

ChatGPT and Bard Performance on the POSCOMP Exam

Mateus Santos Saldanha
Universidade de São Paulo
São Paulo, SP, Brazil
mateusaldanha@usp.br

ABSTRACT

Context: Modern chatbots, built upon advanced language models, have achieved remarkable proficiency in answering questions across diverse fields. **Problem:** Understanding the capabilities and limitations of these chatbots is a significant challenge, particularly as they are integrated into different information systems, including those in education. **Solution:** In this study, we conducted a quantitative assessment of the ability of two prominent chatbots, ChatGPT and Bard, to solve POSCOMP questions. **IS Theory:** The IS theory used in this work is Information processing theory. **Method:** We used a total of 271 questions from the last five POSCOMP exams that did not rely on graphic content as our materials. We presented these questions to the two chatbots in two formats: directly as they appeared in the exam and with additional context. In the latter case, the chatbots were informed that they were answering a multiple-choice question from a computing exam. **Summary of Results:** On average, chatbots outperformed human exam-takers by more than 20%. Interestingly, both chatbots performed better, in average, without additional context added to the prompt. They exhibited similar performance levels, with a slight advantage observed for ChatGPT. **Contributions and Impact in the IS area:** The primary contribution to the field involves the exploration of the capabilities and limitations of chatbots in addressing computing-related questions. This information is valuable for individuals developing Information Systems with the assistance of such chatbots or those relying on technologies built upon these capabilities.

CCS CONCEPTS

- Computing methodologies → Artificial intelligence; • Information systems → Document representation.

KEYWORDS

Large Language Model, ChatBot, Computer Science Examination, ChatGPT, Bard

ACM Reference Format:

Mateus Santos Saldanha and Luciano Antonio Digiampietri. 2024. ChatGPT and Bard Performance on the POSCOMP Exam. In *XX Brazilian Symposium on Information Systems (SBSI '24)*, May 20–23, 2024, Juiz de Fora, Brazil. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3658271.3658320>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SBSI '24, May 20–23, 2024, Juiz de Fora, Brazil

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0996-8/24/05...\$15.00
<https://doi.org/10.1145/3658271.3658320>

Luciano Antonio Digiampietri
Universidade de São Paulo
São Paulo, SP, Brazil
digiampietri@usp.br

1 INTRODUÇÃO

Nas últimas décadas presenciamos diversas grandes mudanças nos sistemas de informação que se tornaram onipresentes no cotidiano das pessoas. A partir dos modernos aparelhos celulares (*smartphones*), as pessoas têm acesso ao e-mail, aplicativos de mensagens, redes sociais online, aplicativos de navegação, entretenimento, entrega de comida e outros itens, etc.

Muitos desses sistemas já possuem algum tipo de inteligência artificial embutida (por exemplo, para a recomendação de músicas ou vídeos), porém, com a nova geração dos Grandes Modelos de Linguagem (*Large Language Models - LLMs*), que surgiram a partir do trabalho de Vaswani et al., 2017 [14] e, especialmente, com a disponibilização para a população em geral do ChatGPT [7] em 2022, uma grande parcela da população teve acesso direto e voluntário a um sistema baseado em inteligência artificial que consegue responder, a princípio, a qualquer pergunta que lhe for feita. Vale ressaltar que as respostas providas pelo ChatGPT podem estar erradas, porém pela qualidade e fluidez do texto, bem como pela capacidade de prover respostas corretas a uma gama muito grande de perguntas sobre diferentes tópicos, esse sistema maravilhou e também chocou muitas pessoas. Adicionalmente, o sistema não apenas responde perguntas no sentido mais tradicional, isto é, perguntas do tipo “o que é isso”, “explique aquilo”, etc. Ele também é capaz de executar diferentes tipos de comandos (por exemplo, “resuma o seguinte texto”, “escreva um programa na linguagem de programação C que faça isso”, etc.). Desta forma, este sistema e outros lançados nos últimos anos se mostraram excelentes sistemas de Inteligência Artificial para uso mais geral, se tornando, também, o estado-da-arte para alguns problemas específicos, em particular aqueles ligados ao processamento de Linguagens Naturais. [3, 10].

As inteligências artificiais desempenham um papel crucial nos sistemas de informação, mas enfrentam desafios fundamentais. A generalização é um desses desafios, pois a IA pode ser excelente em alguns tipos de tarefas e áreas do conhecimento, mas frequentemente falha em aplicar seu “conhecimento” a situações não vistas anteriormente e que envolvem um contexto mais complexo. Essa dificuldade de compreensão é um obstáculo, uma vez que os sistemas de IA frequentemente não conseguem entender nuances e contextos específicos da conversa, o que pode levar a respostas inadequadas.

Neste contexto, é importante avaliar a capacidade desses novos sistemas em resolver problemas específicos ou em responder a perguntas de áreas determinadas de forma a se entender melhor suas capacidades e limitações.

O presente trabalho tem como objetivo avaliar o desempenho da versão de acesso gratuita de dois sistemas que usam grandes modelos de linguagem (ChatGPT¹ e Bard²) em responder questões do

¹ChatGPT: <https://chat.openai.com/>, acessado em 06/10/2023.

²Bard: <https://bard.google.com/>, acessado em 06/10/2023.

Exame Nacional para Ingresso na Pós-Graduação em Computação (POSCOMP)³

Há três questões de pesquisa investigadas no presente trabalho:

1. O desempenho dos *chatbots* Bard e ChatGPT é superior à média dos candidatos que realizaram o POSCOMP?
2. O desempenho dos *chatbots* varia nos diferentes assuntos (ou seções) do POSCOMP?
3. A passagem de um contexto simples no *prompt* dos *chatbots* aumenta o desempenho deles no POSCOMP?

O restante deste artigo está organizado da seguinte forma. A seção 2 apresenta os trabalhos correlatos. Já a seção 3 descreve os materiais e métodos. A seção 4 contém a descrição e análise dos resultados. Por fim, a seção 5 apresenta as conclusões.

2 TRABALHOS CORRELATOS

Desde, ao menos, a década de 1950, tem se discutido a criação de sistemas de software capazes de resolver diferentes problemas e, em especial, sistemas que pudessem ser considerados “inteligentes” (capazes de imitar/simular a “inteligência humana”) [11].

Desde essa década, sistemas foram desenvolvidos ou com objetivos bastante específicos (resolver um único problema usando abordagens estatísticas ou regressões lineares) ou mais ambiciosos, como o desenvolvimento de um Solucionador de Problemas Geral (*General Problem Solver*) [6].

Ao longo das décadas diversos sistemas foram desenvolvidos, alguns com 100% de precisão na execução de suas tarefas (como calculadores científicas) e outros usando inteligência artificial com precisões variando de acordo com o problema tratado e conjunto de dados disponível para treinamento [11].

