa de sucesso de 55%, em comparação com os 10% do Gemini. A disparidade é ainda mais pronunciada dentro dos Fundamentos da Computação, onde o ChatGPT-4 respondeu corretamente a 53,33% das questões, enquanto o Gemini acertou apenas 23,33%. No âmbito da Tecnologia de Computação, o ChatGPT-4 novamente liderou com uma taxa de sucesso de 65%, em contrapartida aos 10% de acerto do Gemini.

Aprofundando-se nos Temas Específicos, obtemos uma visão mais detalhada dos pontos fortes e fracos de cada modelo. Em Matemática, mostrada na Tabela 3, o ChatGPT-4 demonstrou domínio completo em Álgebra Linear e Probabilidade e Estatística, ostentando uma taxa de sucesso perfeita de 100%, enquanto o Gemini não conseguiu acertar nenhuma resposta nessas áreas. O ChatGPT-4 também se destacou em Análise Combinatória e Lógica Matemática, enquanto o Gemini teve dificuldades significativas, indicando uma possível deficiência no manejo de raciocínio lógico complexo e problemas combinatórios.

Tema Específico	ChatGPT	Gemini
Álgebra Linear	5/5	0/5
Geometria Análitica	0/1	0/1
Cálculo Diferencial e Integral	0/3	0/3
Matemática Discreta	0/2	0/2
Lógica Matemática	2/4	0/4
Análise Combinatória	2/3	1/3
Probabilidade e Estatística	2/2	1/2

Tabela 3: Acertos dos Temas Específicos de Matemática POSCOMP 2023 - *prompt* de imagem + "Qual é a resposta?".

Nos Fundamentos da Computação, apresentado na Tabela 4, o ChatGPT-4 mostrou forte desempenho em Análise de Algoritmos, Linguagens de Programação e Organização de Arquivos e Dados, alcançando taxas de sucesso de 100%, 75% e 100%, respectivamente. Os resultados do Gemini foram menos consistentes, indicando uma abordagem menos eficaz ao pensamento algorítmico e aos desafios de programação.

Tema Específico	ChatGPT	Gemini
Teoria dos Grafos	1/3	1/3
Análise de Algoritmos	3/3	2/3
Linguagens de Programação	3/4	1/4
Sistemas Operacionais	2/4	1/4
Circuitos Digitais	2/5	1/5
Técnicas de Programação	0/1	0/1
Organização de Arquivos e Dados	1/1	0/1
Algoritmos e Estrutura de Dados	2/5	1/5
Linguagens Formais, Autômatos e Computabilidade	2/4	0/4

Tabela 4: Acertos dos Temas Específicos de Fundamentos da Computação POSCOMP 2023 - *prompt* de imagem + "Qual é a resposta?".

Em Tecnologia de Computação (Tabela 5), o ChatGPT-4 se destacou em Inteligência Artificial, Processamento de Imagens e Gráficos Computacionais, com taxas de sucesso perfeitas, demonstrando sua capacidade superior no manejo de tópicos tecnologicamente avançados e especializados. Nesse tema geral, o

Gemini foi bastante penalizado pela alucinação ocorrida, acertando, de forma aleatória, duas questões.

Tema Específico	ChatGPT	Gemini
Engenharia de Software	1/3	0/3
Inteligência Artificial	3/3	0/3
Processamento de Imagens	1/1	0/1
Computação Gráfica	2/2	0/2
Compiladores	0/2	0/2
Sistemas Distribuídos	2/4	0/4
Banco de Dados	2/2	2/2
Redes de Computadores	2/3	0/3

Tabela 5: Acertos dos Temas Específicos de Tecnologia de Computação POSCOMP 2023 - *prompt* de imagem + "Qual é a resposta?".

No geral, esses dados sublinham que, enquanto o ChatGPT-4 exibe amplamente uma capacidade superior na resolução de questões do POSCOMP em quase todos os temas, o desempenho do Gemini varia drasticamente, com fraquezas particulares em áreas mais complexas e abstratas da computação.

3.4.4 Alucinação

A alucinação, no contexto da inteligência artificial, refere-se a instâncias em que um modelo gera respostas ou resultados que estão desconectados da realidade ou dos dados nos quais foi treinado [2]. Esse fenômeno ocorre quando a IA interpreta erroneamente os dados de entrada ou gera respostas incorretas ou sem sentido que não estão alinhadas com resultados lógicos ou esperados. Durante esta avaliação, ocorreram casos de alucinação, que serão descritos a seguir.

Para o ChatGPT-4, a maioria das alucinações envolveu interpretações errôneas da formulação da pergunta. Isso levou à criação de respostas que não existiam nas opções definidas ou que não eram relevantes para a pergunta apresentada. Por exemplo, erros foram observados em várias questões onde o modelo ou interpretava a declaração totalmente errado ou divergia significativamente em sua explicação, resultando em respostas que estavam em discordância com as alternativas corretas fornecidas no teste. Isso foi evidente em várias ocasiões, como nas questões 17, 18, 30, entre outras, onde o modelo lia errado a declaração ou fabricava uma escolha não apresentada.

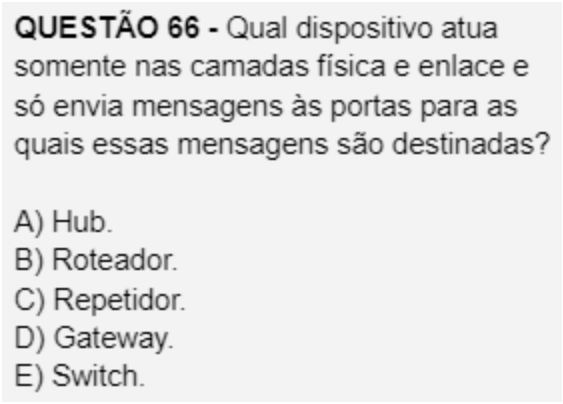

Figura 4: Questão 66 do POSCOMP 2023 (Tema geral: Tecnologia de Computação).

O Gemini exibiu um padrão mais variado de erros. No início do teste, semelhante ao ChatGPT-4, criou alternativas inexistentes e divergiu em suas explicações. No entanto, exclusivo para o Gemini, a partir da questão 31 em diante, ocorreu uma interpretação consistente do formato da pergunta de forma errada, tratando questões de múltipla escolha como verdadeiro ou falso. Esse mal-entendido fundamental do formato do exame persistiu por várias questões. Uma situação problemática foi o comportamento observado da questão 44 até o final do teste na questão 70. Durante este segmento, Gemini não só continuou a interpretar mal o formato da pergunta, mas também começou a gerar respostas quase idênticas para diferentes questões, indicando uma possível falha sistêmica ou 'travamento' em seu processamento. Nessa fase, ocasionalmente marcava respostas corretas ao acaso, apesar do conteúdo de suas respostas serem não relacionadas às perguntas (exemplo apresentado na Figura 5).

Essas alucinações ressaltam as dificuldades que até mesmo modelos avançados como ChatGPT-4 e Gemini enfrentam para interpretar e responder com precisão a formatos de perguntas e conteúdos complexos em ambientes de teste como esta avaliação do POSCOMP. Isso aponta para uma necessidade contínua de melhorias na forma como esses modelos lidam com o contexto e detalhes, garantindo que seus mecanismos de compreensão e resposta sejam robustos o suficiente para lidar com as complexidades de testes padronizados com precisão.

