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

JOÃO VITOR DE MELO CAVALCANTE E SOUZA
ROHIT GHEYI

AVALIANDO MODELOS DE LINGUAGEM GRANDE NA RESOLUÇÃO DE PROBLEMAS DE LÓGICA DA OBI

CAMPINA GRANDE - PB
2024

JOÃO VITOR DE MELO CAVALCANTE E SOUZA

AVALIANDO MODELOS DE LINGUAGEM GRANDE NA RESOLUÇÃO DE PROBLEMAS DE LÓGICA DA OBI

Trabalho de Conclusão de 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

JOÃO VITOR DE MELO CAVALCANTE E SOUZA

AVALIANDO MODELOS DE LINGUAGEM GRANDE NA RESOLUÇÃO DE PROBLEMAS DE LÓGICA DA OBI

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

Hyggo Oliveira de Almeida

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

Modelos de Linguagem Grande (LLM) se tornaram populares, com modelos de fácil acesso e cada vez mais perspicazes, tais como Chat GPT, Gemini, Claude, Mistral, dentre outros. Com uma grande gama de possíveis utilizações em uma era digital, os LLMs tem uma crescente utilização, sendo possíveis suas atuações em diversos campos, inclusive no de aprendizado humano. Mediante isto, é considerável saber e ponderar sobre tais ferramentas e como podem ajudar na resolução de problemas lógicos encontrados na Olimpíada Brasileira de Informática (OBI) para alunos do ensino fundamental. Neste trabalho avaliamos o Chat GPT 3.5 e o Le Chat Mistral com 100 questões de lógica da OBI. O Chat GPT acertou 23% delas, enquanto que o Mistral Le Chat acertou 36% das questões, ambos na primeira tentativa.

EVALUATING LARGE LANGUAGE MODELS ON THE SOLUTION OF LOGIC PROBLEMS IN THE BRAZILIAN INFORMATICS OLYMPIAD

ABSTRACT

Large Language Models (LLM) have become popular, with easy to access and more and more insightful models, such as Chat GPT, Gemini, Claude, Mistral, among others. With a wide range of possible utilizations in a digital era, LLMs have a growing use in the most diverse fields, specially on human learning. For this reason, it's important to study and consider such tools and how they can help in solving logical problems found in the Brazilian Olympiad in Informatics (OBI). In this study, we evaluated Chat GPT 3.5 and Le Chat Mistral with 100 logic questions from OBI. Chat GPT answered 23% of them correctly, while Mistral Le Chat answered 36% of the questions correctly, both on the first attempt.

AVALIANDO MODELOS DE LINGUAGEM GRANDE NA RESOLUÇÃO DE PROBLEMAS DE LÓGICA DA OBI

João Vitor de Melo Cavalcante e Souza
Departamento de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil
joao.vitor.souza@ccc.ufcg.edu.br

Rohit Gheyi
Departamento de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil
rohit@dsc.ufcg.edu.br

RESUMO

Modelos de Linguagem Grande (*LLM*) se tornaram populares, com modelos de fácil acesso e cada vez mais perspicazes, tais como Chat GPT, Gemini, Claude, Mistral, dentre outros. Com uma grande gama de possíveis utilizações em uma era digital, os LLMs tem uma crescente utilização, sendo possíveis suas atuações em diversos campos, inclusive no de aprendizado humano. Mediante isto, é considerável saber e ponderar tais ferramentas e como podem ajudar na resolução de problemas lógicos encontrados na Olimpíada Brasileira de Informática (OBI) para alunos do ensino fundamental. Neste trabalho avaliamos o Chat GPT 3.5 e o Le Chat Mistral com 100 questões de lógica da OBI. O Chat GPT acertou 23% delas, enquanto que o Mistral Le Chat acertou 36% das questões, ambos na primeira tentativa.

Palavras Chave

LLM, Lógica, OBI, Chat GPT, Mistral, Problemas de Lógica, Resolução de Problemas, Olimpíada Brasileira de Informática.

Dados

Avaliando LLMs em Questões da OBI Iniciação

1. INTRODUÇÃO

A capacidade dos Modelos de Linguagem Grande, ou *LLMs*, tem se tornado cada vez mais evidente. Tais ferramentas possuem uma crescente popularidade e acessibilidade, sendo de utilidade nas mais diversas áreas, sejam estas de conhecimento, produção e até mesmo geração, seja de imagens a blocos de código de programação.