Os primeiros sistemas que tinha como objetivo resolver uma grande gama de problemas tinham a limitação de exigir que o usuário definisse de maneira muito detalhada (tipicamente formal) o problema a ser tratado: os dados de entrada e as regras ou operações permitidas [6], o que tornou vários desses sistemas inacessíveis a usuários leigos. Adicionalmente, as abordagens utilizadas na busca por uma solução poderiam sofrer do problema da explosão combinatória [4] ou os dados necessários poderiam ter alta dimensionalidade (o que poderia levar à chamada “maldição da dimensionalidade” (*curse of dimensionality*), o que inviabilizou a execução desses sistemas para vários problemas reais.

Nos últimos anos, a combinação da disponibilização de enormes volumes de dados em formato digital, o grande aumento da capacidade de processamento dos computadores (em especial com o uso de GPUs) e o desenvolvimento de novas abordagens de inteligência artificial e/ou de representação de conhecimento possibilitaram o que pode ser chamada de uma revolução desses sistemas, permitindo não apenas sistemas que conseguem responder a uma gama muito grande de perguntas de diferentes assuntos e que, em especial, conseguem responder a perguntas formuladas por humanos, sem a necessidade de uma representação especial ou treinamento do usuário.

Um dos trabalhos que foi base para o que muitos têm chamado de “revolução”, é o artigo intitulado “Attention Is All You Need.” [14]. Neste artigo é apresentada a arquitetura dos *transformers* que é uma arquitetura de uma rede neural profunda para a criação de modelos

de linguagem inicialmente utilizada para a tradução de texto. Esta arquitetura é baseada na ideia de identificar a atenção de cada palavra de um texto em relação às demais. Por exemplo, na frase: “O menino está triste pois perdeu seu brinquedo.”, a palavra *seu* possui mais “atenção” (está mais relacionada) com as palavras *menino* e *brinquedo*. A premissa dos autores, confirmada em diferentes testes, é que um modelo de linguagem construído utilizando esse tipo de mecanismo de atenção será mais robusto/preciso do que os modelos anteriores que não utilizavam desse tipo de mecanismo.

A arquitetura de *transformers* [14] foi a base para a construção dos grandes modelos de linguagem (do inglês, *LLMs - Large Language Models*) que surgiram nos últimos anos, como GPT [10], BERT [3], LaMDA [12] e Llama [13]. Estes modelos, treinados com grandes volumes de dados, se mostraram capazes de representar o conhecimento de uma forma inédita, atingindo o estado-da-arte na resolução de diferentes problemas de Processamento de Linguagem Natural [3, 10].

Modelos de linguagem são usados há diversas décadas por profissionais de computação ou áreas correlatas. Porém, a grande inovação ou, ao menos a funcionalidade que possibilitou o uso pela população em geral, foi a criação de *chatbots* (programas conversacionais) que utilizam esses modelos de linguagem e a disponibilização via navegador web ou aplicativos. Esses sistemas se destacam pela capacidade de “entender” (ou ao menos passar essa impressão aos usuários) bem as perguntas formuladas por humanos e por produzir respostas em formato de texto com fluidez semelhante a de humanos, respondendo, muitas vezes, de forma correta a pergunta formulada. Destacam-se nessa categoria os *chatbots* ChatGPT da OpenAI⁴ e Bard da Google⁵.

Esses *chatbots* foram desenvolvidos para, a princípio, responder a qualquer pergunta realizada pelo usuário (havendo algumas restrições relacionadas a temas específicos, por exemplo, o sistema não deve ensinar um usuário a construir uma bomba caseira). Apesar da funcionalidade de responder a “qualquer” pergunta, esses sistemas não têm garantia de que a resposta está correta. A forma que o modelo de linguagem é treinado e posteriormente consultado não permite ao sistema saber se a resposta que está sendo produzida está correta ou não.

Nesta conjuntura, é importante avaliar a capacidade desses *chatbots* em responder a perguntas de diferentes áreas. Os trabalhos correlatos apresentam resultados bastante positivos em ciências exatas, evidenciando resultados superiores a de humanos medianos, mas também apresentando algumas respostas incorretas [1, 5, 8, 9].

O trabalho [8] avaliou o desempenho de diferentes versões do GPT 3.5 e 4, bem como da modificação do *prompt* (isto é, da forma em que a pergunta é feita ao sistema) em questões de Fundamentos de Engenharia Ambiental (*Fundamentals of Engineering Environmental*) da Universidade de Iowa. Os autores avaliaram o sistema em diferentes assuntos ligados à área de Fundamentos de Engenharia Ambiental e concluíram que com um uso simples de engenharia de *prompt* o sistema é capaz de passar nas provas. Os autores destacam que estes modelos de linguagem (ou os respectivos *chatbots*) podem se tornar ferramentas úteis para o aperfeiçoamento do aprendizado, porém, as instituições de ensino devem se preocupar em preparar

⁴OpenAI: <https://openai.com/>, acessado em 06/10/2023.

⁵Google: <https://www.google.com/>, acessado em 06/10/2023.

³POSCOMP: <https://www.sbc.org.br/educacao/poscomp>, acessado em 27/09/2023.

provas que não possam ser realizadas por esses sistemas, de forma a efetivamente verificar se os estudantes aprenderam os conteúdos necessários.

Em [5], os autores testaram o uso do ChatGPT em provas relacionadas à área de segurança em computação. O sistema foi testado de diferentes formas para auxiliar a responder as perguntas da prova: desde se copiar diretamente a resposta do sistema, mas também sendo usado como ferramenta auxiliar dos estudantes que interpretavam as respostas do ChatGPT e com base nisso preparavam suas respostas. O artigo destaca que para algumas tarefas o ChatGPT forneceu diretamente respostas, na média, superiores às dos estudantes, porém, para algumas tarefas como a escrita de um “ensaio” (*essay*), na média, os estudantes foram melhores do que o ChatGPT. Os autores ainda destacam que os *chatbots* têm grande potencial como ferramenta de auxílio ao ensino, mas também facilitam a ocorrência de fraudes.

O trabalho [1] avalia o uso do ChatGPT em responder a perguntas da disciplina “Algoritmos e Estruturas de Dados”. A prova incluía questões discursivas que foram transcritas à mão e enviado para correção de um professor, juntamente com provas feitas por estudantes. Tanto o ChatGPT-3.5 quanto o ChatGPT-4 foram capazes de passar na prova, sendo que a versão mais nova do sistema obteve desempenho 17% superior ao da versão inicial.

