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

Lucas Brasileiro Raposo

AVALIAÇÃO DE LLMS NA RESOLUÇÃO DE QUESTÕES DO ENEM

CAMPINA GRANDE - PB
2024

Lucas Brasileiro Raposo

AVALIAÇÃO DE LLMS NA RESOLUÇÃO DE QUESTÕES DO ENEM

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

Orientador : Professor Dr. Fábio Jorge Almeida Morais

CAMPINA GRANDE - PB
2024

Lucas Brasileiro Raposo

AVALIAÇÃO DE LLMS NA RESOLUÇÃO DE QUESTÕES DO ENEM

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

BANCA EXAMINADORA:

Fábio Jorge Almeida Morais
Orientador – UASC/CEEI/UFCG

Carlos Eduardo Santos Pires
Examinador – UASC/CEEI/UFCG

Francisco Vilar Brasileiro
Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 15 de Maio de 2024.

CAMPINA GRANDE - PB

RESUMO

Grandes Modelos de Linguagem (LLMs do inglês, Large Language Models) surgiram como uma quebra de paradigma no uso da Inteligência Artificial (IA) e são amplamente usados em diferentes áreas. Um dos maiores responsáveis pela popularização desse termo é o ChatGPT, desenvolvido pela OpenAI. A partir da ascensão desse, outras empresas, como a Meta e a Google, desenvolveram seus próprios modelos como alternativas ao GPT. Essas ferramentas se apresentam como solução de problemas nos mais variados contextos. Entretanto, pouca atenção é voltada para medir a capacidade de corretude e eficiência de suas respostas. Somado a isso, a maioria dos estudos neste âmbito, se prendem ao contexto da língua inglesa, sem que os modelos sejam efetivamente testados em cenários globalizados. Logo, este estudo propõe submeter os sistemas da Meta, da OpenAI e da Google à avaliações de múltipla escolha objetivas sobre conteúdos de nível médio, por meio das provas do Exame Nacional do Ensino Médio (ENEM). Após colher as respostas dos modelos, análises foram realizadas, comparando desempenho, entre cada uma delas e com médias dos alunos brasileiros, considerando quantidade de acertos por prova. Então, surpreendentemente, este trabalho mostrou que todos os três modelos desempenharam melhor em áreas mais "subjetivas" que em áreas objetivas, indo contra o senso comum.

LLMs’ assessment in question resolution from ENEM

ABSTRACT

Large Language Models (LLMs) have emerged as a paradigm shift in the use of Artificial Intelligence (AI) and are widely employed across various domains. One of the main drivers behind the popularization of this term is ChatGPT, developed by OpenAI. With the rise of ChatGPT, other companies such as Meta and Google have developed their own models as alternatives to GPT. These tools offer solutions to a wide range of problems in diverse contexts. However, little attention is paid to measuring the accuracy and efficiency of their responses. Additionally, most studies in this field are limited to the English language context, without the models being effectively tested in globalized scenarios. Therefore, this study proposes to subject the systems of Meta, OpenAI, and Google to objective multiple-choice evaluations on medium-level content, through the exams of the National High School Examination (ENEM, from Portuguese: Exame Nacional do Ensino Médio). After collecting the models' responses, analyses were conducted, comparing performance among them and with the averages of Brazilian students, considering the quantity of correct answers per test. Surprisingly, this work demonstrated that all three models performed better in more "subjective" areas than in objective ones, contrary to common belief.

Avaliação de LLMs na resolução de questões do ENEM

Lucas Brasileiro Raposo (Aluno)
Departamento de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba - Brasil

Fábio Jorge A. Morais(Orientador)
Departamento de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba - Brasil

RESUMO

Grandes Modelos de Linguagem (LLMs do inglês, *Large Language Models*) surgiram como uma quebra de paradigma no uso da Inteligência Artificial (IA) e são amplamente usados em diferentes áreas. Um dos maiores responsáveis pela popularização desse termo é o ChatGPT, desenvolvido pela OpenAI. A partir da ascensão desse, outras empresas, como a Meta e a Google, desenvolveram seus próprios modelos como alternativas ao GPT. Essas ferramentas se apresentam como solução de problemas nos mais variados contextos. Entretanto, pouca atenção é voltada para medir a capacidade de corretude e eficiência de suas respostas. Somado a isso, a maioria dos estudos neste âmbito, se prendem ao contexto da língua inglesa, sem que os modelos sejam efetivamente testados em cenários globalizados. Logo, este estudo propõe submeter os sistemas da Meta, da OpenAI e da Google à avaliações de múltipla escolha objetivas sobre conteúdos de nível médio, por meio das provas do Exame Nacional do Ensino Médio (ENEM). Após colher as respostas dos modelos, análises foram realizadas, comparando desempenho, entre cada uma delas e com médias dos alunos brasileiros, considerando quantidade de acertos por prova. Então, surpreendentemente, este trabalho mostrou que todos os três modelos desempenharam melhor em áreas mais "subjetivas" que em áreas objetivas, indo contra o senso comum.

Keywords

LLMs, Grandes Modelos de Linguagem, ENEM, ChatGPT, GEMINI, Llama.

1. INTRODUÇÃO

Nos dias de hoje, presenciamos diversos avanços tecnológicos em nossa sociedade, mas, em particular, um dos maiores impactos para a área da computação foi a criação e popularização dos Grandes Modelos de Linguagem (LLMs do inglês, *Large Language Models*)[1]. Esses modelos, fundamentados em inteligência artificial, redes neurais, aprendizagem de máquina e Processamento de Linguagem Natural (PLN), revolucionaram a forma como humanos interagem com a tecnologia e como máquinas processam e compreendem informações passadas em linguagem humana[2].