De acordo com Hore [1], *LLM* ou *Large Language Model* compreende o conjunto de Modelos de Linguagem Grande que, a utilização da técnica de aprendizado *deep learning*, possuem a capacidade de gerar respostas similares às de humanos, compreendendo e as manipulando com base na linguagem natural. Para isso, os modelos são treinados a partir da "leitura" de inúmeras fontes de dados, como conversas, artigos, imagens e sites, visando o acréscimo na precisão do modelo e da cobertura de tópicos dos quais a *LLM* pode formular sua resposta.

Como exemplo de ferramenta da área, vale citar o Chat GPT, lançado em 2022, que consiste em um robô de chat/conversa, com diversos modelos que podem ser utilizados, como o GPT-3.5 e o GPT-4, sendo este último necessário pagamento para utilização.

Desenvolvido pela empresa Open AI, o Chat Bot atraiu até o fim de 2023 cerca de 100 milhões de visitantes por mês [2] e foi capaz de ser aprovado em exames de direito e negócios [3] assim como ser aceito em entrevista de emprego [4].

Portanto, para ocorrer um melhor desenvolvimento na área de inteligência artificial, visando o aprendizado e a compreensão universal com um agente inteligente de mais alto nível, ou em outras palavras, uma inteligência artificial geral, é necessário avaliar a sua capacidade lógica e como tais modelos se comportam mediante problemas lógicos que possivelmente estão envolvidos no cotidiano.

Neste trabalho, será analisada a capacidade de dois Chatbots de propósito geral, sendo eles o Chat GPT (GPT-3.5) e o Le Chat Mistral, em responder questões de lógica ligadas à Olimpíada Brasileira de Informática, nível de iniciação. Foram selecionadas 100 questões dos anos de 2022 e 2023. De todas as questões, o Le Chat respondeu corretamente 36, enquanto que o Chat GPT acertou apenas 23.

Este artigo está organizado da seguinte forma: A Seção 2 trata da Fundamentação Teórica, a Seção 3 trata dos Objetivos deste estudo, a Seção 4 trata da Metodologia enquanto as Seções 5, 6 e 7 tratam, respectivamente, dos Resultados, Trabalhos Relacionados e das Conclusões Formuladas.

2. FUNDAMENTAÇÃO TEÓRICA

Nesta seção serão apresentados conceitos relevantes para o entendimento do estudo realizado.

2.1 *Large Language Models* (*LLM*)

Large Language Models, ou em uma tradução literal, Modelos de Linguagem Grande, são tipos de algoritmos de inteligência artificial que, a partir de um treinamento extensivo, conseguem formular respostas mediante as entradas fornecidas. Utilizando a Técnica de Aprendizagem Profunda, tais modelos são submetidos a treinamento com grande quantidade de informação em um processo complexo para desenvolver a capacidade de interagir com diversos casos de uso e atividades [5]. Tais algoritmos podem ser utilizados na geração de conteúdo (como texto e imagens), assistência virtual, desenvolvimento de aplicações e até no auxílio da educação.

De acordo com Stöffelbauer [6], um engenheiro de dados da Microsoft, o objetivo dos Modelos de Linguagem Grande, em um nível funcional, é identificar padrões na informação submetida para inferir uma relação entre uma entrada e uma saída. Tal processo, em termos de Aprendizado de Máquina, pode ser chamado de problema de classificação. É classificando diversos inputs de diversas áreas, como processamento de linguagem natural e programação, que o modelo aprende um padrão e o aplica a outras situações, efetivamente "prevendo" o resultado.

Dentro do campo de *LLMs*, existem modelos fechados, como o Chat GPT, criado pela Open AI [12], o Gemini, anteriormente

conhecido como Bard, de autoria da Google [13] e o Claude, desenvolvido pela Anthropic [14]. Já referente a *LLMs* abertas, vale citar o Le Chat [16], feito pela Mistral AI e o Llama da Meta [15].

2.2 Olimpíada Brasileira de Informática

A Olimpíada Brasileira de Informática (OBI) é uma competição científica realizada no Brasil pela SBC anualmente desde 1999, reunindo alunos desde o ensino fundamental ao ensino superior. A competição é dividida em duas modalidades, sendo elas iniciação e programação.