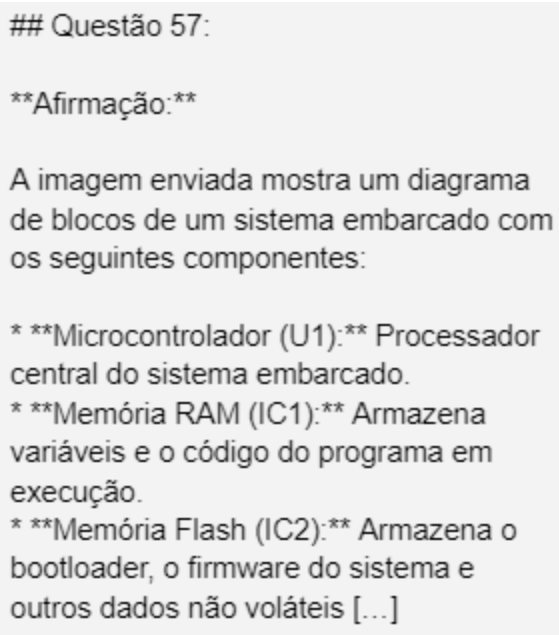

Figura 5: Caso de alucinação: resposta que o Gemini deu à questão 66 do POSCOMP 2023 (Tema geral: Tecnologia de Computação).

3.4.5 Testes Metamórficos

O objetivo do teste metamórfico [11] no contexto de LLMs é tentar diminuir a ameaça à validade do estudo quando o modelo possa já ter sido treinado com as questões e gabaritos do POSCOMP. A ideia é fazer pequenas alterações em cada questão, e ver como os LLMs se comportam.

Para avaliarmos a robustez e adaptabilidade dos modelos, eles foram submetidos a testes metamórficos. O experimento focou em como esses modelos responderam a modificações nas perguntas, que incluíram alterações nos dados, código e no arranjo das opções de resposta, testando assim a capacidade deles de manter a precisão apesar das variações.

Para esse experimento, foi selecionada uma amostra de 5 questões do POSCOMP 2023, todas respondidas de forma correta por ambos os modelos. As questões 20 (mostrada em sua versão original na Figura 6 e modificada na Figura 7), 22, 23, 45 e 53 foram selecionadas para este experimento. As questões foram alteradas da seguinte forma:

- **Questão 20:** no enunciado, a probabilidade de ter uma peça com defeito foi alterada de 0,05 para 0,07 e o conjunto de unidades foi alterado de 10 para 13. A alternativa D foi alterada de 80,0% para 61,0%.
- **Questão 22:** a alternativa A foi alterada de "A complexidade de tempo de um algoritmo recursivo é sempre mais rápida do que a de um algoritmo iterativo equivalente." para "Toda função que puder ser produzida por um computador pode ser escrita como função recursiva sem o uso de iteração; reciprocamente, qualquer função recursiva pode ser descrita através de iterações sucessivas.". A alternativa E foi alterada de "A escolha adequada da estrutura de dados pode reduzir o tempo e o espaço necessários para a execução de algoritmos recursivos." para "A complexidade de tempo de um algoritmo recursivo é sempre mais rápida do que a de um algoritmo iterativo equivalente.".
- **Questão 23:** as variáveis "i" e "j" dos laços na questão que iam de 1 até n e m, respectivamente foram fixados em 1. Na questão original, as alternativas lidavam com diferentes cenários de complexidade assintótica e foram alteradas para tentar descrever a complexidade do código fornecido.
- **Questão 45:** a primeira questão pergunta sobre um tipo de dado que agrupa coleções de constantes nomeadas. A segunda questão pergunta sobre um tipo de dado que possui dois valores, como 0 e 1 ou falso e verdadeiro. As alternativas A e C trocaram de posição entre as questões, a alternativa D (Caracter) é a mesma nas duas questões e a alternativa E (Booleano) também está presente em ambas, mas com um significado diferente conforme a pergunta.
- **Questão 53:** a ordem das afirmações foi alterada, resultando em uma sequência diferente, a adição de uma nova opção de resposta "Todas as alternativas" na segunda questão e mudanças nas combinações de alternativas corretas apresentadas nas opções de resposta.

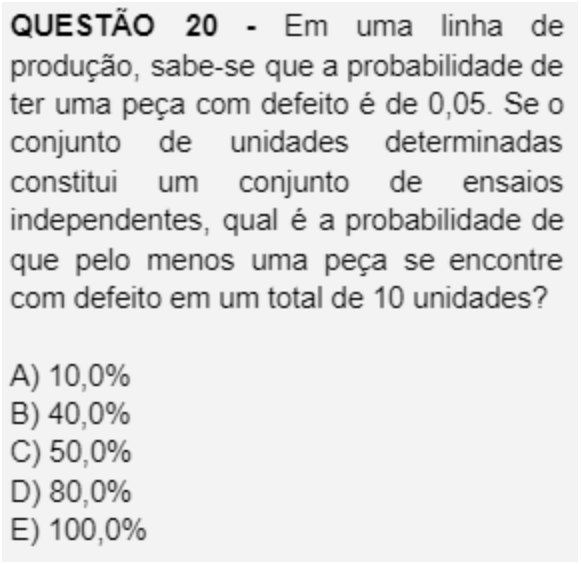

Figura 6: Questão 20 do POSCOMP 2023 em sua versão original.

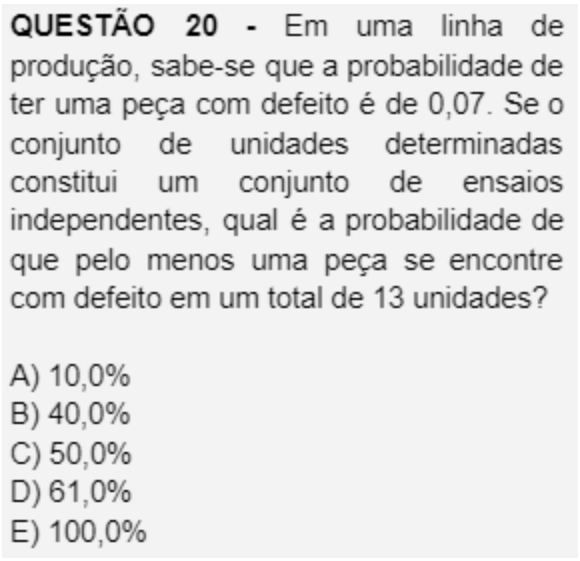

Figura 7: Questão 20 do POSCOMP 2023 em sua versão modificada para o teste metamórfico.

Para assegurar as respostas corretas após a modificação, diferentes métodos foram empregados. Cálculos manuais foram necessários para questões envolvendo mudanças quantitativas, enquanto pesquisas na internet foram usadas para ajustes mais qualitativos ou teóricos. Por exemplo, a Questão 53 exigiu apenas o rastreamento do reposicionamento dos enunciados e das escolhas, tornando a identificação da resposta correta direta.

Os resultados deste teste são apresentados na Tabela 6. O ChatGPT-4 geralmente se adaptou bem às questões alteradas, mantendo a correção na maioria dos casos, exceto pela Questão 23. Esta falha particular foi atribuída à má interpretação do modelo de uma imagem contendo a questão, levando ao processamento incorreto dos dados. Por outro lado, o Gemini, embora fornecesse respostas completas e explicadas, apresentou dificuldades com a interpretação de imagens e a aplicação incorreta de fórmulas em certos casos.

Questão	Continuou acertando? ChatGPT	Gemini
20	Sim	Não
22	Sim	Sim
23	Não	Não
45	Sim	Sim
53	Sim	Não

Tabela 6: Resultados dos testes metamórficos POSCOMP 2023 - *prompt* com imagem.

Este experimento destaca o potencial e as limitações da IA em ambientes de teste dinâmicos e realça a importância da interpretação precisa de dados e da aplicação de fórmulas na obtenção de resultados confiáveis. Tais testes não apenas refinam as capacidades dos modelos, mas também melhoram nosso entendimento de seus limites operacionais e áreas para aprimoramento.