Os LLMs são treinados com uma ampla base de dados, que incluem textos da web, livros, artigos acadêmicos, conversas, entre outros[3]. Essa diversidade de fontes permite que os modelos capturem a complexidade e a riqueza da linguagem humana, aprendendo padrões sintáticos, semânticos e contextuais. Durante o treinamento, os modelos ajustam os pesos das conexões em suas redes neurais para maximizar sua capacidade de prever palavras ou sequências de palavras com base no contexto fornecido. Essa abordagem permite que os modelos internalizem conhecimento e desenvolvam uma compreensão profunda das estruturas linguísticas presentes nos dados de entrada[4] .

Ao capacitar máquinas a compreender e gerar texto de maneira semelhante à humana, os LLMs abriram portas para uma série de aplicações inovadoras em diversas áreas, como tradução automática, geração de conteúdo, assistência virtual, geração de códigos, geração de perguntas e respostas e muito mais. Atualmente, a solução de LLM mais conhecida é o ChatGPT [5] [6], desenvolvido pela OpenAI, que consiste em um modelo conversacional baseado em *Generative pretrained transformer* (GPT). A partir do sucesso do ChatGPT, outras empresas também apresentaram soluções semelhantes, como a Meta[7] e a Google[8], que desenvolveram seus próprios modelos, o Llama[9] (e Llama 2), e o GEMINI[10] (anteriormente conhecido como Google Bard), como alternativas ao GPT [11][12].

Contudo, é importante reconhecer que esses modelos não são infalíveis sendo, portanto, constantemente avaliados quanto às suas capacidades em uma variedade de tarefas, para uma compreensão mais abrangente de suas limitações. As próprias empresas responsáveis pelo desenvolvimento desses, adotam estratégias de avaliação, realizando testes de desempenho de seus produtos, e os resultados desses testes são utilizados como parâmetros de qualidade,como mostrado por OpenAI [13][14].

Entretanto, a maioria dessas avaliações são limitadas à língua inglesa e representam realidades muito divergentes da grande maioria de usuários brasileiros, e em alguns casos, não são capazes de avaliar a precisão de suas respostas em relação às respostas humanas. Desta forma, é fundamental adotar uma abordagem cautelosa e crítica ao utilizar esses modelos de processamento de linguagem natural, levando em consideração a qualidade e a confiabilidade das respostas geradas.

Com base nesse quadro, surge o questionamento de como seria o desempenho de modelos de LLMs, mais especificamente, o Llama 2, o GPT e o GEMINI, em um contexto avaliativo, onde seria exigido deles responder questões objetivas em português.

Nesse contexto, surgiu a ideia de se utilizar o Exame Nacional do Ensino Médio (ENEM)[15]. Cujo objetivo é avaliar o desempenho dos estudantes do ensino médio, além de ser utilizado como critério de seleção para ingresso no ensino superior e fornecer estatísticas sobre a educação do país. O exame é composto por 185 questões, no entanto, devido a escolha de Língua Estrangeira[15] são consideradas 180 questões. Essas questões são distribuídas ao longo de dois dias de provas, com 90 itens por dia. Cada prova é composta por dois blocos de 45 questões, abrangendo quatro áreas distintas: Ciências Humanas, Linguagens, Ciências da Natureza e Matemática, além de uma redação[15].

Diante desse cenário, este trabalho propôs submeter os modelos de linguagens de larga escala, frente ao ENEM de forma a medir seu desempenho quando exposto a questões de múltipla escolha, que simulam situações problemas de alto caráter interpretativo e multidisciplinar. Esta atividade envolveu uma análise minuciosa dos resultados obtidos pelas ferramentas em comparação com o desempenho de alunos brasileiros que fizeram o exame, considerando a quantidade de acertos por prova e pelas macroáreas de conhecimento. O objetivo dessa avaliação foi aprimorar a compreensão do potencial e das limitações dos modelos como ferramentas educacionais, além de explorar seu desempenho em um contexto de avaliação acadêmica.

2. FUNDAMENTAÇÃO TEÓRICA

A presente seção concentra a fundamentação teórica, apresentando os conceitos necessários para a melhor compreensão desta análise.

2.1 Processamento de Linguagem Natural

Processamento de Linguagem Natural é uma subárea da IA e da linguística computacional que, como o nome indica, lida com a interação entre computadores e linguagem humana. Segundo Jurafsky e Martin [16], PLN abrange o desenvolvimento de algoritmos e modelos computacionais para permitir que computadores entendam, interpretem e gerem linguagem humana de maneira semelhante aos seres humanos. Isso envolve uma série de tarefas, como análise morfológica, sintática, semântica e pragmática de texto, além de tarefas mais avançadas, como tradução automática, extração de informações, geração e sumarização de textos.

2.2 LLMs e *Transformers*

Os LLMs são modelos de inteligência artificial capazes de entender e gerar texto em linguagem natural em uma escala massiva. Eles são construídos com arquiteturas de redes neurais profundas e treinados em grandes conjuntos de dados textuais. Esses modelos têm a capacidade de aprender padrões linguísticos complexos e gerar texto coerente e semelhante ao humano em uma variedade de tarefas[2].

