UNIVERSIDADE FEDERAL DO VALE DO SÃO FRANCISCO
CAMPUS SALGUEIRO - PE
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Eládio Leal Alves

Avaliação de Grandes Modelos de Linguagem na Resolução de Questões do ENADE para Cursos de Computação

Salgueiro - PE
2025

Eládio Leal Alves

Avaliação de Grandes Modelos de Linguagem na Resolução de Questões do ENADE para Cursos de Computação

Trabalho de Conclusão de Curso do curso de Bacharelado em Ciência da Computação apresentado ao Colegiado de Ciência da Computação como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.
Orientador: Prof. Zé de Maria Preá, Me.

Salgueiro - PE
2025

UNIVERSIDADE FEDERAL DO VALE DO SÃO FRANCISCO - UNIVASF
Gabinete da Reitoria
Sistema Integrado de Bibliotecas (SIBI)
Av. José de Sá Maniçoba, s/n, Campus Universitário – Centro CEP 56304-917
Caixa Postal 252, Petrolina-PE, Fone: (87) 2101- 6760, biblioteca@univasf.edu.br

Sobrenome do autor, Prenome do autor
* Cutter Título do trabalho / Nome por extenso do autor. - local, ano.
xx (total de folhas antes da introdução em nº romano), 50 f.(total de folhas do trabalho): il. ; (caso tenha ilustrações) 29 cm.(tamanho do papel A4)

Trabalho de Conclusão de Curso (Graduação em nome do curso) - Universidade Federal do Vale do São Francisco, Campus, local, ano

Orientador (a): Prof.(a) titulação e nome do prof(a).

Notas (opcional)

1. Assunto. 2. Assunto. 3. Assunto. I. Título. II. Orientador (Sobrenome, Prenome). III. Universidade Federal do Vale do São Francisco.

* CDD

Ficha catalográfica elaborada pelo Sistema Integrado de Biblioteca SIBI/UNIVASF
Bibliotecário: Nome* e CRB*

* ***Dados inseridos pela biblioteca***

Exemplo:

Souza, José Augusto de
S729c Crianças com dificuldades de aprendizado: estudo nas escolas públicas da cidade de Juazeiro-BA / José Augusto de Souza. – Petrolina - PE, 2009.
xv, 140 f. : il. ; 29 cm.

Trabalho de Conclusão de Curso (Graduação em Psicologia) Universidade Federal do Vale do São Francisco, Campus Petrolina-PE, 2009.

Orientadora: Profª. Drª. Maria de Azevedo.

Inclui referências.

1. Crianças - Ensino. 2. Distúrbios da aprendizagem. 3. Escolas públicas – Juazeiro (BA). I. Título. II. Azevedo, Maria de. III. Universidade Federal do Vale do São Francisco.

370.15

Ficha catalográfica elaborada pelo Sistema Integrado de Biblioteca SIBI/UNIVASF
Bibliotecário: Nome e CRB.

Eládio Leal Alves

Avaliação de Grandes Modelos de Linguagem na Resolução de Questões do ENADE para Cursos de Computação

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do título de Bacharel em Ciência da Computação e aprovado em sua forma final pela banca examinadora.

Salgueiro - PE, 18 de dezembro de 2025.

Prof. Maria Bernadete, Me.
Coordenador do Curso

Banca Examinadora:

Prof. Zé de Maria Preá, Me.
Presidente da Banca

Prof. X Y Z, Me.
Avaliador
Universidade Federal do Vale do São Francisco

Prof. X Y Z, Dr.
Avaliador
Universidade Federal do Vale do São Francisco

AGRADECIMENTOS

Agradeço a meu pai, minha mãe, meu cachorro, minha sogra e por último e menos importante, meu orientador.

RESUMO

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum passages, and more recently with desktop publishing software like Aldus PageMaker including versions of Lorem Ipsum.

Palavras-chave: WebAssembly. Web. Desempenho. Compiladores. Emscripten. Cheerp.

ABSTRACT

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum passages, and more recently with desktop publishing software like Aldus PageMaker including versions of Lorem Ipsum.

Keywords: WebAssembly. Web. Performance. Compilers. Emscripten. Cheerp.

LISTA DE FIGURAS

LISTA DE CÓDIGOS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O uso de grandes modelos de linguagem, do inglês Large Language Model (LLM), já está se tornando cada vez mais comum na sociedade atual, apoiando diversas áreas em que se necessita auxílio em atividades textuais. Recentemente, o desenvolvimento desses modelos impulsionou avanços significativos na área de Processamento de Linguagem Natural (PLN), criando diversas oportunidades de uso tanto em ambientes profissionais quanto educacionais. Os modelos de LLM atuais avançam para além dos modelos de linguagem tradicionais ao integrar datasets maiores e arquiteturas transformer, destacando-se na aprendizagem a partir de dados extensivos e alcançando resultados de ponta em tarefas de PLN. Isso inclui compreensão da linguagem, geração de linguagem, tradução automática, geração de diálogo, análise de sentimentos e sumarização de conteúdo (Mohamed et al., 2024).

Com todo o avanço nessa área da PLN, surge a possibilidade de investigar formas de aplicação dessa tecnologia, e uma das mais promissoras é o apoio à educação. Trabalhos que antes demandavam tempo e esforço humano podem ser automatizados e aprimorados. Assim, essa tecnologia pode ser explorada para a automatização de avaliações e geração de feedbacks que se caracterizam por serem detalhados, oportunos e de apoio , aspectos cruciais para o desenvolvimento e aprendizado do aluno (Liew; Tan, 2024).