O trabalho [9] avaliou o uso do ChatGPT como ferramenta auxiliar na resolução de problemas de programação e estruturas de dados. Os estudantes que participaram do estudo foram divididos em dois grupos, um com acesso a livros e outros materiais didáticos e outro com acesso ao ChatGPT. Na média, estudantes com acesso ao ChatGPT tiveram desempenho superior em todas as tarefas (variando de um aumento de desempenho de 5% a notas mais do que duas vezes maiores do que aquelas dos estudantes que não tiveram acesso à ferramenta. As tarefas envolveram o desenvolvimento de códigos, alguns extensos (centenas de linhas). Os autores observaram que apesar da clara melhoria no desempenho dos estudantes que utilizaram o ChatGPT, o tempo da resolução das tarefas variou bastante: para tarefas simples, o uso do ChatGPT foi bastante eficiente, já para tarefas complexas os estudantes gastaram mais tempo corrigindo/ajustando os códigos (*debugando*) do que o tempo que foi gasto pelos estudantes que não utilizaram a ferramenta.

“O Exame Nacional para Ingresso na Pós-Graduação em Computação (POSCOMP) é um exame aplicado em todas as regiões do País. Em parceria com a Sociedade Peruana de Computação, desde 2006, o Exame passou a ser realizado no Peru. O POSCOMP testa conhecimentos na área de Computação e tem como objetivo específico avaliar os conhecimentos de candidatos a Programas de Pós-Graduação em Computação oferecidos no Brasil. A grande maioria dos Programas de Pós-Graduação no país utiliza, de alguma forma, o resultado do POSCOMP em seu processo seletivo.”⁶ Este exame conta com a participação anual de mais de 3.000 pessoas que realizam a prova dividida em três macro-temas: Matemática (20 questões), Fundamentos da Computação (30 questões) e Tecnologia da Computação (20 questões). As 70 questões são de múltipla escolha, com cinco opções. O desempenho médio no POSCOMP costuma ser de

cerca de 30 questões corretas (isto é, menos de 43% das questões) e o desvio padrão costuma ficar em torno de 7 ou 8 pontos⁷.

Não foi encontrado na literatura nenhum trabalho avaliando o ChatGPT ou o Bard na tarefa de responder a questões do POSCOMP. O presente trabalho tem por objetivo ajudar a preencher esta lacuna.

3 MATERIAIS E MÉTODOS

A coleta de dados para a comparação entre o ChatGPT e o Google Bard envolveu a seleção criteriosa de questões das cinco últimas edições da Prova para Ingresso em Programas de Pós-Graduação em Computação (POSCOMP), referentes aos anos de 2016, 2017, 2018, 2019 e 2022. É importante ressaltar que não houve aplicação da prova nos anos de 2020 e 2021 devido à pandemia de COVID-19. Tanto as provas quanto os gabaritos foram obtidos do site oficial do POSCOMP⁸. Essas questões foram escolhidas para cobrir uma ampla gama de tópicos em Ciência da Computação, proporcionando um desafio significativo para ambos os sistemas de IA.

Os procedimentos adotados desdobram-se em três etapas principais:

1. Seleção de Perguntas do POSCOMP: Durante esta etapa, foram selecionadas questões que podiam ser apresentadas de forma fiel às inteligências artificiais avaliadas. Questões que continham tabelas, fórmulas matemáticas graficamente complexas ou imagens foram excluídas, priorizando questões que pudesse ser totalmente interpretadas com base no texto. O objetivo foi garantir que as respostas dos modelos fossem baseadas apenas no contexto textual da pergunta.
2. Interações Simultâneas: As perguntas selecionadas foram apresentadas simultaneamente ao ChatGPT e ao Google Bard. Essa abordagem permitiu uma avaliação comparativa em tempo real, possibilitando a observação das diferenças de desempenho entre os dois sistemas.
3. Avaliação de Respostas: Após a geração das respostas pelos sistemas, cada resposta foi avaliada quanto à sua assertividade em relação ao gabarito oficial da avaliação, ou seja, os modelos respondiam com a alternativa correta (A, B, C, D e E) e somente isso foi levado em consideração para a avaliação. Com base no conjunto das respostas, foi possível verificar a precisão dos modelos de forma total, mas também em recortes de acordo com o ano da prova e do tema da questão.

A seleção das perguntas buscou amostras representativas e diversas para demonstrar os desafios típicos enfrentados pelos candidatos no POSCOMP, garantindo que os modelos fossem testados em uma variedade de tópicos e níveis de complexidade. Para isto, todas as questões que eram baseadas apenas em conteúdos textuais das últimas cinco edições foram utilizadas, totalizando 271 questões do total de 350 questões das cinco edições do exame consideradas.

Foi possível analisar a assertividade das inteligências artificiais tanto observando o desempenho geral, como com base nos macro-temas da prova (Matemática, Fundamentos da Computação

⁷POSCOMP: médias e desvios padrão: <https://www.sbc.org.br/documentos-da-sbc/summary/185-poscomp/1296-media-desvio-padrao-poscomp>, acessado em 12/10/2023.

⁸POSCOMP - Provas e Gabaritos: <https://www.sbc.org.br/documentos-da-sbc/category/153-provas-e-gabaritos-do-poscomp>, acessado em 27/09/2023.

⁶POSCOMP: <https://www.sbc.org.br/poscomp>, acessado em 06/10/2023

e Tecnologia de Computação). Isso possibilitou uma comparação detalhada da assertividade dos *chatbots* testados.

Para enriquecer a avaliação, uma análise estatística foi realizada, empregando testes de média e desvio padrão. Essas análises têm o propósito de identificar padrões e diferenças significativas nos desempenhos do ChatGPT e do Google Bard, adicionando mais uma dimensão quantitativa à avaliação.

Outro aspecto relevante foi a comparação entre a quantidade de questões corretamente respondidas pelos sistemas avaliados e das notas obtidas por candidatos humanos nas respectivas provas. Isso permitiu uma verificação da média e desvio padrão das respostas entre humanos e inteligências artificiais em relação à prova como um todo e em relação a matérias específicas.

Cada prova do POSCOMP é composta por 70 questões de múltipla escolha, divididas em 20 questões de Matemática, 30 de Fundamentos da Computação e 20 de Tecnologia de Computação. Ao utilizar provas das cinco últimas edições, foram selecionadas 271 perguntas de um total de 350 possíveis (77,4%).

A Tabela 1 apresenta o número de questões de cada macro-tema e para cada uma das cinco edições do POSCOMP utilizadas no presente trabalho.

Destaca-se que, devido à características não textuais presentes em diversas questões, apenas 61% das questões de Matemática puderam ser utilizadas. Por outro lado, 97% das questões de Tecnologia de Computação foram utilizadas no presente trabalho e 75,3% de Fundamentos da Computação.

As médias das respostas dos candidatos e dos modelos, tanto por ano quanto por tema, foram comparadas para identificar eventuais desvios significativos. Isso permitiu determinar se os modelos se aproximam das médias obtidas pelos candidatos humanos. Além disso, testes estatísticos foram aplicados para verificar a significância estatística das diferenças observadas entre as respostas dos modelos e dos candidatos humanos.

O desempenho geral das pessoas que realizaram o POSCOMP, incluindo as médias e desvios padrão, estão disponibilizados no site do exame⁹. O desempenho dos candidatos pode ser observado na Tabela 2.