Transformers, por sua vez, são uma arquitetura de rede neural que se destacou no campo do PLN devido à sua capacidade de lidar com sequências de entrada de comprimento variável e capturar dependências de longo alcance, de forma eficiente e escalável. Eles são compostos por múltiplas camadas de processamento que capturam informações importantes em textos e as relacionam entre si, realizando tudo isso de forma paralela. Os *transformers* revolucionaram o campo do PLN e serviram como base para muitos LLMs de sucesso, como o *Bidirectional Encoder Representations from Transformers* (BERT) e o GPT[4][17].

2.3 Tokens, *prompt* e engenharia de *prompt*

No contexto de NLP e IA, token é um conjunto de caracteres de um texto que possuem valor semântico para o modelo[18]. Esses tokens podem ser palavras, subpalavras, caracteres ou até mesmo símbolos, dependendo do método de tokenização utilizado. O uso de tokens para LLMs permite limitar tamanhos de entradas e saídas do modelos.

Um *prompt* é um texto em linguagem natural que solicita que um modelo de linguagem execute uma tarefa específica. Ele serve como ponto de partida ou contexto para o modelo produzir texto, responder perguntas ou completar tarefas. O prompt pode variar de uma simples pergunta a um parágrafo mais complexo que fornece contexto para a tarefa em questão [19][20].

A engenharia de *prompt* é o processo de design e refinamento de prompts para guiar eficazmente um modelo de linguagem na geração de saídas desejadas. Isso envolve a elaboração de perguntas, declarações ou instruções de forma a otimizar o desempenho da IA, garantindo que ela entenda e responda com precisão à entrada [20].

3. METODOLOGIA

Para desenvolvimento deste estudo, as atividades foram divididas em 3 etapas: Coleta e pré-processamento das questões e dos gabaritos das edições do ENEM entre os anos de 2011 e 2023; Submissão das questões aos modelos e registro das respostas coletadas; e análise dos resultados obtidos. A **Figura 1** ilustra o fluxo de desenvolvimento deste estudo.

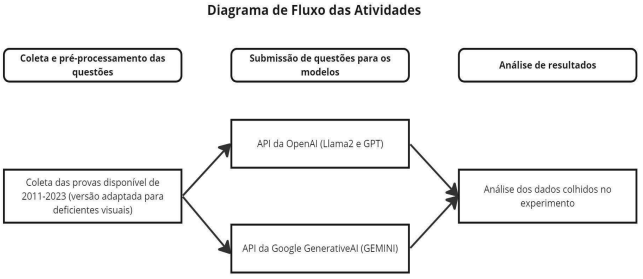

Figura 1: Representação das atividades realizadas.

3.1 Coleta e pré-processamento de dados

A primeira etapa do experimento consistiu na coleta de questões e gabaritos das provas do ENEM a partir da plataforma online do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP)[21] em formato PDF e convertidos em arquivos de textos em formato tabular, descrito na **Tabela 1**.

Como LLMs se baseiam, originalmente, em entradas no formato textual, foram consideradas provas destinadas a deficientes visuais, as quais descrevem imagens e elementos gráficos por meio de texto. Devido a isso, neste estudo foram consideradas as edições de 2011 a 2013, de 2015 a 2020, de 2022 e de 2023 do ENEM, com todas suas 185 questões, sem considerar apenas uma Língua Estrangeira. Apenas os anos de 2014 e 2021 não foram considerados por questões de disponibilidade da prova em versão para deficientes visuais e por problemas na conversão da prova para formato texto, respectivamente.

3.2 Submissão dos exames

Para essa etapa do experimento, foram desenvolvidos algoritmos que buscassem nas tabelas, as questões de uma determinada prova, e as submetessem, uma a uma, por meio de requisições para as respectivas APIs dos modelos. Para os modelos GPT3.5[22] e Llama2[23], ambos foram disponibilizados através API da OpenAI, apenas mudando o parâmetro de qual modelo seria selecionado no momento de submissão das questões. Enquanto o modelo GEMINI , possui sua própria API[24], e por meio dela foram submetidas as questões ao produto da Google.

Foram estabelecidos dois grupos de resultados, estes foram separados em Experimento com temperatura padrão e Experimento com temperatura 0. Haja vista que, o parâmetro de temperatura é responsável por controlar a aleatoriedade das previsões geradas pelo modelo durante a geração de texto. Ambos experimentos obtiveram seus resultados por meio de requisições que foram construídas com parâmetros pré-definidos para que os modelos pudessem retornar a resposta que o modelo considerava ser a correta.

id	*year*	*area*	*body*	*alternatives*	*answer*
Número da questão	Edição da prova	Macroárea que a questão está inserida	Enunciado e textos base das questões	Alternativas de respostas para a questão	Gabarito oficial para a questão

Tabela 1: Representação de partes dos dados contidos nas tabelas geradas após coleta e pré-processamento das questões.

A estrutura base provida de entrada, também chamada de *prompt*, para que os modelos respondessem à questão, foi construída a partir de conceitos de engenharia de *prompt*[25] que resultaram na seguinte estratégia: “usando como base esse enunciado: {enunciado da questão} responda qual a alternativa correta, entre as opções a seguir: {alternativas}”.

Além do *prompt,* foi provido o contexto para o qual o modelo seria utilizado e este foi definido em língua inglesa, para que os modelos compreendessem exatamente o que era necessário para ser retornado. O contexto usado foi esse: "*You are going to answer questions in Portuguese about various topics. Your answers must be composed with ONLY the letter of the correct alternative, do not return anything else to the user, such as explanation or texts, only the letter. Completing the phrase*: A resposta correta é: ". Para o caso do GEMINI, como sua API não possuía um parâmetro específico para contexto do modelo, o contexto citado acima, foi enviado juntamente do *prompt*, como entrada para o modelo.