1.1 QUESTÕES DE PESQUISA

O desenvolvimento desse trabalho foi elaborado com objetivo de responder as seguintes questões de pesquisa:

QP01 O quão acertivo pode ser uma LLM's na resolução de questões em computação?

QP02 Dentre as LLM's selecionadas para o estudo, qual obteve o melhor resultado na resolução de questões?

QP03 Alguma das LLM'S se mostra superior nos acertos em alguma área específica da computação?

1.2 OBJETIVOS

Os objetivos deste trabalho são subdivididos em objetivos gerais e objetivos específicos. Estes são:

1.2.1 Objetivo Geral

Avaliar comparativamente o desempenho de grandes modelos de linguagem na resolução de questões objetivas do ENADE aplicadas a cursos da área de Computação, com base no gabarito oficial.

1.2.2 Objetivos Específicos

Os objetivos específicos são:

- Realizar uma análise da literatura recente sobre grandes modelos de linguagem com foco em tarefas de resposta a perguntas e compreensão de texto;
- Selecionar, com base na literatura analisada, os modelos de linguagem para fins de avaliação comparativa;
- Coletar e organizar questões objetivas de múltipla escolha do ENADE, aplicadas desde 2004, para os cursos da área de Computação;
- Submeter as questões selecionadas aos modelos escolhidos, registrando sistematicamente as respostas geradas;
- Avaliar e comparar o desempenho dos modelos considerando a granularidade de curso, a partir dos gabaritos oficiais disponibilizados pelo INEP.

1.3 JUSTIFICATIVA

O uso de grandes modelos de linguagem (LLMs) em tarefas de processamento de linguagem natural tem se mostrado promissor em diversos contextos, incluindo aplicações na área educacional. Em especial, na área de Computação, muitos processos avaliativos como a correção de questões de múltipla escolha, que frequentemente envolvem conceitos técnicos ainda são realizados manualmente ou com sistemas limitados em flexibilidade e capacidade interpretativa. Nesse cenário, os LLMs se destacam por oferecerem maior adaptabilidade e por possibilitarem a geração de feedbacks mais contextualizados e pedagógicos. No entanto, sua real eficácia nesse tipo de aplicação ainda carece de investigações mais aprofundadas. Embora já existam pesquisas voltadas para a análise do desempenho desses modelos em contextos avaliativos, observa-se uma lacuna no que diz respeito ao uso dos LLMs especificamente em questões do ENADE voltadas para cursos da área de Computação. Assim, este trabalho justifica-se pela necessidade de avaliar o desempenho de diferentes LLMs na resolução e interpretação de questões desse exame, buscando compreender suas limitações, potencialidades e possíveis contribuições para o aprimoramento de processos avaliativos na educação superior.

1.4 ORGANIZAÇÃO DO TRABALHO

Esse trabalho é organizado como segue: No capítulo 2 são apresentados os conceitos base para melhor entendimento das tecnologias abordadas. Portanto, o capítulo apresenta uma introdução sobre os grandes modelos de linguagem (Large Language Models), seguido por apresentar o conceito da tecnologia e uma contextualização histórica sobre a evolução

da área da PLN até a chegada dos transformers. Ao final do capítulo é também listado as principais pesquisas relacionadas. No capítulo seguinte, 3, é descrito os passos necessários para realizar o experimento desejado assim como o ambiente adotado para execução da pesquisa. No capítulo 4 é apresentado os dados coletados no experimento executado, em seguida é feito uma análise dos dados visando responder as questões de pesquisa. Por fim, no capítulo 5 é sintetizado o que foi realizado na pesquisa assim como os resultados obtidos ao final da análise de dados.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os conceitos básicos que envolvem o tema da pesquisa, assim como descreve os trabalhos correlatos a este, para melhor entendimento do contexto em que se encontra as pesquisas em LLMs.

2.1 ENADE

O Exame Nacional de Desempenho dos Estudantes (ENADE) constitui um dos principais instrumentos do Sistema Nacional de Avaliação da Educação Superior (SINAES), instituído pela Lei nº 10.861, de 14 de abril de 2004. O SINAES foi concebido como uma política pública de avaliação capaz de assegurar e promover a qualidade da educação superior brasileira, a partir de uma abordagem formativa, diagnóstica e integradora, que considera dimensões pedagógicas, institucionais e de desempenho discente (Brasil, 2004).

O ENADE tem como principal objetivo avaliar o rendimento dos estudantes em relação aos conteúdos previstos nas Diretrizes Curriculares Nacionais (DCNs) dos respectivos cursos, bem como suas competências para compreender temas externos ao âmbito específico da profissão, situando-se na realidade social, econômica, cultural e política do país. Ao avaliar estudantes ingressantes e concluintes, o exame permite observar a evolução da aprendizagem ao longo da formação acadêmica, sendo seus resultados fundamentais para compor indicadores como o Conceito Preliminar de Curso (CPC) e o Índice Geral de Cursos (IGC), que subsidiam processos de regulação, supervisão e melhoria da qualidade da educação superior no Brasil (Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), 2023).

Sob uma perspectiva teórica, o ENADE também se apresenta como um mecanismo indutor da qualidade, na medida em que seus resultados não apenas retratam o desempenho dos estudantes, mas também impulsionam processos de autorreflexão institucional, revisão de projetos pedagógicos, atualização curricular e aprimoramento da gestão acadêmica. Conforme aponta (Dias Sobrinho, 2003), a avaliação no contexto da educação superior deve ser compreendida como um instrumento de construção coletiva, capaz de orientar transformações qualitativas, evitando uma lógica meramente classificatória ou competitiva. Assim, a efetividade do ENADE depende do engajamento da comunidade acadêmica — docentes, discentes e gestores — que, ao se apropriar dos seus resultados, pode utilizá-los como base para ações de melhoria contínua, alinhadas às exigências da sociedade e do mercado de trabalho.