A modalidade de iniciação, diferentemente da outra modalidade que visa lidar com aspectos mais técnicos e práticos da programação, lida com questões de raciocínio lógico e é direcionada a alunos do ensino fundamental, do quarto ao nono ano. Segundo a própria competição, esta modalidade serve para despertar o interesse pela computação e também pela própria programação [7], atuando em um dos pilares principais que é a lógica.

Ainda sobre a modalidade de iniciação da OBI, é realizada a separação por diferentes níveis de questões, sendo estas o Nível Junior, direcionado a alunos dos 4º e 5º anos do Ensino Fundamental, o Nível 1, direcionado então para alunos dos 6º e 7º anos e por fim, o Nível 2, direcionado para os 8º e 9º anos. Também há a divisão por Fases, da primeira à terceira. As provas podem repetir as questões entre si, mas também existem questões únicas[1] para cada nível.

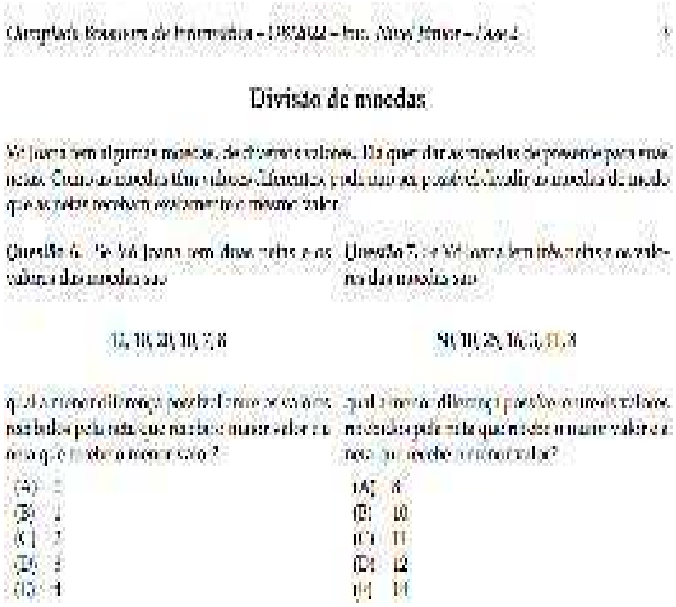

Figura 1. Exemplo de questões selecionadas da OBI 2022

Sobre as Fases e o processo de classificação entre elas, o regulamento[11] situa ainda que apenas os mais altos colocados, com pelo menos ⅓ de acertos, prosseguirão para a próxima etapa.

3. OBJETIVO

A seguir apresentamos o objetivo do trabalho usando GQM [9]. O objetivo deste estudo é avaliar a eficácia dos *Large Language Models* específicos — Chat GPT 3.5 e Le Chat Mistral — na resolução de questões de lógica da OBI, com o propósito de avaliar o potencial dos LLMs no contexto de *prompts* com o enunciado das questões contendo apenas textos em Português.

Para tal, serão respondidas às seguintes questões de pesquisa (QP):

QP_1. Até que ponto o Chat GPT 3.5 é capaz de resolver problemas de lógica da OBI?

QP_2. Até que ponto o Mistral Le Chat é capaz de resolver problemas de lógica da OBI?

Serão contadas as respostas corretas e incorretas fornecidas por cada *LLM* de acordo com o gabarito oficial fornecido pela OBI [10] e terão, subsequentemente, seus respectivos resultados analisados.

4. METODOLOGIA

A fim de quantificar a capacidade de resolução tanto do Chat GPT 3.5, selecionado em detrimento do GPT 4.0 por este requerer assinatura, quanto do Mistral Le Chat, 100 questões das edições de 2022 e 2023 da OBI na modalidade iniciação foram selecionadas e submetidas aos robôs de chat. As questões foram selecionadas de todos os três níveis presentes nas provas: Júnior, 1 e 2. Ambos *LLMs* foram submetidos às mesmas questões e tiveram direito a apenas uma única tentativa.