Entretanto, a estruturação de todos os parâmetros não foi igual. A **Tabela 2** apresenta a representação das configurações dos experimentos, com as datas de última atualização dos modelos considerando até o momento da escrita deste documento.

Modelo	Versão (Última atualização)	Configuração
Gemini	gemini-1.0-pro (Fev. de 2024)	Temperatura: padrão do modelo (0.9) Tokens: Sem limite
GPT	GPT-3.5-turbo (Set. de 2021)	Temperatura: padrão do modelo (1.0) Tokens: 10
Llama2	Llama2-70b-chat (Ago. de 2023)	Temperatura padrão do modelo (1.0) Tokens: 15

Tabela 2: Informações do experimento com parâmetro de temperatura padrão dos modelos.

É válido pontuar que a quantidade de tokens de saída não foi limitada no GEMINI, pois foi o único modelo que com uma quantidade máxima de tokens de saídas definida, apresentava diferenças significativas nos resultados (não respondendo às requisições realizadas). Enquanto os demais necessitavam de uma quantidade menor de tokens para retornar algum resultado possível de análise.

No segundo experimento foi alterado o valor do parâmetro de temperatura para ser igual a 0, visando respostas mais determinísticas e observar como os resultados se apresentam. Assim, após contabilizado esses dois cenários, foram estabelecidas relações comparativas entre os resultados dos dois cenários e de todos os modelos, em um contexto ano a ano e geral.

3.3 Análise dos resultados

Para a última etapa do projeto, comparações iniciais entre os três modelos para cada uma das 11 edições do exame (disponíveis) foram realizadas. Métricas de acertos e erros (brutos e percentuais), análise de desempenho por áreas, resultados generalizados sobre o total de questões respondidas em cada experimento, foram calculadas a fim de identificar qual modelo obteve o melhor desempenho e simulação de notas de acordo com a quantidade de acertos dos modelos.

Vale pontuar que o cálculo da nota do ENEM não considera apenas quantidade de acertos para atribuir uma nota, mas sim, para se obter a nota final do candidato é utilizado da Teoria de Resposta ao Item (TRI), que acaba por pesar questões de acordo com a dificuldade do item em conjunto com os acertos de questões específicas[26].

Até o momento de desenvolvimento deste trabalho ainda não há uma forma de conseguir a nota exata, por meio do TRI, de edições das provas como a nota de um candidato real que realizou a prova. Então, foi feito o uso de um portal online[27] que possibilita simular uma aproximação de notas, considerando o TRI de edições anteriores a 2023, mas não oficial. Logo, com as devidas notas para cada modelo, houve a comparação com médias nacionais brasileiras e com médias de universidades e cursos para a edição 2023.

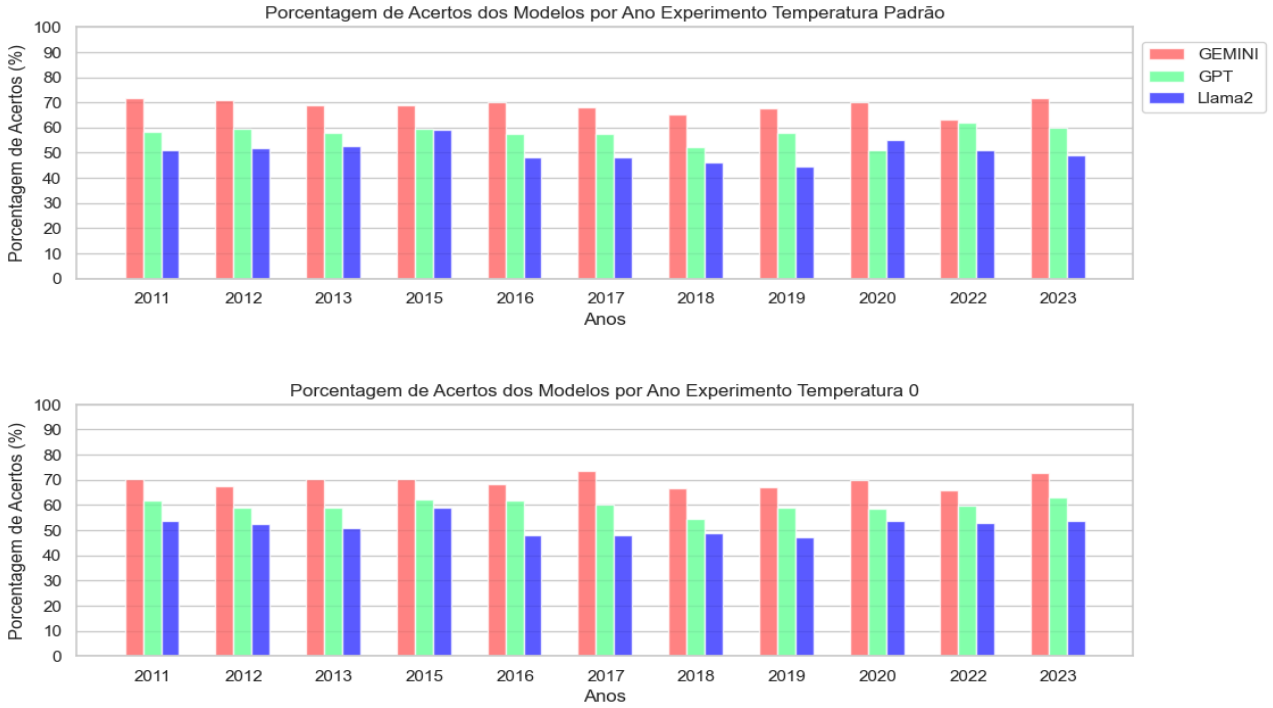

Figura 2: Representação gráfica dos resultados obtidos nos dois cenários de experimentos.