Dessa forma, o ENADE se consolida como um dispositivo fundamental para a promoção de uma cultura avaliativa nas instituições de ensino superior, sendo não apenas uma exigência legal, mas uma oportunidade de aprimoramento dos processos formativos e de fortalecimento do compromisso social da educação superior brasileira.

2.2 LLM - LARGE LANGUAGE MODEL

Grandes Modelos de Linguagem (Large Language Models, LLMs) representam uma quebra de paradigma no uso da Inteligência Artificial (IA)(RAPOSO et al., 2024). Os LLMs aprenderam a entender padrões a partir de grandes quantidades de textos de exemplo e a gerar respostas coerentes. Esses padrões identificam distribuições de probabilidades para sequências de palavras, que podem ser empregadas para gerar textos sintéticos (Peres, 2023).

Uma das características fundamentais dos LLMs reside na sua elevada capacidade computacional, diretamente relacionada à quantidade massiva de parâmetros e ao volume expressivo de dados utilizados em seu treinamento. Esses modelos são projetados para aprender padrões linguísticos complexos por meio de técnicas avançadas de aprendizado profundo (deep learning), particularmente baseadas na arquitetura Transformer, que permite processar e modelar relações contextuais em grandes sequências de texto.

A partir da exposição a extensos conjuntos de dados textuais — provenientes de livros, artigos, sites, fóruns e outras fontes —, os LLMs desenvolvem a habilidade de reconhecer estruturas sintáticas, relações semânticas e padrões discursivos presentes na linguagem natural. Dessa forma, são capazes não apenas de compreender e interpretar o conteúdo textual, mas também de gerar respostas e textos que se assemelham, em termos de coerência e fluidez, à produção humana.

Essas capacidades permitem que os LLMs sejam aplicados em uma ampla gama de tarefas dentro do campo de PLN, como, por exemplo, tradução automática de idiomas, geração de textos originais, elaboração de resumos automáticos, análise de sentimentos, detecção de tópicos e desenvolvimento de sistemas inteligentes de perguntas e respostas (Eze, 2025). Ademais, a precisão e a sofisticação desses modelos em lidar com a linguagem tornam possível sua utilização em contextos cada vez mais desafiadores e especializados, tanto no meio acadêmico quanto no mercado.

2.3 EVOLUÇÃO DA PLN ATÉ A INTRODUÇÃO DAS LLMS

A literatura aborda que o início dos estudos de PLN remonta à década de 50, onde Alan Turing propôs em seu artigo intitulado "Computing Machinery and Intelligence"o tão conhecido como "Teste de Turing"como algo a se definir como critério de inteligência(Jurafsky; Martin, 2023a). Esse teste deu a possibilidade de poder fazer máquinas pensarem e assim trouxe mais motivação aos inicio estudo na PLN.

Entre as decadas de 1950-1960 ouve os primeiro avanços e demonstração de um sistema de tradução automatica. Atráves da colaboração entre IBM e Georgetown University em que teve Léon Dostert como uma das figuras centrais desse projeto que consistiu em fazer uma tradução automatica do russo para o inglês. O estudo feito em pequena escala teve um total de 250 palavras e seis regras gramaticais; ainda assim, foi

considerado um sucesso, pois o sistema demonstrou a capacidade de tradução, mesmo sendo para trechos curtos (Norman, 2025).

Entre as decadas de 1960-1970 no MIT Artificial Intelligence Laboratory Joseph Weizenbaum criou um programa intitulado ELIZA, esse programa tinha a finalidade de se passar por um terapeuta, o programa era simples ao ponto de oferecer respostas pré-definidas aos usuários para fazer eles pensarem que estariam se comunicando com alguém que entendia o que lhe era passado. Considerado o primeiro chatbot, foi um caso inicial ao teste de Turing(Wallace, 2016). Nesse período também houve o SHRDLU criado por Terry Winograd no MIT no Artificial Intelligence Laboratory, o SHRDLU era capaz de compreender e executar comandos complexos, responder a perguntas sobre o estado do mundo, pedir esclarecimento quando necessário, raciocinar sobre possibilidades, aprender novas definições e até mesmo explicar o raciocínio por trás de suas ações(Reihani, 2025).

Entre as décadas de 1980 e 1990, houve uma transição no Processamento de Linguagem Natural (PLN) para abordagens mais estatísticas, baseadas em grandes corpora de textos (Jurafsky; Martin, 2023b).
Destacou-se, nesse período, o uso dos Modelos de Markov Ocultos (Hidden Markov Models – HMMs), especialmente em tarefas como reconhecimento de fala e etiquetagem gramatical (POS tagging) (Jelinek, 1997). Os métodos estatísticos se consolidaram com a expansão dos modelos n-gram e com o desenvolvimento dos Modelos IBM 1 a 5, que formaram a base para os sistemas de tradução automática estatística (SMT) (Brown; Pietra et al., 1993). Além disso, a criação do Penn Treebank, em 1993, forneceu um corpus sintaticamente anotado de grande escala, que se tornou referência para o treinamento e avaliação de parsers probabilísticos (Marcus; Marcinkiewicz; Santorini, 1993). A década também foi marcada pela realização das Message Understanding Conferences (MUCs), que impulsionaram as pesquisas em extração de informação (Information Extraction – IE) e promoveram a padronização de benchmarks para tarefas como reconhecimento de entidades nomeadas (NER) e resolução de co-referência (Grishman; Sundheim, 1996).

