Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
https://sol.sbc.org.br/journals/index.php/rbie

Submission: dd/Mmm/yyyy; 1st round notif.: dd/Mmm/yyyy; New version: dd/Mmm/yyyy; 2nd round notif.: dd/Mmm/yyyy;
Camera ready: dd/Mmm/yyyy; Edition review: dd/Mmm/yyyy; Available online: dd/Mmm/yyyy; Published: dd/Mmm/yyyy;

Avaliação da Capacidade dos Modelos de Linguagem de Grande Escala (LLMs) para Resolver Questões do POSCOMP

André Carvalho
Universidade de São Paulo
afcarvalho@usp.br

Bruno Henrique Ferreira Correia
Universidade de São Paulo
bhcorreia11@usp.br

Guilherme Secundo Santos
Universidade de São Paulo
guilhermesecundo@usp.br

João Gabriel de Senna Lamolha
Universidade de São Paulo
joao.lamolha@usp.br

Resumo
Este projeto tem como objetivo avaliar e comparar o desempenho de diferentes Modelos de Linguagem de Grande Escala (LLMs), incluindo Claude, Gemini, ChatGPT e Deepseek, na resolução de questões de múltipla escolha do exame POSCOMP. A pesquisa investiga cinco questões principais: (1) performance comparativa entre diferentes LLMs; (2) variação de desempenho por área temática; (3) evolução temporal do desempenho em comparação com estudos anteriores; (4) eficácia de recursos multimodais para questões com elementos visuais; (5) impacto de técnicas avançadas de engenharia de prompt no desempenho dos modelos. Os resultados contribuirão para compreender melhor as capacidades e limitações dos LLMs em contextos acadêmicos e avaliativos.
Palavras-chave:

Abstract
This project aims to evaluate and compare the performance of different Large Language Models (LLMs), including Claude, Gemini, ChatGPT, and Deepseek, in solving multiple-choice questions from the POSCOMP exam. The research investigates five main questions: (1) comparative performance among different LLMs; (2) variation in performance by thematic area; (3) temporal evolution of performance compared to previous studies; (4) effectiveness of multimodal resources for questions with visual elements; (5) impact of advanced prompt engineering techniques on model performance. The results will contribute to a better understanding of the capabilities and limitations of LLMs in academic and evaluative contexts.
Keywords:

Resumen
Este proyecto tiene como objetivo evaluar y comparar el rendimiento de diferentes Modelos de Lenguaje de Gran Escala (LLMs), incluyendo Claude, Gemini, ChatGPT y Deepseek, en la resolución de preguntas de opción múltiple del examen POSCOMP. La investigación investiga cinco preguntas principales: (1) rendimiento comparativo entre diferentes LLMs; (2) variación de rendimiento por área temática; (3) evolución temporal del rendimiento en comparación con estudios anteriores; (4) eficacia de recursos multimodales para preguntas con elementos visuales; (5) impacto de técnicas avanzadas de ingeniería de prompt en el rendimiento de los modelos. Los resultados contribuirán a comprender mejor las capacidades y limitaciones de los LLMs en contextos académicos y evaluativos.
Palabras clave:

Cite as: Carvalho, A., Correia, B. H. F., Santos, G. S., & Lamolha, J. G. S. (yyyy). Avaliação da capacidade dos Modelos de Linguagem de Grande Escala (LLMs) para resolver questões do POSCOMP. Revista Brasileira de Informática na Educação, vol, pp-pp. https://doi.org/10.5753/rbie.yyyy.id.

1 Introdução

Nos últimos anos, modelos de linguagem natural (LLMs) têm mostrado avanços significativos em diversas áreas, incluindo resolução de problemas e compreensão de textos complexos.

No cenário acadêmico nacional, o POSCOMP, exame aplicado para candidatos de pós-graduação em computação no Brasil, constitui um desafio ideal para avaliar a capacidade desses modelos. Esta prova é organizada anualmente pela Sociedade Brasileira de Computação (SBC) e aplicada pela FUNDATEC. Ele tem como finalidade avaliar conhecimentos fundamentais de candidatos a programas de mestrado e doutorado em computação no Brasil, funcionando como referência nacional para instituições de ensino superior. A prova é composta por questões de múltipla escolha que abrangem áreas como matemática, lógica, algoritmos, programação, arquitetura de computadores, sistemas operacionais, redes, inteligência artificial, banco de dados e engenharia de software, o que a torna ampla e desafiadora. Por sua relevância acadêmica e variedade temática, o POSCOMP permite investigar não apenas a precisão das respostas das LLMs, mas também a capacidade desses modelos de compreender instruções complexas, interpretar multimodalidade e se adaptar a diferentes estratégias de interação, como variações de prompt.

