



CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

Avaliação do desempenho do ChatGPT na área de Matemática dos testes do ENEM

Juliano Nunes dos Santos

João Pessoa, PB
Novembro – 2023

Juliano Nunes dos Santos

Avaliação do desempenho do ChatGPT na área de Matemática dos testes do ENEM

Monografia apresentada ao curso Ciência da Computação do Centro de Informática, da Universidade Federal da Paraíba, como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação.

Orientadora: Prof. Dra. Thaís Gaudencio do Rêgo

João Pessoa, PB
Novembro - 2023

Catálogo na publicação
Seção de Catalogação e Classificação

S237a Santos, Juliano Nunes Dos.

Avaliação do desempenho do chatgpt na área de
matemática dos testes do ENEM / Juliano Nunes Dos
Santos. - João Pessoa, 2023.

79 f. : il.

Orientação: Thaís Gaudêncio do Rêgo.
TCC (Graduação) - UFPB/CI.

1. Chatbots. 2. Chatgpt. 3. Resolução de Problemas.
4. ENEM. I. Rêgo, Thaís Gaudêncio do. II. Título.

UFPB/CI

CDU 004.8

Elaborado por Michelle de Kássia Fonseca Barbosa - CRB-738



CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

Trabalho de Conclusão de Curso de Ciência de Computação intitulado **Avaliação do desempenho do ChatGPT na área de matemática dos testes do ENEM** de autoria de Juliano Nunes dos Santos, aprovada pela banca examinadora constituída pelos seguintes professores:

Thaís Gaudencio do Rêgo

Prof^ª Dr^ª. Thaís Gaudencio do Rêgo
(Orientadora – Universidade Federal da Paraíba)

Rogéria Gaudencio do Rêgo

Prof^ª Dr^ª. Rogéria Gaudencio do Rêgo
(Examinadora – Universidade Federal da Paraíba)

Yuri de Almeida Malheiros Barbosa

Prof.Dr.Yuri de Almeida Malheiros Barbosa
(Examinador - Universidade Federal da Paraíba)

João Pessoa, 10 de novembro de 2023

AGRADECIMENTOS

Agradeço à meus pais, José Augusto e Edileuza Nunes, pelo apoio e incentivo ao longo desta jornada, integrais no alcance deste objetivo.

Agradeço também aos meus professores que me ajudaram a formar a base necessária para, não só a conclusão deste capítulo da minha vida, mas para que eu me desenvolva continuamente no futuro.

Em especial, agradeço aos professores Thaís Gaudencio, Rogéria Gaudencio e Yuri de Almeida, cujos apoios foram essenciais no desenvolvimento deste trabalho.

Por último, mas não menos importante, agradeço aos colegas de curso que tive a oportunidade de conhecer, conviver com e trabalhar em conjunto neste processo de evolução.

RESUMO

Nos últimos anos houve a popularização dos *chatbots* de uso geral que utilizam modelos de linguagem em larga escala para se comunicarem com seres humanos. Estes sistemas possuem uma abrangência de conhecimentos geralmente satisfatória em diversas áreas devido ao seu treinamento, que utiliza bases de dados massivas compostas de dados da internet, sendo o ChatGPT o mais popular atualmente. Mas uma área que estes *chatbots* frequentemente tem dificuldade em apresentar um raciocínio e respostas corretas é a de resolução de problemas matemáticos, geralmente apresentando desempenho inferior em relação às outras áreas. Por essa razão, esse estudo foi feito para analisar a performance do ChatGPT em questões de matemática de seis provas ENEM (2017 a 2023), a fim de tentar identificar os tipos de dificuldades que o programa encontra ao tentar resolver tais problemas. Foram feitas categorizações em função dos acertos e erros cometidos, como a taxa de cada um, as unidades temáticas e os tipos de representações visuais nas quais se encaixam e as categorias de causa dos erros. Foi observado que o ChatGPT possui dificuldade em responder questões da área, com uma taxa de acerto que varia entre 22,2% e 55,5%, e que tende a ter mais dificuldade em questões relativas à Unidade 2 (Utilizar o conhecimento geométrico para realizar a leitura e a representação da realidade e agir sobre ela) e que possuem representações visuais de “formas geométricas descritas”. As maiores causas de erro encontradas foram interpretações incorretas dos problemas propostos e falta de lógica e coerência nas respostas dadas, além de uma alta frequência de cálculo incorreto de valores. Em trabalhos futuros, esse tipo de avaliação pode ser expandida para outros *chatbots*, e também pode ser feita por pesquisadores e especialistas da área de matemática a fim melhorar a otimização do desempenho deste tipo de ferramenta neste tema.

Palavras-chave: <Chatbots>, <ChatGPT>, <Resolução de Problemas>, <ENEM>

ABSTRACT

In recent years there has been a popularization of general-purpose, large language model chatbots that use artificial intelligence to communicate with human beings. These systems have a generally satisfactory range of knowledge in several areas due to their training, which uses massive databases made up of internet data, with ChatGPT being the most popular currently. But one area in which these chatbots often have difficulty presenting correct reasoning and answers is solving mathematical problems, generally presenting lower performance compared to other areas. For this reason, this study was carried out to analyze ChatGPT's performance in mathematics questions from six ENEM tests (2017 to 2023), in order to try and identify the types of difficulties that the program encounters when trying to solve such problems. Categorizations were made depending on the successes and errors made, such as the rate of each one, the thematic units and the types of visual representations in which they fit and the categories of causes of the errors. It was observed that ChatGPT has difficulty answering questions in the area, with an accuracy rate that varies between 22.2% and 55.5%, and that it tends to have more difficulty in questions pertaining to Unit 2 (Using geometric knowledge to perform the reading and representation of reality and act on it) and that have visual representations of “described geometric shapes”. The biggest causes of error found were incorrect interpretations of the proposed problems and a lack of logic and coherence in the answers given, in addition to a high frequency of incorrect calculation of values. In future works, this type of evaluation can be expanded to other chatbots, and can also be carried out by researchers and experts in the field of mathematics in order to improve the optimization of the performance in this topic of this type of tool.

Key-words: <Chatbots>, <ChatGPT>, <Problems solving>, <ENEM>

LISTA DE ABREVIATURAS

ENEM - Exame Nacional do Ensino Médio

GPT - *Generative pre-trained transformer* (Transformador generativo pré-treinado)

GRU - *Gated Recurrent Unit* (Unidade Recorrente Barrada)

LaMDA - *Language Model for Dialogue Applications* (Modelo de Linguagem para Aplicações de Diálogo)

LTSM - *Long Short-Term Memory* (Memória de Longo Curto-Termo)

MBA - *Master of Business Administration* (Mestre em Administração de Negócios)

NCBE - *National Conference of Bar Examiners* (Conferência Nacional de Exames para Advogados)

OAB - Ordem dos Advogados do Brasil

PaLM - *Pathways Language Model* - (Modelo de Linguagem de Caminhos)

PLN - Processamento de Linguagem Natural

RNNs - *Recurrent Neural Networks* (Redes Neurais Recorrentes)

LISTA DE FIGURAS

1. Diagrama de um processo de Markov de dois estados A e E, no qual os números representam a probabilidade de mudança de estado.	16
2. Arquitetura do modelo Transformer.	20
3. Arquitetura do modelo GPT original, baseada na arquitetura Transformer.	21
4. Exemplo de uso do ChatGPT para responder às questões do ENEM.	29
5. Gráfico de colunas representativo das taxas de acerto do ChatGPT.	35
6. Gráfico de colunas empilhadas representativo das unidades temáticas de todas as questões.	37
7. Gráfico de colunas empilhadas representativo das unidades temáticas das questões acertadas.	38
8. Gráfico de colunas empilhadas representativo das unidades temáticas das questões erradas.	40
9. Gráfico de colunas empilhadas representativo das representações visuais de todas as questões.	41
10. Gráfico de colunas empilhadas representativo das representações visuais das questões acertadas.	42
11. Gráfico de colunas empilhadas representativo das representações visuais das questões erradas.	43
12. Gráfico de colunas empilhadas representativo dos erros cometidos na modalidade com alternativas.	45
13. Gráfico de colunas empilhadas representativo dos erros cometidos na modalidade sem alternativas.	46

LISTA DE TABELAS

1. Comparação dos dados e ferramentas utilizadas neste trabalho com os trabalhos citados neste módulo. 27
2. Medidas das caixas apresentadas e números de potes suportados. 52
3. Diâmetros das pérolas em estoque e diferença de medida em relação à pérola original. 55
4. Listagem e categorização dos erros cometidos pelo ChatGPT e Bard nas questões selecionadas para análise. 72

Sumário

1. INTRODUÇÃO	12
1.1 Motivação	12
1.2 Objetivos	13
1.2.1 Geral	13
1.2.2 Específicos	13
1.3 Estrutura da Monografia	13
2. FUNDAMENTAÇÃO TEÓRICA	15
2.1 Modelo de Linguagem	15
2.2 <i>Transformer</i>	18
2.3 ChatGPT	20
2.4 Avaliação de modelos de linguagem	21
3. TRABALHOS RELACIONADOS	23
3.1 <i>Would ChatGPT get a Wharton MBA?</i>	23
3.2 <i>GPT Takes the Bar Exam</i>	23
3.3 <i>Evaluating GPT-3.5 and GPT-4 Models on Brazilian University Admission Exams</i>	24
3.4 Colocamos Inteligência artificial pra fazer ENEM: Olha no que deu!	26
3.5 ChatGPT faz prova do ENEM e tem resultado melhor que 79% dos candidatos	26
3.6 Tabela de comparação entre os trabalhos	27
4. METODOLOGIA	28
4.1 Base de dados	28
4.2 Ferramentas utilizadas	28
4.3 Categorização de unidade temática das questões respondidas	30
4.4 Categorização dos tipos de representações visuais presentes nas questões respondidas	31

4.5	Categorização dos tipos dos erros cometidos pelo ChatGPT	32
4.6	Experimentos	32
4.7	Métricas de avaliação	33
5.	RESULTADOS E DISCUSSÃO	34
5.1	Taxas de acerto do ChatGPT ao longo do tempo e com e sem alternativas	34
5.2	Taxas de acerto e erro do ChatGPT por unidade temática em cada modalidade	36
5.3	Taxas de acerto e erro do ChatGPT por tipo de representação visual em cada modalidade	40
5.4	Categorização dos erros cometidos pelo ChatGPT em cada modalidade	44
5.5	Análise de erros do ChatGPT e comparação de desempenho com o Bard	47
5.5.1	Categoria “Interpretação”	47
5.5.2	Categoria “Lógica”	49
5.5.3	Categoria “Comparação de números”	53
5.5.4	Categoria “Escolha incorreta de alternativa”	55
5.5.5	Categoria “Elemento de cálculo ignorado”	59
5.5.6	Categoria “Atribuição incorreta de valores”	61
5.5.7	Categoria “Cálculo incorreto de valores”	65
5.5.8	Categoria “Estruturação das equações”	69
5.5.9	Comparação de desempenho entre o ChatGPT e o Bard	72
	CONSIDERAÇÕES FINAIS	74
	TRABALHOS FUTUROS	76
	REFERÊNCIAS	77

1. INTRODUÇÃO

Nos anos recentes os chamados “modelos de linguagem de larga escala”, como o GPT, vêm se consolidando na cultura popular devido à suas capacidades de obter um desempenho, geralmente satisfatório, em várias áreas do conhecimento. A incorporação desses modelos em *chatbots*¹, como o ChatGPT, tem ajudado a torná-los mais acessíveis para a população geral, que pode não possuir, necessariamente, um conhecimento especializado na área de Inteligência Artificial ou às vezes até mesmo no uso de internet e computadores.

O uso desses *chatbots* varia amplamente, desde tarefas mais técnicas como o *debug* de algoritmos de computador, até tarefas triviais, como a obtenção de informações sobre receitas de comida ou conversação em geral. Uma área que estes *chatbots* também vem se destacando é a de resolução de exames/provas/testes, mostrando proficiência em áreas de conhecimento gerais como linguagens (interpretação de texto, regras de escrita e afins), ciências humanas (Filosofia e Sociologia) e também em áreas mais específicas, como finanças e Direito, e frequentemente vem obtendo médias acima das da performance humana, muitas vezes obtendo notas mínimas necessárias para aprovação.

1.1. Motivação

Apesar da performance satisfatória nas áreas do conhecimento listadas anteriormente, foi também notada uma performance abaixo da média em relação à performance humana em áreas relacionadas à Matemática, que diminui cada vez mais à medida em que a área do conhecimento se aproxima da Matemática pura, ou seja, do conhecimento mais abstrato em relação à fórmulas e regras matemáticas. Estudos sobre esse tema em específico e suas causas ainda são escassos, necessitando um estudo mais aprofundado em relação às áreas específicas da Matemática e que dificuldades estas apresentam ao ChatGPT.

¹ <https://pt.wikipedia.org/wiki/Chatterbot>

1.2. Objetivos

1.2.1. Geral

O objetivo deste trabalho é levantar, analisar e categorizar os erros cometidos pelo ChatGPT nas questões de múltipla escolha de Matemática do Exame Nacional do Ensino Médio (ENEM) dos anos de 2017 a 2022, a fim de identificar as dificuldades que o ChatGPT enfrenta ao respondê-las.

1.2.2. Específicos

Os objetivos específicos deste trabalho serão listados a seguir:

- Criar um conjunto de documentos² contendo as perguntas de Matemática das provas ENEM seguidas das respostas dadas pelo ChatGPT;
- Categorizar as questões em termo de acerto e erro (com e sem alternativas);
- Identificar e categorizar os erros que o ChatGPT comete ao responder essas questões.

1.3. Estrutura da Monografia

No Capítulo 2 são fornecidos os fundamentos teóricos necessários para uma melhor compreensão dos próximos capítulos, que são os temas de Modelo de Linguagem, a arquitetura *Transformer*, o ChatGPT e métrica de avaliação de Modelos de Linguagem.

Serão apresentados, no Capítulo 3, trabalhos que se relacionam com o atual, no âmbito acadêmico, sem deixar de citar fontes não necessariamente científicas, como matérias de sites de notícia ou vídeos de internet.

No Capítulo 4 será apresentada a metodologia utilizada na construção da parte prática do trabalho, e isso inclui a construção da base de dados, ferramentas utilizadas, a forma com

² https://drive.google.com/drive/folders/1_sErLS8sJ6CawbXpjoCfr1jjwuVTDxn?usp=sharing

que a categorização foi feita, os experimentos executados e as métricas de avaliação utilizadas.

São discutidos, no Capítulo 5, os resultados obtidos a partir do uso da metodologia especificada no capítulo anterior e um estudo sobre o desempenho obtido pelo ChatGPT.

Finalmente, são trazidas as considerações finais a partir dos resultados obtidos do capítulo anterior e uma discussão sobre o tema, além de sugestões de possíveis trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

Neste Capítulo será apresentada a fundamentação teórica do trabalho, que engloba a definição de modelo de linguagem, da arquitetura *Transformer* utilizada no ChatGPT, do próprio ChatGPT e do modelo de linguagem GPT no qual é baseado, e o modo como é feita a avaliação de modelos de linguagem em relação a lógica de suas saídas.

2.1. Modelo de Linguagem

Um modelo de linguagem é uma distribuição probabilística que determina uma probabilidade a cada palavra em uma sequência de palavras, e essas probabilidades podem ser geradas a partir do treinamento, utilizando bases de dados constituídas por várias sequências (JURAFSKY E MARTIN, 2021). Utilizando esse treinamento, o modelo pode aprender a gerar, prever ou processar texto, reconhecer fala, traduzir texto entre línguas diferentes, reconhecer escrita, entre outras atividades relacionadas. Um grande problema que modelos de linguagem podem ter é o fato de linguagens poderem gerar infinitas sequências válidas, de acordo com a propriedade linguística de Infinitude Digital³, e isso pode acarretar o problema do modelo designar uma probabilidade acima de zero para sequências válidas que podem jamais ser encontradas (CHOMSKY, 1991). Dessa forma, vários modos de modelagem foram arquitetados para lidar com esse problema, como o uso da propriedade de Markov ou de arquiteturas neurais como Redes Neurais Recorrentes e *Transformers*.

Como mencionado anteriormente, a propriedade de Markov (Figura 1) pode ser utilizada para modelar sequências de palavras e consiste em um modelo estocástico que descreve uma sequência de eventos possíveis, no qual a probabilidade de cada evento acontecer depende somente do estado resultante do evento anterior (GAGNIUC, 2017).

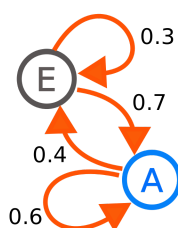
O modelo de linguagem principal que utiliza a propriedade de Markov é o “*n-gram*”, e tem como função principal determinar a próxima palavra a ser gerada, baseada em um valor fixo de palavras anteriores $n-1$ (de onde vem o “*n*” do nome). Um exemplo dessa linguagem seria um *n-gram* com valor de n igual à 2 (também chamado de bigrama) modelando as probabilidades da sentença “Eu vou ali” mostrada a seguir:

³ https://pt.wikipedia.org/wiki/Infinitude_digital

$$P(\text{Eu, vou, ali}) = P(\text{Eu} \mid \{\text{s}\})P(\text{vou} \mid \text{Eu})P(\text{ali} \mid \text{vou})P(\{\text{/s}\} \mid \text{ali}) \text{ (Equação 1)}$$

Onde “{s}” e “{/s}” denominam o começo e fim da sentença. O cálculo dessas probabilidades pode ser estimado de várias maneiras, como pela contagem da frequência de palavras, contagem da frequência de sequências, entre outros modos (JURAFSKY E MARTIN, 2021). Modelos de linguagem baseados em estatística, hoje em dia, não são mais normalmente utilizados, tendo sido substituídos em sua maior parte pelos modelos baseados em aprendizagem profunda.

Figura 1 - Diagrama de um processo de Markov de dois estados A e E, no qual os números representam a probabilidade de mudança de estado.



Fonte: Wikipédia⁴.