Entre as décadas de 2000–2010, com os avanços provenientes da década passada, culminaram também em avanços nos algoritmos utilizados na PLN tem-se uma adoção em larga escala de algoritmos supervisionados como Máquinas de Vetores de Suporte (SVMs), Modelos de Máxima Entropia e principalmente os Conditional Random Fields (CRFs), aplicados com sucesso em tarefas como reconhecimento de entidades nomeadas, POS tagging e parsing sintático (Lafferty; McCallum; Pereira, 2001). Nesse período, também se estabeleceu o domínio da tradução automática estatística baseada em frases, com o desenvolvimento de ferramentas como o Moses (Koehn et al., 2007), que substituíram os antigos modelos palavra-a-palavra (word-based). Paralelamente, houve uma inovação no uso de representações vetoriais para palavras: modelos como o Latent Semantic Analysis (LSA) já vinham sendo utilizados, mas o grande marco foi o trabalho de Bengio et al. (2003), que introduziu o primeiro modelo de linguagem neural, propondo a ideia de treinar

representações distribuídas de palavras em redes neurais. Esses avanços foram suportados pelo crescimento dos corpora anotados como o Penn Treebank e a organização de desafios como CoNLL-2003 e SemEval, que padronizaram benchmarks para tarefas como NER e análise de sentimentos. Com isso, a década de 2000 solidificou as bases estatísticas do PLN e iniciou a transição para abordagens neurais mais sofisticadas que viriam na década seguinte.

Entre as décadas de 2010–2017, tiveram novas revoluções significativas com a adoção crescente de técnicas de aprendizado profundo. Modelos de redes neurais recorrentes (RNNs), especialmente as variações LSTM(Long Short-Term Memory) e GRU(Gated Recurrent Unit), passaram a ser amplamente utilizados em tarefas sequenciais como análise de sentimentos, tradução automática e resposta a perguntas (**tang2015lstm** ; Cho et al., 2014). O modelo Word2Vec, proposto por Mikolov et al. (2013), introduziu embeddings capazes de capturar relações semânticas complexas, seguido pelo GloVe, de Pennington, Socher e Manning (2014), que combinava estatísticas globais de coocorrência com propriedades locais do texto. Na tradução automática, os métodos estatísticos deram lugar aos sistemas baseados em redes neurais, especialmente após a introdução do modelo de atenção por Bahdanau, Cho e Bengio (2015).

A partir de 2017 até o presente momento, a PLN passou por uma transformação profunda com o surgimento da arquitetura Transformer, apresentada por (Vaswani et al., 2017) no artigo "Attention is All You Need". Ao eliminar o uso de redes recorrentes e basear o processamento em mecanismos de atenção, os Transformers permitiram maior paralelização no treinamento e obtiveram resultados superiores em diversas tarefas de PLN, como tradução, classificação e resposta a perguntas. Esse avanço abriu caminho para a era dos modelos pré-treinados em larga escala. Em 2018, o BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) introduziu um novo paradigma: pré-treinamento em grandes volumes de texto seguido de ajuste fino (fine-tuning) para tarefas específicas. BERT e suas variantes como RoBERTa, XLNet e ALBERT superaram modelos anteriores em benchmarks como GLUE e SQuAD, consolidando o uso de embeddings contextuais profundos. A partir de 2019, os modelos generativos ganham destaque com o GPT-2 (Radford; Wu; Child et al., 2019), seguido por GPT-3 (Brown; Mann; Ryder et al., 2020), que popularizou o conceito de few-shot learning, permitindo resolver tarefas complexas apenas com instruções em linguagem natural. O GPT-3, com 175 bilhões de parâmetros, tornou-se um marco em geração de texto, raciocínio e aplicações comerciais.

2.4 TRABALHOS CORRELATOS

Nunes et al. (2023) em seu trabalho "Evaluating GPT-3.5 and GPT-4 Models on Brazilian University Admission Exams"analisou o uso de LLM's na resolução de questões do exame nacional do ensino médio, utilizando técnicas diferentes de prompts, nesse trabalho foram utilizados os exames das edições de 2009 até 2019 e também há 2022. As

técnicas de prompt utilizadas foram a zero-shot, few-shot e few-shot com Chain-of-Thought (CoT). O modelo GPT-4 obteve um resultado significativo com o uso da técnica CoT, e, analisando pelo campo da matemática, obteve um aumento mais expressivo quando se analisa as outras áreas utilizando essa técnica. No modelo GPT 3.5, o uso do CoT também refletiu uma melhoria nas resoluções de questões de matemática, no entanto, no GPT 4, outras áreas não foram afetadas; porém, nesse modelo, o uso do CoT acompanhou um declínio nas resoluções das questões de ciências da natureza e linguagens. Portanto, o estudo mostrou que o modelo GPT 4 possui uma alta capacidade de resolver as questões do Enem e retornar *insigts* váliosso em sua resposta. Essa capacidade é vista como uma ferramenta educacional promissora, pois pode aprimorar a compreensão dos alunos sobre conceitos complexos e apoiar seu processo de aprendizagem, oferecendo respostas mais transparentes e informativas para questões desafiadoras.