Este trabalho busca avançar além dos trabalhos correlatos, como o de (Saldanha & Digiampietri, 2024), ao realizar uma análise comparativa expandida envolvendo quatro modelos de destaque: Claude, Gemini, ChatGPT e Deepseek em múltiplas edições do POSCOMP. Especificamente, investigamos cinco questões de pesquisa fundamentais: (1) a performance comparativa entre diferentes LLMs; (2) a variação de desempenho por área temática; (3) a evolução temporal do desempenho em comparação com estudos anteriores; (4) a eficácia de recursos multimodais para questões com elementos visuais; e (5) o impacto de técnicas avançadas de engenharia de prompt no desempenho dos modelos.

2 Fundamentos Teóricos

O campo do Processamento de Linguagem Natural (PLN), que busca capacitar máquinas a compreender e gerar a linguagem humana, tem suas raízes nos anos 1950, com abordagens simbólicas baseadas em regras gramaticais codificadas manualmente. Inicialmente, as abordagens eram baseadas em regras *handcrafted* e sistemas especialistas, que dependiam fortemente da codificação manual de conhecimento linguístico por especialistas, sendo de escopo limitado, como ilustrado pelo programa ELIZA (Weizenbaum, 1966). Este programa, um marco inicial dessa era, simulava conversas usando um conjunto simples de regras de correspondência e substituição, demonstrando tanto o potencial quanto as óbvias limitações dessa abordagem.

Conceitualmente, o PLN divide-se em duas grandes subáreas: Interpretação (ou Compreensão) de Linguagem Natural (NLU) e Geração de Linguagem Natural (NLG) (Caseli & Nunes, 2024). Uma transição significativa no campo ocorreu com o advento de métodos estatísticos e de *machine learning*, onde modelos passaram a aprender probabilidades de ocorrência e associação de palavras a partir de grandes volumes de texto, evoluindo das técnicas puramente baseadas em regras. Essa evolução deu-se com a transição para modelos neurais profundos, que aprenderam representações distribuídas de palavras (*word embeddings*), e culminou com o surgimento revolu-

cionário da arquitetura Transformer (Vaswani et al., 2023), que abandonou as restrições de processamento sequencial de modelos como RNNs e LSTMs. Essa inovação substituiu o processamento sequencial por um mecanismo de atenção totalmente paralelizável, tornando computacionalmente viável treinar modelos em volumes de dados massivos e na escala de parâmetros que define os LLMs modernos.

A Arquitetura Transformer, apresentada em 2017, é um tipo de arquitetura de rede neural projetada para processar sequências de dados, como frases em texto. Ela se tornou a base para modelos de linguagem avançados como o GPT. Seu principal diferencial reside no mecanismo de auto-atenção (*self-attention*), que capacita o modelo a ponderar a importância de cada palavra em relação a todas as outras na mesma frase. Por exemplo, na sentença "O banco do parque estava vazio", a auto-atenção permite que o modelo interprete "banco" como um assento, e não como uma instituição financeira, ao focar no contexto fornecido pelas palavras "parque" e "vazio". Essa arquitetura é composta por um Codificador (*Encoder*), que processa a entrada para criar uma representação contextual rica, e um Decodificador (*Decoder*), que, com base nessa representação, gera a saída palavra por palavra. Diferentemente das redes neurais tradicionais que processam informações sequencialmente, o Transformer as processa em paralelo, utilizando a auto-atenção para entender as relações entre todas as palavras de uma vez só, o que a torna significativamente mais eficiente e poderosa para tarefas como tradução e geração de conteúdo.

Os Modelos de Linguagem de Grande Escala (LLMs) diferenciam-se qualitativamente de outros modelos por duas características fundamentais: sua escala massiva, com arquiteturas que tipicamente possuem mais de um bilhão de parâmetros (Zhao, 2023), podendo alcançar centenas de bilhões, e seu enquadramento como pilar central da Inteligência Artificial Generativa (Caseli & Nunes, 2024). Nesse contexto, isso os posiciona como ferramentas de propósito geral para processamento de linguagem.

Sua tarefa de pré-treinamento fundamental é a previsão da próxima palavra em sequências textuais massivas. Ao otimizar esta tarefa em escala monumental, os LLMs internalizam padrões complexos de linguagem, conhecimento factual e relações lógicos, desenvolvendo capacidades emergentes como execução de tarefas *zero-shot* (sem exemplos) e *few-shot* (com poucos exemplos), o que os torna adequados para aplicações diversificadas desde resolução de questões complexas até geração de código e texto criativo.