Modelos de linguagem neurais se baseiam em aprendizagem profunda e redes neurais e utilizam uma representação contínua, ou incorporação de palavras para fazer as previsões (KAPARTHY, 2015). O uso de redes neurais evita o crescimento da "maldição da dimensionalidade", que ocorre quando a máquina é treinada em bases de dados de texto cada vez maiores e mais complexas, o que aumenta o número de sequências possíveis exponencialmente e causa um problema de esparsidade dos dados. As redes neurais evitam esse problema ao representar as palavras como uma combinação não-linear de pesos (BENGIO, 2008).

O treinamento deste tipo de modelo é geralmente feito com classificação probabilística, feita de forma similar ao modelo “*n-gram*” onde é calculada a probabilidade de predição da palavra atual em relação ao contexto, que pode ser tanto de um conjunto de *n* palavras passadas, como uma combinação entre *n* palavras anteriores e posteriores, de acordo com as seguintes fórmulas (Equações 2, 3 e 4):

⁴ https://en.wikipedia.org/wiki/Markov_chain

$$P(w_t | \text{contexto}) \forall t \in V \text{ (Equação 2)}$$

Onde w_t representa o *token* (*token* é uma *string* com um significado atribuído a esta, neste caso sendo uma palavra utilizada nestes modelos) atual; contexto representa o conjunto de *tokens* anteriores, posteriores ou uma amalgamação dos dois dependendo de qual tipo de contexto será utilizado; e V representa a sequência de *tokens*.

$$P(w_t | w_{t-k}, \dots, w_{t-1}) \text{ (Equação 3)}$$

Onde w_t representa o *token* atual e $(w_{t-k}, \dots, w_{t-1})$ representa o contexto citado anteriormente, formado apenas de *tokens* predecessores ao atual, e K representa o alcance do número de *tokens* do contexto a serem considerados no cálculo da probabilidade.

$$P(w_t | w_{t-k}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+k}) \text{ (Equação 4)}$$

Onde $(w_{t-k}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+k})$ representa o contexto formado de K tokens predecessores e posteriores (BENGIO, 2008).

A partir do ano de 2018, começaram a surgir e se popularizar os modelos de linguagem de larga escala, que possuem uma quantidade de parâmetros na casa dos bilhões (GOLED, 2021). Um dos principais responsáveis por esse surgimento foi o GPT (sigla para *Generative pre-trained transformer*, do português “Transformador generativo pré-treinado”) modelo 1, lançado pela empresa OpenAI, que foi treinado em 985 milhões de parâmetros (ZHU ET AL., 2015). Estes tipos de modelos se destacam por sua capacidade de obter um bom desempenho em várias áreas gerais do conhecimento, além de serem capazes de emular de forma satisfatória a sintaxe e semântica da linguagem humana (MANNING, 2015). A arquitetura *Transformer*, que serve como base para a arquitetura GPT, é um modelo de aprendizagem profunda e será discutido na próxima seção.

2.2. Transformer

Transformer é um modelo de aprendizagem profunda introduzido inicialmente pelo Google Brain (time de pesquisas de aprendizagem profunda da Google) e foi proposto como um modelo a ser usado principalmente nas áreas de processamento de linguagem natural (PLN) (VASWANI *ET AL.*, 2017) e visão computacional (HE, 2020). Este modelo vem sendo cada vez mais usado nas duas áreas citadas anteriormente, substituindo modelos de Redes Neurais Recorrentes (do inglês, *Recurrent Neural Networks* - RNNs) como a Memória de Longo Curto-Prazo (do inglês *Long Short-Term Memory* - LSTM) e a Unidade Recorrente Barrada (do inglês, *Gated Recurrent Unit* - GRU), devido à sua capacidade de processar suas entradas de uma vez só. Graças a seu mecanismo de auto-atenção, que provê contexto a qualquer posição da sequência de entrada, é possível uma paralelização maior do que as RNNs, reduzindo tempo de treinamento (RADFORD *ET AL.*, 2019).

O mecanismo de auto-atenção citado anteriormente resolveu um problema que as RNNs possuíam em relação ao processamento sequencial de *tokens*, já que ao processar o *token* de número n , os modelos combinavam o estado de todos os *tokens* anteriores em um só. Em seguida, estes eram combinados com o estado do *token* atual, o que pode ser um problema em sequências grandes, pois o contexto de *tokens* anteriores é diluído e o sentido original pode ser perdido.

Essa dependência da computação dos resultados dos *tokens* anteriores para processar o atual também dificulta a paralelização de processos, o que pode tornar o treinamento destes modelos ineficiente. O mecanismo de auto-atenção permite ao modelo buscar o estado em qualquer ponto anterior da sequência e pesá-los de acordo com uma medida de relevância aprendida no treinamento, o que possibilita o fornecimento de informações relevantes sobre *tokens* distantes em relação ao estado atual (VASWANI *ET AL.*, 2017).

A arquitetura do modelo *Transformer* é composta por três componentes: 1) Entrada; 2) Codificador/Decodificador e 3) Unidades de atenção escalonada de produto escalar. A entrada tem o formato de texto e é convertido em *tokens* através de um “tokenizador de codificação de pares de bytes”, e em seguida cada *token* é convertido em um vetor através da técnica de “*embedding* de palavras”, junto com a informação de suas posições (VASWANI *ET AL.*, 2017).

O codificador é composto por camadas que processam a entrada e as passam para a camada seguinte, gerando codificações contendo informações sobre a relevância contextual

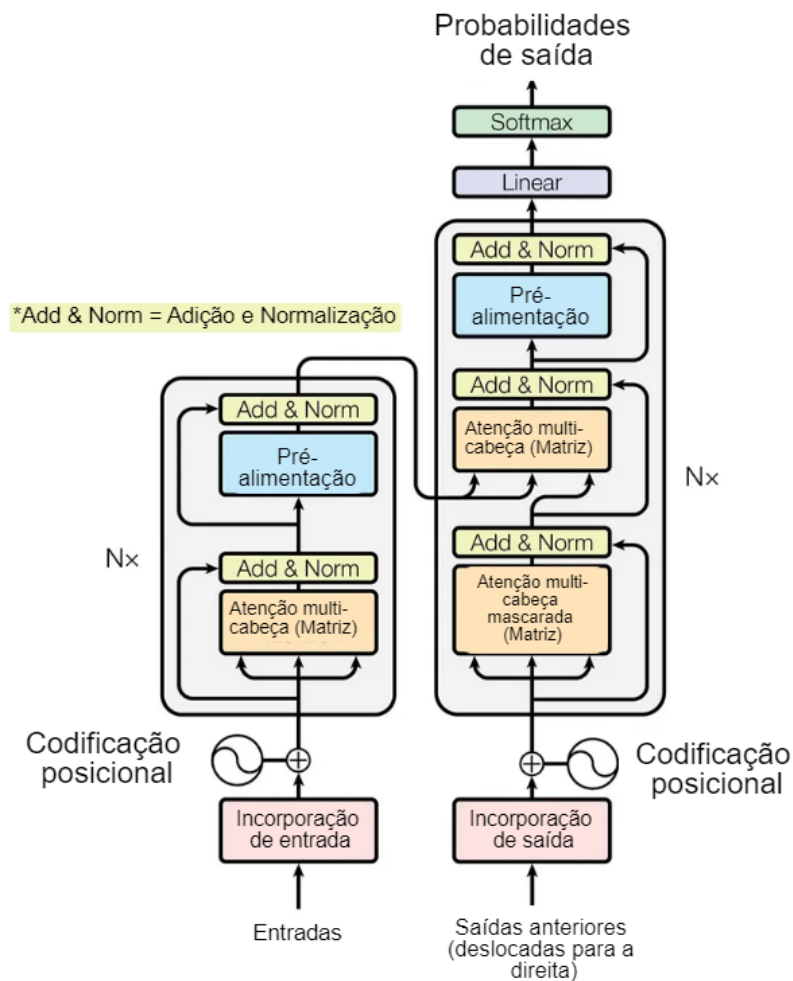
dos elementos da sequência em relação aos outros. O decodificador faz o oposto, juntando as informações contextuais geradas pelo codificador para gerar uma sequência de saída (VASWANI *ET AL.*, 2017).

As “unidades de atenção escalonada de produto escalar” são responsáveis pelo cálculo da auto-atenção para todos os *tokens* da sequência, e esse cálculo pode ser representado por uma matriz que utiliza a função *softmax*, que é útil no treinamento devido a otimização que possui para o cálculo de matrizes, o que agiliza o processo. A fórmula é mostrada na Equação 5:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \text{ (Equação 5)}$$

Onde Q é uma matriz que representa a palavra atual; K é uma matriz que representa as chaves das palavras restantes da sequência; V representa os valores dessas palavras; d_k representa a dimensão dos vetores de K (E a matriz K^T que representa a transposta de K) (VASWANI *ET AL.*, 2017). A seguir será mostrada a arquitetura do modelo *Transformer* em uma imagem (Figura 2):

Figura 2 - Arquitetura do modelo *Transformer*.



Fonte: Artigo no site Medium⁵.

Na próxima seção será discutido o *chatbot* ChatGPT, que utiliza uma arquitetura baseada no modelo *Transformer*.

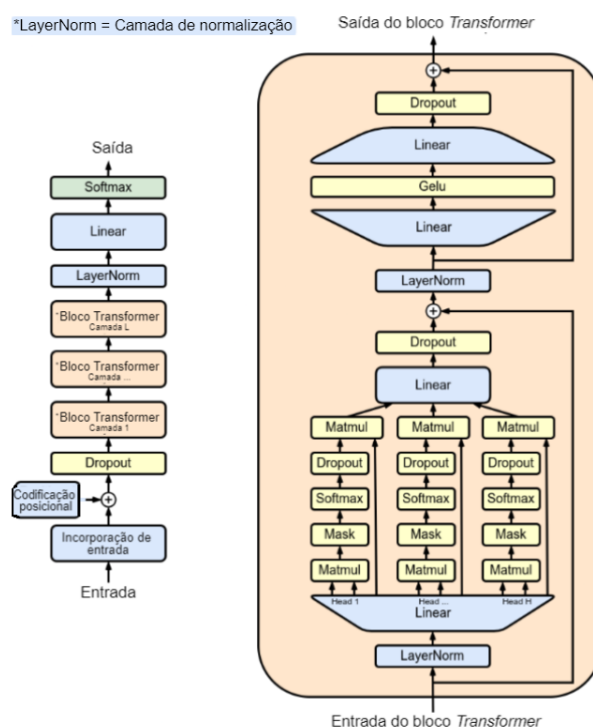
2.3. ChatGPT

ChatGPT é um *chatbot* que utiliza inteligência artificial, criado pela empresa OpenAI e lançado em 30 de novembro de 2022, e utiliza os modelos de linguagem GPT 3.5 e GPT 4 também criados pela OpenAI (o primeiro modelo mencionado está disponível de graça, e o segundo requer pagamento de inscrição). O “Chat” no nome remete ao fato de ser um *chatbot*, e GPT significa “*Generative Pre-Trained Transformer*”, referente ao modelo de

⁵ <https://medium.com/inside-machine-learning/what-is-a-transformer-d07dd1fbec04>

linguagem já discutido anteriormente. A principal diferença entre o GPT e a arquitetura Transformer é a falta de codificação posicional em seus blocos *Transformer*, que é feita apenas uma vez na entrada inicial (RADFORD *ET AL.*, 2018). A seguir será mostrada uma imagem contendo a arquitetura do modelo GPT original (Figura 3).

Figura 3 - Arquitetura do modelo GPT original, baseada na arquitetura *Transformer*.



Fonte: Wikipédia⁶.

O ChatGPT é considerado um “GPT de atividade específica”, devido ao seu foco na conversação gerado a partir do uso de treinamento supervisionado em larga escala, no qual os pesquisadores criavam perguntas e respostas e as mostravam para a máquina, e treinamento por reforço, no qual os pesquisadores ranqueavam as respostas geradas pela máquina em conversas anteriores (SCHULMAN *ET AL.*, 2023).

O modelo tem sido treinado principalmente como um *bot* de conversas, mas também tem provado possuir habilidades relativamente consistentes em outras áreas, como na de programação, escrita de poemas, poesias e músicas, tradução, etc., além de possuir a capacidade de responder a testes (LOCK, 2022). Apesar disso, foi notada a capacidade do ChatGPT de responder questões de forma superficialmente plausível mas que possui lógica falha, chegando a conclusões aparentemente de forma aleatória (VINCENT, 2022).

⁶ https://en.wikipedia.org/wiki/Generative_pre-trained_transformer

2.4. Avaliação de modelos de linguagem

No quesito de avaliação de modelos de linguagem, dois componentes são essenciais para sua realização: o método de avaliação a ser utilizado e a estrutura da base de dados a ser utilizada na avaliação para fins de teste. O método de avaliação mais comumente utilizado é o de Perplexidade, que mede a capacidade de uma distribuição probabilística prever uma amostra (JURAFSKY E MARTIN, 2021). Nos próximos parágrafos serão discutidos esse método e as estruturas de bases de dados geralmente utilizadas.

Como mencionado anteriormente, “Perplexidade” é um método da área de teoria da informação, que mede a capacidade de um modelo probabilístico prever determinadas amostras, como a próxima palavra de uma frase, por exemplo. A Equação 6 define a perplexidade utilizada na área de PLN, sendo calculada em cima de cada palavra:

$$PP(W) = P(w_1 w_2 \dots w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1 w_2 \dots w_N)}} \text{ (Equação 6)}$$

Onde as variáveis w se referem às palavras da sentença sendo processada e N representa o número de palavras da sentença (JURAFSKY E MARTIN, 2021).

As bases de dados utilizadas neste tipo de avaliação geralmente são construídas para se avaliar a performance do modelo em áreas específicas, como na área de processamento de linguagem natural, na qual se utiliza comumente bases de dados como a GLUE, SuperGLUE e decaNLP (HUYEN, 2019).

Em relação ao ChatGPT, o uso da perplexidade é adequado, mas no problema específico deste trabalho, como será avaliada a capacidade de acerto do ChatGPT na área de questões de múltipla escolha, será necessário apenas uma comparação de acerto e erro com os gabaritos oficiais.

3. TRABALHOS RELACIONADOS

Neste Capítulo serão apresentados trabalhos, artigos e dissertações sobre experimentos efetuados avaliando a ferramenta ChatGPT e suas capacidades para obter um bom desempenho em testes com áreas específicas de conhecimento, e a implicação do seu poder atual de emulação do comportamento humano nessas atividades e na área de ensino.

3.1. *Would Chat GPT Get a Wharton MBA?*

De autoria do professor Christian Terwiesch, da escola de Wharton na Universidade da Pensilvânia, o artigo "*Would Chat GPT Get a Wharton MBA?*" (2023) relata uma avaliação feita utilizando ChatGPT (não é especificada sua versão), em relação à sua capacidade de passar no teste final para obtenção do diploma de Mestre em Administração de Negócios (*Master of Business Administration* (MBA), em inglês) e assim refletir sobre as implicações que o uso de tal ferramenta tem e terá no futuro do ensino. O artigo foi dividido em três partes: 1) O fornecimento de sete questões da prova MBA para que o ChatGPT às responda; 2) O *prompt* solicitando ao ChatGPT que formule duas questões de prova do MBA; 3) A avaliação e reflexões do professor acerca dos resultados obtidos.

As conclusões obtidas pelo professor foram que o ChatGPT tem um bom desempenho no gerenciamento de operações básicas e na análise de processamento, sendo capaz de dar boas respostas, de acordo com a análise das respostas recebidas. No entanto, também é propenso a cometer erros de Matemática simples, o que pode levar a uma cascata de erros maiores, além de não conseguir lidar com questões mais complexas e nem ser capaz de gerar questões que façam muito sentido, precisando de um humano para corrigi-las depois. Apesar dos problemas, o ChatGPT também mostrou possuir uma boa adaptabilidade, respondendo bem a sugestões humanas sobre os seus erros.

3.2. *GPT takes the Bar Exam*

O artigo "*GPT takes the Bar Exam*", de autoria dos professores Michael J. Bommarito e Daniel Martin Katz, avaliou, em 2022, o desempenho do GPT 3.5 na seção de múltipla escolha do *Bar Exam*, um exame requerido para profissionais da área de Direito poderem

exercer sua profissão, que é similar ao exame aplicado no Brasil pela Ordem dos Advogados do Brasil (OAB). Este exame, por pertencer à área de Direito, é altamente dependente de uma linguagem complexa e repleta de nuances na sua estrutura, o que a torna difícil até mesmo para os estudantes da área, que possuem uma taxa média de acerto nas questões do teste de 68%, e pouco mais de um quinto falha na primeira tentativa.

O exame é formado por três componentes principais: 1) Teste de múltipla escolha; 2) Teste escrito; e 3) Teste de performance baseado em cenários, com o módulo de múltipla escolha, compondo 50% da nota do teste, e sendo dividido em oito categorias, com 25 questões cada (200 questões no total).

O banco de dados utilizado é formado pela seção de questões de múltipla escolha incluídas em um material de preparação para o teste que é vendido pela organização responsável pela sua formulação, a NCBE (sigla para *National Conference of Bar Examiners*, do português Conferência Nacional de Exames para Advogados), que custa 300 dólares. A razão pela escolha deste material se deu ao fato de ele não estar disponível ao público, o que diminui a chance deste ter sido usado como base de treinamento para o GPT-3.5.

A avaliação foi feita utilizando o método *zero-shot*, que consiste em apenas enviar a questão sem nenhum exemplo de questões e respostas demonstrados à máquina, e o envio de questões foi feito de sete formas: 1) Seleção única de resposta; 2) Seleção única de resposta com explicação; 3) Seleção das duas respostas mais corretas; 4) Seleção das duas respostas mais corretas com explicação; 4) Seleção das duas respostas mais corretas e reenviar a pergunta; 5) Seleção das duas respostas mais corretas com reenvio; 6) Ranquear as respostas da mais correta à mais errada; 7) Seleção das três respostas mais corretas. Os resultados ficaram entre 42% e 46% de acurácia para as seis primeiras formas, mas para a sétima foi significativamente maior, ficando na casa dos 50%.