4. RESULTADOS E DISCUSSÕES

Nesta seção, serão discutidos os resultados dos experimentos executados, observando o desempenho dos modelos sob diversos aspectos. Ambos experimentos consideraram as respostas dos três modelos para 2030 questões (5 questões foram anuladas e logo desconsideradas para este estudos), de quatro macroáreas:

- Ciências da Natureza e suas Tecnologias - CN;
- Ciências Humanas e suas Tecnologias - CH;
- Linguagens, códigos e suas Tecnologias - LC;
- Matemática e suas Tecnologias - MT.

4.1 Análise de acertos geral

A **Figura 2** mostra o percentual de acerto dos modelos estudados por edição do exame nos dois experimentos realizados, o primeiro com a temperatura padrão e o segundo com temperatura 0. Para a temperatura padrão, o modelo GEMINI apresentou o melhor desempenho comparado com os demais em todas as edições do ENEM consideradas.

Em contrapartida, o GPT ficou com o segundo maior percentual de acertos em 10 das 11 edições, nas provas de 2020, na qual não obteve o segundo melhor resultado, acabou sendo o modelo que acertou menos questões (com 93 acertos, aproximadamente 50,82%). Enquanto o GEMINI acertou 128 questões e o Llama2 101 questões, em um universo de 183 questões, para esse ano, correspondendo a 69,94% e 55,19% respectivamente. A **Tabela 3** mostra os resultados dos modelos considerando todas as edições do ENEM no experimento com temperatura padrão. De forma geral, o modelo GEMINI obteve os melhores resultados, seguido dos modelos GPT, e Llama2.

Modelo	Acertos	Erros	Sem Resposta
GEMINI	1395 (68.72%)	612 (30,15%)	23 (1,13%)
GPT	1167 (57.49%)	863 (42,51%)	0 (0%)
Llama2	1026 (50.54%)	1004 (49,46%)	0 (0%)

Tabela 3: Resultados quantitativo bruto e percentual, do experimento com temperatura padrão.

Considerando o experimento com temperatura zero, foi evidenciado que todos os modelos obtiveram uma leve melhora em seus desempenhos. Os resultados são apresentados na **Tabela 4**, com o número de acertos e erros de cada modelo. Também foi analisado que este experimento mostrou uma maior constância nos resultados. Tendo em vista o que foi apresentado na **Figura 2** e observando que o modelo GEMINI alcançou a maior quantidade de acertos, o modelo GPT apresentou o segundo melhor desempenho e o modelo Llama2 a mais baixa quantidade de acertos, em todas as edições consideradas no estudo.

Modelo	Acertos	Erros	Sem Resposta
GEMINI	1407 (69.31%)	609 (30%)	14 (0,69%)
GPT	1214 (59,80%)	816 (40,20%)	0 (0%)
Llama2	1048 (51.63%)	982 (48,37%)	0 (0%)

Tabela 4: Resultados quantitativo bruto e percentual, do experimento com temperatura 0.

Apesar da melhora na quantidade de acertos, o percentual não é considerado significativo para concluir que quando se usa temperatura 0 em LLMs, os modelos apresentam resultados superiores quando encaram questões-problemas de múltipla escolha, assim como apontado em Renze e Guven [28]. Haja vista que, em algumas edições, como 2012 e 2016 o resultado de pelo menos um dos modelos era inferior ao resultado obtido pelo mesmo modelo, na mesma edição prova, mas no experimento anterior, com a temperatura padrão.

É válido ressaltar que o modelo GEMINI foi o único modelo a não responder algumas questões. Acredita-se que isso se deve ao fato que o modelo em questão possui “Configurações de Segurança" que segundo o portal da Google, *AI for Developers*[24], essas configurações bloqueiam conteúdo com “probabilidade média ou maior de não ser seguro em qualquer dimensão”[29]. Logo, quando as questões abordavam algum tema sensível, o modelo era alertado quanto a conteúdo hostil e não retornava resposta alguma.

4.2 Análise de acertos por macroáreas de estudo

Para esta subseção foram considerados os resultados do experimento com temperatura 0, devido aos resultados mais otimistas (maior número de acertos). A **Tabela 5** mostra que os modelos apresentaram padrões semelhantes de desempenho. Onde, a área de Ciências Humanas obteve as maiores porcentagens de acertos, seguida de perto por Linguagens, com Ciências da Natureza em terceiro lugar e Matemática em último.

Macroárea do Enem	Modelos								
	GEMINI			GPT			Llama2		
	Acertos (%)	Erros (%)	Sem resposta (%)	Acertos (%)	Erros (%)	Sem resposta (%)	Acertos (%)	Erros (%)	Sem resposta (%)
CH	84,04	15,56	0,40	75,56	24,44	0	66,26	33,74	0
CN	64,78	34,41	0,81	53,24	46,76	0	44,94	55,06	0
LC	83,27	15,64	1,09	73,82	26,18	0	62,91	37,09	0
MT	31,77	67,82	0,41	26,27	73,73	0	23,83	76,17	0

Tabela 5: Resultados quantitativo dos percentuais de acertos aproximados (com grau de precisão de duas casas decimais) do experimento com temperatura 0, por macroárea do exame.