3.4.6 Comparação com os Alunos

Comparado aos 617 alunos que fizeram a prova, o ChatGPT-4, com 40 das 70 perguntas respondidas corretamente, ficou na 118ª posição, colocando-se no *top* 22,17% de melhores resultados. Especificamente, na seção de Matemática, ChatGPT-4 marcou 11 de 20, ficando na 160ª posição, que é o *top* 45,95% melhores resultados de Matemática, uma colocação que reflete uma competência moderada. Em Fundamentos da Computação, ele obteve 16 das 30 respostas corretas, alcançando o *top* 31,39% dos melhores alunos. O ChatGPT-4 se destacou em Tecnologia de Computação, onde respondeu corretamente a 13 de 20 perguntas, ficando no 14º lugar e colocando-se entre os 5,83% melhores, indicando um forte domínio nesta área especializada.

Por outro lado, o desempenho de Gemini foi bem diferente, destacando limitações potenciais em sua configuração atual para lidar com esse tipo de avaliação e a interpretação de *prompt* por imagem. Com apenas 11 de 70 respostas corretas, Gemini ficou em último lugar entre todos os participantes. Esse resultado drástico continuou nos Temas Gerais. Em Matemática, Gemini acertou apenas 2 de 20 perguntas, ficando novamente em último lugar e caindo para os 0,97% inferiores. Seu desempenho em Fundamentos da Computação foi um pouco melhor, com 7 das 30 respostas corretas, mas ainda assim o colocou nos 7,44% inferiores. O desempenho ruim se estendeu à Tecnologia de Computação, onde o Gemini acertou apenas 2 de 20 perguntas, ficando na 597ª posição, entre os 3,4% inferiores.

Esses resultados não apenas destacam as diferenças marcantes na performance entre ChatGPT-4 e Gemini em ambientes de avaliação, mas também revelam como eles se comparam aos estudantes em contextos acadêmicos. Enquanto ChatGPT-4 mostrou uma competência notável, especialmente em Tecnologia de Computação, colocando-se bem acima da média dos estudantes e alcançando uma das melhores colocações, o desempenho de Gemini indicou uma necessidade urgente de melhorias. Ficando consistentemente entre os últimos lugares, Gemini teve dificuldades significativas em se equiparar ao nível dos estudantes participantes, o que sugere áreas cruciais para desenvolvimento e ajustes no modelo para futuros exames.

3.5 Respostas às Questões de Pesquisa

A seguir respondemos as questões de pesquisa.

QP_1. Até que ponto o ChatGPT-4 é capaz de resolver questões do POSCOMP?

O ChatGPT-4 demonstrou uma forte capacidade para resolver questões do POSCOMP, alcançando uma taxa de sucesso de aproximadamente 57,1% (40 de 70 questões). Mostrou proficiência em várias matérias, com pontos fortes notáveis em Tecnologia da Computação, Fundamentos da Computação e áreas específicas de Matemática como Álgebra Linear e Probabilidade e Estatística. O desempenho do ChatGPT-4 colocou-o bem dentro do grupo de elite de todos os participantes do teste, evidenciando sua robustez e adaptabilidade no manuseio de uma gama diversificada de tópicos dentro do campo da computação.

QP_2. Até que ponto o Gemini 1.0 Advanced é capaz de resolver questões do POSCOMP?
O Gemini 1.0 Advanced teve significativas dificuldades com as questões do POSCOMP, garantindo apenas 11 acertos de 70, o que se traduz em uma taxa de sucesso de cerca de 15,7%. O modelo teve um desempenho ruim em todas as matérias, com uma taxa de sucesso particularmente baixa em Matemática e Tecnologia da Computação. Os desafios do Gemini foram agravados por problemas de alucinações durante o teste, levando a respostas incorretas ou sem sentido. Sua classificação geral no fim da lista de participantes destaca lacunas substanciais em sua capacidade de interpretar e responder com precisão a conteúdos educacionais complexos.

4. AVALIAÇÃO 2: TEXTO

A seguir apresentamos a avaliação considerando *prompts* usando apenas enunciados das questões em texto.

4.1 GQM

A seguir estruturamos a nossa avaliação usando GQM [10]. O objetivo deste estudo é avaliar a eficácia dos *Large Language Models* específicos — ChatGPT-4, Gemini 1.0 Advanced, Claude 3 Sonnet e Le Chat Mistral — na resolução de questões dos exames POSCOMP 2022 e 2023, com o propósito de identificar os domínios e tipos de perguntas nos quais esses modelos se sobressaem ou apresentam deficiências no contexto de *prompts* com o enunciado das questões contendo apenas textos em Inglês. Para isso, serão respondidas às seguintes questões de pesquisa (QP):

- **QP_1.** Até que ponto o ChatGPT-4 é capaz de resolver problemas do POSCOMP?
- **QP_2**. Até que ponto o Gemini 1.0 Advanced é capaz de resolver problemas do POSCOMP?
- **QP_3**. Até que ponto o Claude 3 Sonnet é capaz de resolver problemas do POSCOMP?
- **QP_4.** Até que ponto o Le Chat Mistral é capaz de resolver problemas do POSCOMP?

Serão contadas as respostas corretas e incorretas fornecidas por cada LLM de acordo com o gabarito oficial fornecido pelo POSCOMP.

4.2 Planejamento

Nós avaliamos 4 LLMs: ChatGPT-4, Gemini 1.0 Advanced, Claude 3 Sonnet, e Le Chat Mistral. Além da prova do POSCOMP de 2023, os LLMs também foram testados com a prova de 2022. Foi levantada a hipótese de que os resultados da Seção 3 foram prejudicados pela capacidade limitada dos LLMs de processar e entender informações visuais dentro das capturas de tela. Portanto, nesta avaliação, os LLMs receberam traduções para o inglês do texto da questão e das opções de resposta, transcritos via Google Lens e traduzidos com DeepL. A decisão de usar traduções para o inglês decorreu da ideia de que os LLMs geralmente são treinados em conjuntos de dados maiores em inglês, o que potencialmente melhoraria seu desempenho [1]. As imagens de apoio às questões (diagramas de classe, circuitos e autômatos) foram incluídas quando necessário, especificamente, na prova de 2022, questões 27, 31, 40 e 43; na prova de 2023, questões 17, 31, 32 e 34. Le Chat Mistral, devido à sua incapacidade de processar imagens, recebeu apenas o texto traduzido. Esta avaliação ocorreu entre 16 de março de 2024 e 22 de março de 2024. A avaliação considerou *zero-shot prompt*, onde nenhum exemplo é apresentado para o LLM [9].

4.3 Resultados

A seguir apresentamos os resultados da avaliação. Sobre a prova de 2022 (Tabela 7), o ChatGPT-4 aparece como o líder em acertos, alcançando a pontuação total de 57 acertos dentre 70 questões. Tal desempenho é impulsionado pela pontuação em Matemática (18 acertos dentre 20 questões) e Fundamentos da Computação (23 acertos dentre 30 questões). O Gemini 1.0 Advanced foi o segundo com mais acertos, com 49 de 70, ficando atrás do ChatGPT-4 em todas as áreas. Le Chat Mistral e Claude 3 Sonnet obtiveram desempenho comparável ao Gemini, com 48 e 44 acertos, respectivamente. No entanto, o Le Chat Mistral se destaca em Tecnologia de Computadores, obtendo 18 de 20 acertos e superando todos os outros modelos nesse domínio.

Tema Geral	Chat GPT	Gemini	Claude	Mistral
Matemática	18/20	14/20	11/20	12/20
Fundamentos da Computação	23/30	20/30	18/30	18/30
Tecnologia de Computação	16/20	15/20	15/20	18/20
TOTAL	57/70	49/70	44/70	48/70

Tabela 7: Acertos POSCOMP 2022 - *prompt* de texto em inglês.

Sobre a prova de 2023, apresentada na Tabela 8, o ChatGPT-4 novamente aparece com a maior pontuação de acertos, com o total de 59 acertos dentre 70 questões. O Gemini mais uma vez demonstra competência consistente em todas as áreas, com uma pontuação total de 53 acertos dentre 70 questões. Le Chat Mistral atingiu 50 acertos, enquanto o Claude 3 Sonnet ficou um pouco para trás com 45 acertos.