3.3. Evaluating GPT-3.5 and GPT-4 Models on Brazilian University Admission Exams

Cinco profissionais da Universidade de São Paulo, em conjunto com a Universidade de São Francisco, dos Estados Unidos, relataram, em 2023, no artigo de título "*Evaluating GPT-3.5 and GPT-4 Models on Brazilian University Admission Exams*", o desempenho das versões do GPT 3.5 e 4 (usando parâmetros customizados) em responder questões do ENEM (sigla para Exame Nacional do Ensino Médio). Foram usadas duas bases de dados: a

primeira, chamada de “ENEM *Challenge Dataset*” criada pela Universidade de São Paulo, possui em sua totalidade 1754 questões das provas ENEM de 2009 a 2017, que são divididas em seis categorias: 1) Compreensão de texto; 2) Conhecimento enciclopédico; 3) Compreensão de imagens; 4) Conhecimento de domínio específico; 5) Raciocínio matemático e 6) Elementos químicos. Foram eliminadas as categorias 3 (já que o GPT não analisa imagens), 5 e 6. Além dessa base, foi também usada a prova do ENEM de 2022.

As questões foram enviadas ao GPT de três formas: *zero-shot*, já mencionada na Seção 3.1, *few-shot*, que mostra à máquina alguns exemplos de questões e como devem ser respondidas seguido da pergunta real, e *few-shot* com *Chain of Thought*, que acompanha as perguntas com explicações de como devem ser respondidas para que a máquina quebre o problema em problemas menores com mais efetividade.

No ENEM 2022, utilizando o método *zero-shot* no GPT-4, obteve-se uma acurácia média de 79,66%, com a performance de acertos em questões de Ciências Humanas de 94,59%, mas, em Matemática, 40,91%. O uso do método *few-shot* aumentou a média em apenas 1%. Com o uso do método *few-shot* com *Chain of Thought*, a média teve um aumento de 6%, com Matemática tendo o maior aperfeiçoamento (50,00% → 72,73%), seguido de Ciências Naturais (76,92% → 88,46%), o que mostra a efetividade de incentivar o programa a quebrar problemas em partes menores. O uso deste último método com o GPT-3.5 resultou em um aumento na performance em questões de Matemática, mas, em contrapartida, resultou em um declínio na performance nas demais áreas, resultando em uma acurácia média menor.

Para a base de dados ENEM *Challenge*, o modelo GPT-3.5 obteve uma acurácia média no método *zero-shot* de 77,41% a 82,88%, enquanto que o GPT-4 obteve uma performance melhor, de 94,56%. Também foi testada a performance *zero-shot* do ENEM 2022 utilizando as regras aplicadas ao ENEM *Challenge*, e os resultados obtidos para o GPT-3.5 foram similares aos da base ENEM *Challenge*, o que reduziu suspeitas de contaminação de dados em seu treinamento. O GPT-4, porém, apresentou performance inferior nessa mesma relação, o que pode indicar a memorização de questões do ENEM *Challenge*, já que essa base se encontra disponível online.

Com esse estudo, foi observado o GPT-4 como o melhor modelo de linguagem para se usar para responder o ENEM, juntamente com o método *few-shot* com *Chain of Thought*, que permitiu um aumento significativo na acurácia das respostas, principalmente na área de Matemática.

3.4. Colocamos Inteligência artificial pra fazer ENEM: Olha no que deu!

O vídeo do Youtube, de autoria do canal “Manual do Mundo”⁷, relata o experimento de colocar o ChatGPT para responder as 180 questões do ENEM 2022. Para isso, a prova foi dividida em cinco partes entre o apresentador e seus editores, já que caso somente uma pessoa fizesse o envio de perguntas, demoraria mais ou menos seis horas para terminar.

Em algumas questões, o ChatGPT se recusou a escolher alternativas ao alegar que não haviam informações suficientes, assim sendo consideradas erradas, ou escrevia uma resposta ignorando as alternativas, sendo necessário guiar a máquina a escolher uma das alternativas. Em Linguagens e Códigos, a taxa de acerto foi de 32/45, ou 71,1%; em Ciências Humanas, a taxa de acerto foi de 40/45, ou 88,8%; em Ciências Naturais, a taxa de acerto foi de 21/45, ou 46,6%; e finalmente, na área de Matemática, a taxa de acerto foi de 12/44 (uma questão foi anulada), ou 27,2%.

A partir do resultado, vemos que a máquina tem uma maior facilidade nas áreas de Linguagens e Ciências Humanas, enquanto que possui mais dificuldade nas áreas de Matemática e Ciências Naturais. Após isso, então, o apresentador discorre sobre o impacto do ChatGPT no setor de educação e suas implicações, já que essa ferramenta torna possível o seu uso por alunos em atividades e trabalhos escolares, dificultando a identificação do autor real. Os autores supõem que, após um certo tempo de adaptação, o ChatGPT se tornará mais uma ferramenta utilizada no aprendizado escolar, como ocorreu com as escritas de trabalho pelo computador, onde havia as mesmas preocupações acerca de autoria.

3.5. ChatGPT faz prova do ENEM e tem resultado melhor que 79% dos candidatos

Em um estudo realizado pelo jornal investigativo DeltaFolha⁸, o ChatGPT-3.5 foi colocado para responder 1290 questões das provas do ENEM de 2017 a 2021 e obteve uma média de 612,3 pontos, o que representa uma nota maior que 78,9% dos candidatos reais. Além disso, se saiu melhor que 98,9% dos alunos na área de Ciências Humanas, 95,3%, na área de Linguagens e Códigos e 94,6%, na área de Ciências da Natureza. Já em Matemática,

⁷ <https://www.youtube.com/@manualdomundo>

⁸

<https://guiadoestudante.abril.com.br/noticia/chatgpt-faz-prova-do-enem-e-tem-nota-melhor-que-79-dos-candidatos/>

o resultado foi abaixo da média de 527,1 dos candidatos humanos, e obteve uma taxa de acerto entre 13,6% e 27,3% apenas, reforçando a dificuldade que a ferramenta tem ao lidar com questões matemáticas.

3.6. Tabela de comparação entre os trabalhos

A seguir será feita uma comparação de dados e ferramentas utilizadas em nosso trabalho com os trabalhos citados neste módulo, que inclui três artigos, uma reportagem de site de notícias e um vídeo no *Youtube*.

Tabela 1. Comparação dos dados e ferramentas utilizadas neste trabalho com os trabalhos citados neste módulo.

Trabalho	Ferramentas utilizadas e suas versões	Alvo de testes
Nosso trabalho	ChatGPT 3.5 e Bard	Provas do ENEM 2017 a 2022
Would Chat GPT Get a Wharton MBA?	ChatGPT (Não foram especificadas versões)	Teste de Mestre em Administração de Negócios
GPT takes the Bar Exam	GPT 3.5	<i>Bar Exam</i>
Evaluating GPT-3.5 and GPT-4 Models on Brazilian University Admission Exams	GPT 3.5 e 4	Base de dados contendo questões das provas ENEM 2009 a 2017; prova do ENEM de 2022
Colocamos Inteligência artificial pra fazer ENEM: Olha no que deu!	ChatGPT (Não foram especificadas versões)	Prova do ENEM 2022
ChatGPT faz prova do ENEM e tem resultado melhor que 79% dos candidatos	ChatGPT 3.5	Provas do ENEM 2017 a 2021

4. METODOLOGIA

Neste Capítulo será apresentada a metodologia usada nos experimentos realizados, e isto inclui a base de dados e ferramentas utilizadas, bem como a categorização feita nas questões utilizadas para entrada, experimento, e forma de avaliação considerada para medir o desempenho do ChatGPT nas respostas dadas às questões.

4.1. Base de dados

A construção da base de dados⁹ desse trabalho foi realizada através da transcrição e formatação de questões de Matemática das provas ENEM (localizadas em sua íntegra no site do INEP em formato .pdf) que foram aplicadas entre 2017 e 2022. Esse período foi selecionado devido à mudança na estrutura de aplicação das provas que, no passado (2009-2016), eram aplicadas as questões de Ciências Humanas e Ciências da Natureza no sábado, com 4h30 de duração, e Linguagens, Matemática e Redação no domingo seguinte, com 5h30 de duração, formato que mudou em 2017, quando o ENEM passou a ser aplicado em dois domingos consecutivos, sendo no primeiro aplicadas as provas de Linguagens, Ciências Humanas e Redação, com 5h30 de duração, e no segundo as provas de Matemática e Ciências da Natureza, com 4h30 de duração.

As edições escolhidas para uso foram as direcionadas à inscritos portadores de deficiência visual, deficiência intelectual, autismo, déficit de atenção ou dislexia, pois as questões dessas provas possuem descrição para figuras apresentadas, e isso é um detalhe importante, uma vez que o ChatGPT não possui suporte à leitura de imagens. Cada uma das seis provas contém 45 questões da área de Matemática, o que totaliza em 270 questões para o ChatGPT responder.

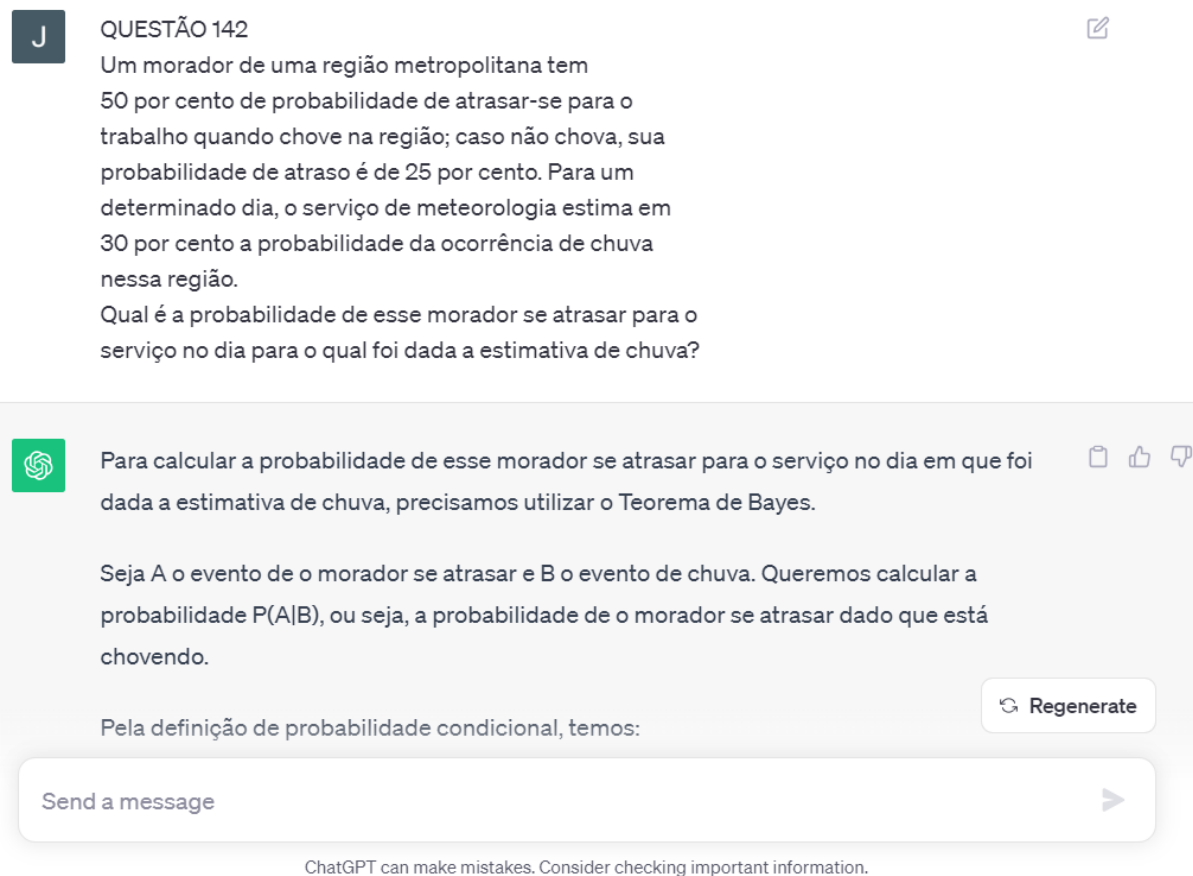
4.2. Ferramentas utilizadas

A ferramenta utilizada no processamento das questões neste experimento foi o ChatGPT-3.5, um *chatbot* de Inteligência Artificial, de uso geral. A ferramenta é baseada no

⁹ https://drive.google.com/drive/folders/1_sErLS8sJ6CawbXpjoCfr1jjwuVTDxn?usp=sharing

envio de *prompts* em formato de (e somente de) texto, que envia de volta uma resposta após um tempo de processamento, como mostra a Figura 4.

Figura 4 - Exemplo de uso do ChatGPT para responder as questões do ENEM.



Fonte: Autoria própria.

Foi também utilizada a plataforma Google Docs para armazenamento das questões, que precisaram de alguns ajustes de formatação antes de serem enviadas ao ChatGPT, como a retirada de espaços inadequados e a digitação manual de palavras que não podiam ser copiadas e coladas, e o armazenamento também das respostas coletadas, bem como o registro de questões certas e erradas, de categorias para atribuir às questões e o motivo para os erros cometidos pelo ChatGPT.

Em alguns casos, a cópia das questões não era possível, então foi utilizada uma plataforma online de conversão de imagens para texto chamada “Invertexto” para obtenção da questão desejada, seguida de correção manual de erros encontrados.

Além disso, ocorreram ocasiões em que o ChatGPT gerou fórmulas matemáticas em formato LaTeX bruto, o que dificultou o entendimento destas, sendo necessário utilizar um

editor LaTeX online para gerar as fórmulas formatadas, capturar suas imagens em formato .jpg e colar nos documentos onde elas se encontravam.

Para o armazenamento da listagem de questões, as unidades temáticas às quais as questões pertencem, e motivos dos erros cometidos pelo ChatGPT, foi utilizada a plataforma Google Sheets, que gerencia arquivos de planilhas.

Para a análise das resoluções efetuadas pelo ChatGPT, foram utilizadas as resoluções comentadas disponíveis no site do Curso Objetivo, escolhido por sua popularidade e acervo extenso de correções; para questões não encontradas neste site, foram utilizadas resoluções do Youtube, mais especificamente dos canais “Professor Caju¹⁰”, “Explicae¹¹” e “Equaciona com Paulo Pereira¹²”, escolhidos pela alta disponibilidade de questões respondidas.

4.3. Categorização de unidade temática das questões respondidas

A categorização das questões respondidas foi feita utilizando a matriz de referência¹³ do ENEM, que divide a área de Matemática em sete competências, listadas a seguir:

1. Construir significados para os números naturais, inteiros, racionais e reais;
2. Utilizar o conhecimento geométrico para realizar a leitura e a representação da realidade e agir sobre ela;
3. Construir noções de grandezas e medidas para a compreensão da realidade e a solução de problemas do cotidiano;
4. Construir noções de variação de grandezas para a compreensão da realidade e a solução de problemas do cotidiano;
5. Construir noções de variação de grandezas para a compreensão da realidade e a solução de problemas do cotidiano;
6. Interpretar informações de natureza científica e social obtidas da leitura de gráficos e tabelas, realizando previsão de tendência, extrapolação, interpolação e interpretação;
7. Compreender o caráter aleatório e não-determinístico dos fenômenos naturais e sociais e utilizar instrumentos adequados para medidas, determinação de amostras e cálculos de probabilidade para interpretar informações de variáveis apresentadas em uma distribuição estatística.

¹⁰ <https://www.youtube.com/@profcaju/featured>

¹¹ <https://www.youtube.com/@ExplicaeEnem/featured>

¹² <https://www.youtube.com/@equacionamatematica/featured>

¹³ https://download.inep.gov.br/download/enem/matriz_referencia.pdf

4.4. Categorização dos tipos de representações visuais presentes nas questões respondidas

A categorização das questões respondidas, em relação à presença (ou ausência) de representações visuais e o tipo destas, foi feita de acordo com as categorias de autoria própria listadas abaixo:

1. Sem representações visuais de dados: a questão somente possui texto a descrevendo.
2. Figuras(s): a questão possui uma ou mais figuras ilustrativas (como por exemplo a imagem de uma imagem relevante para a questão);
3. Fórmula(s): a questão apresenta uma ou mais fórmulas a serem utilizadas na resolução;
4. Matriz(es): a questão apresenta uma ou mais matrizes a serem utilizadas na resolução;
5. Tabela(s): a questão apresenta uma ou mais tabelas a serem utilizadas na resolução;
6. Forma(s) geométrica(s) descrita(s): a questão apresenta uma ou mais formas geométricas relevantes para a resolução da questão;
7. Gráfico(s) de linha: a questão apresenta um ou mais gráficos de linha relevantes para a resolução da questão;
8. Gráfico(s) de barra: a questão apresenta um ou mais gráficos de barra relevantes para a resolução da questão;
9. Gráfico(s) em rede: a questão apresenta um ou mais gráficos em rede relevantes para a resolução da questão;
10. Gráfico(s) de pontos: a questão apresenta um ou mais gráficos de pontos relevantes para a resolução da questão;
11. Gráfico(s) com figuras (pictogramas): a questão apresenta um ou mais gráficos que utilizam figuras para representar os dados relevantes para a resolução da questão;
12. Gráfico(s) com formas geométricas: a questão apresenta um ou mais gráficos nos quais estão representados figuras geométricas a serem utilizadas na resolução da questão;
13. Gráfico(s) de coluna: a questão apresenta um ou mais gráficos de coluna relevantes para a resolução da questão;
14. Gráfico(s) de setor: a questão apresenta um ou mais gráficos de setor relevantes para a resolução da questão.

4.5. Categorização dos tipos de erros cometidos pelo ChatGPT

A categorização das questões respondidas pelo ChatGPT de forma incorreta, em relação aos tipos de erros cometidos, foi feita de acordo com as categorias de autoria própria listadas abaixo:

1. Interpretação: houve uma interpretação errônea em relação às características de determinada questão (como por exemplo, a interpretação incorreta de gráficos e seus valores, das características de formas geométricas apresentadas, ou de que tipo de raciocínio uma determinada questão exige);
2. Lógica: a forma proposta para resolução não possui sentido algum (como por exemplo, definir uma fórmula sem sentido que utiliza variáveis desnecessárias);
3. Comparação de números: em uma questão de escolher o maior ou menor valor dentre dois ou mais valores, faz uma comparação incorreta e, por consequência, escolhe o valor incorreto (como por exemplo, ao escolher o maior valor entre os números 1 e 2, o ChatGPT escolhe o valor 1);
4. Escolha incorreta de alternativa: apresenta um raciocínio correto na hora de resolver a questão, inclusive chegando no valor correto a ser selecionado, mas, por alguma razão, escolhe uma alternativa que não condiz com a conclusão obtida;
5. Elemento de cálculo ignorado: ignora variáveis apresentadas pela questão na hora do cálculo e, por consequência, apresenta um raciocínio e conclusão incorretos;
6. Atribuição incorreta de valores: atribui valores diferentes e incorretos à variáveis apresentadas na questão;
7. Cálculo incorreto de valores: faz um ou mais cálculos incorretos de expressões em seu raciocínio (por exemplo, $1 + 1 = 3$);
8. Estruturação das equações: apresenta um raciocínio correto, mas erra na hora de estruturar equações, chegando a uma conclusão incorreta.