Tendo isso em vista, as questões foram submetidas com seus enunciados, evitando a seleção de questões com imagens visto que ambos Chat GPT 3.5 e Le Chat não suportam suas análises. A resposta fornecida por estes, então, é avaliada mediante o gabarito fornecido pela OBI, como correta ou incorreta.

Um fator importante a ressaltar é que as provas de diferentes níveis podem repetir questões entre si, sendo observável questões do Nível Júnior nos níveis posteriores. Portanto, foram selecionadas apenas a primeira aparição de cada questão pela ordem do nivelamento das provas, evitando a repetição de tais.

Visando o melhor aproveitamento do formato das questões, as consultas, fornecidas igualmente ao Chat GPT e ao Le Chat, foram levemente adaptadas para dar um senso de continuidade entre as mesmas questões dentro de um mesmo enunciado, como demonstrado na Figura 2.

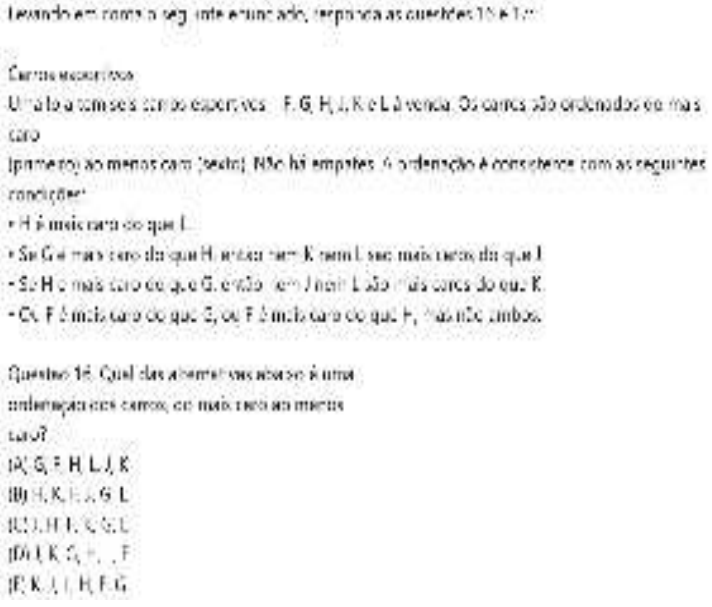

Figura 2. Exemplo de *prompt* utilizado para o Mistral Le Chat e o Chat GPT 3.5.

[1] Os diferentes níveis da avaliação podem repetir questões entre si. Tendo isso em vista, e buscando evitar a repetição de questões, foram apenas consideradas, na ordem de Júnior a Nível 2, a primeira aparição destas.

Para ambos robôs de chat foram fornecidas apenas uma chance. Tanto as consultas enviadas quanto suas respectivas respostas foram armazenadas em uma planilha para subsequente análise e verificação de suas corretude.

5. RESULTADOS E DISCUSSÃO

Os resultados gerais deste trabalho foram separados e ilustrados na Figura 3, sendo catalogados os números de acertos por cada um dos dois modelos analisados, Chat GPT e Mistral Le Chat.

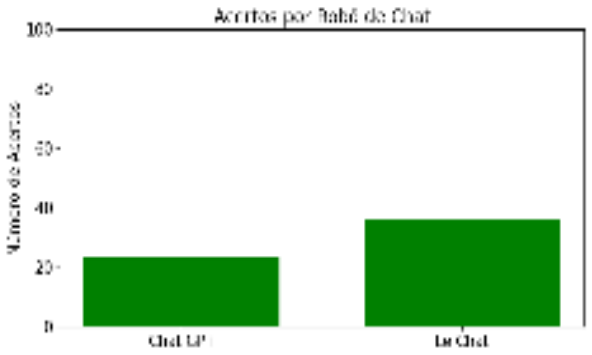

Figura 3. Número de acertos do Chat GPT e do Mistral Le Chat.

Foi observado que o Chat GPT 3.5 acertou 23 questões das 100 propostas, enquanto Le Chat Mistral, mesmo apresentando um melhor desempenho, ainda sim acertou apenas 36 questões do total.

O Le Chat da Mistral, possuindo uma taxa de acerto média geral de 36%, teve a sua distribuição de questões corretas descritas na Tabela 1 a seguir.