Macroárea do Enem	Modelos								
	GEMINI			GPT			Llama2		
	Min. de acertos	**Max. de acertos**	**Mediana de acertos**	**Min. de acertos**	**Max. de acertos**	**Mediana de acertos**	**Min. de acertos**	**Max. de acertos**	**Mediana de acertos**
CN	63,64%	75,56%	71,11%	51,11%	68,89%	53,33%	37,78%	56,82%	48,89%
CH	75,56%	100%	88,89%	68,89%	91,11%	82,22%	57,78%	82,22%	71,11%
LC	74%	88%	86%	60%	84%	74%	48%	76%	66%
MT	22,22%	40%	31,11%	18,18%	34,09%	25%	11,30%	36,36%	22,22%

Tabela 6: Resultados do mínimo, do máximo e da mediana de percentuais da quantidade de acertos, por modelos em cada área.

Tendo isso como base, foi possível perceber que os melhores desempenhos foram em áreas com muitos textos para serem usados de base, bastante interpretação de texto e conteúdos históricos. Assim, foi hipotetizado que como diversos modelos são treinados com base de dados de diversos acervos históricos, dados com alto volume de textos [3], esse desempenho de quase 20% a mais de acertos, das duas melhores áreas em relação às demais (em **Tabela 5**), tenha se dado devido a este fator.

A **Tabela 6** apresenta percentuais mínimos e máximos de acertos por área de estudo para cada um dos modelos. Com base nela, e reduzindo o escopo para a área com melhor desempenho, Ciências Humanas, foi percebido que os valores mínimos de acerto para esta área foram todos da mesma edição (2012). Então, foi levantada uma nova hipótese, que a prova desta área para essa edição foi a mais difícil entre as 11 analisadas, tendo em vista que essa situação ocorreu apenas para essa área e apenas nesta métrica. Entretanto, não foi encontrada nenhuma notícia, postagem oficial ou trabalho acadêmico que reforçasse tal conjectura. Destaca-se também o fato do modelo GEMINI ter conseguido alcançar a acurácia de 100%, nessa área, em uma das edições, fato não repetido por nenhum outro modelo.

Em contraste com isso, tem-se as macroáreas de índice de corretude menor, Ciências da Natureza e suas Tecnologias e a de Matemática e suas Tecnologias. Para estes casos, acredita-se que além de serem áreas de estudo com definições mais específicas, o ENEM possui a característica de propor a junção entre interpretação de situações problemas e conceitos específicos (como fórmulas matemáticas)[30], e esses fatores podem ter sido problemáticos para os modelos. Uma vez que, muito provavelmente, LLMs não são treinados com foco para esse tipo de abordagem, onde até para questões pouco subjetivas, a interpretação da situação problema, se faz fundamental para resolução da questão.

Nesse contexto, chamou a atenção o fato da porcentagem máxima de acertos em matemática, onde o máximo de acertos do Llama2 superou o máximo do GPT, fato evidenciado apenas para esta área do conhecimento e esta métrica, enquanto nas demais, o GPT seguiu superando o percentual de acerto do Llama2.

4.3 Análise aprofundada dos resultados no ENEM 2023

Nesta subseção, foi observado apenas o contexto do ENEM 2023, no experimento com temperatura 0. Devido a ser a edição mais recente, até o momento da escrita deste documento, e o cenário de experimento com a maior quantidade de acertos.

Modelo	Acertos por área do ENEM 2023 em %				
	Total	CH	CN	LC	MT
GEMINI	72,83	100	68,89	86	34,09
GPT	63,04	86,67	51,11	78	34,09
Llama2	53,80	71,11	40	74	27,27

Tabela 7: Resultados gerais em porcentagem do ENEM 2023 para todos os modelos, no experimento com temperatura 0.

Na **Tabela 7**, pode-se ver os percentuais de acurácia dos modelos na edição 2023, e é válido reforçar o melhor desempenho dos modelos nas áreas de Ciências Humanas e Linguagens, o terceiro melhor resultado em Ciências da Natureza e o pior na área de Matemática. Chama atenção, também, o fato de que o modelo Llama2 obteve um percentual de acerto melhor na área de linguagem (74% de acertos) que na área que em Ciências Humanas (71,11%), divergindo do GEMINI e do GPT.

A fim de criar mais um cenário de comparação entre os modelos e com os participantes do exame, foi utilizada uma plataforma online[27] capaz de calcular uma aproximação da nota do participante para o ENEM 2023. Essa plataforma usa como base, notas de edições anteriores e a quantidade de acertos, por área da prova. Entretanto, devido a não ser capaz de calcular as notas exatas, a plataforma estipula valores mínimos e máximos de possíveis notas para cada área. Junto com isso, foram consideradas as notas e os resultados gerais divulgados pelo INEP, sobre os dados oficiais da edição 2023[31].

Área do Enem	Modelos			Médias Nacionais
	GEMINI	GPT	Llama2	
	Notas			
CN	691,9	622,9	564,5	**497,4**
CH	823	749,2	659,8	**522**
LC	719,2	705,4	671	**516,2**
MT	595,3	595,3	525,7	**534,9**
Média	707,35	668,2	607,75	-

Tabela 8: Notas simuladas para os modelos, por área e médias nacionais obtidas em [31].

Foram utilizados os valores médios das possíveis notas, considerando um cenário neutro, e assim foram obtidas as notas representadas na **Tabela 8.** Vale salientar que esses desempenhos dos modelos em uma situação real, onde há a aplicação do TRI, como é o caso das médias nacionais, essas notas poderiam aumentar ou diminuir, além de que como neste estudo não foi considerada a redação ainda haveria a influência do desempenho desta etapa da prova em uma nota real.