Tema Geral	Chat GPT	Gemini	Claude	Mistral
Matemática	17/20	13/20	9/20	14/20
Fundamentos da Computação	24/30	23/30	22/30	21/30
Tecnologia de Computação	18/20	17/20	14/20	15/20
TOTAL	59/70	53/70	45/70	50/70

Tabela 8: Acertos POSCOMP 2023 - *prompt* de texto em inglês.

4.4 Discussão

As seções a seguir discutem os resultados da avaliação.

4.4.1. Corretude

Na prova de 2022, um total de 31 questões receberam a mesma resposta correta de todos os modelos, enquanto que, na prova de 2023, foram 36 questões. Isso indica uma leve melhora no desempenho coletivo dos modelos de um ano para o outro. No entanto, algumas questões permaneceram desafiadoras, como

evidenciado por 5 questões em 2022 e 6 em 2023 que nenhum dos modelos conseguiu resolver.

Como mostrado na Tabela 9, o ChatGPT-4 conseguiu responder 6 questões sozinho, em que nenhuma LLM acertou, na prova de 2022 e conseguiu repetir o feito na de 2023. Nesse mesmo quesito, o Gemini conseguiu responder sozinho 2 questões em 2022 e 2 em 2023. O Claude respondeu 1 questão sozinho na prova de 2022.

Modelo	2022	2023
ChatGPT	6	6
Gemini	2	2
Claude	1	0
Mistral	0	0

Tabela 9: Número de questões respondidas corretamente por apenas um modelo.

Duas questões foram anuladas no gabarito oficial, uma na prova de 2022 e outra na prova de 2023. Mesmo tendo sido anuladas, ambas as questões foram avaliadas pelos modelos. A questão anulada da prova de 2022 foi a 40, do tema geral Fundamentos da Computação. O ChatGPT-4, Claude e Mistral apontaram C como resposta correta para essa questão, enquanto o Gemini apontou A. Na prova de 2023, a questão anulada foi a 17, do tema geral Matemática. O ChatGPT-4, Claude e Mistral apontaram B como resposta correta para essa questão, enquanto o Gemini apontou C.

Os modelos forneceram mais de uma resposta para uma questão em algumas questões (Tabela 10), ou falharam em fornecer qualquer resposta (Tabela 11). Por exemplo, Claude frequentemente selecionou várias respostas em ambos os testes, sugerindo uma tendência a ser menos decisivo ou talvez mais exploratório em sua estratégia de resposta. Em contraste, Mistral, ChatGPT-4 e Gemini tenderam a deixar de apontar qualquer alternativa em várias questões em 2022, indicando possíveis lacunas em seu conhecimento ou cautela em sua abordagem de resposta.

Modelo	2022	2023
ChatGPT	0	0
Gemini	1	3
Claude	3	5
Mistral	1	5

Tabela 10: Número de questões onde os modelos apontaram mais de uma alternativa como corretas.

Modelo	2022	2023
ChatGPT	5	1
Gemini	4	2
Claude	1	1
Mistral	6	3

Tabela 11: Número de questões onde os modelos não apontaram alternativa correta.

4.4.2. Explicação

Os dados apresentados nas Tabelas 12 e 13 mostram o número de questões do POSCOMP 2022 e 2023, respectivamente, usando os seguintes critérios, baseados em suas respostas:

- **Explica sobre o tema:** significa que a resposta do LLM contém alguma explicação sobre o tema da questão.
- **Explica sobre a alternativa apontada:** significa que a resposta do LLM contém explicação sobre a alternativa que foi apontada como correta pelo modelo.
- **Explica sobre cada alternativa:** significa que a resposta do LLM contém explicação sobre as outras alternativas que não foram apontadas como correta (por exemplo, porque elas são incorretas).
- **Apenas a alternativa:** significa que o modelo escreveu apenas a alternativa correta em sua resposta, sem explicar adicional ou o porquê.

Explica:	Chat GPT	Gemini	Claude	Mistral
sobre o tema	38/70	52/70	50/70	38/70
sobre a alternativa apontada	22/70	48/70	68/70	63/70
sobre cada alternativa	12/70	29/70	19/70	5/70
apenas a alternativa	10/70	0/70	0/70	2/70

Tabela 12: Respostas dos modelos ao POSCOMP 2022 de acordo com o nível de explicação (*prompt* de texto).

Em 2022, o Gemini se destaca por frequentemente explicar tanto o tópico (52/70) quanto a alternativa correta (48/70), significativamente mais do que o ChatGPT-4, Claude e Mistral em categorias semelhantes. Gemini e Claude não recorreram a responder apenas com a alternativa correta, sugerindo uma abordagem voltada para a profundidade em suas respostas. Claude, notavelmente forte em explicar a alternativa correta (68/70), indica uma clareza focada em suas respostas. Em contraste, Mistral e ChatGPT-4 ocasionalmente optaram pela resposta menos detalhada, com o ChatGPT-4 fornecendo apenas a alternativa correta em 10 de 70 casos e o Mistral em 2 de 70 casos.

Explica:	Chat GPT	Gemini	Claude	Mistral
sobre o tema	49/70	52/70	63/70	34/70
sobre a alternativa apontada	30/70	31/70	63/70	58/70
sobre cada alternativa	19/70	32/70	25/70	12/70
apenas a alternativa	3/70	0/70	0/70	5/70

Tabela 13: Respostas dos modelos ao POSCOMP 2023 de acordo com o nível de explicação (*prompt* de texto).

Já na prova de 2023, o Claude mostra uma consistência notável, particularmente em suas explicações sobre o tópico (63/70) e a alternativa correta (63/70). Essa consistência sublinha a saída confiável e completa de Claude ao lidar com consultas. O desempenho do Gemini permanece forte na explicação do tópico, mas mostra uma leve queda na detalhação da alternativa correta em comparação com o ano anterior. O ChatGPT-4 mostra melhoria ao explicar o tópico, passando de 38/70 para 49/70, indicando um aprimoramento na oferta de contexto. Mistral, no entanto, mostra um desempenho modesto em geral, com ênfase em explicar a alternativa correta.

Essas tendências destacam diferentes forças e estratégias empregadas pelos modelos. Enquanto alguns modelos como Claude e Gemini consistentemente visam explicações abrangentes, outros como Mistral e ChatGPT-4 podem focar mais em respostas diretas sob certas condições. Essa variação na abordagem pode refletir o design subjacente e a aplicação pretendida de cada modelo, atendendo a diferentes necessidades dos usuários por explicação e detalhamento.

4.4.3. Temas

Ao avaliar a prova de 2022 usando *prompts* de texto em inglês, o ChatGPT-4 continuou a mostrar uma alta taxa de sucesso, mas foi seguido de perto por outros modelos. O ChatGPT-4 alcançou 18/20 acertos em Matemática, 23/30 acertos em Fundamentos da Computação e 16/20 acertos em Tecnologia da Computação, resultando em um total de 57/70 acertos. Gemini, Claude e Mistral tiveram um bom desempenho neste teste, com taxas de sucesso não muito atrás do ChatGPT-4: 49/70, 44/70 e 48/70, respectivamente.

Na prova de 2023 com *prompts* de texto em inglês, o ChatGPT-4 liderou novamente com 59/70 acertos, mostrando sua versatilidade e consistência nos temas gerais. No entanto, Gemini, Claude e Mistral também demonstraram bom desempenho, marcando 53/70 , 45/70 e 50/70 acertos, respectivamente.