Mistral Le Chat	**Nível Júnior**	**Nível 1**	**Nível 2**
Correta	17	18	1
Incorreta	37	22	5
Total	54	40	6

Tabela 1. Distribuição de acertos e erros por nível de questão do Mistral Le Chat

Levando em conta a taxa de acerto para os Níveis Júnior de 31,5% e 45% para o Nível 1, o Mistral Le Chat cumpre com o requisito presente na regulação, referente à classificação para uma próxima etapa, de acerto de ⅓ das questões. Contudo, para convocação efetiva de um colocado para próxima fase, também é necessária a boa classificação nos resultados finais da prova. De fato, o resultado obtido não se assemelha ao de alunos de nível fundamental, o qual é o público alvo da competição, sendo bastante aquém a estes referidos [17].

Em uma análise mais aprofundada, é perceptível que a *LLM* possui dificuldade na resolução de problemas lógicos. Levando em conta o conjunto de todas as respostas incorretas fornecidas pelo Le Chat, há 5 questões, às quais 4 são de Nível Júnior e 1 de Nível 2, em que o Le Chat não conseguiu atribuir uma solução dentre as alternativas das próprias questões, se abstendo de fornecer qualquer resposta.

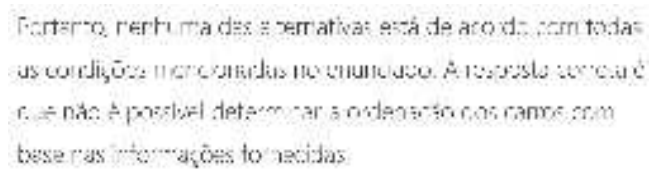
Portanto, nenhuma das alternativas está de acordo com todas as condições mencionadas no enunciado. A resposta correta é que não é possível determinar a ordenação dos ramos com base nas informações fornecidas.

Figura 4. Trecho de *prompt* demonstrando indecisão do Mistral Le Chat sobre a alternativa correta.

Em pelo menos mais 3 questões ocorreu uma indecisão por parte do Le Chat, o que, no entanto, resultou em um acerto. Também foram notadas em duas ocasiões a seleção de múltiplas alternativas como resposta correta. A OBI, exceto em casos seletos, tem suas questões como única escolha.

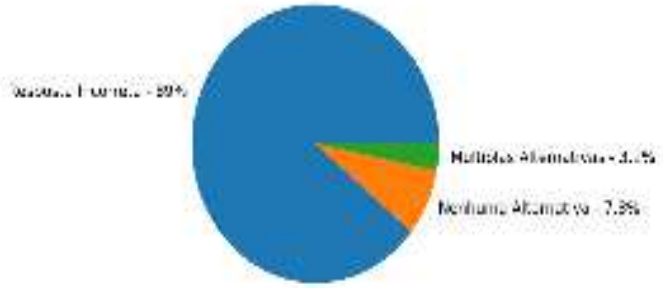

Figura 5. Gráfico contendo a identificação das 64 respostas do Mistral Le Chat avaliadas como incorretas.

Em relação ao Chat GPT 3.5, uma taxa de acerto média geral de 23% foi registrada, com a distribuição de acertos abaixo.

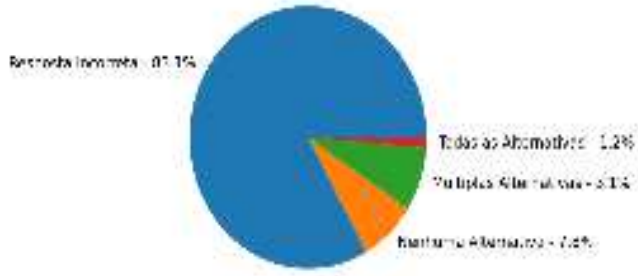

Figura 6. Gráfico contendo a identificação das 77 respostas do Chat GPT avaliadas como incorretas.

Esta *LLM*, com a supramencionada baixa taxa de acerto, demonstrou um decréscimo de performance de aproximadamente 36% quando comparada ao Mistral Le Chat. Visualizando então os acertos por nível de questão, nota-se que ocorreu acerto em apenas 22,2% em questões de Nível Júnior e 27,5% em questões de Nível 1. Já em questões exclusivas do Nível 2, não ocorreu nenhum acerto.