Comparando as notas de cada área dos modelos com as médias nacionais [31], os modelos GEMINI e GPT superaram as médias em todas as 4 áreas avaliadas, enquanto o Llama2, superou em 3 áreas e ficou abaixo da média em Matemática. Com o foco nas médias gerais, foi observado que, baseado nos dados do Sistema de Seleção Unificada (SISU) de 2023 [31], os modelos foram superiores a mais de 37,1% dos participantes do ENEM 2023 e estariam aprovados em mais de 20% de todos cursos no SISU em todo o Brasil.

A fim de reduzir o escopo, foi restringido o universo comparativo para a Universidade Federal de Campina Grande(UFCG). Então, fazendo uso da base de dados divulgados do SISU 2024 a partir do portal Comprov da UFCG[32] e de um blog online[33], filtrando para o campus sede, e vagas para ampla concorrência, foi percebido que o modelo GEMINI seria aprovado em diversos cursos da instituição, os únicos cursos nos quais não seria aprovado na chamada regular seriam esses: Psicologia, Enfermagem, Medicina e Ciência da Computação, sem adicionar à nota do modelo o bônus regional, que acrescenta de 5% a 10% da nota do participante, dependendo do curso, e caso o participante seja natural da federação da universidade.

A situação do GPT é parecida, sem considerar a cota regional, o modelo não passaria na chamada regular em seis cursos, nos mesmos quatro cursos que o modelo anterior, e, em Design e em Engenharia Civil. Já para o caso do Llama2, o desempenho do modelo permitiria ingressar em menos cursos que os outros modelos, mas alguns dos cursos aos quais ele passaria, independente de bônus regional, seriam as Engenharias de Minas, de Matérias e de Alimentos, Letras Português, Geografia, Filosofia, Física, Matemática e Ciências Sociais.

Com a adição da bonificação regional de 5% para os modelos, o GEMINI não seria aprovado apenas em Medicina e em Ciências da Computação. Enquanto o GPT não seria aprovado nos mesmos quatro cursos que o GEMINI, sem a adição do bônus.

5. CONCLUSÕES E TRABALHOS FUTUROS

Este estudo apresentou resultados quantitativos da avaliação de três LLMs diferentes frente ao ENEM, não limitando o universo de questões apenas a questões sem imagens, cartazes ou gráficos; e assim esse desafio que trabalhos anteriores, como Santos e Campelo[34], enfrentaram em relação a textos não-verbais foi contornado.

Este estudo apontou uma superioridade no desempenho do modelo GEMINI, em relação aos outros dois modelos. Entretanto, é importante pontuar que isso pode ter tido grande influência devido a este ser o modelo com atualização mais recente dos três. E, devido à restrição orçamentária, não foi viável realizar o comparativo com o modelo mais recente da OpenAI, o GPT 4.

Também foi possível realizar um comparativo com notas simuladas dos modelos e notas reais de participantes do exame. Com isso, pôde-se perceber que os modelos se assemelham aos alunos brasileiros quando apresentaram dificuldade, por meio de resultados inferiores, nas áreas de Matemática e de Ciências da Natureza, e uma facilidade nas demais áreas. As áreas de maior dificuldade, devido ao TRI, possuem uma pontuação máxima maior que as outras[31], pois essa metodologia recompensa um aluno que desempenha bem áreas onde a maioria não se destaca.

Esses resultados sugerem que, à medida que a IA avança exponencialmente, é provável que LLMs superem cada vez mais seres humanos em testes de conhecimento, como o ENEM, o Exame de Ordem dos Advogados do Brasil, vestibulares de instituições de ensino superior e outros desafios similares.

Com base no trabalho desenvolvido, há várias oportunidades para pesquisas futuras. Uma delas é a reutilização do código criado para acesso às APIs dos modelos, coleta e análise de dados, disponível no *github* do autor [35], oferecendo uma base sólida para estudos subsequentes. Outra possibilidade é treinar os modelos com as edições do ENEM anteriores a 2011 e depois repetir o experimento realizado neste trabalho, para observar como se comportam os resultados. Além disso, testar modelos mais recentes, como o GPT-4 e o Maritaca AI[36], em comparação com resultados do ENEM, pode fornecer insights valiosos sobre o progresso da IA em tarefas de avaliação de conhecimento. Essas abordagens promissoras podem contribuir para avanços significativos na área, orientando o desenvolvimento e aprimoramento de modelos de LLMs e expandindo o conhecimento sobre sua aplicabilidade em diferentes contextos educacionais e profissionais.

6. AGRADECIMENTOS

Primeiramente gostaria de agradecer aos meus pais Arão e Valéria, pela educação que me permitiram ter, pelas constantes orientações em diversas áreas da vida. A minha irmã Maria Clara que foi e é uma inspiração diária para mim. Agradeço também a meu orientador, Fábio, pelo suporte prestado, ao professor Fubica, e a meus amigos e colegas de curso por ouvirem minhas constantes falas entusiasmadas sobre o trabalho e sempre possuírem sugestões e incentivos.

7. REFERÊNCIAS BIBLIOGRÁFICAS

[1] Singularitynet Ambassadors. 2023. The revolution of large language models (LLMS) in Artificial Intelligence. *Medium*.