No entanto, na prova de 2022 com *prompts* de texto em inglês, houve uma maior paridade entre os modelos em temas específicos. O ChatGPT-4 se destacou em muitas áreas, mas Gemini, Claude e Mistral também alcançaram altas taxas de sucesso. Por exemplo, em Geometria Analítica, Cálculo Diferencial e Integral e Inteligência Artificial, todos os modelos tiveram um desempenho forte, com taxas de sucesso acima de 60%. No entanto, diferenças surgiram em áreas específicas como Matemática Discreta e Linguagens Formais, Autômatos e Computabilidade, onde o ChatGPT-4 superou outros modelos.

Na prova de 2023 com *prompts* de texto em inglês, o ChatGPT-4 manteve sua liderança em vários temas específicos, marcando 100% em Álgebra Linear, Geometria Analítica, Lógica Matemática e outras áreas. O Gemini e outros modelos tiveram sucessos mistos, com algumas altas pontuações em temas específicos como Análise Combinatória, Probabilidade e Estatística e Teoria dos Grafos.

Em Matemática como um todo, o Claude teve um fraco desempenho quando comparado aos outros modelos, com 50% de acerto nas duas provas. Esse desempenho inferior pode resultar de fatores como a falta de treinamento em conceitos específicos da matemática ou menor ênfase na resolução rigorosa de problemas matemáticos nos seus dados de treinamento.

No geral, esses resultados sugerem que, o ChatGPT-4 geralmente supera outros modelos em temas gerais. As taxas de sucesso em temas específicos variam, indicando que diferentes modelos se destacam em áreas distintas.

4.4.4. Alucinação

No exame de 2022 com um *prompt* de texto em inglês, o ChatGPT-4 demonstrou precisão, sem alucinações em suas respostas. No entanto, o Gemini alucinou em 2 casos, um em que apresentou um cenário que correspondia à resposta correta, mas apontou para a alternativa errada, e outro em que escreveu que todas as assertivas estavam corretas, mas indicou a alternativa onde uma delas não estava, sugerindo inconsistências na interpretação. O Claude mostrou alucinações em 6 casos, geralmente envolvendo contradições entre o conteúdo da resposta e sua conclusão. O Mistral teve 3 alucinações, envolvendo a geração de alternativas inexistentes, confundindo a resposta correta com outra e também o caso citado anteriormente da questão 19, com o loop de texto repetido e erros matemáticos que durou 24 minutos.

Já no exame de 2023 com um *prompt* de texto em inglês, as alucinações do ChatGPT-4 foram relativamente poucas, com 2 casos envolvendo alternativas inventadas e alterações nas opções originais da questão. O Gemini também alucinou em 5 casos, frequentemente envolvendo mudanças nas alternativas da questão. O Claude mostrou um padrão semelhante de alucinações, totalizando 5 casos, com confusão e alternativas alteradas. Mistral teve 2 casos de alucinação, um envolvendo uma má interpretação de uma questão assertiva e outro em que identificou o resultado correto, mas apontou para a alternativa errada.

No geral, o Gemini parece ter os problemas mais significativos com alucinações, especialmente ao interpretar *prompts* de imagem ou entradas de texto complexas. Outros modelos, como Claude e Mistral, mostram diferentes graus de alucinação, enquanto o ChatGPT-4 demonstra maior precisão, especialmente com *prompts* de texto em inglês.

4.4.5 Testes Metamórficos

Assim como foi feito na Seção 3.4.5, testes metamórficos [11] foram realizados a fim de avaliar a robustez e adaptabilidade dos modelos nas condições desta avaliação.

Para esse experimento, foi selecionada uma amostra de 10 questões, 5 do POSCOMP 2022 e 5 do POSCOMP 2023, todas respondidas de forma correta por ambos os modelos. Na prova de 2022, foram selecionadas as questões 10, 20, 22 (mostrada em sua versão original na Figura 8 e modificada na Figura 9), 30 e 57. Na prova de 2023, foram selecionadas as questões 5, 19, 21, 39 e 65. As questões da prova de 2022 foram alteradas da seguinte forma:

- **Questão 10:** o numerador $((2x+3)^3(3x-2)^2)$ foi alterado para $((2x+3)^6(3x-2)^2)$, o denominador x^5+5 foi alterado para x^8+5. As alternativas, que eram A)72, B)19, C)9, D)8 e E)0 foram alteradas para A)1024, B)720, C)576, D)12 e E)0.
- **Questão 20:** os valores de t foram alterados de 2, 3, 4, 5, 6, 7 para 3, 4, 5, 6, 7, 8. As alternativas foram alteradas de A)4.5 s, B)5.0 s, C)1.0 s, D)0.9 s, E)5.4 s para A)0.9 s, B)5.4 s, C)1.0 s, D)6.4 s, E)5.0 s.
- **Questão 22:** no enunciado, f3(n) = O(2^n) foi alterada para f3(n) = O(2^(3^n)). As alternativas B e C foram trocadas de lugar com as alternativas D e E.
- **Questão 30:** afirmativa III alterada de "Tipos inteiros são utilizados para armazenar valores que pertencem ao conjunto dos números naturais (sem a parte fracionária)." para "O tipo booleanos só armazena o valor "False" e números fracionários". Alternativas A) Apenas I., B) Apenas II. e C) Apenas III. alteradas para A) Apenas II., B) Apenas III. e C) Apenas I e II.
- **Questão 57:** afirmativa I alterada de "O mapeamento de imagens como textura (textura de superfície) é uma técnica que utiliza um sistema de coordenadas 2D." para "O Ray Tracing é a técnica utilizada para gerar texturas traçando o percurso de círculos de sombra através de um plano de imagem.". Alternativa B) Apenas III. alterada para B) Apenas II.

QUESTÃO 22 - Considere as funções a seguir:

f1(n) = O(n)
f2(n) = O(n!)
f3(n) = O(2^n)
f4(n) = O(n²)

A ordem dessas funções, por ordem crescente de taxa de crescimento, é:

A) f2 - f1 - f3 - f4.
B) f3 - f2 - f4 - f1.
C) f1 - f4 - f3 - f2.
D) f1 - f4 - f2 - f3.
E) f4 - f3 - f1 - f2.

Figura 8: Questão 22 do POSCOMP 2022 em sua versão original.

Na prova de 2023, as questões foram alteradas da seguinte forma:

- **Questão 5:** número de regiões alterado de 10 para 8. Alternativas alteradas de A) 1024, B) 10, C) 100, D) 512 e E) 20 para A) 80, B) 64, C) 8, D) 1024 e E) 256
- **Questão 19:** número de famílias alterado de 5, 6, 8, 4 e 2 para 65, 43, 9, 73 e 12. Alternativas A) 1,12, C) 2,11 e E) 3,21 alteradas para A) 1,62, C) 2,71 e E) 3,02.
- **Questão 21:** alternativa A trocada de lugar com a alternativa C, alternativa D trocada de lugar para a alternativa B e alternativa B trocada de lugar para a alternativa E. Nova alternativa D adicionada: "Se o tempo exigido por um algoritmo em todas as entradas de tamanho n for, no máximo, $5n^3 + 3n$, a complexidade assintótica do tempo é $O(n^3)$.".
- **Questão 39:** as afirmativas "I. A leitura dos registros na ordem dos valores da chave de ordenação é mais eficiente se comparada à leitura desses registros em arquivos heap." e a "III. Para acelerar o acesso a um registro baseado no valor de uma chave em arquivos ordenados, a melhor técnica de pesquisa é a técnica de hash." foram trocadas de lugar. Na alternativa "II. Permite atender de forma eficiente condições de pesquisa sobre o campo de ordenação no formato <chave = valor> ou condição de intervalo (isto é, a chave estar no intervalo entre o valor1 e valor2)." a palavra "eficiente" foi trocada pela palavra "INEFICIENTE".
- **Questão 65:** a máscara de sub-rede do enunciado da qustão foi alterada de 255.255.255.128 para 255.255.255.64. As alternativas A) 126, B) 128, D) 255.255.255.128 e E) 256 foram alteradas para A) 128, B) 62, D) 255.255.255.64 e E) 255.255.255.128.