Vale destacar que o aproveitamento médio dos candidatos por ano considera as 70 questões do POSCOMP e não apenas as questões selecionadas neste trabalho, pois o desempenho individual de cada pessoa em cada questão não é uma informação disponível publicamente. Apesar desta diferença, acreditamos que a comparação de desempenho seja suficientemente justa, pois as questões selecionadas correspondem à mais de 77,4% de todas as questões do POSCOMP.

Para cada questão selecionada do POSCOMP, dois *prompts* de consulta foram realizados para cada um dos *chatbots* utilizados. No primeiro, apenas a questão era apresentada. Já no segundo, um contexto simples era apresentado junto à questão, conforme indicado a seguir: “Estou montando o gabarito para questões de múltipla escolha sobre computação. Cada questão tem um enunciado e cinco alternativas (de A até E). Você consegue me dizer qual é a alternativa correta?”.

⁹POSCOMP: médias e desvios padrão: <https://www.sbc.org.br/documents-dasbc/summary/185-poscomp/1296-media-desvio-padrao-poscomp>, acessado em 12/10/2023.

The screenshot shows two separate AI-generated responses from different models. Both responses are in Portuguese and correctly identify option B as the answer to a question about the Singleton design pattern. The first response is a detailed explanation of the Singleton pattern's implementation, while the second is a brief summary. Both responses include a 'Regenerate' button at the bottom right.

Assinale a alternativa correta em relação ao padrão de projeto Singleton.

A) Possui apenas 2 classes.
B) É instanciado através da chamada de um método público e estático.
C) Possui um membro privado não estático da própria classe.
D) Tem que ter o construtor público para funcionar.
E) Não é um padrão de criação.

A alternativa correta em relação ao padrão de projeto Singleton é a opção B) *É instanciado através da chamada de um método público e estático.* O Singleton é um padrão de projeto que garante que uma classe tenha uma única instância e fornece um método público e estático para acessar essa instância única. Isso garante que apenas uma instância da classe seja criada e que ela seja acessível globalmente.

Assinale a alternativa correta em relação ao padrão de projeto Singleton.

A) Possui apenas 2 classes.
B) É instanciado através da chamada de um método público e estático.
C) Possui um membro privado não estático da própria classe.
D) Tem que ter o construtor público para funcionar.
E) Não é um padrão de criação.

A resposta correta é (B). O padrão de projeto Singleton é instanciado através da chamada de um método público e estático. Esse método é responsável por criar uma instância da classe Singleton, se ela ainda não existir, ou retornar a instância já criada.

Figure 1: Questão em que ambos os modelos acertaram

Apenas para ilustrar questões e respectivas respostas, as Figuras 1 e 2 apresentam, respectivamente, um exemplo de *prompt* em que ambos os *chatbots* acertaram e outro em que ambos erraram.

Todos os *prompts* utilizados neste trabalho estão disponíveis de forma online para permitir a reprodução ou extensão do presente trabalho¹⁰.

4 RESULTADOS

A análise realizada neste trabalho tem por objetivo obter uma visão aprofundada do desempenho dos modelos ChatGPT e Google Bard em relação aos candidatos humanos que participaram da Prova para Ingresso em Programas de Pós-Graduação em Computação (POSCOMP). Os resultados contribuem para a compreensão das capacidades e limitações das inteligências artificiais no contexto de avaliações acadêmicas em Ciência da Computação.

Ao longo das próximas seções, são apresentados os desempenhos do ChatGPT e do Bard em relação a cada uma das últimas cinco edições do POSCOMP, bem como em relação a cada um dos macro-temas da prova.

Inicialmente será apresentada um análise do desempenho geral, seguida por uma análise dos macro-temas.

4.1 Desempenho Geral

A Tabela 3 e a Figura 3 apresentam o aproveitamento geral dos candidatos do POSCOMP e dos quatro modelos avaliados neste

¹⁰Sítio com os *prompts* omitido por causa do processo de revisão às cegas.

Table 1: Questões selecionadas do POSCOMP

Ano	Matemática	Porcentagem	Fundamentos de Computação	Porcentagem	Tecnologia da Computação	Porcentagem
2016	11	55,0%	20	66,7%	18	90,0%
2017	11	55,0%	20	66,7%	20	100,0%
2018	11	55,0%	25	83,3%	19	95,0%
2019	13	65,0%	23	76,7%	19	95,0%
2022	15	75,0%	25	83,3%	19	95,0%
Total	61	61,0%	113	75,3%	97	97,0%

Table 2: Médias de Pontos e Aproveitamento Médio dos Candidatos por Ano e por Macro-tema

Ano	Média de Pontos	Desvio Padrão	Aprov. Médio	Média Matemática	Aprov. Matemática	Média Fundamentos de Computação	Aprov. Fundamentos de Computação	Média Tecnologia da Computação	Aprov. Tecnologia da Computação
2016	29,62	7,17	42,31%	8,1	40,50%	14,31	47,70%	7,21	36,05%
2017	32,2	6,79	46,00%	8,23	41,15%	15,37	51,23%	8,6	43,00%
2018	28,8	7,62	41,14%	7,51	37,55%	12,37	41,23%	8,92	44,60%
2019	28,67	8,35	40,96%	8,52	42,60%	13,54	45,13%	6,61	33,05%
2022	28,94	8,81	41,34%	8,36	41,80%	14	46,67%	6,58	32,90%

Em sistemas de arquivos distribuídos, o requisito no qual os programas clientes não devem conhecer a distribuição de arquivos, sendo que um único conjunto de operações é fornecido para acesso a arquivos locais e remotos, é denominado transparência de

- A) acesso.
- B) desempenho.
- C) escala.
- D) localização.
- E) mobilidade.

O requisito no qual os programas clientes não devem conhecer a distribuição de arquivos, sendo que um único conjunto de operações é fornecido para acesso a arquivos locais e remotos, é denominado transparência de **localização**. Portanto, a resposta correta é a opção

- D) localização.**

Em sistemas de arquivos distribuídos, o requisito no qual os programas clientes não devem conhecer a distribuição de arquivos, sendo que um único conjunto de operações é fornecido para acesso a arquivos locais e remotos, é denominado transparência de

- A) acesso.
- B) desempenho.
- C) escala.
- D) localização.
- E) mobilidade.

Acessar outros rascunhos ▾

A resposta correta é (D), transparência de localização. A transparência de localização visa ocultar dos usuários e aplicações a localização de um recurso, ou seja, o conhecimento de sua localização física ou na rede. Nesse caso, os programas clientes não devem conhecer a distribuição de arquivos, sendo que um único conjunto de operações é fornecido para acesso a arquivos locais e remotos.

Figure 2: Questão em que ambos os modelos erraram. A alternativa correta é a alternativa A

trabalho. Destaca-se que, para os candidatos (humanos), o desempenho é relativo à prova toda (70 questões) enquanto que para os *chatbots* o aproveitamento considera apenas as questões totalmente textuais (em torno de 77% das questões).