[2] Radford, A., Sutskever, I., Amodei, D., Luan, D., Child , R., and Wu, J. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

[3] Brown T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G.,Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G.,Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu,J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M.,Gray, S., Chess, B., Clark, J., Berner, C., McCandlish,S., Radford, A., Sutskever, I., Amodei, D.. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

[4] Chang, M., Devlin, J., Lee, K., and Toutanova, K.. 2019. BERT: Pre-Training of deep bidirectional Transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[5] OpenAI. ChatGPT. https://chat.openai.com/. Acessado em: 2024-01-13.

[6] OpenAI. ChatGPT Overview. https://openai.com/chatgpt. Acessado em: 2024-01-13.

[7] Meta. Meta Company webpage. https://about.meta.com/. Acessado em: 2024-01-13.

[8] Google. Google: Saiba mais sobre o que fazemos. https://about.google/intl/ALL_br/. Acessado em 2024-01-13.

[9] Meta. Llama webpage. https://llama.meta.com/. Acessado em: 2024-01-13.

[10] Google. Gemini webpage. https://gemini.google.com/. Acessado em: 2024-01-13.

[11] Reuters.2023. LLaMA: Meta anuncia 'rival' do ChatGPT para pesquisadores de inteligência artificial. *g1- O portal de notícias da Globo.*

[12] Silva, V. H. 2023. Google lança Gemini, sua inteligência artificial mais poderosa; veja como ela funciona. *g1- O portal de notícias da Globo.*

[13] OpenAI. 2023. GPT-4 Technical Report. *OpenAI blog*. Disponível em https://openai.com/research/gpt-4.

[14] Kifleswing. 2023. OpenAI announces GPT-4, claims it can beat 90% of humans on the SAT. CNBC. Disponível em: https://www.cnbc.com/2023/03/14/openai-announces-gpt-4-says-beats-90percent-of-humans-on-sat.html. Acessado em: 2023-11-25.

[15] Ministério da Educação - MEC , INEP. 2023. EDITAL ENEM 2023. Diário Oficial da União, Edição: 86, Seção: 3, Pág. 66.

[16] Jurafsky, D., Martin, J. H. (2009). Speech and Language Processing. Prentice Hall.

[17] Gomez, A. N., Jones, L., Kaiser, Ł.,Parmar, N., Polosukhin, I., Shazeer, N., Uszkoreit, J., Vaswani, A. (2017). Attention Is All You Need. *arXiv preprint arXiv:1706.03762*.

[18] IBM. 2024. IBM documentation Tokens and tokenization. Disponível em: https://www.ibm.com/docs/en/watsonx/saas?topic=solutions-tokens.

[19] Beard C. L. 2024. What is a Prompt with AI?. *Medium*. Disponível em: https://medium.com/brainscriblr/what-is-a-prompt-with-ai-c542669f2b4b

[20] Amazon. O que é engenharia por prompt? .Disponível em: https://aws.amazon.com/pt/what-is/prompt-engineering/

[21] INEP. 2023. Provas e Gabaritos ENEM. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem/provas-e-gabaritos. Acessado em: 2023-12-07.

[22] OpenAI. 2021. OpenAI API Documentation. Acessado em: 2024-01-16.

[23] Meta. Llama API Documentation Quickstart. Acessado em: 2024-01-17.

[24] Google. 2024. Google for Developers Documentation. Acessado em: 2024-02-02.

[25] Ba., J.,Chan, H., Han, Z., Muresanu, A. I., Paster, K., Pitis, S., Zhou, Y. 2023. Large Language Models are Human-Level prompt engineers. *arXiv preprint arXiv:2211.01910*.

[26] INEP 2021. Entenda a sua nota no Enem:guia do participante. gov.br.Brasília, DF, Brasil. Disponível em: https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/entenda_a_sua_nota_no_enem_guia_do_participante.pdf. Acessado em: 2024-03-17.

[27] Me Salva. Calculadora de Nota do ENEM. Disponível em: https://www.mesalva.com/nota-enem. Acessado em: 2024-04-05.

[28] Guven, E., Renze, M. 2024. The Effect of Sampling Temperature on Problem Solving in Large Language Models. *arXiv preprint arXiv:2402.05201*.

[29] Google. 2023. Google for Developers Documentation: Safety Settings. Acessado em: 2024-03-09.

[30] INEP. 2020. Matriz de referência ENEM. gov.br. Disponível em: https://download.inep.gov.br/download/enem/matriz_referencia.pdf. Acessado em: 2024-03-17.

[31] INEP. 2024. ENEM 2023: Resultados. gov.br. Disponível em: https://download.inep.gov.br/enem/resultados/2023/apresentacao_resultados.pdf. Acessado em 2024-03-10.

[32] Comissão de Processo de Vestibulares-Comprov Universidade Federal de Campina Grande. 2024. Disponível em: https://comprov.ufcg.edu.br/graduacao.html. Acessado em: 2024-03-10.

[33] Wesley, J. 2024. Notas de corte SISU na UFCG: todos os cursos. Disponível em: https://blogdoenem.com.br/ufcg-notas-de-corte-sisu/. Acessado em: 2024-03-10.

[34] Santos, M. L. O., Campelo, C. 2023. Avaliação de grandes modelos de linguagem quantizados na resolução de questões do ENEM. *SISTEMOTECA-Sistema de Bibliotecas da UFCG, Biblioteca Digital de Teses e Dissertações*.

[35] Raposo L. B. 2024. Repositório utilizado no desenvolvimento do experimento. GitHub. Disponível em: https://github.com/LucasBrasileiroRaposo/TCC.

[36] MARITACA AI. Maritaca AI webpage. Acessado em: 2024-04-12.