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 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.ª Me. Débora da Conceição Araújo

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. Me. Ricardo Azevedo Moreira da Silva
Coordenador do Curso

Banca Examinadora:

Prof.ª Me. Débora da Conceição Araújo
Presidente da Banca

Prof. Me. Marcos Vinicius Bião Cerqueira
Avaliador
Universidade Federal do Vale do São Francisco

Prof. Dr. Walter Felipe dos Santos
Avaliador
Universidade Federal do Vale do São Francisco

AGRADECIMENTOS

Agradeço a meu pai, a minha mãe, a meu cachorro e a minha orientadora.

RESUMO

Este trabalho apresenta uma avaliação comparativa do desempenho dos Grandes Modelos de Linguagem (LLM's) GPT, Gemini e DeepSeek na resolução de questões objetivas do Exame Nacional de Desempenho dos Estudantes (ENADE) voltadas aos cursos da área de Computação. Por meio de uma abordagem quantitativa e experimental, a pesquisa utilizou uma plataforma web desenvolvida especificamente para submeter questões disponíeis aos modelos, aplicando variações de temperatura entre 0.0 e 2.0 para analisar a precisão, o recall e a estabilidade das respostas em comparação aos gabaritos oficiais. Os resultados indicaram que, embora os modelos convirjam em desempenho na temperatura 0.8, o GPT consolidou-se como a ferramenta mais robusta e consistente, apresentando baixa sensibilidade a variações de parâmetros, enquanto o Gemini e o DeepSeek demonstraram superioridade em nichos específicos, como Segurança da Informação e Estruturas de Dados, respectivamente, apesar de serem mais instáveis em temperaturas elevadas. Conclui-se que não existe uma LLM universalmente superior para a Computação, mas sim perfis complementares, sendo observada uma limitação comum a todos os modelos na interpretação de questões visuais e diagramáticas, notadamente na área de Sistemas Operacionais.

Palavras-chave: Grandes Modelos de Linguagem (LLM). Inteligência Artificial. ENADE. Processamento de Linguagem Natural (PLN). Avaliação Educacional. Ensino de Computação.

ABSTRACT

This work presents a comparative evaluation of the performance of the Large Language Models (LLMs) GPT, Gemini, and DeepSeek in solving multiple-choice questions from the National Student Performance Exam (ENADE) aimed at Computing-related undergraduate programs. Through a quantitative and experimental approach, the research employed a web platform developed specifically to submit questions to the models, applying temperature variations between 0.0 and 2.0 to analyze accuracy, recall, and response stability in comparison to the official answer keys. The results indicated that although the models converge in performance at a temperature of 0.8, GPT established itself as the most robust and consistent tool, showing low sensitivity to parameter variations, while Gemini and DeepSeek demonstrated superiority in specific niches such as Information Security and Data Structures, respectively despite being more unstable at higher temperatures. It is concluded that there is no universally superior LLM for Computing; instead, the models exhibit complementary profiles, with a limitation shared by all of them in interpreting visual and diagrammatic questions, especially in the area of Operating Systems.

Keywords: Large Language Models (LLM's). Artificial Intelligence. ENADE. Natural Language Processing (NLP). Educational Assessment. Computing Education.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API	Interface de Programação de Aplicação (do inglês, *Application Programming Interface*)
BERT	Representações de Codificador Bidirecional de Transformadores (do inglês, *Bidirectional Encoder Representations from Transformers*)
BNCC	Base Nacional Comum Curricular
CoT	Cadeia de Pensamento (do inglês, *Chain-of-Thought*)
CPC	Conceito Preliminar de Curso
CRF	Campos Aleatórios Condicionais (do inglês, *Conditional Random Fields*)
DCN	Diretrizes Curriculares Nacionais
DS	DeepSeek (Sigla utilizada nas tabelas de resultados)
ENADE	Exame Nacional de Desempenho dos Estudantes
ENEM	Exame Nacional do Ensino Médio
GMN	Gemini (Sigla utilizada nas tabelas de resultados)
GPT	Transformador Pré-treinado Generativo (do inglês, *Generative Pre-trained Transformer*)
GPU	Unidade de Processamento Gráfico (do inglês, *Graphics Processing Unit*)
GRU	Unidade Recorrente com Portão (do inglês, *Gated Recurrent Unit*)
HMM	Modelo Oculto de Markov (do inglês, *Hidden Markov Model*)
IA	Inteligência Artificial
IE	Extração de Informação (do inglês, *Information Extraction*)
IGC	Índice Geral de Cursos
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
LLM	Grande Modelo de Linguagem (do inglês, *Large Language Model*)
LSA	Análise Semântica Latente (do inglês, *Latent Semantic Analysis*)
LSTM	Memória de Longo e Curto Prazo (do inglês, *Long Short-Term Memory*)
MEC	Ministério da Educação
MIT	Instituto de Tecnologia de Massachusetts (do inglês, *Massachusetts Institute of Technology*)
MoE	Mistura de Especialistas (do inglês, *Mixture-of-Experts*)
NER	Reconhecimento de Entidades Nomeadas (do inglês, *Named Entity Recognition*)
NLP	Processamento de Linguagem Natural (do inglês, *Natural Language Processing*)
PLN	Processamento de Linguagem Natural
POS	Etiquetagem Gramatical (do inglês, *Part-of-Speech Tagging*)