Para cada linha na Tabela 3, a maior porcentagem de aproveitamento está destacada em negrito. Observa-se que todos os resultados dos *chatbots* foram superiores à média dos candidatos humanos,

porém, apenas nos anos de 2018 e 2022 os melhores resultados dos *chatbots* foram capazes de superar a média dos candidatos acrescida de um desvio padrão do respectivo ano (dados do desvio padrão são apresentados na Tabela 2). Destaca-se que o melhor resultado de 2022 (ChatGPT sem contexto) obteve um resultado 58% superior à média dos candidatos.

Ao analisar cinco edições do POSCOMP, constata-se que o ChatGPT e o Google Bard tiveram desempenhos relevantes. O ChatGPT foi superior nas edições de 2016 e 2022, enquanto o Google Bard nas edições de 2017 e 2018. Em 2019, ambos apresentaram resultados equivalentes. Considerando a análise de dez resultados de cada modelo, levando em conta diferentes contextos, verificou-se que ambos tiveram desempenhos similares. Na média, o ChatGPT alcançou 51,94% e o Google Bard, 51,52%. Essas médias são aproximadamente 25% superiores às médias gerais dos candidatos.

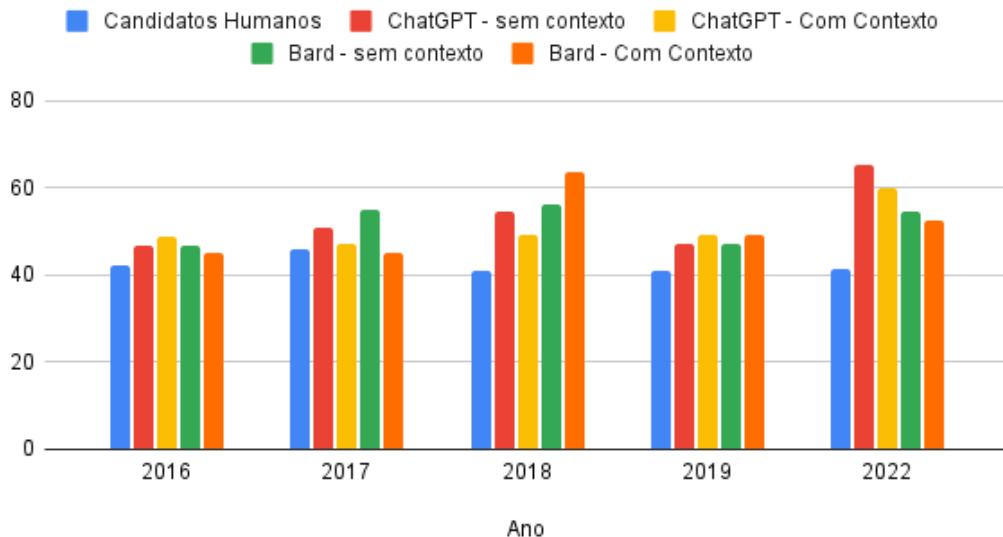
Ao analisar o uso ou não do contexto simplificado, observa-se que em três das cinco edições do POSCOMP o melhor resultado utilizou o contexto e em dois não. Porém, ao analisar a média de desempenho de cada *chatbot* nas cinco edições, os resultados sem contexto foram, na média, superiores (53,03% contra 50,84% para o ChatGPT e 51,97% contra 51,07% para o Google Bard). A contextualização, dependendo de como é estruturada, pode não apenas não contribuir, mas potencialmente confundir o modelo, levando a resultados menos precisos. Essa observação indica a importância de otimizar a maneira como o contexto é apresentado, ajustando-o de forma a maximizar a relevância e a clareza para o *chatbot*, o que poderia variar significativamente os resultados obtidos.

Apesar da pequena quantidade de valores de desempenho (cinco para cada modelo analisado), o teste-t indicou diferenças significativas de desempenho entre o resultado dos candidatos humanos e dos *chatbots*. Obteve-se confiança acima 95% para o ChatGPT com e sem contexto e para o Bard sem contexto. Já para o Bard com contexto a confiança ficou acima de 90% (a menor estabilidade dos

Table 3: Aproveitamento Geral no POSCOMP

Ano	Candidatos Humanos	ChatGPT Sem Contexto	ChatGPT Com Contexto	Bard Sem Contexto	Bard Com Contexto
2016	42,31%	46,93%	48,97%	46,93%	44,89%
2017	46%	50,98%	47,05%	54,90%	45,09%
2018	41,14%	54,54%	49,09%	56,36%	63,63%
2019	40,96%	47,27%	49,09%	47,27%	49,09%
2022	41,34%	65,45%	60%	54,38%	52,63%

Aproveitamento Médio no POSCOMP

**Figure 3: Aproveitamento total no POSCOMP**

resultados para o Bard com contexto acarretou em um resultado menos significativo).

Do ponto de vista estatístico, ao se comparar o desempenho geral entre os *chatbots* usando ou não informação adicional de contexto, não há diferença significativa entre os resultados.

4.2 Desempenho nos Macro-temas

Conforme apresentado, o exame POSCOMP é organizado em três macro-temas: Matemática (20 questões), Fundamentos da Computação (30 questões) e Tecnologia da Computação (20 questões). Esta subseção apresenta e analisa o desempenho dos *chatbots* em cada um deles.

4.2.1 Desempenho em Matemática. A Tabela 4 e a Figura 4 apresentam o aproveitamento em Matemática dos candidatos do POSCOMP e dos quatro modelos avaliados. Destaca-se que apenas 61% das questões de Matemática foram apresentadas aos *chatbots*, pois as demais possuem elementos gráficos necessários ao entendimento da questão.

Destaca-se, a partir da Tabela 4, que em duas edições (2016 e 2019) os candidatos humanos obtiveram desempenhos médios superiores aos *chatbots* em Matemática. Para as demais edições, observa-se empates nos melhores desempenhos, exceto em 2022.

O ChatGPT apresentou o melhor desempenho em três edições, sem o uso de contexto, mas empatando ou com o Google Bard (em 2017 e 2018) ou com ele mesmo, mas usando contexto (em 2018).

Ao considerarmos as médias de todos os resultados para Matemática nas cinco edições analisadas do POSCOMP, o melhor desempenho foi do ChatGPT sem uso de contexto (40,94%), porém apenas 0,5% superior à média dos aproveitamentos dos candidatos (40,72%). Para as questões de Matemática, o desempenho médio do ChatGPT foi mais de 25% superior ao desempenho do Google Bard.