A engenharia de prompt é uma área que surge dentro do contexto de LLMs e tem por objetivo descrever e projetar entradas mais refinadas para os modelos de linguagem a fim de obter alguma melhoria desejada na resposta. Alguns exemplos de técnicas são o "Zero-Shot Prompting", o "Few-Shot Prompting", e o "Chain-of-Thought Prompting". O primeiro consiste em perguntar diretamente para o modelo e esperar uma resposta (Exemplo 1).

```
Prompt:
Classifique o texto em neutro, negativo ou positivo.
Texto: Acho que as férias estão boas.
Sentimento:

Resposta:
Positivo
```

Code 1: Exemplo de Zero-Shot Prompting..

A técnica de "Few-Shot" consiste em apresentar uma quantidade pequena de exemplos antes de pedir a resposta (Exemplo 2).

```
Prompt:
Isso é incrível! // Positivo
Isto é mau! // Negativo
Uau, esse filme foi muito legal // Positivo
Acho que as férias estão boas. //

Resposta:
Positivo
```

Code 2: Exemplo de Few-Shot Prompting..

A técnica de "Chain-of-Thought Prompting" se baseia em descrever o passo de raciocínio para resolver algum problema (Exemplo 3).

```
Prompt:
Pergunta: Bruce tem 5 bolas de tênis e comprou
mais 2 latas de bolas. Cada lata tem 3 bolas.
Quantas bolas de tênis ele possui agora?

Resposta: Bruce tinha no início 5 bolas, 2 latas
com 3 bolas cada são 6 bolas. 5 + 6 = 11. A
resposta é 11.

Pergunta: Um malabarista consegue fazer
malabarismos com 16 bolas. Metade delas são
bolas de golfe e a outra metade são azuis.
Quantas bolas de golfe azuis existem?

Resposta:
O malabarista consegue fazer malabarismos com 16
bolas. Metade das bolas são bolas de golfe.
Portanto, há 16/2 = 8 bolas de golfe. Metade das
bolas de golfe são azuis. Portanto, há 8/2 = 4
bolas de golfe azuis. A resposta é 4.
```

Code 3: Exemplo de Chain-of-Thought Prompting..

3 Trabalhos Relacionados

No contexto nacional, o trabalho de (Saldanha & Digiampietri, 2024) avaliou o desempenho do ChatGPT e Bard no exame POSCOMP, comparando-o com a média de candidatos humanos. Os autores verificaram que ambos os modelos performaram, em média, melhor sem a adição de um contexto simples ao prompt. No entanto, o estudo limitou-se a dois modelos e a questões textuais, sem explorar técnicas avançadas de prompt. Nosso projeto expande essa investigação ao incluir modelos mais recentes (Claude, Gemini, Deepseek), incorporar questões com elementos visuais e analisar técnicas de engenharia de prompt.

O estudo de (Ashrafimoghari et al., 2024) apresenta uma comparação entre os modelos da família GPT, Gemini e Claude para resolver questões do GMAT (*Graduate Management Admission Test*), um exame para a admissão em pós-graduações em negócios. O estudo demonstrou que a maioria dos modelos superou os candidatos humanos nas seções de raciocínio verbal e quantitativo do exame. O modelo que obteve melhores resultados foi o GPT-4 Turbo, que se colocou entre os 1% superior dos candidatos humanos. O estudo também revela que os modelos mais recentes (GPT-4 Turbo, Claude 2.1 e Gemini 1.0 Pro) possuem melhores resultados em tarefas de raciocínio. A maior parte dos LLMs testados se destacou na parte de compreensão textual e teve mais dificuldade com tarefas de suficiência de dados, que são questões cujo objetivo não é responder ao problema, mas sim dizer se a quantidade de informação fornecida é suficiente para responder à pergunta.

Um dos trabalhos mais relevantes que evidenciam o potencial dos modelos de linguagem em avaliações padronizadas é o de (Kung et al., 2023), que investigou o desempenho do ChatGPT em exames do USMLE (*United States Medical Licensing Examination*). Esse exame, composto por questões de múltipla escolha e considerado altamente desafiador, é utilizado para aferir a proficiência de profissionais da área médica nos Estados Unidos. Os autores verificaram que o modelo foi capaz de alcançar desempenho comparável ao de candidatos humanos, respondendo corretamente a uma parcela significativa das questões. Os resultados sugerem que os LLMs podem atuar como ferramentas auxiliares no ensino médico, tanto no processo de aprendizagem de estudantes quanto no apoio à tomada de decisão clínica. O caráter multimodal e interpretativo exigido pelo USMLE o torna um caso de estudo interessante, aproximando-se do desafio proposto pelo POSCOMP, no qual a interpretação de código, fórmulas e representações visuais também desempenha papel central.