4.6. Experimentos

As questões foram fornecidas ao ChatGPT de duas formas: primeiro, com as cinco alternativas de respostas, já disponíveis para os estudantes, que é a forma padrão das questões do ENEM (forma que será referenciada como “modalidade com alternativas” daqui em diante); em seguida, foram retiradas as alternativas de forma a deixar as questões abertas

(será referenciada como “modalidade sem alternativas”). O ChatGPT se recusou a responder algumas questões alegando falta de informações, e também respondeu algumas questões de forma irregular, ao ignorar as alternativas (no formato de questões padrão mencionado anteriormente) e criar uma resposta própria, que diferia das existentes; nestes dois casos, foi necessário informar ao ChatGPT para processar novamente a questão e, em alguns casos, repetir este passo várias vezes antes de finalmente obter uma resposta satisfatória (que informe uma alternativa ou que informe um valor ou conclusão, no caso do experimento sem alternativas).

Além disso, foram selecionadas oito questões aleatórias pertencentes a cada categoria de erro (Seção 4.5) respondidas pelo ChatGPT, para uma análise mais detalhada da construção da solução e do erro ocorrido. Por fim, essas mesmas questões foram processadas no Bard, para comparação do desempenho dos dois *chatbots*. O *chatbot* Bard é baseado na família de modelos de linguagem LaMDA (sigla para *Language Model for Dialogue Applications*, do português Modelo de Linguagem para Aplicações de Diálogo) e posteriormente PaLM (sigla para *Pathways Language Model*, do português Modelo de Linguagem de Caminhos), desenvolvido pelo Google e lançado inicialmente no dia 21 de março de 2023.

4.7. Métricas de avaliação

Após as provas serem completamente respondidas, foi feita a verificação de quais questões o ChatGPT acertou e errou, de modo geral, e a taxa percentual obtida; a verificação da taxa de acerto e erro, em relação às categorias de divisão das provas de Matemática do ENEM; a verificação da presença (ou ausência) de elementos visuais e, finalmente, uma avaliação dos tipos de erros cometidos e suas frequências para a detecção de padrões nas respostas do ChatGPT. Em relação aos elementos visuais e os tipos de erros, foram criadas categorias, que serão ilustradas nos próximos pontos.

5. RESULTADOS E DISCUSSÃO

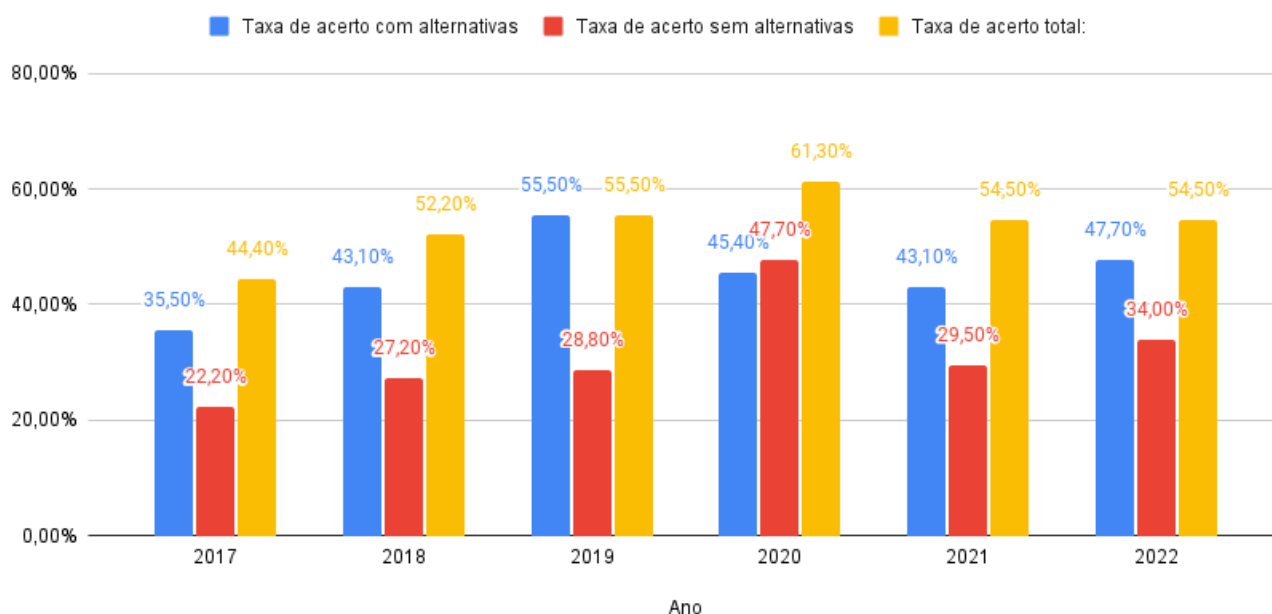
Neste Capítulo serão apresentados os resultados obtidos a partir dos experimentos e avaliações feitas. A Seção 5.1 contém a análise com as taxas de acerto nas modalidades com alternativa, sem alternativa, e as taxas de acerto total (união das duas modalidades mencionadas sem repetição). A Seção 5.2 contém as discussões sobre as unidades temáticas de todas as questões e das unidades temáticas das questões acertadas e erradas, divididas entre as modalidades com alternativas e sem alternativas. A Seção 5.3 faz o mesmo da Seção 5.2, só que para os tipos de representação visual das questões. A Seção 5.4 trata das questões em relação à categorização dos erros cometidos nas modalidades com alternativas e sem alternativas. A Seção 5.5 contém uma avaliação dos erros cometidos pelo ChatGPT em questões pertencentes à cada categoria de erro, selecionadas aleatoriamente; também é feita uma análise de performance do *chatbot* Bard nessas mesmas questões que em seguida é comparada com a do ChatGPT em termos de acerto ou erro e uma comparação das categorias de erro nas quais se encaixam.

5.1. Taxas de acerto do ChatGPT ao longo do tempo e com e sem alternativas

O gráfico de colunas (Figura 5) mostra as taxas percentuais de acerto do ChatGPT em cada prova, dividido por ano e pelas seguintes modalidades: com alternativas, sem alternativas e a taxa de acerto total, que representa uma união sem repetição das modalidades já citadas.

Figura 5 - Gráfico de colunas representativo das taxas de acerto do ChatGPT.

Taxas de acerto



Fonte: Autoria própria.

De acordo com o gráfico, podemos observar um padrão de crescimento nas taxas de acerto até o ano de 2019, onde a categoria com alternativas cresce de 35,5% a 55,5% e a sem alternativas cresce de 22,2% a 28,8%, sempre se mantendo em um valor inferior em relação à com alternativas. Em 2020, a modalidade sem alternativas obtém uma taxa de acerto maior que a com alternativas, o que não se repete no resto do gráfico. Em 2021 e 2022, as duas modalidades retornam ao seu comportamento anterior de crescimento, com a modalidade com alternativas obtendo novamente os maiores valores, em relação à sem alternativa.

Na terceira categoria, taxa de acerto total, podemos ver que apresenta praticamente o mesmo comportamento de crescimento de 2017 a 2020, seguido de uma queda e estabilização em 2021 e 2022. Podemos também observar que a taxa de interseção entre as duas categorias já mencionadas anteriormente é alta ou até completa, no caso de 2019.

No geral, as médias obtidas foram: 45,05% de acerto na modalidade com alternativas, 31,57% na modalidade com alternativas e 53,73% na taxa de acerto total.

A performance deste experimento quando comparado com alguns dos trabalhos relacionados mostra que o artigo “Evaluating GPT-3.5 and GPT-4 Models on Brazilian University Admission Exams” (Seção 3.3) e que utiliza o método *zero-shot* no GPT-4 sob os dados da prova de Matemática do ENEM 2022, obteve uma taxa de acurácia de 40,91%,

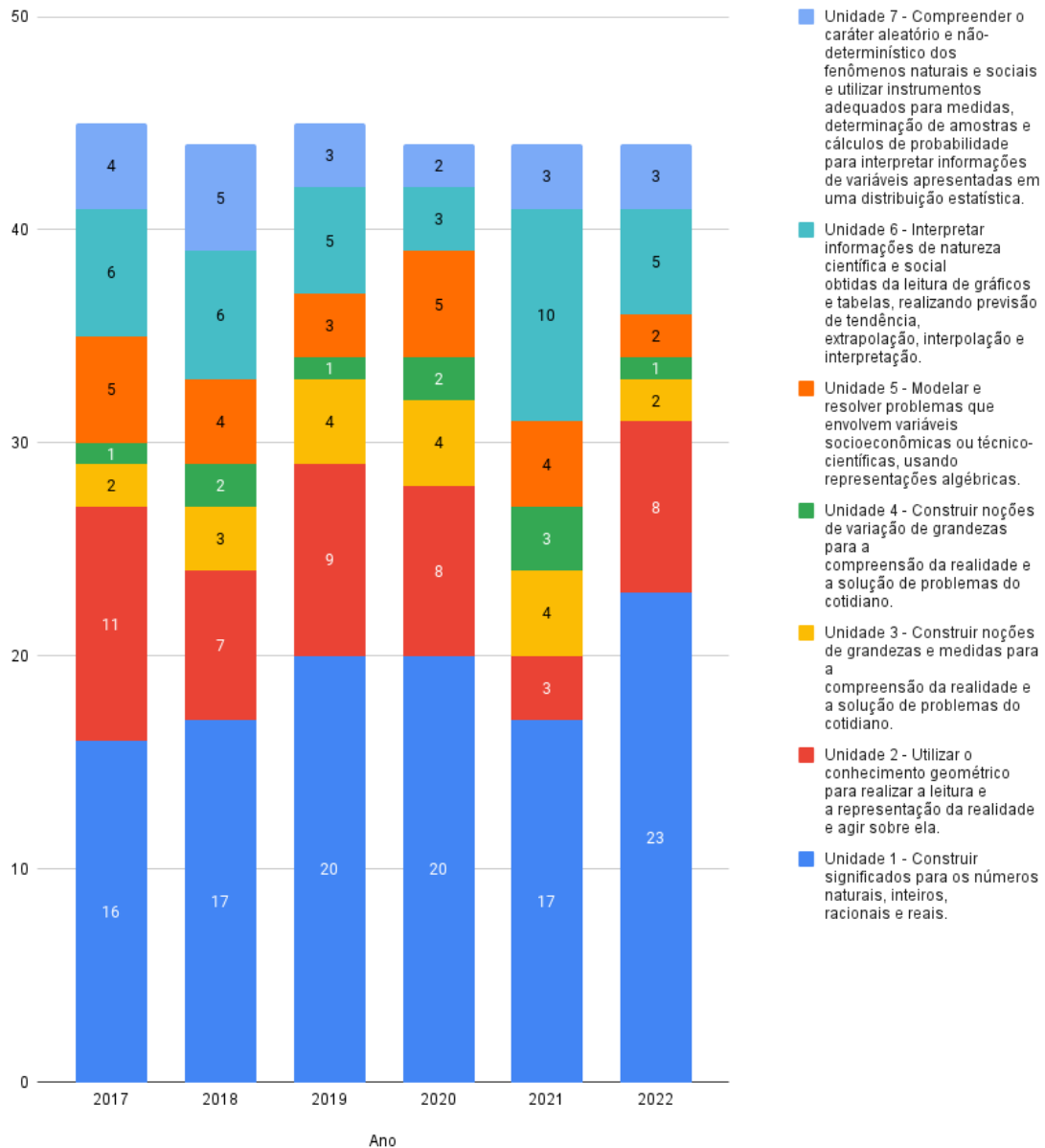
diferente dos 47,4% obtidos neste trabalho. Esta média também foi consideravelmente maior que a obtida no vídeo do youtube “Colocamos Inteligência artificial pra fazer ENEM: Olha no que deu!” (Seção 3.4), no qual foi obtida uma taxa de 27,3%. Isso também foi observado na publicação “ChatGPT faz prova do ENEM e tem resultado melhor que 79% dos candidatos”, onde foi obtida uma taxa de acerto entre 13,6% e 27,3%. Como não há uma quantidade satisfatória de experimentos realizados em cima desta prova, não é possível criar uma hipótese sobre o porquê da diferença entre estas performances.

5.2. Taxas de acerto e erro do ChatGPT por unidade temática em cada modalidade

A seguir serão mostrados os gráficos de coluna empilhada para categorização das unidades temáticas das questões das provas do ENEM avaliadas neste trabalho, nas modalidades com e sem alternativas (Figuras 6, 7 e 8).

Figura 6 - Gráfico de colunas empilhadas representativo das unidades temáticas de todas as questões.

Unidades temáticas de todas as questões



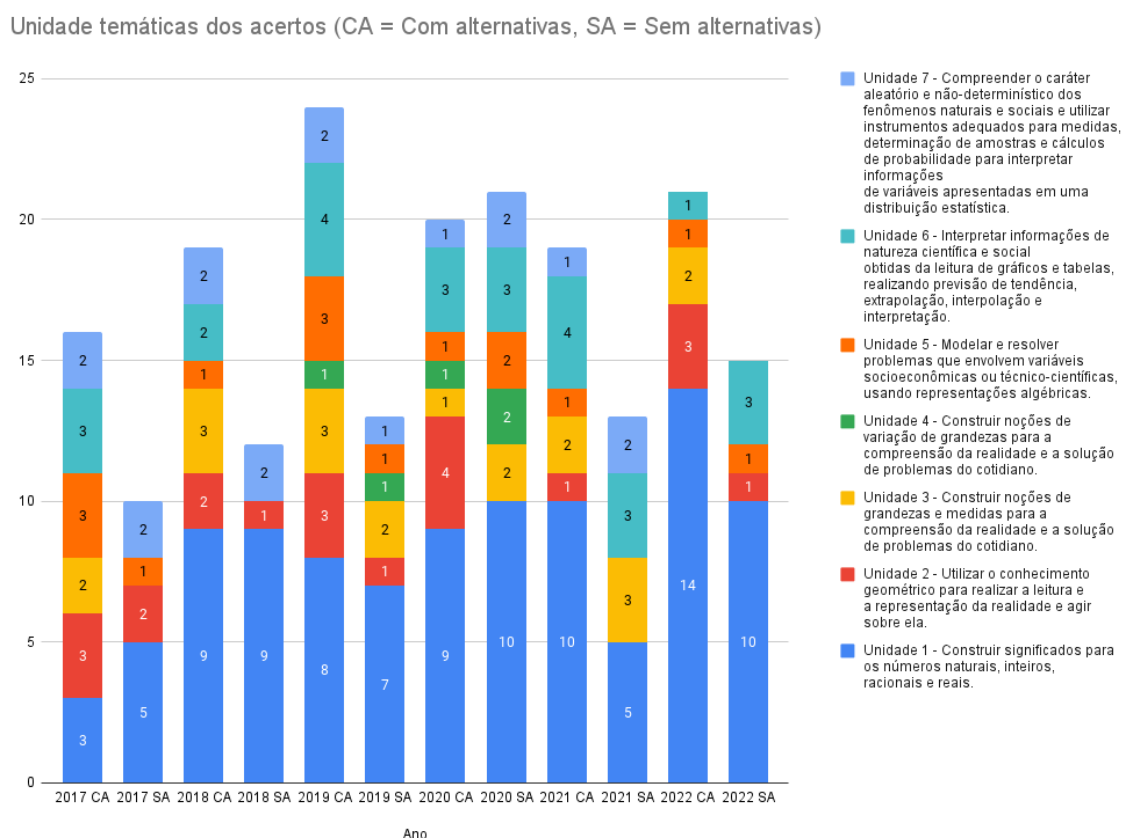
Fonte: Autoria própria.

Como vemos no gráfico da Figura 6, a maior parte das questões são categorizadas na Unidade 1 (Construir significados para os números naturais, inteiros, racionais e reais) (113 questões no total), seguido da Unidade 2 (Utilizar o conhecimento geométrico para realizar a leitura e a representação da realidade e agir sobre ela) (46 questões) e da Unidade 6

(Interpretar informações de natureza científica e social obtidas da leitura de gráficos e tabelas, realizando previsão de tendência, extrapolação, interpolação e interpretação) (35 questões), em termos de frequência. As unidades que são menos frequentes são as 3 (Construir noções de grandezas e medidas para a compreensão da realidade e a solução de problemas do cotidiano) (19 questões) e 4 (Construir noções de variação de grandezas para a compreensão da realidade e a solução de problemas do cotidiano) (10 questões).

No gráfico da Figura 7 vemos que as questões acertadas são categorizadas em sua maior parte também na Unidade 1, com exceção da prova de 2017 com alternativas, onde vemos que a esta unidade possui o mesmo número de questões das Unidades 2, 5 (Construir noções de variação de grandezas para a compreensão da realidade e a solução de problemas do cotidiano) e 6.

Figura 7 - Gráfico de colunas empilhadas representativo das unidades temáticas das questões acertadas.



Fonte: Autoria própria.