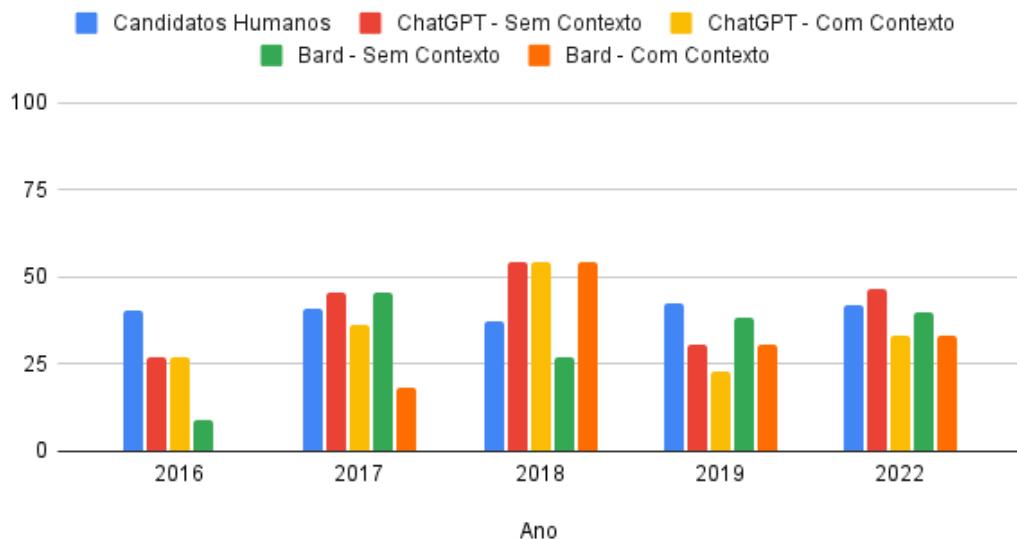
Em geral, para as questões de Matemática, o uso de um único contexto simplificado para todas as questões do POSCOMP prejudicou o desempenho dos *chatbots*. Isso destaca a capacidade do *chatbots* em responder a perguntas de Matemática sem depender de informações adicionais.

Do ponto de vista estatístico, ao se comparar o desempenho médio dos candidatos e o desempenho entre os *chatbots* usando ou

Table 4: Aproveitamento em Matemática no POSCOMP

Ano	Candidatos Humanos	ChatGPT Sem Contexto	ChatGPT Com Contexto	Bard Sem Contexto	Bard Com Contexto
2016	40,5%	27,27%	27,27%	9,09%	0,00%
2017	41,15%	45,45%	36,36%	45,45%	18,18%
2018	37,55%	54,54%	54,54%	27,27%	54,54%
2019	42,6%	30,76%	23,07%	38,46%	30,76%
2022	41,8%	46,66%	33,33%	40,00%	33,33%

Aproveitamento em Matemática

**Figure 4: Aproveitamento em Matemática**

não informação adicional de contexto, não foi encontrada diferença significativa entre os resultados para o macro-tema Matemática.

4.2.2 Desempenho em Fundamentos de Computação. A Tabela 5 e a Figura 6 apresentam o aproveitamento em Fundamentos de Computação dos candidatos do POSCOMP e dos quatro modelos avaliados. A cada ano, no POSCOMP, há 30 questões sobre este macro-tema e foi possível utilizar 75% destas questões no presente trabalho (113 de 150).

Diferentemente do que aconteceu para Matemática, em Fundamentos de Computação todos os melhores desempenhos (destacados em negrito na Tabela 6) foram alcançados pelos *chatbots*. O ChatGPT apresentou o melhor desempenho, utilizando contexto, em 2019 e 2022, o Google Bard, sem contexto, foi melhor em 2016 e 2018 e em 2017 os dois *chatbots* empataram, ambos sem utilizar contexto.

Considerando os quatro conjuntos de resultados explorados neste trabalho (dois *chatbots* com e sem contexto), na média das cinco edições do POSCOMP, os *chatbots* tiveram um desempenho de 20,0% (Bard com contexto) a 30,1% (Bard sem contexto) superiores à média

do desempenho dos humanos. Porém, se olharmos para cada um os resultados de cada uma das edições do POSCOMP, o maior destaque ocorre com o ChatGPT com contexto no exame de 2022, o qual teve um desempenho 77% superior à média dos candidatos humanos.

O uso de um contexto simples adicionado a cada questão, para Fundamentos de Computação, foi prejudicial, na média, para o Google Bard (o aproveitamento nas questões foi reduzido de 60,34% para 55,68%). Já para o ChatGPT o uso de contexto foi levemente positivo (aumento de 56,82% para 57,89%, na média dos resultados).

Devido à grande variação de valores e ao fato de diferentes modelos terem se sobressaído em diferentes anos, a única diferença de resultados com significância estatística¹¹ ocorreu entre o Bard sem informação de contexto e os candidatos humanos.

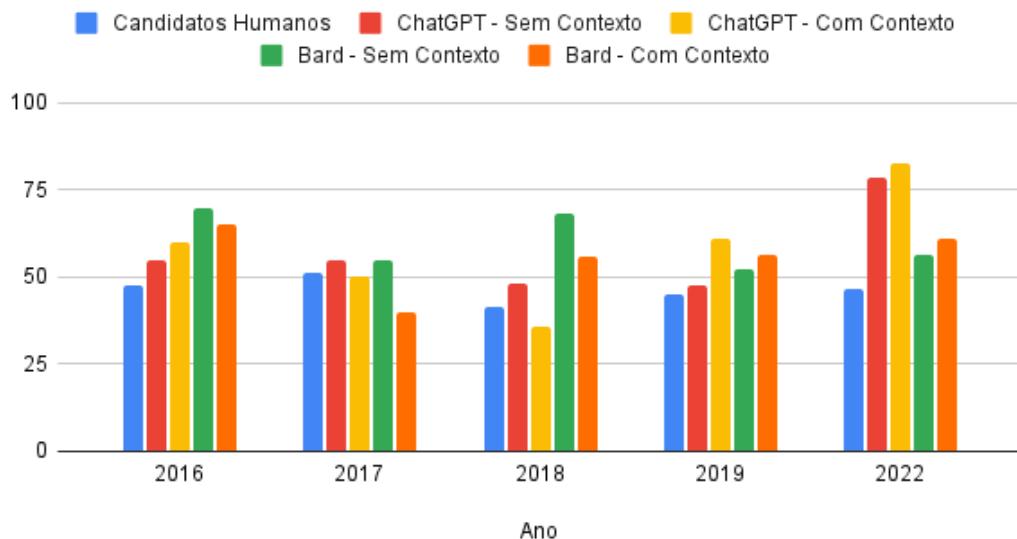
4.2.3 Desempenho em Tecnologia da Computação. A Tabela 6 e a Figura 5 apresentam o aproveitamento em Tecnologia da Computação dos candidatos do POSCOMP e dos quatro modelos avaliados. Das 100 questões deste macro-tema presentes nas cinco últimas edições do POSCOMP, foi possível utilizar 97 no presente

¹¹Confiança acima de 95%.