Chat GPT	Nível Júnior	Nível 1	Nível 2
Correta	12	11	0
Incorreta	42	29	6
Total	54	40	6

Tabela 2. Distribuição de acertos e erros por nível de questão do Chat GPT 3.5.

Durante a realização dos testes, o Chat GPT veio a informar sobre erros na própria formulação da questão, indicando uma má interpretação do que é pedido.

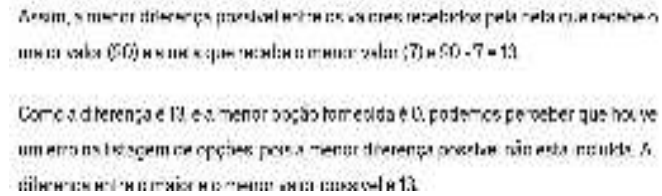

Figura 7. *Prompt* contendo interpretação equivocada do Chat GPT acerca da própria questão.

Com pelo menos três situações ocorreu uma má interpretação de questões que utilizavam da expressão "exceto" em seus enunciados, duas delas de Nível 2 e uma de Nível Júnior, o que resultou em raciocínios errôneos e respostas incorretas.

Em mais de seis ocasiões, a *LLM* retornou como corretas mais de uma alternativa em questões de única resposta. Como demonstrado na Figura 7, durante a resolução da questão 37 de Nível 1 na Fase 3, ano de 2022, o Chat GPT conseguiu separar, dentre as alternativas fornecidas, três possíveis candidatos de respostas, incluindo a alternativa correta.

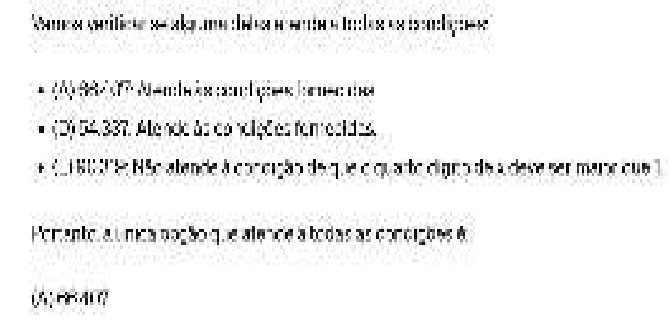

Figura 8. *Prompt* contendo resposta incorreta gerada pelo Chat GPT 3.5.

Após a remoção da alternativa (E), graças à sua comparação com as diretivas fornecidas pela respectiva questão, ocorreu a escolha sem qualquer justificativa da alternativa restante incorreta. O Chat GPT também realizou interpretações errôneas em palavras chave dos enunciados das questões. Um caso que vale a pena destacar é a interpretação desconsiderando o termo "exceto", de suma importância para selecionar a alternativa correta.

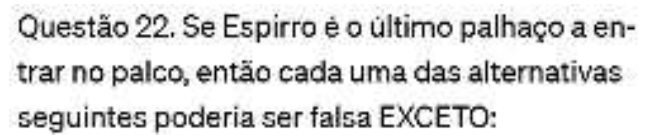

Figura 9. Enunciado da questão em que é solicitada a única alternativa correta, utilizando do jogo de linguagem com a palavra "EXCETO".

Durante a resolução da questão acima, a lógica utilizada pelo Chat GPT interpretou que todas as alternativas eram corretas e que, contraditoriamente, deveria selecionar alguma incorreta, indo totalmente contra o enunciado proposto.