No trabalho de Viegas et al. (2024), em que se investigou a capacidade das LLMs em igualar ou superar o desempenho humano no POSCOMP (Exame Nacional para Ingresso na Pós-Graduação em Computação), utilizando os modelos ChatGPT-4, Gemini 1.0 Advanced, Claude 3 Sonnet e Le Chat Mistral Large – utilizando as edições de 2022 e 2023 dos exames. A pesquisa foi dividida em duas etapas, a primeira em analisar os modelos na resolução das questões por meio de interpretação de imagens das questões, e a segunda maneira foi por meio da resolução das questões por meio de prompt textual convertido para o idioma do inglês. As alucinações foram menos frequentes na avaliação baseada em texto. Os modelos variaram em seus níveis de explicação; o Gemini e o Claude consistentemente ofereceram explicações mais abrangentes, enquanto o ChatGPT-4 e o Mistral ocasionalmente optaram por respostas mais diretas. Por fim, o estudo concluiu que os modelos LLM's têm boas e consideráveis respostas ao resolver questões do POSCOMP, o modelo que mais se destacou foi o ChatGPT-4, pois ele superou outros modelos em ambos os tipos de testes na metodologia. Porém, para ambos os modelos, a interpretação de imagens ainda é um desafio a ser melhorado em versões futuras dessas LLM's.

No trabalho de RAPOSO et al. (2024) investiga a capacidade de Grandes Modelos de Linguagem (LLMs) em responder a questões objetivas do Exame Nacional do Ensino Médio (ENEM). Os LLMs surgiram como uma "quebra de paradigma"na Inteligência Artificial (IA) e são amplamente utilizados, com o ChatGPT (OpenAI) sendo um dos principais responsáveis pela sua popularização. O estudo ressalta que a maioria das avaliações de LLMs se limita ao contexto da língua inglesa, sem testes eficazes em cenários globalizados como o brasileiro. Nessa pesquisa utilizou-se as LLMs: Llama 2, GPT-3.5 e GEMINI 1.0 Pro em questões de múltipla escolha do ENEM de 11 edições (2011-2013, 2015-2020, 2022, 2023). Na metodologia desse trabalho buscou fazer uma integração do envio das questões por meio de API das LLMs e também fazer váriações na temperatura das respostas entregues pela as mesmas. No geral dos acertos o Gemini obteve um percentual maior nas questões com a temperatura definida como a padrão das LLMs, resultados

melhores que o GPT-3.5 e o Llama 2. Com a temperatura do modelo calibrada em 0 para dar respostas mais deterministicas, todos os modelos obtiveram melhorias na quantidad de acertos, no entando, o Gemini ainda se saiu superior que os demais. Portanto o estudo mostrou que no contexto de responder as questões do Enem, o LLM da Google (Gemini) mostrou uma capacidade superior que o GPT-3.5 e o Llama 2. Porém todos os LLMs mostraram maior dificuldade em Matemática e Ciências da Natureza.

Rodrigues et al. (2025) em seu estudo aborda a capacidade das LLMs em sistemas de resposta a perguntas educacionais, que facilitam o aprendizado adaptativo e respondem às dúvidas dos alunos. O estudo investiga como as LLMs podem ser integrados eficientemente em sistemas educacionais de perguntas e respostas para atender a diversas necessidades educacionais. Nesse estudo foram utilizados os modelos GPT-4 o modelo mais atual da OpenIA e também o Sabiá (um LLM de código aberto, otimizado especificamente para o português brasileiro, baseado nos modelos LLaMA) e o uso dela visa abordar a carência de consideração de LLMs nativos na pesquisa. Foi criado um questionario de 70 questões em que foram baseadas na Base Nacional Comum Curricular (BNCC) do Brasil com 40 delas sendo de Matématico e 30 da Lingua Portuguesa, e destinadas a alunos do terceiro ano do ensino fundamental, representando o momento em que os alunos completam a alfabetização. Para esse trabalho foram utilizadas questões além das de multipla escolha, como as de preencher lacunas no texto e dissertativas curtas. Ambos os modelos demonstraram um desempenho forte e confiável, com uma pontuação média geral de 9,79 de 10. Portanto o estudo conclui que os LLMs, tanto o GPT-4 quanto o Sabiá, demonstram fortes capacidades na resolução de questões educacionais em português brasileiro. Essa consistência indica que ambos os modelos são hábeis em lidar com questões em português, refletindo um desempenho confiável até mesmo com questões em outro idioma não-inglês.

3 DELINEAMENTO METODOLÓGICO

A presente seção descreve detalhadamente os procedimentos metodológicos adotados ao longo da pesquisa. São apresentados o tipo de pesquisa conduzida, os critérios para seleção dos modelos de linguagem, os métodos utilizados para coleta de dados, bem como o procedimento de aplicação e avaliação dos modelos. Além disso, são descritas as ferramentas e recursos tecnológicos empregados no desenvolvimento da plataforma utilizada para a simulação dos questionários e integração com os modelos de linguagem. Cada etapa foi cuidadosamente planejada para garantir uma análise rigorosa e comparativa do desempenho das LLMs na resolução de questões do ENADE relacionadas à área de Computação.

3.1 TIPO DE PESQUISA

Trata-se de uma pesquisa aplicada, uma vez que utiliza o conhecimento técnico acerca de Grandes Modelos de Linguagem (LLMs) com o intuito de avaliar quais modelos apresentam melhor desempenho no contexto da área de Computação, sendo capazes de resolver, de forma precisa, questões relacionadas a esse domínio. A abordagem adotada é quantitativa, pois os resultados produzidos pelos modelos são mensurados por meio de dados numéricos, especialmente pela taxa de acertos obtida em comparação com os gabaritos oficiais da prova do ENADE. Adicionalmente, a pesquisa possui caráter avaliativo e experimental, tendo em vista que envolve a execução controlada de experimentos com diferentes modelos de linguagem, com o objetivo de observar, comparar e analisar o desempenho desses modelos em um conjunto específico de questões. Por concentrar-se nas edições do ENADE direcionadas aos cursos de Computação, a investigação também pode ser caracterizada como um estudo de caso, voltado à identificação dos modelos que apresentam melhor desempenho nesse contexto específico de aplicação.