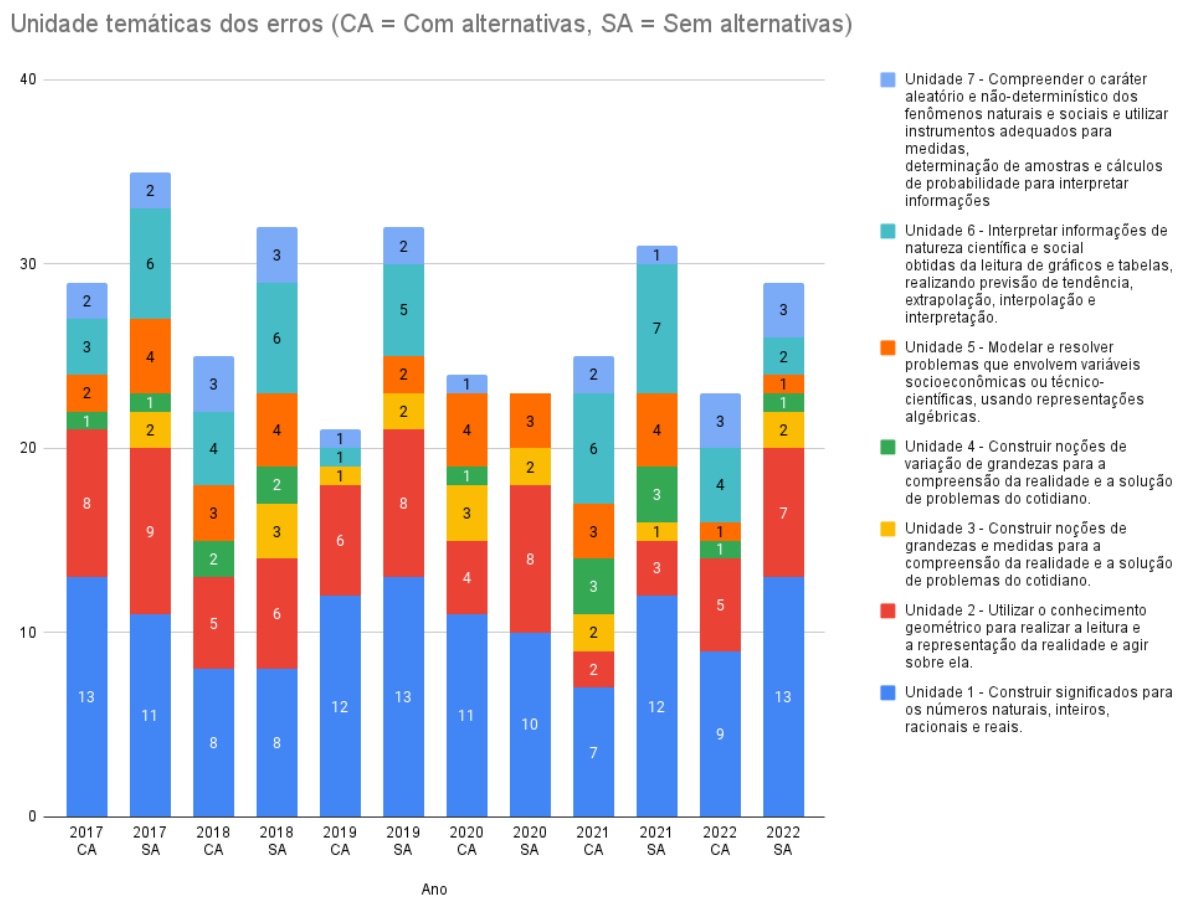
Na modalidade com alternativas, há uma variedade na segunda, terceira e quarta colocação da frequência de unidades, que são compostas em sua maior parte das Unidades 2, 3, 5 ou 6, que variam em quantidade de 1 a 4 questões, com a Unidade 6 mantendo uma maior frequência. As unidades desta modalidade com menor número de acertos são as Unidades 4 e 7. A unidade 4 somente aparece nas provas de 2019 e 2020, enquanto que a Unidade 7 aparece com consistência no acerto das provas, com exceção da prova de 2022.

Na modalidade sem alternativas há, geralmente, uma variedade menor na quantidade de unidades presentes, provavelmente devido ao número menor de acertos; o que não ocorre nas provas de 2019 e 2020, onde a variedade de unidades com acertos aumenta consideravelmente (ambas provas com 6 unidades). Esta modalidade mantém a alta frequência de acertos na Unidade 1, seguido de frequências consideravelmente menores para as unidades restantes.

Como podemos ver pelo gráfico, no geral, a frequência de unidades dos acertos tende a se igualar a frequência de unidades de todas as questões da Figura 6, o que pode significar que o ChatGPT não está tendo uma maior facilidade com unidades específicas da matriz de referência da prova.

O gráfico da Figura 8 mostra um comportamento similar ao gráfico de acertos, com exceção de um aumento na frequência de questões da Unidade 2, com um total de 30 na modalidade com alternativas e 41, para a modalidade sem alternativas (como demonstrado no gráfico da Figura 6 - esta unidade possui 46 questões no total); o que pode significar que o ChatGPT obteve uma maior dificuldade em questões desta unidade.

Figura 8 - Gráfico de colunas empilhadas representativo das unidades temáticas das questões erradas.



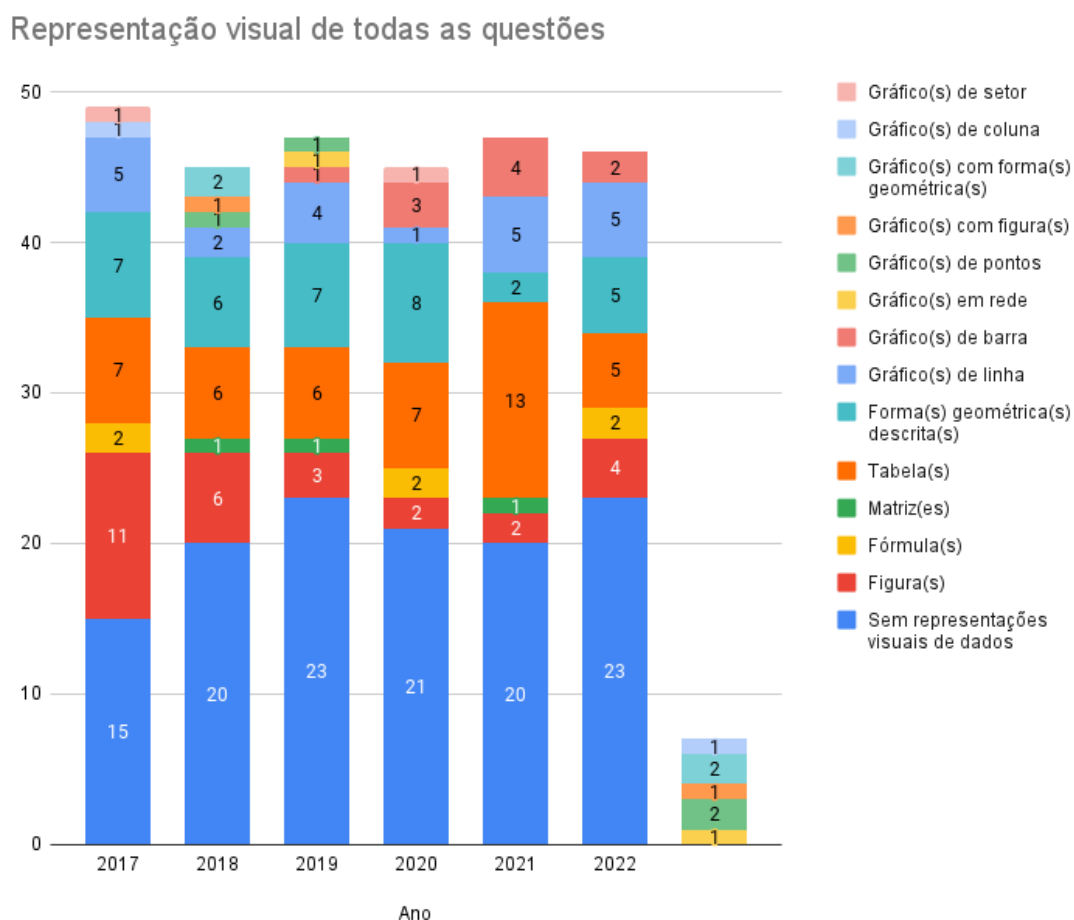
Fonte: Autoria própria.

Não foi possível efetuar uma comparação com os resultados obtidos nos trabalhos relacionados pois não foram feitas análises similares à esta.

5.3. Taxas de acerto e erro do ChatGPT por tipo de representação visual em cada modalidade

A seguir serão mostrados os gráficos de coluna empilhada para categorização da representação visual das questões das provas do ENEM avaliadas neste trabalho, nas modalidades com e sem alternativas (Figuras 9, 10 e 11).

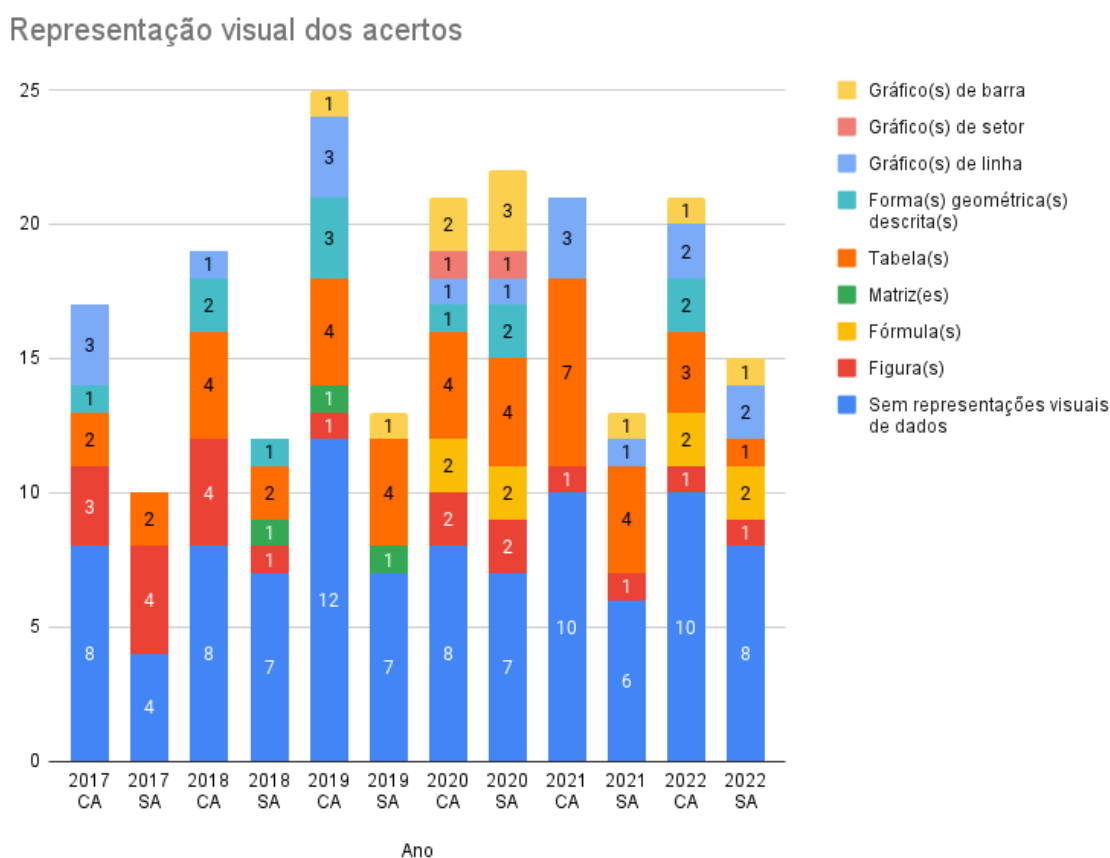
Figura 9 - Gráfico de colunas empilhadas representativo das representações visuais de todas as questões.



Fonte: Autoria própria.

Como vemos no gráfico da Figura 9, questões sem representação visual de dados compõem a maior parte das provas (122 questões no total), seguidos de tabelas (44 questões) e formas geométricas descritas (35 questões). Outras representações com frequências notáveis são as de figuras (28 questões) e gráficos de linha (22 questões).

Figura 10 - Gráfico de colunas empilhadas representativo das representações visuais das questões acertadas.



Fonte: Autoria própria.

No gráfico da Figura 10 vemos que a maior parte das questões com mais acertos não possuem representação visual, com exceção da prova de 2017 sem alternativas, onde questões com figuras possuem a mesma quantidade das sem representação, ambas com quatro. Podemos também ver que somente nove das 14 categorias de representação estão presentes no gráfico, faltando as categorias de “Gráficos em rede” (1 questão no total), “Gráficos de pontos” (2 questões), “Gráficos com figuras” (1 questão), “Gráficos com formas geométricas” (2 questões) e “Gráficos de coluna” (1 questão). Essa taxa nula de acerto nestas categorias poderia indicar uma maior dificuldade em responder questões com esses tipos de representações, mas como a quantidade de amostras de cada categoria é baixa, não é possível tirar uma conclusão.

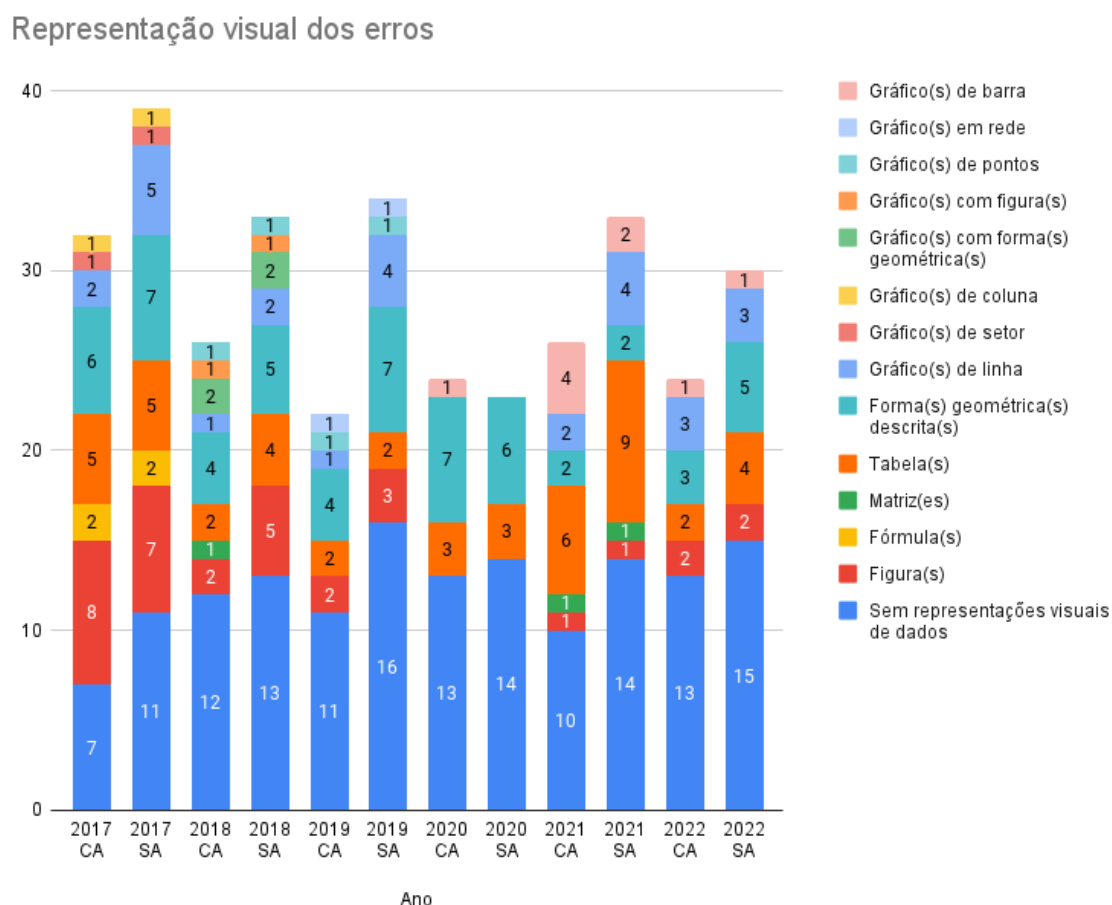
Em geral, não há uma diferença significativa na frequência de representações entre as modalidades sem e com alternativa, com a categoria “Tabelas” tomando o segundo lugar,

sendo ultrapassado pelas categorias “Gráficos de linha” e “Figuras” na prova de 2017 com alternativas, “Figuras” na prova de 2017 sem alternativas e estando empatado com “Figuras” na prova de 2018 com alternativas.

Como os gráficos da Seção 5.2, os gráficos de acerto desta seção não apresentam uma grande variabilidade em relação ao gráfico de todas as questões (Figura 9), o que pode significar que o ChatGPT não está tendo uma facilidade maior com questões que possuem representações visuais específicas.

Na Figura 11 temos o gráfico das questões respondidas erroneamente pelo ChatGPT e as representações visuais correspondentes nelas presentes.

Figura 11 - Gráfico de colunas empilhadas representativo das representações visuais das questões erradas.



Fonte: Autoria própria.

O gráfico da Figura 11 mostra um comportamento similar ao gráfico de acertos, com exceção de um aumento na frequência de questões com formas geométricas descritas, com um total de 26 questões para a modalidade com alternativas e 32, para a modalidade sem alternativas (como demonstrado no gráfico da Figura 9 - esta unidade possui 35 questões no total). Isso pode significar que o ChatGPT obteve uma maior dificuldade em questões com formas geométricas descritas, provavelmente devido ao alto nível de abstração requerido para interpretar corretamente este tipo de figura.