O artigo de (Bordt & von Luxburg, 2023) oferece uma base metodológica e analítica de grande valor para a nossa pesquisa. A relevância desse estudo reside em sua abordagem rigorosa e na clareza com que apresenta a avaliação quantitativa do desempenho de um LLM em um domínio de conhecimento especializado. A pesquisa estabelece um precedente importante por sua análise de desempenho e modelo de comparação, demonstrando a diferença de performance entre o GPT-3.5 e o GPT-4, indicando que o modelo mais recente obteve uma melhoria significativa. Essa comparação direta entre diferentes versões de um mesmo LLM é um componente crucial para a nossa pesquisa, que busca justamente quantificar e analisar as diferenças de desempenho entre modelos distintos no exame POSCOMP. Além disso, a análise dos autores utiliza um ponto de referência crucial para contextualizar o desempenho dos modelos: eles comparam a pontuação da IA com a performance do estudante médio, mostrando que o modelo mais avançado alcançou um nível de desempenho similar ao humano. Essa abordagem de comparar a performance da IA com um *baseline* humano é um componente fundamental para a nossa pesquisa.

4 Método

A metodologia deste trabalho está estruturada nas seguintes etapas:

1. **Seleção das provas:** Serão coletadas provas completas de diferentes anos do POSCOMP, disponibilizadas pela FUNDATEC. A escolha desse exame se justifica por sua abrangência em tópicos fundamentais da ciência da computação e por representar um instrumento consolidado de avaliação. As provas serão organizadas em um banco de dados digital, categorizando cada questão segundo as áreas temáticas definidas pelo exame (Matemática, Fundamentos da Computação, Tecnologia da Computação, etc.).

2. **Execução dos testes:** Cada questão será submetida a quatro modelos de linguagem de última geração (Claude, Gemini, ChatGPT e DeepSeek). Para cada modelo, serão testados dois cenários de prompt: (i) envio da questão de forma direta; (ii) após analise prelimiar do resultado do teste (i) será realizado um teste posterior considerando técnicas de engenharia de prompt, como "Chain-of-Thought" e "Few-Shot", para o modelo com pior desempenho.

3. **Coleta e organização dos dados:** Todas as respostas serão registradas em planilhas no formato CSV, contendo os seguintes elementos: (a) identificação da questão e gabarito oficial;

(b) modelo de IA utilizado; (c) resposta fornecida; (d) indicadores binários sobre o uso de recursos complementares (imagem, gráfico, fórmula ou código).

4. **Análise e visualização dos resultados:** Será realizada uma comparação de desempenho entre modelos de forma geral e por área temática da prova. Também será avaliado o efeito da engenharia de prompt no desempenho das IAs. Os resultados serão sistematizados por meio de tabelas, gráficos e estatísticas descritivas, de modo a oferecer uma visão clara e objetiva das tendências observadas.

5 Cronograma

O cronograma previsto para a execução do projeto está detalhado na Tabela 1.

Table 1: Cronograma de atividades do projeto..

Etapa	Atividade	Período
1	Coleta de dados: download das provas do POSCOMP no site da FUNDATEC	Até 30/08
2	Envio das questões aos modelos de linguagem (Claude, Gemini, ChatGPT, Deepseek), em dois cenários de prompt	01/09 – 17/10
3	Organização dos resultados em CSV	Paralelo à etapa 2
4	Análise de dados: estatísticas, gráficos e tabelas	18/10 – 31/10
5	Escrita do artigo: redação dos capítulos	01/11 – 14/11
6	Revisão e entrega final	15/11 – 17/11

References

Ashrafimoghari, V., Gürkan, N., & Suchow, J. W. (2024). Evaluating large language models on the gmat: Implications for the future of business education.

Bordt, S., & von Luxburg, U. (2023). Chatgpt participates in a computer science exam.

Caseli, H. M., & Nunes, M. G. V. (Eds.). (2024). *Processamento de linguagem natural: Conceitos, técnicas e aplicações em português* (2nd ed.). BPLN.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., et al. (2023). Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLOS Digit Health*, *2*(2), e0000198.

Saldanha, M. S., & Digiampietri, L. A. (2024). Chatgpt and bard performance on the poscomp exam. *Proceedings of the 20th Brazilian Symposium on Information Systems*, 1–10.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

Zhao, W. X. e. a. (2023). A survey of large language models.