3.2 SELEÇÃO DOS MODELOS DE LINGUAGEM

A seleção dos modelos de linguagem adotados nesta pesquisa baseou-se em critérios de relevância tecnológica, representatividade no estado da arte e acessibilidade para experimentação e nos modelos que tiveram melhores resultados nos trabalhos correlatos. Foram escolhidos quatro modelos amplamente utilizados e reconhecidos pela comunidade científica e pelo mercado: GPT, da OpenAI; Gemini, desenvolvido pelo Google DeepMind; LLaMA, da Meta; e DeepSeek, de código aberto e com crescente adoção em aplicações de geração e compreensão de linguagem natural. A escolha por esses modelos visa representar diferentes abordagens arquiteturais e estratégias de treinamento, possibilitando uma análise comparativa mais abrangente em relação à capacidade de resolução de questões do ENADE na área de Computação. Adicionalmente, considerou-se a viabilidade de acesso

às interfaces de inferência dos modelos, por meio de APIs públicas de tais modelos.

3.3 COLETA DE DADOS

Para a etapa de coleta de dados, foram selecionadas questões objetivas provenientes do ENADE, abrangendo os cursos de Ciência da Computação, Engenharia da Computação e Sistemas de Informação. A escolha dessas áreas deve-se à sua proximidade em termos de conteúdo programático e à relevância dos temas abordados para a formação em Computação. As questões foram extraídas de edições anteriores do exame, disponíveis publicamente por meio do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), garantindo a legitimidade e a padronização dos dados utilizados. Optou-se por questões de múltipla escolha com gabarito oficial disponível, o que possibilita a análise objetiva do desempenho dos modelos de linguagem por meio da comparação direta entre as respostas geradas e as respostas corretas fornecidas pelos exames.

3.4 PROCEDIMENTO DE APLICAÇÃO

O procedimento de aplicação foi realizado por meio de uma plataforma web desenvolvida exclusivamente para esta pesquisa, com o objetivo de simular a interação entre usuários e modelos de linguagem natural após a realização de atividades, podendo requisitar os serviços para fazerem as correções. Essa plataforma foi projetada para apresentar questionários compostos por questões extraídas do ENADE, de forma sequencial e interativa, reproduzindo a experiência de um ambiente de avaliação tradicional. Cada questionário contém um conjunto de perguntas organizadas por área do conhecimento — Ciência da Computação, Engenharia da Computação e Sistemas de Informação — permitindo a aplicação controlada dos testes para os modelos selecionados. A interface simula o comportamento de um usuário humano, submetendo cada questão individualmente, o que proporciona uma avaliação mais próxima da realidade de uso desses modelos em contextos educacionais.

A plataforma está integrada às APIs REST dos modelos de linguagem por meio de um módulo de comunicação que realiza o envio das perguntas e o recebimento das respostas de forma automatizada. Cada requisição contém o enunciado da questão e as alternativas de resposta, formatadas de acordo com os requisitos de cada modelo. Após a obtenção das respostas geradas pelas LLMs, os dados são armazenados em um banco estruturado, permitindo a análise posterior quanto à precisão, coerência e taxa de acerto em relação ao gabarito oficial. Essa abordagem garante reprodutibilidade, rastreabilidade e padronização na aplicação dos testes, além de facilitar a comparação entre os diferentes modelos avaliados sob as mesmas condições experimentais. Dessa forma, é possível conduzir uma análise sistemática e confiável do desempenho das LLMs em tarefas que envolvem raciocínio e conhecimento técnico na área de Computação.

3.5 PROCEDIMENTO DE AVALIAÇÃO

A avaliação dos modelos de linguagem nesta pesquisa foi realizada com base em um processo sistemático e controlado, que visa mensurar com precisão a capacidade dos LLMs de resolver questões de múltipla escolha do ENADE nas áreas de Ciência da Computação, Engenharia da Computação e Sistemas de Informação. Inspirado por metodologias adotadas em trabalhos anteriores, como os de Nunes et al. (2023), Viegas et al. (2024) e RAPOSO et al. (2024), o procedimento foi adaptado à realidade e às características do ENADE.

As respostas fornecidas por cada modelo foram comparadas com os gabaritos oficiais disponibilizados pelo INEP, permitindo a avaliação com base na acurácia, definida como a razão entre o número de acertos e o total de questões respondidas. Essa métrica foi utilizada como principal indicador de desempenho dos modelos. Para garantir a padronização, foram consideradas apenas questões de múltipla escolha com alternativas claras. Questões com elementos visuais essenciais à sua resolução foram excluídas do conjunto avaliado, garantindo a equidade entre os modelos, sobretudo os que não processam imagens.

Além da acurácia global, o desempenho também foi analisado por curso de origem da questão (Ciência da Computação, Engenharia da Computação e Sistemas de Informação), permitindo observar possíveis variações no desempenho dos modelos em diferentes subdomínios da Computação. As análises estatísticas foram realizadas com base em ferramentas de estatística descritiva (como média, desvio padrão e distribuição de acertos), possibilitando uma compreensão mais detalhada do comportamento dos modelos.

Por fim, a metodologia adotada permite não apenas identificar o modelo com melhor desempenho geral, mas também compreender as forças e limitações de cada LLM no contexto de avaliação educacional técnica. O uso de uma plataforma própria, associada a um procedimento padronizado e reprodutível, assegura a validade, consistência e confiabilidade dos resultados obtidos ao longo da pesquisa.