Table 5: Aproveitamento em Fundamentos de Computação no POSCOMP

Ano	Candidatos Humanos	ChatGPT Sem Contexto	ChatGPT Com Contexto	Bard Sem Contexto	Bard Com Contexto
2016	47,7%	55%	60%	70%	65%
2017	51,23%	55%	50%	55%	40%
2018	41,23%	48%	36%	68%	56%
2019	45,13%	47,82%	60,86%	52,17%	56,52%
2022	46,67%	78,26%	82,6%	56,52%	60,86

Aproveitamento em Fundamentos de Computação

**Figure 5: Aproveitamento em Fundamentos de Computação**

trabalho, pois apenas 3% dessas questões possuem elementos gráficos necessários ao seu entendimento.

Assim como ocorreu para Fundamentos de Computação, todos os melhores desempenhos para Tecnologia da Computação foram atingidos pelos *chatbots*. Para o ano de 2016, três modelos apresentaram o mesmo desempenho, acertando 50% da prova: ChatGPT com e sem contexto e Bard com contexto. Para o exame de 2019, ChatGPT sem contexto apresentou o melhor desempenho. Nos três exames restantes, o Google Bard se destacou com os melhores desempenhos utilizando contexto em 2017 e 2018 e sem usar contexto em 2022. Destaca-se o resultado de 2018, 77% superior à média de desempenho dos candidatos humanos.

Ao se analisar a média de desempenho nos cinco anos de cada um dos quatro modelos, verifica-se que o Bard com contexto se sobressaiu (com uma média de acertos de 60,89%, o que é 60,58% superior ao desempenho médio dos humanos. Em seguida temos o ChatGPT sem contexto praticamente empatado com o Bard sem contexto (55,79% e 55,62%, respectivamente) e por último o ChatGPT com contexto, cujo aproveitamento médio foi de 52,63%, valor 38,79% superior à média dos resultados dos candidatos.

Para Tecnologia da Computação, os desempenhos dos quatro modelos se mostraram estatisticamente superiores aos dos candidatos humanos. Porém, ao se comparar cada um dos modelos par a par, não houve nenhuma diferença estatisticamente significativa, considerando o conjunto das cinco edições do POSCOMP.

5 CONCLUSÕES E TRABALHOS FUTUROS

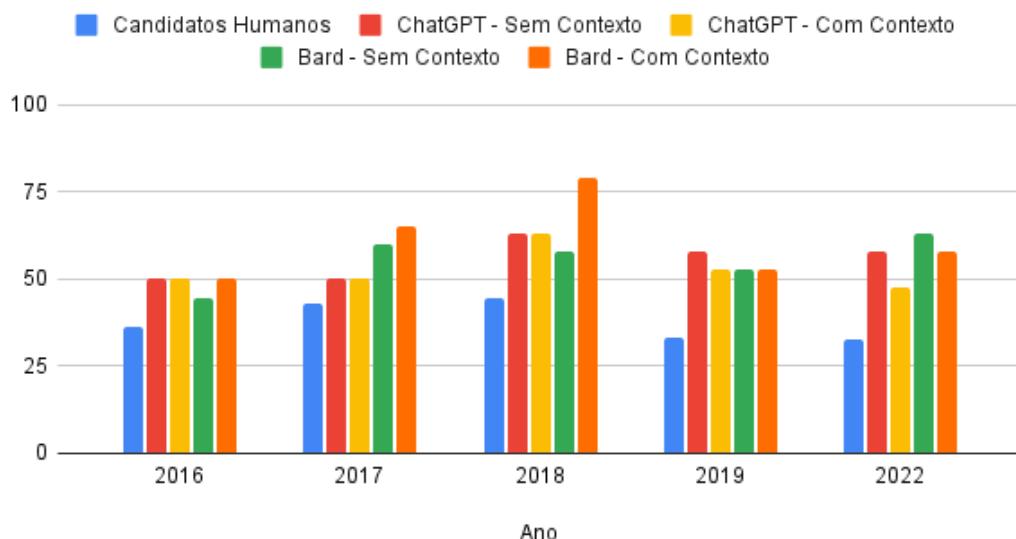
O presente trabalho comparou de forma detalhada o desempenho dos *chatbots* ChatGPT e Google Bard com o desempenho médio dos candidatos humanos que participaram da Prova para Ingresso em Programas de Pós-Graduação em Computação (POSCOMP). A avaliação foi conduzida com base em uma seleção das questões das edições mais recentes do exame, abrangendo os anos de 2016, 2017, 2018, 2019 e 2022. Foram utilizadas as 271 questões (de um total de 350) que não faziam uso de aspectos gráficos para o entendimento e resolução da questão.

A coleta de dados e a avaliação foram estruturadas em várias etapas, incluindo a seleção das perguntas com base na capacidade de representação pelas inteligências artificiais avaliadas, interações simultâneas entre os modelos e as questões da prova, bem como

Table 6: Aproveitamento em Tecnologia da Computação no POSCOMP

Ano	Candidatos Humanos	ChatGPT Sem Contexto	ChatGPT Com Contexto	Bard Sem Contexto	Bard Com Contexto
2016	36,05%	50%	50%	44,44%	50%
2017	43%	50%	50%	60%	65%
2018	44,6%	63,15%	63,15%	57,89%	78,94%
2019	33,05%	57,89%	52,63%	52,63%	52,63%
2022	32,9%	57,89%	47,36%	63,15%	57,89%

Aproveitamento em Tecnologia da Computação

**Figure 6: Aproveitamento em Tecnologia da Computação**

a avaliação da assertividade das respostas geradas pelos sistemas em relação ao gabarito oficial do POSCOMP. Essa abordagem proporcionou um exame detalhado do desempenho dos modelos em um cenário que abrangeu múltiplos tópicos e níveis de complexidade, replicando o desafio típico enfrentado pelos candidatos no POSCOMP.

Observou-se que o desempenho dos *chatbots* usando ou não a adição de um contexto simples a cada pergunta foi significativamente superior à média dos resultados humanos, ao considerar a prova como um todo. No geral, o ChatGPT sem uso de contexto teve o maior aproveitamento médio, 25% superior ao dos candidatos humanos e, se observarmos apenas a edição de 2022 deste exame, o desempenho foi 58% superior ao dos candidatos.

Considerando cada um dos macro-temas, ambos os *chatbots* não foram capazes de se destacar nas questões de Matemática. Já para Fundamentos de Computação, considerando as cinco edições do exame, o Bard sem uso de contexto apresentou o melhor desempenho (30,1% superior à média dos candidatos), mas ao se observar cada edição do exame, o ChatGPT com uso de contexto se destacou atingindo um aproveitamento 77% superior à média dos candidatos

em 2022. Por fim, para Tecnologia da Computação, no geral, o Bard, com uso de contexto, se destacou com um desempenho, na média, 60,6% superior ao dos candidatos humanos.