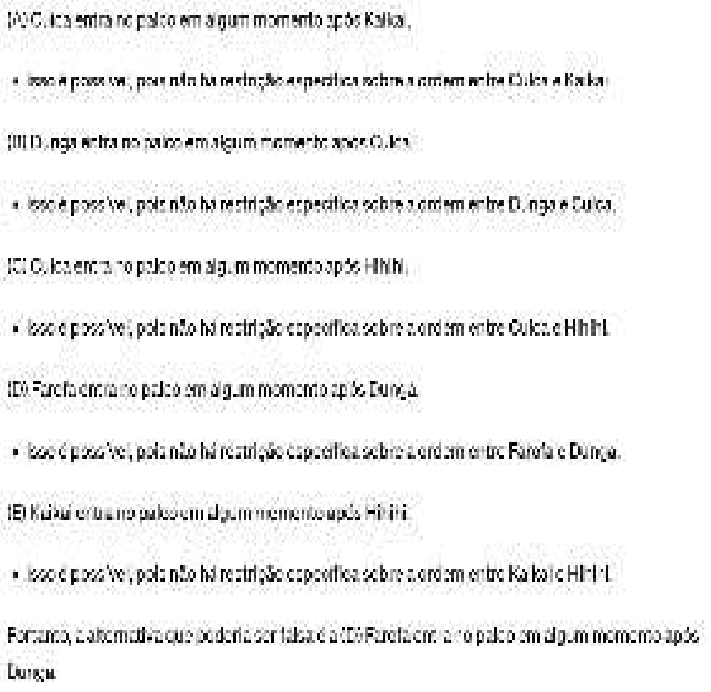

Figura 10. Parte do *prompt* resposta à questão da Figura 8.

Analogamente ao Mistral Le Chat, o Chat GPT 3.5 se mostrou inadequado ao ter uma baixíssima taxa de respostas corretas e, levando em conta que a primeira *LLM* não teria um êxito em uma boa classificação na competição, pode-se afirmar um pior resultado referente a esta.

Portanto, é verossímil afirmar que, mesmo com a performance do Mistral Le Chat superior ao Chat GPT, ambos robôs de chat movidos por *LLM* possuem dificuldades para responder questões lógicas e que, em determinados casos, não iriam prosseguir para uma próxima etapa dentro da Olimpíada Brasileira de Informática.

6. TRABALHOS RELACIONADOS

Souza e Gheyi [8] investigaram a capacidade do Chat GPT-3.5, um chatbot de Modelo de Linguagem de Grande Escala, de resolver problemas de programação. De um total de 100 problemas submetidos, o Modelo de Linguagem Grande resolveu corretamente 71 problemas em 3 tentativas, sendo 50 da plataforma LeetCode e 21 da plataforma BeeCrowd. Nestre trabalho, avaliamos 2 *LLMs*, sendo elas o Mistral Le Chat e também o Chat GPT-3.5, na resolução de 100 questões de lógica das provas da Olimpíada Brasileira de Informática.

Pires et al. [18] avaliaram a capacidade de parte da família de Modelos de Linguagem Grande do Chat GPT, como o Chat GPT-3.5 Turbo e o Chat GPT-4.0 Vision, na compreensão visual, especificamente na resolução de questões do Exame Nacional do Ensino Médio (ENEM). Tendo como base as questões dos exames dos anos de 2022 e 2023, os experimentos reuniram resultados com precisão entre 67% e 91% de acertos, levando os autores a concluir que a compreensão visual expande a aplicabilidade das *LLMs*, garantindo um melhor entendimento das questões por parte do modelo. No entanto, foi notado um espaço para melhorias e aperfeiçoamento, visto a superioridade da utilização de legendas descrevendo as imagens ao invés do uso das mesmas. Por fim, os autores concluem afirmando o potencial que tais ferramentas

possuem no campo educacional, como na preparação e identificação de novas questões, consequentemente colaborando para o auxílio oferecido aos alunos.

7. CONCLUSÃO

Neste trabalho, avaliamos a capacidade de duas LLMs (Chat GPT 3.5 e Mistral Le Chat) na resolução de 100 questões de lógica da Olimpíada Brasileira de Informática de categorias destinadas a alunos do ensino fundamental. Como demonstrado, ambos Modelos de Linguagem Grande utilizados no Chat GPT 3.5 e no Mistral Le Chat são insuficientes na resolução de problemas de lógica, com resultados obtidos que não condizem com o desempenho de alunos competidores da Olimpíada Brasileira de Informática. Nenhum dos *LLMs* iria conseguir conquistar nenhum prêmio nas categorias Iniciação da Olimpíada Brasileira de Informática nas edições 2022 e 2023.

Uma explicação para tal pode ser fundada na falha das *LLMs* em compreender problemas lógicos, na interpretação destes e na captação de um contexto, o que dificulta na resolução dos problemas. As questões utilizadas para este trabalho faziam o forte uso de hipóteses e da abstração, com a recorrente criação de contextos que eram cruciais para a resolução dos problemas, assim como também eram constantemente revisitados ao decorrer das questões. Tais aspectos compõem uma área que o imaginário humano é excelente em interagir e que, no entanto, se apresentam como um desafio para as *LLMs*. Em outras palavras, a lógica é uma área que se demonstra de grande complexidade para estas novas ferramentas, o que resulta em raciocínios não satisfatórios e consequentemente, más consultas.