QUESTION 22 - Consider the following functions:

f1(n) = O(n)
f2(n) = O(n!)
f3(n) = O(2^(3^n))
f4(n) = O(n²)

The order of these functions, in ascending order of growth rate, is:

A) f2 - f1 - f3 - f4.
B) f1 - f4 - f2 - f3.
C) f4 - f3 - f1 - f2.
D) f3 - f2 - f4 - f1.
E) f1 - f4 - f3 - f2.

Figura 9: Questão 22 do POSCOMP 2022 em sua versão traduzida e modificada para o teste metamórfico.

Para assegurar as respostas corretas após a modificação, diferentes métodos foram empregados. Cálculos manuais foram necessários para questões envolvendo mudanças quantitativas, enquanto

pesquisas na internet foram usadas para ajustes mais qualitativos ou teóricos.

Para a prova de 2022, como apresentado na Tabela 14, tanto ChatGPT quanto Gemini demonstraram desempenho impecável, alinhando-se perfeitamente com as respostas corretas em todas as cinco questões. Claude e Mistral, no entanto, exibiram algumas discrepâncias. Especificamente, Claude respondeu incorretamente à questão 10. O Mistral, além de ter errado a questão 10, também mostrou erros nas questões 22 e 57. Esses erros estão relacionados a erros de interpretação dos modelos em suas respostas.

	Continuou acertando?			
Questão	ChatGPT	Gemini	Claude	Mistral
10	Sim	Sim	Não	Não
20	Sim	Sim	Sim	Sim
22	Sim	Sim	Sim	Não
30	Sim	Sim	Sim	Sim
57	Sim	Sim	Sim	Não

Tabela 14: Resultados dos testes metamórficos POSCOMP 2022 - *prompt* de texto em inglês.

Na transição para o exame de 2023, mostrado na Tabela 15, observou-se um padrão consistente de precisão para ChatGPT-4 e Gemini, mantendo sua taxa de sucesso de 100% da prova anterior. Ambos os modelos corresponderam com sucesso às respostas corretas em todas as cinco questões. Em contraste, enquanto Claude respondeu corretamente à maioria das questões, ele falhou na questão 21 ao marcar tanto 'A' quanto 'D' como respostas corretas. Mistral espelhou esse erro, indicando uma falha potencial em seu processamento ou um mal-entendido dos requisitos da questão.

	Continuou acertando?			
Questão	ChatGPT	Gemini	Claude	Mistral
5	Sim	Sim	Sim	Sim
19	Sim	Sim	Sim	Sim
21	Sim	Sim	Não	Não
39	Sim	Sim	Sim	Sim
65	Sim	Sim	Sim	Sim

Tabela 15: Resultados dos testes metamórficos POSCOMP 2023 - *prompt* de texto em inglês.

Os resultados gerais destacaram que ChatGPT-4 e Gemini são robustos em sua adaptabilidade e precisão no contexto desses exames, mantendo um desempenho consistente ao longo dos dois exames. O desempenho menos consistente de Claude e Mistral, particularmente no exame de 2022 e na questão 21 em 2023, sugere que, embora geralmente apresentem bom desempenho, eles poderiam se beneficiar de refinamentos no tratamento de cenários de questões ambíguas ou complexas.

4.4.6 Comparação com os Alunos

No POSCOMP de 2022, as pontuações médias nas disciplinas gerais para os estudantes foram relativamente modestas. Por exemplo, em Matemática, a pontuação média foi de 8,63 de 20, enquanto em Fundamentos da Computação e Tecnologia da Computação, os estudantes obtiveram uma média de 14 de 30 e 6,58 de 20, respectivamente. Em contraste, os LLMs geralmente superaram essas médias significativamente. O ChatGPT-4 liderou o grupo com pontuações impressionantes em todas as categorias, notavelmente alcançando 18 de 20 em Matemática e 23 de 30 em Fundamentos da Computação.

A tendência dos modelos de IA de se destacarem continuou no POSCOMP de 2023. O ChatGPT-4 não apenas superou o desempenho do ano anterior, mas também ultrapassou todos os participantes ao marcar 59 de 70, posicionando-se como o melhor resultado nesse exame. Essa realização notável incluiu pontuações máximas em Tecnologia da Computação e boas performances em Matemática e Fundamentos da Computação. Da mesma forma, Gemini e Mistral demonstraram desempenhos robustos, garantindo lugares dentro dos 4% melhores participantes. Claude, embora atrás de seus colegas, ainda conseguiu uma colocação respeitável no top 10%.

Notavelmente, na edição de 2023, os modelos de IA demonstraram suas forças em áreas específicas. A dominância do ChatGPT-4 em Tecnologia da Computação com uma pontuação quase perfeita de 18 de 20 foi espelhada pela excelência de Gemini na mesma categoria. Sua proficiência sugere que esses modelos são particularmente bem adaptados para lidar com consultas e problemas técnicos complexos, uma característica que é indicativa de seu treinamento subjacente e foco algorítmico.

4.5 Respostas às Questões de Pesquisa

A seguir respondemos as questões de pesquisa.

QP_1. Até que ponto o ChatGPT-4 é capaz de resolver problemas do POSCOMP?

O ChatGPT-4 alcançou a maior pontuação entre os modelos testados, com 57 de 70 acertos no teste de 2022 e 59 de 70 no teste de 2023. Ele se destacou particularmente em Matemática e Fundamentos da Computação. Seu desempenho o colocou como o melhor resultado no exame de 2023, indicando sua eficácia em lidar com uma ampla gama de tópicos, incluindo questões técnicas complexas.

QP_2. Até que ponto o Gemini 1.0 Advanced é capaz de resolver problemas do POSCOMP?

O Gemini 1.0 Advanced também mostra forte capacidade de resolver problemas do POSCOMP, embora ligeiramente atrás do ChatGPT-4. Ele marcou 49 de 70 no teste de 2022 e 53 de 70 no teste de 2023. O Gemini demonstrou consistência em diferentes matérias e manteve um desempenho robusto nas explicações, o que demonstra sua abordagem abrangente para a resolução de problemas.

QP_3. Até que ponto o Claude 3 Sonnet é capaz de resolver problemas do POSCOMP?

Claude 3 Sonnet teve sucesso moderado na resolução de problemas do POSCOMP, com pontuações de 44 de 70 em 2022 e 45 de 70 em 2023. Apesar de mostrar potencial nas explicações e compreensão, o desempenho de Claude foi menos consistente em comparação com o ChatGPT-4 e o Gemini. Ele teve algumas dificuldades em certas matérias como Matemática, sugerindo uma área potencial para melhoria no tratamento de tópicos acadêmicos específicos.

QP_4. Até que ponto o Le Chat Mistral é capaz de resolver problemas do POSCOMP?

Le Chat Mistral teve um desempenho comparável ao de Gemini e Claude, mas foi limitado por sua incapacidade de processar imagens, uma desvantagem significativa para questões baseadas em visualizações. Ele marcou 48 de 70 no teste de 2022 e 50 de 70 no teste de 2023. O desempenho do Mistral indica uma competência razoável, mas destaca a importância das capacidades de processamento de imagens para alcançar maior precisão nesses exames.

5. AMEAÇAS À VALIDADE

Algumas ameaças à validade podem surgir no contexto do uso de LLMs [7]. Os resultados são baseados na performance dos LLMs em um exame específico, o POSCOMP. Assim, a generalização para outras avaliações ou contextos em ciência da computação pode não ser direta. Variações nos formatos de teste, nos tipos de perguntas ou nos tópicos abordados em outros exames podem levar a desempenhos diferentes. No nosso trabalho, tentamos minimizar essa ameaça avaliando duas edições do POSCOMP.