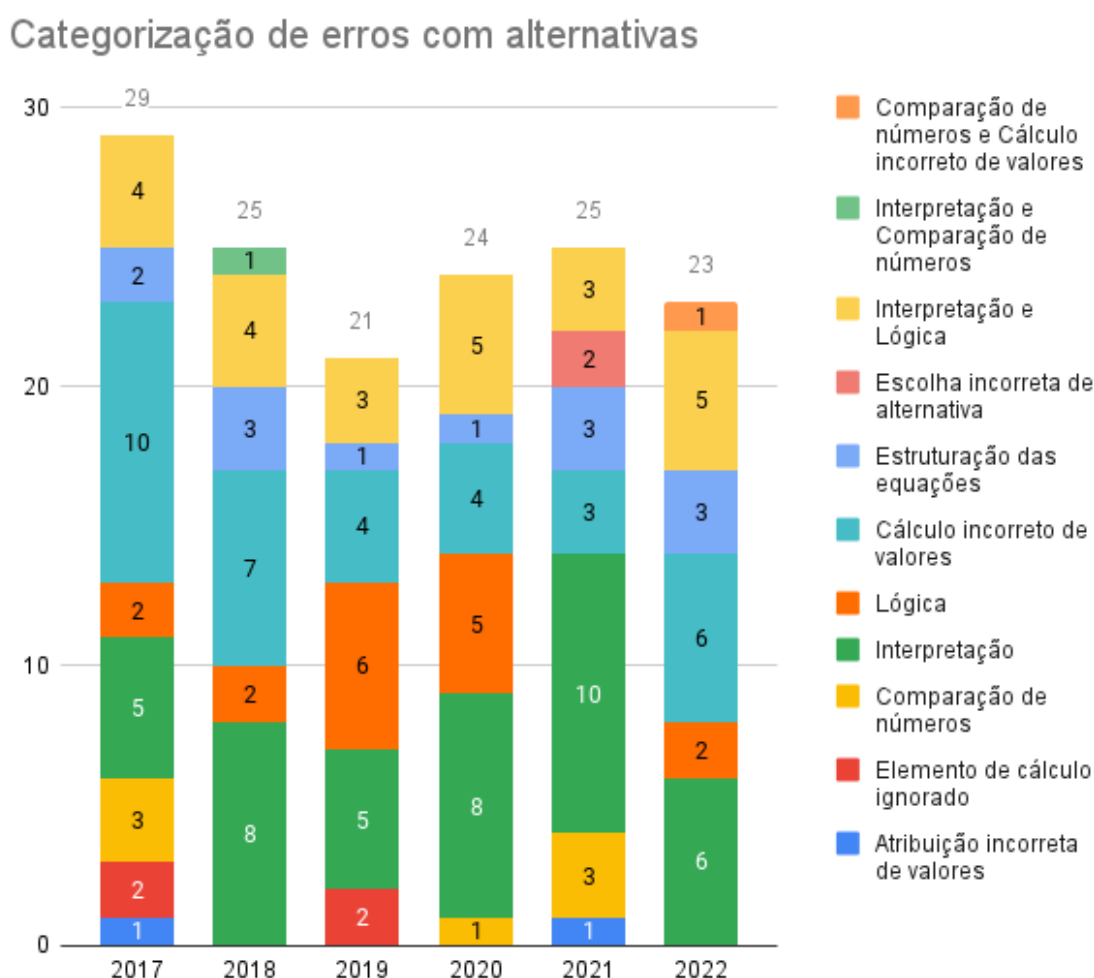
Não foi possível efetuar uma comparação com os resultados obtidos nos trabalhos relacionados pois não foram feitas análises similares à esta.

5.4. Categorização dos erros cometidos pelo ChatGPT em cada modalidade

A seguir serão mostrados gráficos (Figuras 12 e 13) de coluna empilhada para categorização dos erros cometidos nas provas do ENEM avaliadas neste trabalho, nas modalidades com e sem alternativas.

No gráfico da Figura 12 vemos que as categorias com maiores frequências são: “Interpretação”, com 42 questões no total; “Cálculo incorreto de valores”, com 34 questões; “Interpretação e Lógica”, com 24 questões; e “Lógica”, com 17 questões. Se integrarmos a categoria de “Interpretação e Lógica” às suas respectivas categorias singulares, acabamos tendo 66 questões na categoria de “Interpretação” e 41 questões na categoria de “Lógica”.

Figura 12 - Gráfico de colunas empilhadas representativo dos erros cometidos na modalidade com alternativas.



Fonte: Autoria própria.

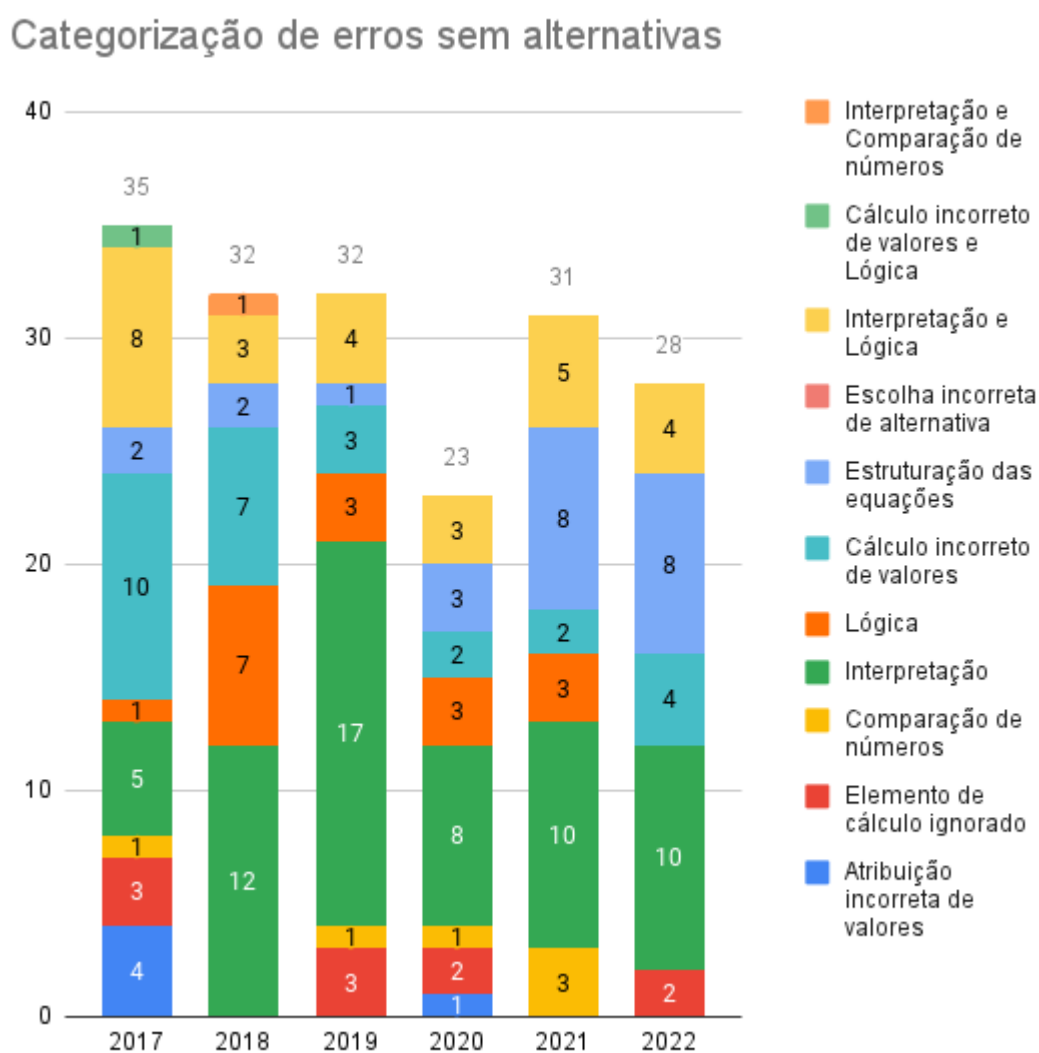
Este resultado da categoria de “Interpretação” mostra que o ChatGPT na maior parte das vezes tem dificuldade de analisar o que as questões propõem (provavelmente devido ao nível de abstração, frequentemente requerido para questões de Matemática). Em relação à categoria de “Lógica”, o ChatGPT também demonstra dificuldade ao formular uma solução que siga a lógica do problema proposto, soluções que podem possuir múltiplos passos, fórmulas e raciocínios que são utilizados sem uma sequência única.

Finalmente, temos a categoria de “Cálculo incorreto de valores”, na qual o ChatGPT mostra dificuldade na resolução de fórmulas (como por exemplo, $1 + 1 = 3$), sugerindo que o programa cria soluções misturando diferentes fontes usadas para seu treinamento e não

procura realmente processar o cálculo recebido como entrada, ou criada no seu processo de raciocínio da resolução.

No gráfico da Figura 13 vemos que a frequência das categorias principais se mantém em relação ao gráfico da modalidade com alternativas (Figura 12), com poucas mudanças: Vemos um aumento considerável na frequência da categoria “Interpretação”, que passou de 42 questões para 62 (um aumento de aproximadamente 50%); e um aumento na frequência da categoria “Elemento de cálculo ignorado”, que passou de quatro elementos para 10 (aumento de 150%). Nas demais, temos variações mínimas de frequências.

Figura 13 - Gráfico de colunas empilhadas representativo dos erros cometidos na modalidade sem alternativas.



Fonte: Autoria própria.

Este aumento na frequência da categoria “Interpretação” provavelmente se dá pela falta de alternativas nesta modalidade, que muitas vezes servem de modelo e “guia” para o ChatGPT identificar o objetivo a ser alcançado e que tipo de saída gerar.

Não foi possível efetuar uma comparação com os resultados obtidos nos trabalhos relacionados pois não foram feitas análises similares à esta.

5.5. Análise de erros do ChatGPT e comparação de desempenho com o Bard

Nesta seção foram selecionadas aleatoriamente oito questões pertencentes a cada categoria de erro para a análise do processo que levou o ChatGPT a errá-las, e em seguida, foi feita uma análise de performance do Bard, *chatbot* (que também utiliza modelos de linguagem de larga escala) de uso geral desenvolvido pelo Google, nestas mesmas questões, para fins de comparação com o desempenho do ChatGPT.

5.5.1. Categoria “Interpretação”

Para a primeira categoria foi selecionada a Questão 136 na modalidade sem alternativas da prova do ENEM 2017.

- Enunciado:

Os congestionamentos de trânsito constituem um problema que aflige, todos os dias, milhares de motoristas brasileiros. O gráfico ilustra a situação, representando, ao longo de um intervalo definido de tempo, a variação da velocidade de um veículo durante um congestionamento.

Descrição do gráfico: Gráfico de linha com o eixo horizontal definido pelo tempo (minuto), de zero a dez, e o eixo vertical definido pela velocidade. A variação da velocidade é descrita nos seguintes intervalos:

- de zero a um, a velocidade permanece constante acima de zero;
- de um a dois, a velocidade aumenta, atingindo o seu ponto mais alto;
- de dois a três, a velocidade diminui, mantendo-se acima da velocidade inicial;
- de três a quatro, a velocidade permanece constante, acima da velocidade inicial;
- de quatro a cinco, a velocidade aumenta, abaixo do ponto mais alto;
- de cinco a seis, a velocidade diminui até zero;
- de seis a oito, a velocidade permanece constante em zero;
- e de oito a dez, a velocidade volta a aumentar, abaixo do ponto mais alto.

Quantos minutos o veículo permaneceu imóvel ao longo do intervalo de tempo total analisado?

- Resposta do ChatGPT:

De acordo com a descrição do gráfico, podemos identificar que o veículo ficou imóvel durante os minutos compreendidos entre 5 e 6, além de ter se mantido com velocidade zero nos minutos entre 6 e 8. Portanto, o veículo permaneceu imóvel por um total de 3 minutos (do minuto 5 ao minuto 8).

- Avaliação da resposta do ChatGPT:

O motivo do erro ter sido cometido nesta questão foi uma interpretação errônea da seção “de cinco a seis, a velocidade diminui até zero”, provavelmente assumindo que a velocidade foi zero em todo esse intervalo. A interpretação correta do gráfico seria que o veículo permaneceu imóvel dos minutos 6 à 8, totalizando um tempo total de 2 minutos.

- Resposta do Bard:

De acordo com a descrição do gráfico, o veículo permaneceu imóvel durante o intervalo de tempo de seis a oito minutos, ou seja, 2 minutos. A velocidade do veículo é representada pela linha do gráfico. O veículo está em repouso nos instantes em que a velocidade é igual a zero. No gráfico, a velocidade é igual a zero de seis a oito minutos. Portanto, o veículo permaneceu imóvel durante esse intervalo de tempo. A resposta é 2 minutos.

Explicação detalhada (Fornecida pelo Bard)

A velocidade do veículo é representada pela linha do gráfico. O veículo está em repouso nos instantes em que a velocidade é igual a zero. No gráfico, a velocidade é igual a zero de seis a oito minutos. De zero a um minuto, a velocidade permanece constante acima de zero. De um a dois minutos, a velocidade aumenta, atingindo o seu ponto mais alto. De dois a três minutos, a velocidade diminui, mantendo-se acima da velocidade inicial. De três a quatro minutos, a velocidade permanece constante, acima da velocidade inicial. De quatro a cinco minutos, a velocidade aumenta, abaixo do ponto mais alto. De cinco a seis minutos, a velocidade diminui até zero. De seis a oito minutos, a velocidade permanece constante em zero. De oito a dez minutos, a velocidade volta a aumentar, abaixo do ponto mais alto. Portanto, o veículo permaneceu imóvel durante o intervalo de tempo de seis a oito minutos, ou seja, 2 minutos.

- Avaliação da resposta do Bard:

Vemos que o Bard acertou esta questão e, além disso, forneceu duas explicações: uma breve, na qual explicita somente as partes do gráfico nas quais a velocidade do veículo é igual a zero, e uma detalhada, onde percorre todo o gráfico, demonstrando sua interpretação completa da variação de velocidade.

5.5.2. Categoria “Lógica”

Para a segunda categoria foi selecionada a Questão 158 na modalidade com alternativas da prova do ENEM 2018.

- Enunciado:

Um artesão possui potes cilíndricos de tinta cujas medidas externas são 4 centímetros de diâmetro e 6 centímetros de altura. Ele pretende adquirir caixas organizadoras para armazenar seus potes de tinta, empilhados verticalmente com tampas voltadas para cima, de forma que as caixas possam ser fechadas. No mercado, existem cinco opções de caixas

organizadoras, com tampa, em formato de paralelepípedo reto retângulo, vendidas pelo mesmo preço, possuindo as seguintes dimensões internas:

Descrição do quadro:

Quadro com as informações:

Modelo, Comprimento, Largura e Altura, em centímetro:

Modelo I: Comprimento 8; Largura 8; Altura 40.

Modelo II: Comprimento 8; Largura 20; Altura 14.

Modelo III: Comprimento 18; Largura 5; Altura 35.

Modelo IV: Comprimento 20; Largura 12; Altura 12.

Modelo V: Comprimento 24; Largura 8; Altura 14.

Qual desses modelos o artesão deve adquirir para conseguir armazenar o maior número de potes por caixa?

- A. I
- B. II
- C. III
- D. IV
- E. V

- Resposta do ChatGPT:

Para empilhar os potes, o espaço necessário em cada caixa deve ser a altura dos potes mais uma folga para a tampa. Portanto, o espaço necessário em cada caixa é de $6+4=10$ centímetros de altura.

Calculando o volume interno de cada modelo de caixa, temos:

Modelo I: $8 \times 8 \times 40 = 2560$

Modelo II: $8 \times 20 \times 14 = 2240$

Modelo III: $18 \times 5 \times 35 = 3150$

Modelo IV: $20 \times 12 \times 12 = 2880$

Modelo V: $24 \times 8 \times 14 = 2688$

Assim, a caixa que comporta o maior número de potes é a que possui maior volume interno, que é o modelo III. Portanto, a resposta correta é a letra C.

- Avaliação da resposta do ChatGPT:

O primeiro erro cometido pelo ChatGPT é assumir que “o espaço necessário em cada caixa deve ser a altura dos potes mais uma folga para a tampa”, visto que, em nenhum momento, a espessura da tampa é mencionada. O segundo erro foi assumir que os potes não podem ser empilhados uns em cima dos outros. O terceiro erro foi calcular a altura necessária do espaço dentro da caixa somando o diâmetro e a altura dos potes. Finalmente, foram calculados os volumes das diferentes caixas apresentadas e escolhida a caixa com maior volume, sem considerar as dimensões dos potes de tinta (Resposta correta: D) IV).

- Resposta do Bard:

A resposta correta é (E), modelo V. Para calcular o número de potes que cabem em uma caixa, precisamos calcular o volume da caixa e dividir pelo volume de um pote. O volume de um pote cilíndrico é dado por:

$$V = \pi r^2 h$$

$$V = \pi (2 \text{ cm})^2 (6 \text{ cm})$$

$$V = 72\pi \text{ cm}^3$$

O volume de cada caixa é dado por:

$$V = lwh$$

$$V = 24 \text{ cm} \times 8 \text{ cm} \times 14 \text{ cm}$$

$$V = 2784 \text{ cm}^3$$

Portanto, o número de potes que cabem em cada caixa é dado por:

$$N = V_{\text{caixa}} / V_{\text{pote}}$$

$$N = 2784 \text{ cm}^3 / 72\pi \text{ cm}^3$$

$$N = 39,15 \text{ potes}$$

O modelo V tem o maior volume interno, portanto, é o modelo que consegue armazenar o maior número de potes. Na Tabela 1, observamos o número de potes que cabem em cada modelo.

Tabela 2. Medidas das caixas apresentadas e números de potes suportados.

Modelo	Comprimento (cm)	Largura (cm)	Altura (cm)	Número de potes
I	8	8	40	21,42
II	8	20	14	15,03
III	18	5	35	14,44
IV	20	12	12	12,19
V	24	8	14	39,15

Portanto, a resposta é (E), modelo V.

- Avaliação da resposta do Bard:

Vemos que nesta resposta, o Bard omite inicialmente os cálculos dos volumes das caixas, apesar de na tabela final mostrar todos. Além disso, comete um erro ao não considerar as limitações causadas pelas dimensões das caixas e dos potes, optando por considerar apenas o volume total das caixas dividido pelas dimensões de um pote, o que gera a resposta incorreta (Resposta correta: D. IV).

5.5.3. Categoria “Comparação de números”

Para a terceira categoria foi selecionada a Questão 164 na modalidade com alternativas da prova do ENEM 2017.

- Enunciado:

Uma pessoa ganhou uma pulseira formada por pérolas esféricas, na qual faltava uma das pérolas. Ela levou a jóia a um joalheiro que verificou que a medida do diâmetro dessas pérolas era 4 milímetros. Em seu estoque, as pérolas do mesmo tipo e formato, disponíveis para reposição, tinham diâmetros iguais a: 4,025 milímetros; 4,100 milímetros; 3,970 milímetros; 4,080 milímetros e 3,099 milímetros. O joalheiro então colocou na pulseira a pérola cujo diâmetro era o mais próximo do diâmetro das pérolas originais. A pérola colocada na pulseira pelo joalheiro tem diâmetro, em milímetro, igual a

- A. 3,099.
- B. 3,970.
- C. 4,025.
- D. 4,080.
- E. 4,100.