Como trabalhos futuros, pretendemos avaliar outros *LLMs* como o Chat GPT 4, Gemini e Llama 3. Além de avaliar mais questões de outras provas da Olimpíada Brasileira de Informática, pretende-se fazer um *fine tunning* de alguns modelos para ver se tem impacto na performance dos modelos. Iremos também avaliar outros estilos de *prompt* para ver o impacto.

8. AGRADECIMENTOS

Gostaria de agradecer a Deus por ter me acompanhado durante a jornada da graduação e por ter me infundido de resiliência durante os anos. Agradeço igualmente aos docentes, técnicos e todos demais funcionários pela dedicação com que agem em relação ao Curso de Ciências da Computação, em especial ao professor Dr. Rohit Gheyi, pelo seu persistente esforço no acompanhamento e na orientação deste trabalho. Também agradeço à minha família por ter me dado todo apoio necessário para prosseguir com meus estudos. Agradeço aos amigos que conheci durante meus estudos, em especial a João Pedro Silva de Melo e Ezequias de Oliveira Rocha, por todo auxílio na elaboração deste artigo. Por fim, agradecimentos especiais ao finado Terrence Andrew *Davis*, cujo legado e inspiração ainda nos guia. Sua dedicação e persistência, mesmo em meio a seus tantos problemas e desvios, serão eternamente lembradas.

REFERÊNCIAS

[1] Hore, Suvojit. 2023, What are Large Language Models(LLMs)? <https://www.analyticsvidhya.com/blog/2023/03/an-introduction-to-large-language-models-llms/> Acessado em Abril de 2024.

[2] Must-Know Chat GPT-4 Statistics <https://gitnux.org/chat-gpt-statistics/> Acessado em Abril de 2024.

[3] What Exams Has ChatGPT Passed? <https://www.bestcolleges.com/news/what-exams-has-chatgpt-passed/> Acessado em Abril de 2024.

[4] Recruitment team unwittingly recommends ChatGPT for job interview <https://news.sky.com/story/recruitment-team-unwittingly-recommends-ChatGPT-for-job-interview-12788770> Acessado em Abril de 2024.

[5] What are large language models (LLMs)? <https://www.ibm.com/topics/large-language-models> Acessado em Abril de 2024.

[6] How Large Language Models work <https://medium.com/data-science-at-microsoft/how-large-language-models-work-91c362f5b78f> Acessado em Abril de 2024.

[7] Olimpíada Brasileira de Informática <https://olimpiada.ic.unicamp.br/info/>. Acessado em Abril de 2024.

[8] 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.

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

[10] OBI Anos Anteriores <https://olimpiada.ic.unicamp.br/passadas/> Acessado em Maio de 2024..

[11] Regulamento <https://olimpiada.ic.unicamp.br/info/regulamento/> Acessado em Maio de 2024.

[12] Introducing ChatGPT <https://openai.com/index/chatgpt> Acessado em Maio de 2024.

[13] Introducing Gemini: our largest and most capable AI model <https://blog.google/technology/ai/google-gemini-ai/> Acessado em Maio de 2024.

[14] Introducing Claude <https://www.anthropic.com/news/introducing-claude> Acessado em Maio de 2024.

[15] Build the future of AI with Meta Llama 3 <https://llama.meta.com/llama3> Acessado em Maio de 2024.

[16] Le Chat: Our assistant is now in beta access, demonstrating what can be built with our technology. <https://mistral.ai/news/le-chat-mistral/> Acessado em Maio de 2024.

[17] Quadro de Medalhas: Modalidade Iniciação Nível Júnior <https://olimpiada.ic.unicamp.br/passadas/OBI2022/qmerito/ij/> Acessado em Maio de 2024.

[18] Ramon Pires, Thales Sales Almeida, Hugo Abonizio, and Rodrigo Nogueira. 2023. Evaluating GPT-4's Vision Capabilities on Brazilian University Admission Exams <https://arxiv.org/abs/2311.14169>. Acessado em Maio de 2024.