O contexto simples utilizado não foi capaz de produzir resultados, na média, superiores àqueles que não utilizaram contexto. É possível que contextos mais sofisticados e/ou específicos a cada macro-tema poderão produzir resultados melhores.

Assim, verificou-se que tanto para as questões de Fundamentos de Computação quanto de Tecnologia da Computação os *chatbots* testados foram capazes de superar a média do desempenho humano, destacando-se que o exame POSCOMP costuma ser realizado por alunos formados ou que estão prestes a se formar em cursos superiores de computação principalmente no Brasil, mas também no restante da América Latina.

Neste contexto e considerando que muitos dos exames da área de computação têm sido aplicados de forma online, é importante que os gestores desses exames considerem a facilidade atual existente do uso dessas ferramentas que poderiam dar vantagens indevidas a candidatos mal intencionados.

Seria importante discutir questões éticas relacionadas à segurança e integridade dos exames, bem como propor diretrizes para garantir a equidade e a confiabilidade dos processos de avaliação.

Por outro lado, vale destacar que esses *chatbots* são sistemas “especialistas” em conversar (responder a perguntas, mas sem um compromisso de fornecer respostas corretas). Apesar do bom desempenho apresentado neste trabalho em relação às médias obtidas pelos candidatos humanos, é importante destacar que o desempenho geral dos *chatbots* em cada uma das edições do POSCOMP nunca superou dois terços da prova.

Os resultados do presente trabalho são limitados ao contexto de sua aplicação, isto é, dois *chatbots* que foram avaliados, cujos *prompts* utilizaram ou não um contexto simples e geral, juntamente com 271 questões de cinco edições diferentes do POSCOMP. Porém, assim como apresentado por diversos trabalhos atuais sobre as capacidades e limitações da nova geração de *chatbots*, acreditamos que o conhecimento aqui adquirido pode ser, com os devidos cuidados, generalizado para diferentes áreas da computação. Esses resultados podem ser úteis para orientar futuras pesquisas e desenvolvimentos em uma variedade de áreas da computação, envolvendo outros exames de múltipla escolha na área.

Uma limitação potencial do trabalho é que eventualmente os grandes modelos de linguagem utilizados pelos *chatbots* já poderiam ter entrado em contato com as questões do POSCOMP durante seu treinamento. Esta limitação é bastante atenuada pelo fato das questões estarem em português e não fazerem parte da lista das bases de dados oficialmente utilizadas no treinamento de seus modelos de linguagem. Adicionalmente, se esta exposição estivesse enviesando os resultados, provavelmente haveria um desempenho inferior no exame de 2022 (devido à data de treinamento dos modelos), coisa que não ocorreu. Adicionalmente, o desempenho utilizado dos candidatos humanos foi aquele obtido nas provas completas de cada macro-tema, enquanto os *chatbots* foram avaliados considerando as questões com conteúdos exclusivamente textuais. É possível que essa diferença adicione certo viés nos resultados.

Vislumbramos diversas direções para trabalhos futuros. Algumas das versões atuais dos *chatbots* já permitem receber como entrada fotos/imagens então seria possível passar para estes modelos todas as questões da prova, ressaltando-se que o processamento ou a “interpretação” dessas imagens ainda não tem um desempenho tão grande quanto o de processamento de texto. Há também novos modelos de linguagens e *chatbots* sendo lançados mensalmente, assim é possível estender a análise comparativa para outros sistemas.

Seria interessante também examinar o desempenho dos *chatbots* nas diferentes subáreas da computação (apenas algumas edições do POSCOMP têm subáreas associadas a cada questões), a fim de identificar padrões de desempenho em áreas específicas. Adicionalmente, um processo de *engenharia de prompt*[2] poderia ser utilizado para tentar melhorar o desempenho dos *chatbots*.

Essas adições ajudariam a fornecer uma visão mais completa do impacto dos *chatbots* em avaliações acadêmicas e das implicações mais amplas dessas tecnologias na educação e em processos de avaliação. Além disso, essas considerações podem orientar futuras pesquisas e desenvolvimentos nessa área em constante evolução. Por fim, é possível aprofundar o estudo do desempenho desses sistemas para outras áreas, diferentes da computação, ou mesmo para as sub-áreas da computação.

REFERENCES

- [1] Sebastian Bordt and Ulrike von Luxburg. 2023. ChatGPT Participates in a Computer Science Exam. arXiv:2303.09461 [cs.CL]
- [2] Felipe de Fonseca, Ivandro Paraboni, and Luciano Digiampietri. 2023. Contextual stance classification using prompt engineering. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana* (Belo Horizonte/MG). SBC, Porto Alegre, RS, Brasil, 33–42. <https://doi.org/10.5753/stil.2023.233708>
- [3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- [4] Burton A. Leland, Bradley D. Christie, James G. Nourse, David L. Grier, Raymond E. Carhart, Tim Maffett, Steve M. Welford, and Dennis H. Smith. 1997. Managing the Combinatorial Explosion. *Journal of Chemical Information and Computer Sciences* 37, 1 (1997), 62–70. <https://doi.org/10.1021/ci960088t>
- [5] Kamil Malinka, Martin Peresini, Anton Firc, Ondrej Hujnák, and Filip Janus. 2023. On the Educational Impact of ChatGPT: Is Artificial Intelligence Ready to Obtain a University Degree?. In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1* (Turku, Finland) (ITiCSE 2023). Association for Computing Machinery, New York, NY, USA, 47–53. <https://doi.org/10.1145/3587102.3588827>
- [6] A. Newell, J.C. Shaw, and H.A. Simon. 1959. Report on a general problem-solving program. In *Proceedings of the International Conference on Information Processing*. I.K International Publishing House, Paris, France, 256–264.
- [7] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., New Orleans, LA, USA, 27730–27744. https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf
- [8] Vinay Purnani, Yusuf Sermet, and Ibrahim Demir. 2023. Performance of ChatGPT on the US Fundamentals of Engineering Exam: Comprehensive Assessment of Proficiency and Potential Implications for Professional Environmental Engineering Practice. arXiv:2304.12198 [cs.CY]
- [9] Basit Qureshi. 2023. Exploring the Use of ChatGPT as a Tool for Learning and Assessment in Undergraduate Computer Science Curriculum: Opportunities and Challenges. arXiv:2304.11214 [cs.CY]
- [10] Alex Radford and Karthik Narasimhan. 2018. Improving Language Understanding by Generative Pre-Training. <https://api.semanticscholar.org/CorpusID:49313245>
- [11] Stuart J. Russell and Peter Norvig. 2009. *Artificial Intelligence: a modern approach* (3 ed.). Pearson, London, England.
- [12] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixing Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichen Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Arroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Jon Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. arXiv:2201.08239 [cs.CL]
- [13] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambo, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL]
- [14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc., Long Beach, CA, USA. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fb053c1c4a845aa-Paper.pdf