3.6 FERRAMENTAS E RECURSOS UTILIZADOS

Para o desenvolvimento da plataforma de simulação de questionários utilizada nesta pesquisa, foram empregadas tecnologias modernas e consolidadas no desenvolvimento de aplicações web. A camada de backend foi construída utilizando o framework Java Spring Boot, devido à sua robustez, escalabilidade e suporte à arquitetura de microsserviços. A interface do sistema foi desenvolvida com React, biblioteca JavaScript amplamente utilizada para construção de interfaces dinâmicas e responsivas, proporcionando uma experiência interativa ao simular a resolução de questões pelo usuário. Para armazenamento dos dados, como o histórico das interações, questões, respostas e resultados, foi utilizado o MySQL, sistema gerenciador de banco de dados relacional que oferece confiabilidade e desempenho adequado para aplicações transacionais.

Além disso, foi desenvolvido um microsserviço específico para integração com os modelos de linguagem utilizados na pesquisa. Esse serviço foi implementado em Python, linguagem escolhida por sua extensa compatibilidade com bibliotecas de ciência de dados e inteligência artificial, além de sua facilidade de integração com APIs externas. O microsserviço atua como um worker assíncrono, responsável por receber as requisições da plataforma principal, enviar os prompts aos modelos via API REST e processar as respostas recebidas. Essa arquitetura desacoplada contribui para maior escalabilidade e permite que a avaliação dos modelos ocorra de forma eficiente e paralela, sem comprometer o desempenho da aplicação principal.

3.7 CRONOGRAMA

Tabela 1 – Cronograma de Execução do TCC (Julho a Dezembro)

Atividade	**Jul**	**Ago**	**Set**	**Out**	**Nov**	**Dez**
Reuniões de Orientação	X	X	X	X	X	X
Análise da literatura sobre LLMs para seleção de modelos	X					
Revisão e refinamento da Introdução e Fundamentação Teórica	X			X		
Preparação e organização do conjunto de questões do ENADE	X					
Revisão e refinamento da Metodologia		X				
Aplicação dos modelos às questões selecionadas		X	X			
Comparação das respostas com os gabaritos oficiais			X			
Análise dos resultados e discussão por curso e tema			X	X		
Redação da seção de resultados e considerações finais				X	X	
Revisão geral do texto e adequação às normas					X	
Preparação da apresentação (slides, roteiro etc.)					X	
Entrega da versão final do TCC						X
Apresentação e defesa do TCC						X

4 RESULTADOS

No capítulo atual, é apresentado tabelas com os dados coletados, os dados abrangem o tamanho do binário emitido pelos compiladores e quantidade de memória utilizada ao executa-los. Ademais, é apresentado estatísticas descritivas sobre o tempo de execução coletado. Por fim, nesse capítulo é analisado os dados apresentados visando responder as questões de pesquisa formuladas no capítulo inicial.

4.1 ANÁLISE DO TEMPO DE EXECUÇÃO

A análise do tempo de execução será realizada de forma semelhante a análise feita na seção anterior, isto é, para cada tripla(algoritmo, navegador e tamanho de entrada) será calculado a razão entre o tempo de execução apresentado pelo Cheerp sobre o tempo apresentado pelo seu rival.

Tabela 2 – Razões do tempo de execução

	Tempo Exec. (entrada grande)	Tempo Exec. (entrada média)
Média	1.578	2.095
Desvio p.	0.637	1.009
Min.	0.978	0.778
1° quartil	1.043	1.199
2° quartil	1.312	2.064
3° quartil	1.859	2.915
Max.	3.180	5.000

4.2 QUESTÕES DE PESQUISA

Diante da análise realiza, há informações necessárias para responder as questões de pesquisa enunciadas no capítulo inicial. Portanto, a seguir será respondido cada uma delas, utilizando as conclusões obtidas nesse capítulo.

QP01 Qual dos dois compiladores estudados emite um binário com tamanho menor?

Independente do tamanho da entrada, na média o Cheerp apresentou um binário 10% menor que o binário emitido pelo Emscripten. Ademais, a variação desse percentual foi muito pequena, logo, em todos os algoritmos utilizados esse resultado se mostrou verdadeiro.

QP02 Entre os dois, qual produz um binário que utiliza menos memória, considerando o tamanho inicial da memória igual para ambos?

5 CONCLUSÕES

Nessa monografia foi realizado uma comparação entre dois compiladores para a plataforma WebAssembly, são os compiladores Cheerp e Emscripten. A pesquisa teve objetivo de comparar a performance das duas ferramentas através de uma análise do tamanho do binário emitido pelos compiladores, do uso de memória e tempo de execução.

5.1 TRABALHOS FUTUROS

Tem um monte de coisa para fazer ainda, mas eu quero é meu canudo.

REFERÊNCIAS

BAHDANAU, Dzmitry; CHO, Kyunghyun; BENGIO, Yoshua. Neural Machine Translation by Jointly Learning to Align and Translate. In: INTERNATIONAL Conference on Learning Representations (ICLR). [S.l.*:* s.n.], 2015.

BENGIO, Yoshua et al. A neural probabilistic language model. **Journal of machine learning research**, v. 3, p. 1137–1155, 2003.

BRASIL. **Lei nº 10.861, de 14 de abril de 2004. Institui o Sistema Nacional de Avaliação da Educação Superior – SINAES e dá outras providências**. [S.l.*:* s.n.], 2004. `https://www.planalto.gov.br/ccivil_03/_ato2004-2006/2004/lei/l10.861.htm`. Acesso em: 2 jun. 2025.

BROWN, Peter F; PIETRA, Stephen A. Della et al. The mathematics of statistical machine translation: Parameter estimation. **Computational linguistics**, v. 19, n. 2, p. 263–311, 1993.

BROWN, Tom; MANN, Benjamin; RYDER, Nick et al. Language Models are Few-Shot Learners. **Advances in Neural Information Processing Systems**, 2020.