A seleção de apenas quatro modelos de LLM pode não representar a gama completa de capacidades dos modelos de linguagem atualmente disponíveis. Modelos mais recentes ou configurados de maneira diferente podem apresentar desempenhos distintos. Minimizamos essa ameaça a partir da escolha de 4 LLMs que são bem populares na comunidade.

O desempenho dos modelos de LLM depende fortemente dos dados utilizados durante o treinamento. Diferenças na qualidade, quantidade, e diversidade dos dados de treinamento podem influenciar significativamente os resultados, podendo não refletir a capacidade real dos modelos em compreender e processar novas informações. Para minimizar essa ameaça, realizamos teste metamórfico alterando o enunciado de algumas questões. O ChatGPT-4 continuou acertando a maior parte das questões.

6. TRABALHOS RELACIONADOS

Pires et al. [12] avaliaram o ChatGPT-4 Vision e ChatGPT-4 Text em dois exames do ENEM. O *prompt* apenas com texto e legendas descrevendo imagens foi melhor do que a versão contendo apenas imagens. Usaram *few-shot* com *Chain-of Thought*. O LLM teve mais dificuldades na área de matemática. No nosso trabalho, nós avaliamos 4 LLMs na resolução de duas edições do POSCOMP usando *zero-shot prompt*. Além disso, o ChatGPT-4 teve um bom desempenho em questões de matemática.

Zhang et al. [13] avaliaram os modelos ChatGPT-3.5, ChatGPT-4 e o ERNIE-Bot (também sua versão Turbo) com questões Exame de Admissão às Universidades Chinesas (GAOKAO) de 2010 a 2022, excluindo questões contendo imagens. Foi usado *zero-shot prompt* e avaliação humana. Os resultados mostraram que os modelos tiveram um bom desempenho em questões baseadas no conhecimento, mas existiram dificuldades em certos tipos de raciocínio lógico e problemas matemáticos, bem como na compreensão da leitura de textos mais longos em chinês. No nosso trabalho, nós avaliamos 4 LLMs na resolução de duas edições do POSCOMP usando *zero-shot prompt*. Além disso, o ChatGPT-4 teve um bom desempenho em questões de matemática.

Guillen-Grima et al. [14] avaliaram o desempenho dos modelos ChatGPT-3.5 e ChatGPT-4 no exame de acesso à Residência Médica Espanhola. O exame avaliado foi o de 2022, na língua nativa e nas traduções para o inglês, e obtiveram um desempenho ligeiramente superior com a última. No nosso trabalho, avaliamos 4 LLMs na resolução de duas edições do POSCOMP. Após a avaliação com *prompt* de imagem na língua nativa, fizemos a tradução do texto das questões para o inglês e também obtivemos um desempenho superior em todos os modelos.

Bommarito e Katz [15] avaliaram o desempenho do ChatGPT-3.5 na parte *Multistate Bar Examination* do modelo de Exame de Ordem da *National Conference of Bar Examiners*, um exame que é pré-requisito para a prática jurídica nos Estados Unidos. O ChatGPT-3.5 atingiu a taxa de aprovação no Exame e resultados que se equiparam aos examinados humanos. Na nossa avaliação, dois dos modelos testados no POSCOMP 2023 atingiram nota superior aos examinados humanos.

Bommarito et al. [16] avaliaram o desempenho do ChatGPT-3.5 em dois testes baseados nos *Blueprints* do Exame Uniforme do Contador Público Certificado do Instituto Americano de Contadores Públicos Certificados, um exame de contabilidade dos Estados Unidos. Na primeira avaliação, que envolvia raciocínio quantitativo, o modelo, utilizando o text-davinci-003, atingiu uma modesta precisão de 14,4%. Na segunda avaliação, focada em habilidades fundamentais sem raciocínio quantitativo, alcançou 57% de precisão, significativamente acima do acaso e aproximando-se do desempenho humano relatado. No nosso trabalho, nós avaliamos 4 LLMs na resolução de duas edições do POSCOMP usando *zero-shot prompt*. Além disso, o ChatGPT-4 teve um bom desempenho em questões de matemática.

Joshi et al. [17] examinaram a utilização do ChatGPT-4 como ferramenta educativa em graduandos em Ciência da Computação. Ao analisar o seu desempenho em vários tipos de perguntas, incluindo múltipla escolha, codificação e resposta curta/longa, o estudo encontrou imprecisões significativas, destacando potenciais riscos para a aprendizagem e a integridade acadêmica. Apesar destes desafios, o documento oferece recomendações para que estudantes e professores utilizem o ChatGPT-4 de forma construtiva para melhorar o ensino e a experiência educativa em geral. No nosso trabalho, nós avaliamos 4 LLMs na resolução de duas edições do POSCOMP e apesar de também termos encontrado imprecisões em casos específicos, a boa taxa de acerto dos modelos e suas explicações podem guiar estudantes em sua preparação.

Espejel et al. [18] avaliaram a capacidade de raciocínio dos modelos ChatGPT-3.5, ChatGPT-4 e BARD em tarefas de Processamento de Linguagem Natural. Os resultados demonstram que o ChatGPT-4 supera tanto o ChatGPT-3.5 como o BARD num cenário de *zero-shot prompt* em quase todas as tarefas. Apesar dos seus pontos fortes, os três modelos apresentam uma proficiência limitada nas tarefas de raciocínio indutivo e matemático. No nosso trabalho, nós avaliamos 4 LLMs na resolução de duas edições do POSCOMP usando *zero-shot prompt*. Além disso, o ChatGPT-4 teve um bom desempenho em questões de matemática.

Toyama et al. [19] avaliou o desempenho do ChatGPT-3.5, ChatGPT-4 e Google Bard foi avaliado através da resposta a 103 perguntas do Japan Radiology Board Examination. O ChatGPT-4

respondeu corretamente a 65,0% das perguntas, superando significativamente o ChatGPT-4 (40,8%) e o Google Bard (38,8%). O ChatGPT-4 se destacou em categorias que requerem pensamento de ordem inferior e em questões complexas de radiologia clínica. No nosso trabalho, nós avaliamos 4 LLMs na resolução de duas edições do POSCOMP usando *zero-shot prompt*. Além disso, o ChatGPT-4 teve um desempenho melhor do que o Gemini, modelo sucessor do Bard..

Nunes et al. [20] avaliaram os modelos ChatGPT-3.5 e ChatGPT-4 no Exame Nacional do Ensino Médio (ENEM). Analisando questões de 2009-2017 e 2022, o estudo utilizou várias estratégias de *prompt*, incluindo *Chain-of-Thought* (CoT). O ChatGPT-4 com CoT alcançou uma precisão de 87% no exame de 2022, superando o ChatGPT-3.5 em 11 pontos. No nosso trabalho, nós avaliamos 4 LLMs na resolução de duas edições do POSCOMP usando *zero-shot prompt*. O ChatGPT-4 usando texto obteve o melhor desempenho, inclusive acertando mais questões que todos os alunos que realizaram a prova do POSCOMP 2023.

Souza e Gheyi [6] investigaram a capacidade do ChatGPT-3.5, um chatbot de modelo de linguagem em grande escala, de resolver problemas de programação. De um total de 100 problemas submetidos, o modelo de linguagem resolveu corretamente 71 problemas em 3 tentativas, sendo 50 da plataforma LeetCode e 21 da plataforma BeeCrowd. No nosso trabalho, avaliamos 4 LLMs na resolução de questões do POSCOMP.

7. CONCLUSÕES

Neste trabalho apresentamos avaliamos até que ponto 4 LLMs conseguem resolver questões das edições de 2022 e 2023 do POSCOMP. Identificamos que modelos de linguagem de grande escala (LLMs), particularmente o ChatGPT-4, possuem um bom desempenho em resolver problemas dos exames do POSCOMP. O ChatGPT-4 consistentemente superou outros modelos, alcançando pontuações que o colocaram com o melhor resultado da edição de 2023. Seu desempenho superior, especialmente em Matemática e Fundamentos da Computação, sublinha seu potencial como uma ferramenta formidável para estudantes que se preparam para esses exames competitivos.