- Resposta do ChatGPT:

Para escolher a pérola mais adequada, precisamos calcular a diferença entre o diâmetro das pérolas disponíveis e o diâmetro das pérolas originais, e escolher aquela com menor diferença.

A diferença entre o diâmetro da primeira pérola disponível (4,025 mm) e o diâmetro original (4 mm) é de 0,025 mm. A diferença entre o diâmetro da segunda pérola disponível (4,100 mm) e o diâmetro original é de 0,100 mm. A diferença entre o diâmetro da terceira pérola disponível (3,970 mm) e o diâmetro original é de 0,030 mm. A diferença entre o diâmetro da quarta pérola disponível (4,080 mm) e o diâmetro original é de 0,080 mm. A

diferença entre o diâmetro da quinta pérola disponível (3,099 mm) e o diâmetro original é de 0,901 mm.

A menor diferença é encontrada na segunda pérola disponível, cujo diâmetro é de 4,100 mm. Portanto, a pérola colocada na pulseira pelo joalheiro tem diâmetro de 4,100 mm.

Resposta: E, 4,100.

- Avaliação da resposta do ChatGPT:

Nesta resposta vemos que o ChatGPT fez os cálculos de forma correta, mas na hora de comparar os valores finais encontrados, o programa escolhe erroneamente o valor “0,100 mm”, que não é o que representa a menor diferença de diâmetro, sendo o valor correto “0,025mm”, que é correspondente à letra C, 4,025 mm.

- Resposta do Bard:

A resposta correta é (C), 4,025 milímetros. A pérola que tem o diâmetro mais próximo do diâmetro das pérolas originais é a pérola com diâmetro 4,025 milímetros. A diferença entre esse diâmetro e o diâmetro das pérolas originais é de 0,025 milímetros, que é a menor diferença entre todos os diâmetros das pérolas disponíveis. As respostas (A), (B), (D) e (E) estão incorretas porque as pérolas correspondentes têm diâmetros que estão mais distantes do diâmetro das pérolas originais. Portanto, a resposta é (C), 4,025 milímetros.

Explicação detalhada (Fornecida pelo Bard)

O diâmetro das pérolas originais é 4 milímetros. A diferença entre o diâmetro de cada pérola disponível e o diâmetro das pérolas originais está descrita na Tabela 3.

Tabela 3. Diâmetros das pérolas em estoque e diferença de medida em relação à pérola original.

Pérola	Diâmetro	Diferença
3,099 milímetros	3,099 milímetros	0,901 milímetros
3,970 milímetros	3,970 milímetros	0,030 milímetros
4,025 milímetros	4,025 milímetros	0,025 milímetros
4,080 milímetros	4,080 milímetros	0,080 milímetros
4,100 milímetros	4,100 milímetros	0,100 milímetros

Como a diferença entre o diâmetro da pérola “4,025 milímetros” e o diâmetro das pérolas originais é a menor, a resposta é (C), 4,025 milímetros.

- Avaliação da resposta do Bard:

Vemos que o Bard acertou esta questão e repetiu o ocorrido da Seção 5.5.1, na qual fornece uma explicação simples e outra mais completa.

5.5.4. Categoria “Escolha incorreta de alternativa”

Para a quarta categoria foi selecionada a Questão 137 na modalidade com alternativas da prova do ENEM 2017.

- Enunciado:

Uma das bases mais utilizadas para representar um número é a base decimal. Entretanto, os computadores trabalham com números na base binária. Nessa base, qualquer número natural é representado usando apenas os algarismos 0 e 1. Por exemplo, as representações dos números 9 e 12, na base binária, “são abre parêntese um zero zero um fecha parêntese” e “abre parêntese um um zero zero fecha parêntese”, respectivamente. A

operação de adição, na base binária, segue um algoritmo similar ao utilizado na base decimal, como detalhado no quadro:

Descrição da imagem:

O quadro apresenta as seguintes informações: “a, b e a mais b”.

Quando a é igual a zero e b é igual a zero, a mais b é igual a zero.

Quando a é igual a zero e b é igual a um, a mais b é igual a um.

Quando a é igual a um e b é igual a zero, a mais b é igual a um.

Quando a é igual a um e b é igual a um, a mais b é igual a abre parêntese um zero fecha parêntese.

Por exemplo, na base binária, a soma dos números abre parêntese um zero fecha parêntese e abre parêntese um zero fecha parêntese é abre parêntese um zero zero fecha parêntese, como apresentado:

Descrição da imagem:

Adição: abre parêntese um zero fecha parêntese mais abre parêntese um zero fecha parêntese é igual a abre parêntese um zero zero fecha parêntese.

Considerando as informações do texto, o resultado da adição 9 mais 12 será representado, na base binária, por

- A. abre parêntese um zero um fecha parêntese.
- B. abre parêntese um um zero um fecha parêntese.
- C. abre parêntese um um um um fecha parêntese.
- D. abre parêntese um zero um zero um fecha parêntese.
- E. abre parêntese um um zero zero um fecha parêntese.

- Resposta do ChatGPT:

Para realizar a adição dos números 9 e 12 na base binária, precisamos escrevê-los em notação binária primeiro:

$$9 = 1001$$

$$12 = 1100$$

Agora, podemos usar o algoritmo de adição apresentado no quadro:

1 <-- *carry* (vai 1)

1001

1100

10101

Assim, o resultado da adição $9 + 12$ na base binária é 10101. Convertendo de volta para notação decimal, temos:

$$10101 = 1x2^4 + 0x2^3 + 1x2^2 + 0x2^1 + 1x2^0 = 16 + 0 + 4 + 0 + 1 = 21$$

Portanto, a resposta correta é a letra A, (0101) em notação binária, que equivale a 21 em notação decimal.

- Avaliação da resposta do ChatGPT:

O ChatGPT calcula e responde a questão corretamente, mas por alguma razão acaba escolhendo a alternativa errada, desconsiderando o valor da letra A e errando a conversão em notação binária do número 21 ao mesmo tempo.

- Resposta do Bard:

A resposta correta é (B), abre parêntese um um zero um fecha parêntese. Para realizar a adição de 9 e 12 na base binária, precisamos primeiro converter esses números para sua representação binária. 9 é igual a 1001 em binário e 12 é igual a 1100 em binário.

9 em binário = 1001

12 em binário = 1100

Agora, podemos realizar a adição usando o algoritmo apresentado no texto.

$$1001 + 1100$$

$$1 + 0 = 1$$

$$0 + 1 = 1$$

$$0 + 0 = 0$$

$$1001 + 1100 = 1101$$

Portanto, o resultado da adição 9 mais 12 será representado, na base binária, por abre parêntese um um zero um fecha parêntese. As respostas (A), (C), (D) e (E) estão incorretas porque representam resultados diferentes da adição de 9 e 12 na base binária.

Explicação detalhada (Fornecida pelo Bard)

A adição de 9 e 12 na base binária é realizada da seguinte forma:

$$1001 + 1100$$

$$1 + 0 = 1$$

$$0 + 1 = 1$$

$$0 + 0 = 0$$

$$1001 + 1100 = 1101$$

O resultado da adição é “abre parêntese um um zero um fecha parêntese”.

- Avaliação da resposta do Bard:

Nesta questão vemos que o Bard erra a soma binária, que deveria ter o resultado “10101”, mais especificamente na soma binária de “1 + 1”, que resulta no valor “10”, e que o Bard considerou como sendo “1” (Resposta correta: D. 10101).

5.5.5. Categoria “Elemento de cálculo ignorado”

Para a quinta categoria foi selecionada a Questão 150 na modalidade sem alternativas da prova do ENEM 2017.

- Enunciado:

Uma empresa especializada em conservação de piscinas utiliza um produto para tratamento da água cujas especificações técnicas sugerem que seja adicionado 1,5 mililitros desse produto para cada 1 000 litros de água da piscina. Essa empresa foi contratada para cuidar de uma piscina de base retangular, de profundidade constante igual a 1,7 metro, com largura e comprimento iguais a 3 metros e 5 metros, respectivamente. O nível da lâmina d’água dessa piscina é mantido a 50 centímetros da borda da piscina. A quantidade desse produto, em mililitro, que deve ser adicionada a essa piscina de modo a atender às suas especificações técnicas é

- Resposta do ChatGPT:

Primeiramente, vamos calcular o volume da piscina:

Volume = largura x comprimento x profundidade

Volume = 3 x 5 x 1,7

Volume = 25,5 metros cúbicos

Convertendo metros cúbicos para litros, temos:

1 metro cúbico = 1000 litros

25,5 metros cúbicos = 25500 litros

Portanto, para tratar a água dessa piscina, devemos adicionar:

1,5 mililitros / 1000 litros x 25500 litros = 38,25 mililitros

Logo, a quantidade de produto que deve ser adicionada é de 38,25 mililitros.

- Avaliação da resposta do ChatGPT:

O ChatGPT teria respondido a questão de forma correta, mas ao ignorar a variável “nível da lâmina d’água”, que corresponde a uma subtração de 0,5 metros da altura da água da piscina, o programa acaba chegando a uma resposta incorreta.

- Resposta do Bard:

A quantidade de água na piscina é dada pelo volume da piscina, que é igual a:

$$V = lwh$$

$$V = 3 \text{ m} \times 5 \text{ m} \times 1,7 \text{ m}$$

$$V = 25,5 \text{ m}^3$$

$$V = 25.500 \text{ litros}$$

Portanto, a quantidade de produto que deve ser adicionada à piscina é de:

$$Q = 1,5 \text{ ml} / 1\,000 \text{ litros} \times 25\,500 \text{ litros}$$

$$Q = 38,25 \text{ ml}$$

Resposta: 38,25 ml.

Explicação detalhada (Fornecida pelo Bard)

O volume da piscina é dado pelo produto da largura, do comprimento e da profundidade.

$$V = lwh$$

$$V = 3 \text{ m} \times 5 \text{ m} \times 1,7 \text{ m}$$

$$V = 25,5 \text{ m}^3$$

$$V = 25\,500 \text{ litros}$$

A quantidade de produto que deve ser adicionada à piscina é calculada dividindo-se o volume da piscina pelo volume de água tratado por miligrama do produto.

$$Q = 1,5 \text{ ml} / 1\,000 \text{ litros} \times 25\,500 \text{ litros}$$

$$Q = 38,25 \text{ ml}$$

Portanto, a quantidade de produto que deve ser adicionada à piscina é de 38,25 ml.

- Avaliação da resposta do Bard:

O Bard, assim como o ChatGPT, teria respondido a questão de forma correta, mas ao ignorar a variável “nível da lâmina d’água”, que corresponde a uma subtração de 0,5 metros da altura da água da piscina, acaba chegando a uma resposta incorreta.

5.5.6. Categoria “Atribuição incorreta de valores”

Para a sexta categoria foi selecionada a Questão 175 na modalidade sem alternativas da prova do ENEM 2017.

- Enunciado:

Pivô central é um sistema de irrigação muito usado na agricultura, em que uma área circular é projetada para receber uma estrutura suspensa. No centro dessa área, há uma tubulação vertical que transmite água através de um cano horizontal longo, apoiado em torres de sustentação, as quais giram, sobre rodas, em torno do centro do pivô, também chamado de base, conforme mostram as figuras. Cada torre move-se com velocidade constante.

Descrição das figuras: A primeira figura apresenta uma fotografia de um pivô central irrigando uma área de plantação de soja. A segunda figura apresenta três círculos concêntricos, tendo o centro representado pela base do pivô com a primeira região delimitada pelo círculo com raio traçado por T1, a segunda região delimitada pelo círculo com raio traçado por T2 e a terceira região delimitada pelo círculo com raio traçado por T3.

Um pivô de três torres (T1, T2 e T3) será instalado em uma fazenda, sendo que as distâncias entre torres consecutivas bem como da base à torre T1 são iguais a 50 metros. O fazendeiro pretende ajustar as velocidades das torres, de tal forma que o pivô efetue uma volta completa em 25 horas. Use 3 como aproximação para π . Para atingir seu objetivo, as velocidades das torres T1, T2 e T3 devem ser, em metro por hora, de

- Resposta do ChatGPT:

Para calcular as velocidades das torres, é necessário conhecer a circunferência da área irrigada pelo pivô central. Essa área é dada pelo círculo com raio traçado por T3 na segunda figura.

O raio desse círculo é igual a 3 vezes a distância entre a base e uma das torres consecutivas, ou seja, $3 \times 50 = 150$ metros.

Logo, a circunferência é igual a $2 \times \pi \times \text{raio} = 2 \times 3,14 \times 150 = 942$ metros.

Como o pivô deve dar uma volta completa em 25 horas, a velocidade angular das torres é dada por:

velocidade angular = $2 \times \pi / \text{tempo de uma volta completa}$

velocidade angular = $2 \times 3,14 / 25$

velocidade angular = 0,2512 rad/hora

A velocidade linear de uma torre é dada por:

velocidade linear = velocidade angular x raio da área irrigada

Para a torre T1, que está a 50 metros da base, temos:

velocidade linear de T1 = $0,2512 \times 50 = 12,56$ m/h

Para a torre T2, que está a 100 metros da base, temos:

velocidade linear de T2 = $0,2512 \times 100 = 25,12$ m/h

Para a torre T3, que está a 150 metros da base, temos:

velocidade linear de T3 = $0,2512 \times 150 = 37,68$ m/h

Portanto, as velocidades das torres T1, T2 e T3 devem ser, em metro por hora, de 12,56 m/h, 25,12 m/h e 37,68 m/h, respectivamente.

- Avaliação da resposta do ChatGPT:

Aqui vemos que a resposta dada pelo ChatGPT estaria correta, se não fosse pelo fato do programa ignorar a instrução da questão de utilizar o valor 3 como aproximação para π . A resposta correta, de acordo com a demanda da questão seria 12, 24 e 36 metros por hora, respectivamente.

- Resposta do Bard:

A resposta correta é (A), 5,28, 2,64 e 1,32. Para que o pivô efetue uma volta completa em 25 horas, a distância total percorrida por ele deve ser igual a $2 \times \pi \times \text{raio}$, onde *raio* é a distância entre a base e a torre T1.

Distância total = $2 \times \pi \times 50 = 314,15$ metros

Portanto, a velocidade total do pivô deve ser de:

Velocidade total = $314,15 / 25 = 12,56$ metros por hora

Como as torres T2 e T3 estão mais próximas da base, elas devem se mover com velocidades maiores do que a torre T1. A velocidade de cada torre deve ser proporcional ao seu raio.

Velocidade T1 = $12,56 / (50 / 50) = 5,28$ metros por hora

Velocidade T2 = $12,56 / (25 / 50) = 2,64$ metros por hora

Velocidade T3 = $12,56 / (12,5 / 50) = 1,32$ metros por hora

Portanto, as velocidades das torres T1, T2 e T3 devem ser de 5,28, 2,64 e 1,32 metros por hora, respectivamente.

Explicação detalhada (Fornecida pelo Bard)

Para que o pivô efetue uma volta completa em 25 horas, a distância total percorrida por ele deve ser igual a $2 \times \pi \times \text{raio}$, onde raio é a distância entre a base e a torre T1.

Distância total = $2 \times \pi \times 50 = 314,15$ metros

Portanto, a velocidade total do pivô deve ser de:

Velocidade total = $314,15 / 25 = 12,56$ metros por hora

Como as torres T2 e T3 estão mais próximas da base, elas devem se mover com velocidades maiores do que a torre T1. A velocidade de cada torre deve ser proporcional ao seu raio.

Velocidade T1 = $12,56 / (50 / 50) = 5,28$ metros por hora

Velocidade T2 = $12,56 / (25 / 50) = 2,64$ metros por hora

Velocidade T3 = $12,56 / (12,5 / 50) = 1,32$ metros por hora

Portanto, as velocidades das torres T1, T2 e T3 devem ser de 5,28, 2,64 e 1,32 metros por hora, respectivamente.

- Avaliação da resposta do Bard:

Nesta resposta vemos que o primeiro erro do Bard é não considerar o valor 3 como aproximação para π . O segundo erro foi interpretar que o comprimento total do pivô é de 50 metros. O terceiro erro foi interpretar que as torres T2 e T3 estão mais próximas da base que T1, quando na verdade é o contrário. Após isso, as três operações realizadas para T1, T2 e T3

não fazem sentido, pois além de resultarem em valores errados para cada expressão (5,28, 2,64 e 1,32 metros), o Bard assume distâncias incorretas para as torres (50 metros para T1, 25 metros para T2 e 12,5 metros para T3). A resposta correta dessa questão é “12, 24 e 36 metros” respectivamente.

5.5.7. Categoria “Cálculo incorreto de valores”

Para a sétima categoria foi selecionada a Questão 140 na modalidade com alternativas da prova do ENEM 2017.

- Enunciado:

Uma empresa construirá sua página na internet e espera atrair um público de aproximadamente um milhão de clientes. Para acessar essa página, será necessária uma senha com formato a ser definido pela empresa. Existem cinco opções de formato oferecidas pelo programador, descritas no quadro, em que “L” e “D” representam, respectivamente, letra maiúscula e dígito.

Descrição do quadro: O quadro apresenta cinco opções e um formato para cada uma delas.

Opção I: Formato: LDDDDD;

Opção II: Formato: DDDDDD;

Opção III: Formato: LLDDDD;

Opção IV: Formato: DDDDD;

Opção V: Formato: LLLDD.

As letras do alfabeto, entre as 26 possíveis, bem como os dígitos, entre os 10 possíveis, podem se repetir em qualquer das opções. A empresa quer escolher uma opção de formato cujo número de senhas distintas possíveis seja superior ao número esperado de clientes, mas que esse número não seja superior ao dobro do número esperado de clientes. A opção que mais se adequa às condições da empresa é

A. I

- B. II
- C. III
- D. IV
- E. V

- Resposta do ChatGPT:

Para determinar a opção que mais se adequa às condições da empresa, é preciso calcular o número de senhas distintas possíveis para cada opção de formato e compará-las com o número esperado de clientes.

Opção I: Formato LDDDDD

O primeiro caractere pode ser escolhido entre 26 letras. Os demais cinco caracteres podem ser escolhidos entre os 10 dígitos. Portanto, o número de senhas distintas possíveis é $26 \times 10^5 = 26.000.000$.