CHO, Kyunghyun et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In: PROCEEDINGS of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.*:* s.n.], 2014. p. 1724–1734.

DEVLIN, Jacob et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **arXiv preprint arXiv:1810.04805**, 2018.

DIAS SOBRINHO, José. **Avaliação da educação superior: democracia e construção da autonomia**. São Paulo: Cortez, 2003.

ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP), Instituto Nacional de. **Guia ENADE 2023**. [S.l.*:* s.n.], 2023. `https://www.gov.br/inep/pt-br/assuntos/avaliacao-e-exames-superiores/enade`. Acesso em: 2 jun. 2025.

EZE, Lucky. **O que são Large Language Models (LLM)? — bureauworks.com**. [S.l.*:* s.n.], 2025. `https://www.bureauworks.com/pt/blog/o-que-e-large-language-models-llm`. [Acessado em 04-06-2025].

GRISHMAN, Ralph; SUNDHEIM, Beth. Message Understanding Conference—6: A Brief History. In: COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics. [S.l.*:* s.n.], 1996.

JELINEK, Frederick. **Statistical Methods for Speech Recognition**. [S.l.]: MIT press, 1997.

JURAFSKY, Daniel; MARTIN, James H. **Speech and Language Processing**. 3. ed. [S.l.*:* s.n.], 2023. `https://web.stanford.edu/~jurafsky/slp3/`. (draft). Disponível em: `https://web.stanford.edu/~jurafsky/slp3/`. Acesso em: 9 jun. 2025.

JURAFSKY, Daniel; MARTIN, James H. **Speech and Language Processing**. 3. ed. [S.l.*:* s.n.], 2023. Draft version. Disponível em: `https://web.stanford.edu/~jurafsky/slp3/`. Acesso em: 9 jun. 2025.

KOEHN, Philipp et al. Moses: Open source toolkit for statistical machine translation. In: PROCEEDINGS of the ACL 2007 Demo and Poster Sessions. [S.l.*:* s.n.], 2007. p. 177–180.

LAFFERTY, John; MCCALLUM, Andrew; PEREIRA, Fernando CN. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: PROCEEDINGS of the 18th International Conference on Machine Learning (ICML). [S.l.*:* s.n.], 2001. p. 282–289.

LIEW, Pei Yee; TAN, Ian KT. On Automated Essay Grading using Large Language Models. In: PROCEEDINGS of the 2024 8th International Conference on Computer Science and Artificial Intelligence. [S.l.*:* s.n.], 2024. p. 204–211.

MARCUS, Mitchell P; MARCINKIEWICZ, Mary Ann; SANTORINI, Beatrice. Building a large annotated corpus of English: The Penn Treebank. **Computational linguistics**, v. 19, n. 2, p. 313–330, 1993.

MIKOLOV, Tomas et al. Efficient Estimation of Word Representations in Vector Space. **arXiv preprint arXiv:1301.3781**, 2013.

MOHAMED, Kareem et al. Hands-on analysis of using large language models for the auto evaluation of programming assignments. **Information Systems**, Elsevier, p. 102473, 2024.

NORMAN, Jeremy M. **The First Public Demonstration of Machine Translation Occurs : History of Information — historyofinformation.com**. [S.l.*:* s.n.], 2025. `https://www.historyofinformation.com/detail.php?id=666`. [Accessed 09-06-2025].

NUNES, Desnes et al. Evaluating GPT-3.5 and GPT-4 models on Brazilian university admission exams. **arXiv preprint arXiv:2303.17003**, 2023.

PENNINGTON, Jeffrey; SOCHER, Richard; MANNING, Christopher D. GloVe: Global Vectors for Word Representation. In: PROCEEDINGS of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.*:* s.n.], 2014. p. 1532–1543.

PERES, Rodrigo Silva. **Grandes Modelos de Linguagem na resolução de questões de vestibular: o caso dos institutos militares brasileiros**. 2023. Diss. (Mestrado).

RADFORD, Alec; WU, Jeffrey; CHILD, Rewon et al. Language Models are Unsupervised Multitask Learners. **OpenAI**, 2019.

RAPOSO, Lucas Brasileiro et al. Avaliação de LLMS na resolução de questões do ENEM. Universidade Federal de Campina Grande, 2024.

REIHANI, Ali A. **Understanding SHRDLU: A Pioneering AI in Language and Reasoning**. [S.l.*:* s.n.], 2025. `https://cryptlabs.com/understanding-shrdlu-a-pioneering-ai-in-language-and-reasoning/`. [Accessed 09-06-2025].

RODRIGUES, Luiz et al. LLMs Performance in Answering Educational Questions in Brazilian Portuguese: A Preliminary Analysis on LLMs Potential to Support Diverse Educational Needs. In: PROCEEDINGS of the 15th International Learning Analytics and Knowledge Conference. [S.l.*:* s.n.], 2025. p. 865–871.

TANG, Duyu; QIN, Bing; LIU, Ting. Effective LSTMs for Target-Dependent Sentiment Classification. In: PROCEEDINGS of COLING 2015. [S.l.*:* s.n.], 2015. p. 3298–3307.

VASWANI, Ashish et al. Attention is All You Need. **Advances in Neural Information Processing Systems**, 2017.

VIEGAS, Cayo Vinicíus et al. Avaliando a capacidade de LLMS na resolução de questões do POSCOMP. Universidade Federal de Campina Grande, 2024.

WALLACE, Michael. **Eliza, a chatbot therapist — web.njit.edu**. [S.l.*:* s.n.], 2016. `https://web.njit.edu/~ronkowit/eliza.html`. [Accessed 09-06-2025].