Outros modelos como o Gemini 1.0 Advanced e Le Chat Mistral também mostraram capacidade na resolução de problemas em várias disciplinas, embora não tenham atingido os patamares do ChatGPT-4. O Claude 3 Sonnet, embora promissor em explicação e compreensão, ficou um pouco atrás, com dificuldades particulares em Matemática. Identificamos que cada LLM possui áreas em que possuem pontos fortes e fracos.

O desempenho do ChatGPT-4 sugere que ele possui uma base de treinamento mais abrangente e possivelmente mecanismos mais sofisticados para interpretar e responder a consultas diversas e complexas em comparação com o Gemini 1.0 Advanced. Embora ambos os modelos tenham enfrentado problemas com alucinações e interpretações incorretas, estes foram mais pronunciados e prejudiciais no Gemini.

Os dados apontam a necessidade de melhorias contínuas nos Large Language Models, particularmente na compreensão do contexto e no manuseio eficaz de ambientes de teste dinâmicos e variados. Melhorias futuras poderiam focar no refinamento da interpretação de dados, raciocínio lógico e geração de respostas para alinhar mais estreitamente com as capacidades de compreensão e resolução de problemas semelhantes às humanas.

Para trabalhos futuros, planejamos expandir nossa análise para incluir outras versões do POSCOMP e incorporar LLMs mais diversos, como o Llama 3. Esse teste abrangente, juntamente com a exploração de diferentes estilos de *prompts* [9, 3], ajudará a entender melhor o potencial total dos LLMs em contextos educacionais. Por meio desses esforços, visamos refinar as habilidades dos modelos e possivelmente descobrir novas maneiras pelas quais eles podem auxiliar em ambientes educacionais, ampliando assim sua aplicabilidade e eficácia. Adicionalmente, é importante considerar outros estilos de *prompts* [9, 3] para avaliar se as LLMs conseguem ter um desempenho ainda melhor, e realizar um teste metamórfico com mais questões.

REFERÊNCIAS

[1] Yotam Intrator, Matan Halfon, Roman Goldenberg, Reut Tsarfaty, Matan Eyal, Ehud Rivlin, Yossi Matias, Natalia Aizenberg. 2024. Breaking the Language Barrier: Can Direct Inference Outperform Pre-Translation in Multilingual LLM Applications?. doi: 10.48550/arXiv.2403.04792.

[2] Yue Zhang et al. 2023. Siren's song in the AI ocean: a survey on hallucination in large language models. (2023). arXiv: 2309.01219 [cs.CL].

[3] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in Large Language Models. In Advances in Neural Information Processing Systems.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, 5998–6008.

[5] Junjie Wang, Yuchao Huang, Chunyang Chen, Zhe Liu, Song Wang, and Qing Wang. 2024. Software testing with Large Language Models: survey, landscape, and vision. doi: 10.48550/arXiv.2307.07221.

[6] Debora Souza and Rohit Gheyi. 2023. Estudo de caso: uso do ChatGPT para resolução de problemas de programação. In Brazilian Symposium on Software Engineering, CTIC, 80–89

[7] June Sallou, Thomas Durieux, and Annibale Panichella. 2024. Breaking the silence: the threats of using LLMs in software engineering. In ACM/IEEE 46th International Conference on Software Engineering - New Ideas and Emerging Results. ACM/IEEE.

[8] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. Deep Learning. MIT Press.

[9] Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, Li Li, Xiapu Luo, David Lo, John Grundy, Haoyu Wang. 2023. Large Language Models for software engineering: A systematic literature review. doi: 10.48550/ARXIV.2308.10620.

[10] Victor R. Basili, Gianluigi Caldiera, and H. Dieter Rombach. 1994. The Goal Question Metric Approach, 528–532.

[11] Chen, Tsong Yueh; Kuo, Fei-Ching; Liu, Huai; Poon, Pak-Lok; Towey, Dave; Tse, T. H.; and Zhou, Zhi Quan. "Metamorphic Testing: A Review of Challenges and Opportunities" (2018). Faculty of Engineering and Information Sciences - Papers: Part B. 975.

[12] Ramon Pires, Thales Sales Almeida, Hugo Queiroz Abonizio, Rodrigo Frassetto Nogueira: Evaluating GPT-4's Vision Capabilities on Brazilian University Admission Exams. CoRR abs/2311.14169 (2023)

[13] Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, Xipeng Qiu. 2023. Evaluating the Performance of Large Language Models on GAOKAO Benchmark. doi: 10.48550/arXiv.2305.12474.

[14] Francisco Guillen-Grima, Sara Guillen-Aguinaga, Laura Guillen-Aguinaga, Rosa Alas-Brun, Luc Onambele, Wilfrido Ortega, Rocio Montejo, Enrique Aguinaga-Ontoso, Paul Barach, and Ines Aguinaga-Ontoso. Evaluating ChatGPT efficacy in navigating the spanish medical residency entrance examination (mir): A new horizon for AI in clinical medicine. Clin. Pract. 2023, 13(6), 1460-1487; https://doi.org/10.3390/clinpract13060130

[15] Michael James Bommarito and Daniel Martin Katz. GPT Takes the Bar Exam. Available at SSRN 4314839, 2022.

[16] Jillian Bommarito, Michael James Bommarito, Jessica Katz, and Daniel Martin Katz. GPT as Knowledge Worker: A Zero-Shot Evaluation of (AI)CPA Capabilities. SSRN Electronic Journal, 2023.

[17] Ishika Joshi, Ritvik Budhiraja, Harshal Dev, Jahnvi Kadia, M. Osama Ataullah, Sayan Mitra, Harshal D. Akolekar, and Dhruv Kumar. 2018. ChatGPT in the Classroom: An Analysis of Its Strengths and Weaknesses for Solving Undergraduate Computer Science Questions. In Proceedings of Technical Symposium (SIGCSE TS'24). ACM, New York, NY, USA, 8 pages. https://doi.org/10.48550/arXiv.2304.14993

[18] Jessica López Espejel, El Hassane Ettifouri, Mahaman Sanoussi Yahaya Alassan, El Mehdi Chouham, Walid Dahhane. GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts. Natural Language Processing Journal, 5:100032, 2023

[19] Yoshitaka Toyama, Ayaka Harigai, Mirei Abe, Mitsutoshi Nagano, Masahiro Kawabata, Yasuhiro, Seki, Kei Takase. Performance evaluation of ChatGPT, GPT-4, and Bard on the official board examination of the Japan Radiology Society. Japanese Journal of Radiology, pages 1–7, 2023.

[20] Desnes Nunes et al. Evaluating GPT-3.5 and GPT-4 Models on Brazilian University Admission Exams. arXiv preprint arXiv:2303.17003, 2023.

[21] POSCOMP, Sociedade Brasileira de Computação. https://www.sbc.org.br/index.php?option=com_content&view=article&layout=edit&id=458 Acessado em Maio de 2024.

[22] ChatGPT, OpenAI, https://chat.openai.com Acessado em Maio de 2024

[23] Gemini, Google, https://gemini.google.com Acessado em Maio de 2024

[24] Claude, Anthropic, https://claude.ai Acessado em Maio de 2024

[25] Le Chat Mistral, Mistral AI, https://chat.mistral.ai/chat Acessado em Maio de 2024

[26] Cayo Viegas and Rohit Gheyi. Avaliando a Capacidade de LLMs na Resolução de Questões do POSCOMP - Dados https://docs.google.com/spreadsheets/d/1AaitdorAJX7SXzhBv36c_EAFii52yC_ihi7OjzPEXEg/edit?usp=sharing 2024.