Opção II: Formato DDDDDD

Todos os seis caracteres podem ser escolhidos entre os 10 dígitos. Portanto, o número de senhas distintas possíveis é $10^6 = 1.000.000$.

Opção III: Formato LLDDDD

Os dois primeiros caracteres podem ser escolhidos entre as 26 letras. Os demais cinco caracteres podem ser escolhidos entre os 10 dígitos. Portanto, o número de senhas distintas possíveis é $26^2 \times 10^5 = 67.600.000$.

Opção IV: Formato DDDDD

Todos os cinco caracteres podem ser escolhidos entre os 10 dígitos. Portanto, o número de senhas distintas possíveis é $10^5 = 100.000$.

Opção V: Formato LLLDD

Os três primeiros caracteres podem ser escolhidos entre as 26 letras. Os dois últimos caracteres podem ser escolhidos entre os 10 dígitos. Portanto, o número de senhas distintas possíveis é $26^3 \times 10^2 = 175.760.000$.

O número esperado de clientes é de um milhão. O dobro desse número é dois milhões.

Assim, a opção que mais se adequa às condições da empresa é a opção III, cujo número de senhas distintas possíveis é 67.600.000, que é superior a um milhão e inferior a dois milhões. Portanto, a resposta é a letra C.

- Avaliação da resposta do ChatGPT:

Nesta questão o ChatGPT errou a maior parte dos cálculos e no final errou a comparação entre o valor obtido e o intervalo proposto pela questão. O primeiro cálculo incorreto foi o da “Opção I: Formato LDDDDD”, no qual calcula que “ $26 \times 10^5 = 26$ milhões”, sendo 2,6 milhões o valor correto. O segundo cálculo incorreto foi o da “Opção III: Formato LLDDDD”, no qual assume que o tamanho da senha é igual a 7 (6 é o tamanho correto) e calcula que “ $26^2 \times 10^5 = 67,6$ milhões”, sendo “ $26^2 \times 10^4 = 6,76$ milhões” o cálculo correto. O terceiro cálculo incorreto foi o da “Opção V: Formato LLLDD”, no qual calcula que “ $26^3 \times 10^2 = 175,76$ milhões”, sendo 1,7576 milhões o valor correto. O quarto e último erro foi cometido na conclusão do raciocínio, em que determina que o valor 67,6 milhões está contido no intervalo entre 1 e 2 milhões, quando na verdade é um valor maior.

- Resposta do Bard:

A resposta correta é (C), III. O número de senhas distintas possíveis para cada opção de formato é calculado da seguinte forma:

Opção I: $26 \times 10 = 2600$

Opção II: $26 \times 10 = 2600$

Opção III: $26 \times 26 \times 10 = 67600$

Opção IV: $26 \times 10 = 2600$

Opção V: $26 \times 26 \times 10 = 67600$

Como a empresa espera atrair um público de aproximadamente um milhão de clientes, a opção de formato que mais se adequa às condições da empresa é a III, pois o número de senhas distintas possíveis é superior a um milhão e não é superior a dois milhões.

Explicação detalhada (Fornecida pelo Bard)

O número de senhas distintas possíveis para cada opção de formato é calculado da seguinte forma:

Opção I:

Número de letras possíveis = 26

Número de dígitos possíveis = 10

Número de senhas distintas possíveis = $26 * 10 = 2600$

Opção II:

Número de letras possíveis = 26

Número de dígitos possíveis = 10

Número de senhas distintas possíveis = $26 * 10 = 2600$

Opção III:

Número de letras possíveis = 26

Número de dígitos possíveis = 10

Número de senhas distintas possíveis = $26 * 26 * 10 = 67600$

Opção IV:

Número de letras possíveis = 26

Número de dígitos possíveis = 10

Número de senhas distintas possíveis = $26 * 10 = 2600$

Opção V:

Número de letras possíveis = 26

Número de dígitos possíveis = 10

Número de senhas distintas possíveis = $26 * 26 * 10 = 67600$

Portanto, como a empresa espera atrair um público de aproximadamente um milhão de clientes, a opção de formato que mais se adequa às condições da empresa é a III, pois o número de senhas distintas possíveis é superior a um milhão e não é superior a dois milhões.

- Avaliação da resposta do Bard:

Na resposta do Bard vemos que as fórmulas utilizadas para calcular o número de senhas possíveis para cada opção não fazem sentido, pois não consideram o número correto de posições que cada opção tem (seis para as opções 1, 2, e 3, e cinco para as opções 4 e 5), bem como a disponibilidade de símbolos (números de 0 a 9 ou letras do alfabeto). Além disso, a resposta final também está incorreta, visto que esta considera o número 67600 como sendo um valor entre 1 e 2 milhões. A resposta correta é a letra E. V.

5.5.8. Categoria “Estruturação das equações”

Para a oitava categoria foi selecionada a Questão 145 na modalidade sem alternativas da prova do ENEM 2017.

- Enunciado:

Para realizar a viagem dos sonhos, uma pessoa precisava fazer um empréstimo no valor de R\$5.000,00. Para pagar as prestações, dispõe de, no máximo, R\$400,00 mensais. Para esse valor de empréstimo, o valor da prestação (P) é calculado em função do número de prestações (n) segundo a fórmula P é igual a fração: no numerador 5 000 vezes abre parêntese 1,013 elevado a n fecha parêntese vezes 0,013, no denominador abre parêntese 1,013 elevado a n fecha parêntese menos 1. Se necessário, utilize 0,005 como aproximação para $\log 1,013$; 2,602 como aproximação para $\log 400$; 2,525 como aproximação para $\log 335$. De acordo com a fórmula dada, o menor número de parcelas cujos valores não comprometem o limite definido pela pessoa é

- Resposta do ChatGPT:

Substituindo os valores dados na fórmula, temos:

$$400 = [5000 \times (1,013^n) \times (0,013)] / [1,013^n - 1]$$

Multiplicando ambos os lados da equação por $1,013^n - 1$ e dividindo por 0,013, obtemos:

$$30\,769,23 = 5000 \times (1,013^n)$$

Dividindo ambos os lados da equação por 5000, temos:

$$1,013^n = 6,15384615$$

Tomando logaritmo natural em ambos os lados, temos:

$$\ln 1,013^n = \ln 6,15384615$$

$$n \ln 1,013 = \ln 6,15384615$$

$$n = \ln 6,15384615 / \ln 1,013$$

$$n \approx 29,3$$

O menor número de parcelas é 30. Verificando se o valor da prestação é menor que R\$400,00 para 30 parcelas:

$$P = [5000 \times (1,013^{30}) \times (0,013)] / [1,013^{30} - 1]$$

$$P \approx \text{R\$ } 176,39$$

Portanto, a menor quantidade de parcelas que não compromete o limite de R\$400,00 é 30.

- Avaliação da resposta do ChatGPT:

Neste raciocínio o ChatGPT gera uma expressão resultante incorreta ao multiplicar e dividir ambos os lados da equação “ $400 = [5000(1,013^n)(0,013)] / [1,013^n - 1]$ ” por “ $1,013^n - 1$ ” e “0,013”, respectivamente, pois a expressão resultante correta seria “ $30\,769,23 =$

$[5000(1,013^n)] / [1,013^n - 1]$ ”, ao invés de “ $30\,769,23 = 5000(1,013^n)$ ”. Por causa disso, o ChatGPT acaba chegando na resposta incorreta de 30 parcelas (Resposta correta: 16 parcelas).

- Resposta do Bard:

A resposta correta é (A), 12 parcelas. Para o valor da prestação não comprometer o limite definido pela pessoa, o valor da prestação deve ser igual ou inferior a R\$400,00.

$$P \leq 400$$

$$5000 \times (1,013)^n \times 0,013 \leq 400$$

$$(1,013)^n \leq 400 / 5000 \times 1000 / 13$$

$$(1,013)^n \leq 23,077$$

$$\log (1,013)^n \leq \log 23,077$$

$$n \log 1,013 \leq \log 23,077$$

$$n \leq \log 23,077 / \log 1,013$$

$$n \leq 12,80 \quad n \leq 12$$

Portanto, o menor número de parcelas cujos valores não comprometem o limite definido pela pessoa é 12 parcelas.

Explicação detalhada (Fornecida pelo Bard)

Para o valor da prestação não comprometer o limite definido pela pessoa, o valor da prestação deve ser igual ou inferior a R\$400,00.

$$P \leq 400$$

$$5000 \times (1,013)^n \times 0,013 \leq 400$$

$$(1,013)^n \leq 400 / 5000 \times 1000 / 13$$

$$(1,013)^n \leq 23,077$$

$$\log (1,013)^n \leq \log 23,077$$

$$n \log 1,013 \leq \log 23,077$$

$$n \leq \log 23,077 / \log 1,013$$

$$n \leq 12,80$$

$$n \leq 12$$

Como o número de parcelas deve ser um número inteiro, o menor número de parcelas é 12 parcelas.

- Avaliação da resposta do Bard:

Nesta resposta o Bard chega ao valor incorreto pois ignora o denominador da fórmula apresentada na questão, além disso, o programa apresenta a resposta como uma alternativa (letra A), que não existe, pois a questão foi dada como aberta. A resposta correta é “16 parcelas”.

5.5.9. Comparação de desempenho entre o ChatGPT e o Bard

Na Tabela 4, podemos ver que o desempenho do Bard não foi muito diferente do ChatGPT, acertando apenas duas questões das oito apresentadas. Vemos também que nas seis questões que o Bard errou, só duas apresentaram o mesmo tipo de erro do ChatGPT.

Tabela 4. Listagem e categorização dos erros cometidos pelo ChatGPT e Bard nas questões selecionadas para análise.

Questão	Que tipo de erro o ChatGPT cometeu?	O Bard errou?	Se errou, que tipo de erro cometeu?
1 - Questão 136 na modalidade sem alternativas - ENEM 2017	Interpretação	Não	-
2 - Questão 158 na modalidade com alternativas - ENEM 2018	Lógica	Sim	Elemento de cálculo ignorado
3 - Questão 164 na modalidade com	Comparação de números	Não	-

alternativas - ENEM 2017			
4 - Questão 137 na modalidade com alternativas - ENEM 2017.	Escolha incorreta de alternativa	Sim	Cálculo incorreto de valores
5 - Questão 150 na modalidade sem alternativas - ENEM 2017.	Elemento de cálculo ignorado	Sim	Elemento de cálculo ignorado
6 - Questão 175 na modalidade sem alternativas - ENEM 2017.	Atribuição incorreta de valores	Sim	Elemento de cálculo ignorado / Interpretação / Cálculo incorreto de valores
7 - Questão 140 na modalidade com alternativas - ENEM 2017.	Cálculo incorreto de valores	Sim	Lógica / Comparação de números
8 - Questão 145 na modalidade sem alternativas - ENEM 2017.	Estruturação das equações	Sim	Estruturação das equações

No geral, ambas as ferramentas cometem erros nos cálculos e ignoram elementos nestes, ou mesmo na estruturação das equações: apresentam problemas de lógica e de comparação de números; de entendimento sobre a questão; e até erros primários, como na escolha incorreta de alternativa.

CONSIDERAÇÕES FINAIS

De acordo com os resultados obtidos neste trabalho, foi evidenciado que o ChatGPT, para uma avaliação com questões de matemática, no ENEM, de 2017 a 2023, apresentou dificuldade para responder corretamente às questões. Foi observada uma taxa de acerto que varia entre 35,5% e 55,5% para a modalidade com alternativas, com uma tendência de crescimento de 2017 a 2019 (35,5%, 43,1% e 55,5%, respectivamente), que posteriormente voltou a cair de 2019 a 2021 (55,5%, 45,4% e 43,1%, respectivamente), crescendo novamente em 2022 (47,7%).

Para a modalidade sem alternativas foi observada uma taxa de acerto que varia entre 22,2% à 47,7%, com uma tendência de crescimento de 2017 a 2020 (22,2%, 27,2%, 28,8% e 47,7%, , respectivamente), sendo destacado o ano de 2020 (47,7%) com a taxa de acerto maior do que as dos anos restantes, e maior do que a da modalidade com alternativas do mesmo ano. A tendência de queda retornou de 2020 a 2021 (47,7% e 29,5%, respectivamente), voltando a crescer novamente em 2022 (34%).

Os resultados de acerto e erro em relação às unidades temáticas das questões estavam correlacionadas, em sua maior parte, com as unidades temáticas do total de questões, com exceção de um aumento notável na frequência de questões da Unidade 2 (Utilizar o conhecimento geométrico para realizar a leitura e a representação da realidade e agir sobre ela) para os erros (30/46 questões incorretas para a modalidade com alternativas, 41/46 questões incorretas para a modalidade sem alternativas), o que poderia indicar um maior nível de dificuldade nesses tipos de questões.

Sobre a presença de representações visuais, também não houve uma diferença considerável entre os resultados de acerto e erro em relação à representação visual do total de questões, com exceção de um aumento na frequência de questões com “formas geométricas descritas” nos resultados dos erros (26/35 questões para a modalidade com alternativas e 32/35, para a modalidade sem alternativas), o que pode significar que o ChatGPT obteve uma maior dificuldade em questões com formas geométricas descritas, provavelmente devido ao alto nível de abstração requerido para interpretar corretamente este tipo de figura.

Em relação às categorias de erro, foi notada uma dominância das categorias de “Interpretação”, ocasionados por problemas no entendimento dos requerimentos dos problemas propostos, e “Lógica”, onde a resposta gerada pelo ChatGPT não segue uma linha de raciocínio consistente e que não possui sentido, além de uma frequência

consideravelmente alta da categoria de “Cálculo incorreto de valores”. Vale notar que nos erros da modalidade sem alternativa, houve um aumento considerável no número de questões de “Interpretação” (42 questões na modalidade com alternativas para 62 na modalidade sem alternativa) e este aumento provavelmente se dá pela falta de alternativas nesta modalidade, que muitas vezes servem de modelo e “guia” para o ChatGPT identificar o objetivo a ser alcançado e que tipo de saída gerar. A frequência desses tipos de erros e o tamanho das respostas dadas (corretas ou não) indicam que o programa tenta emular ações semelhantes relacionadas ao tema a partir da amalgamação do que aprendeu nas bases de dados utilizadas em seu treinamento, com o objetivo de parecer coerente sem necessariamente considerar a qualidade do conteúdo gerado.

Também vale notar que os dados de performance obtidos neste trabalho apresentaram valores relativamente constantes e com poucos *outliers* em termo dos acertos e erros, e que reforçam a suposição de que sites que possuem resoluções para estas provas (como o site utilizado neste trabalho para checagem das soluções¹⁴) não foram usados em seu treinamento, dando assim uma maior credibilidade para este trabalho em termos de avaliação de performance real do do ChatGPT nessa área.

¹⁴ https://www.curso-objetivo.br/vestibular/resolucao_comentada/enem.asp

TRABALHOS FUTUROS

No futuro, outros *chatbots* e ferramentas de inteligência artificial para uso geral podem ser analisadas da mesma forma como foi feito neste trabalho. O ChatGPT-4, por exemplo, é a versão paga mais recente do ChatGPT, e possui uma base de dados mais atualizada e um modelo mais complexo que o utilizado aqui, ou mesmo o Bard, utilizado na Seção 5.5 deste documento, pode ser avaliado de forma mais detalhada.

Poderiam ser feitas ainda análises mais profundas por pesquisadores e especialistas da área de Matemática e correlatas sobre o desempenho do ChatGPT com problemas desse tipo, havendo assim uma maior concentração de estudos sobre as limitações que o programa possui, possibilitando identificar fragilidades existentes em níveis mais baixos das camadas do software, além de facilitar a criação de modelos de linguagem e bases de treinamento mais adequadas para lidar com questões e conversas que envolvam raciocínio matemático.

REFERÊNCIAS

BENGIO, Y. Neural net language models. Scholarpedia, v. 3, n. 1, p. 3881, 2008. Disponível em: <https://www.scholarpedia.org/article/Neural_net_language_models>.

BOMMARITO, M. J.; KATZ, D. M. GPT Takes the Bar Exam. Disponível em: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4314839>.

CHOMSKY, N. Linguistics and Cognitive Science: Problems and Mysteries. Disponível em: <<https://philpapers.org/rec/CHOLAC>>.

GAGNIUC, P. A. Markov Chains: From Theory to Implementation and Experimentation. Disponível em: <<https://www.wiley.com/en-it/Markov+Chains:+From+Theory+to+Implementation+and+Experimentation-p-9781119387558>>.

HE, C. Transformer in CV. Disponível em: <<https://towardsdatascience.com/transformer-in-cv-bbdb58bf335e>>.

HUYEN, C. Evaluation Metrics for Language Modeling. Disponível em: <<https://thegradient.pub/understanding-evaluation-metrics-for-language-models>>.

JURAFSKY, D.; MARTIN, J. H. Speech and Language Processing. Disponível em: <<https://web.stanford.edu/~jurafsky/slp3>>.

KAPARTHY, A. The Unreasonable Effectiveness of Recurrent Neural Networks. Disponível em: <<https://web.archive.org/web/20201101215448/http://karpathy.github.io/2015/05/21/rnn-effectiveness>>.

MANNING, C. D. Human Language Understanding & Reasoning. Disponível em: <<https://www.amacad.org/publication/human-language-understanding-reasoning>>.

NUNES, D. et al. Evaluating GPT-3.5 and GPT-4 Models on Brazilian University Admission Exams. [s.l: s.n.]. Disponível em: <<https://arxiv.org/pdf/2303.17003.pdf>>.

RADFORD, A. et al. Improving Language Understanding by Generative Pre-Training. [s.l: s.n.]. Disponível em: <https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf>.

RADFORD, A. et al. Language Models are Unsupervised Multitask Learners. [s.l: s.n.]. Disponível em: <https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf>.

TERWIESCH, C. Would Chat GPT Get a Wharton MBA? A Prediction Based on Its Performance in the Operations Management Course. [s.l: s.n.]. Disponível em: <<https://mackinstitute.wharton.upenn.edu/wp-content/uploads/2023/01/Christian-Terwiesch-Chat-GTP-1.24.pdf>>.

VASWANI, A. et al. Attention Is All You Need. Disponível em: <<https://arxiv.org/abs/1706.03762>>.

ZHU, Y. et al. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. [s.l: s.n.]. Disponível em: <https://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Zhu_Aligning_Books_and_ICCV_2015_paper.pdf>.