POSCOMP	Exame Nacional para Ingresso na Pós-Graduação em Computação
QP	Questão de Pesquisa
REST	Transferência de Estado Representacional (do inglês, *Representational State Transfer*)
RLHF	Aprendizado por Reforço com Feedback Humano (do inglês, *Reinforcement Learning from Human Feedback*)
RNN	Rede Neural Recorrente (do inglês, *Recurrent Neural Network*)
SINAES	Sistema Nacional de Avaliação da Educação Superior
SMT	Tradução Automática Estatística (do inglês, *Statistical Machine Translation*)
SVM	Máquina de Vetores de Suporte (do inglês, *Support Vector Machine*)

SUMÁRIO

1 INTRODUÇÃO

A evolução da Inteligência Artificial (IA) tem transformado paradigmas em diversos setores da sociedade, migrando de sistemas baseados em regras rígidas para modelos capazes de gerar conteúdo e interpretar nuances humanas (K; Bhadani; Pathak, 2025). Nesse cenário, o Processamento de Linguagem Natural (PLN) ganhou destaque com o surgimento dos Grandes Modelos de Linguagem, do inglês *Large Language Models* (LLM). Esses modelos, fundamentados em arquiteturas *Transformer* e treinados com volumes massivos de dados, superaram as limitações das tecnologias anteriores, demonstrando capacidades avançadas de compreensão, tradução e geração de texto coerente. Tal avanço permitiu que a IA deixasse de ser apenas uma ferramenta de classificação para se tornar um instrumento de apoio cognitivo em tarefas complexas (Mohamed et al., 2024).

Dentro desse espectro, o Processamento de Linguagem Natural (PLN) consolida-se como uma subárea interdisciplinar, situada na interseção entre a Ciência da Computação, a Inteligência Artificial e a Linguística (Tilton; Arnold, 2016). O objetivo primordial do PLN é capacitar sistemas computacionais a compreender, interpretar e manipular a linguagem humana de forma significativa, superando a barreira entre a comunicação natural e o código de máquina. Diferentemente das linguagens de programação, que são estruturadas e inequívocas, a linguagem natural é inerentemente complexa, repleta de ambiguidades semânticas, variações sintáticas e contextos implícitos. Historicamente, essa complexidade limitava os sistemas tradicionais, que dependiam de regras manuais rígidas e falhavam em capturar as nuances da comunicação humana real.

A superação dessas barreiras ocorreu com o advento de técnicas de Aprendizado Profundo (*Deep Learning*) e, mais recentemente, com a popularização das LLM's(Vaswani et al., 2017). Fundamentados majoritariamente na arquitetura *Transformer*, esses modelos utilizam mecanismos de atenção para ponderar a relevância de diferentes elementos em uma sequência textual, permitindo a compreensão de dependências de longo prazo e contextos complexos (Vaswani et al., 2017). Ao serem treinados em *datasets* massivos que abrangem grande parte do conhecimento disponível na internet, os LLM's adquiriram uma capacidade de generalização inédita. Isso permitiu que a tecnologia evoluísse de tarefas simples de classificação para a geração autônoma de texto, raciocínio lógico e adaptação a diferentes domínios de conhecimento sem a necessidade de retreinamento específico (Zhao et al., 2023).

Com todo o avanço na 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 ou correções que se caracterizam por serem detalhados, oportunos e de apoio, aspectos cruciais para o desenvolvimento e aprendizado do aluno (Liew; Tan, 2024).

O uso de LLM's 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 (Das et al., 2021). Nesse cenário, os LLM's 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.

Estudos recentes têm explorado a competência de LLM's em domínios de conhecimento especializados na área de educação médica(Kung et al., 2023), observa-se um crecente aumento no uso de LLM's para resolução de provas do Exame Nacional do Ensino Médio (ENEM), também o seu uso no Exame Nacional para Ingresso na Pós-Graduação em Computação (POSCOMP) e na medicina por resolução de questões de múltipla escolha nessa área (Superbi et al., 2024; Viegas; Gheyi; Ribeiro, 2025; Grévisse, 2024), no entanto há uma lacuna no que diz respeito ao uso dos LLM's especificamente em questões do Exame Nacional de Desempenho de Estudantes (ENADE) voltadas para cursos da área de Computação. Assim, este trabalho justifica-se pela necessidade de avaliar o desempenho de diferentes LLM's 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.

Para validar a eficácia dos modelos nesse nível de exigência, o ENADE apresenta-se como um instrumento de referência ideal. Instituído pelo Ministério da Educação (MEC), o exame transcende a simples verificação de memorização, sendo desenhado para avaliar o desenvolvimento de competências, habilidades e a capacidade de síntese dos graduandos frente a problemas reais da profissão (Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), 2023). Utilizar o ENADE como parâmetro de teste para as LLM's é relevante, pois suas questões frequentemente interdisciplinares e contextualizadas impõem um desafio de interpretação superior ao de exercícios convencionais. Então, verificar se esses modelos conseguem resolver satisfatoriamente tais questões é um passo decisivo para atestar sua viabilidade como ferramentas de apoio pedagógico no ensino superior de Computação.

Portanto, torna-se evidente que os LLM's representam uma oportunidade relevante para o aprimoramento de processos avaliativos no ensino superior, especialmente em áreas que exigem interpretação técnica e contextual, como a Computação. A escolha do ENADE como objeto de análise possibilita investigar o desempenho desses modelos em um cenário real, complexo e alinhado às competências esperadas de futuros profissionais. Assim, este trabalho propõe avaliar a eficácia de diferentes LLM's na resolução de questões do exame, buscando identificar suas potencialidades, limitações e contribuições para aplicações educacionais. A partir dessa análise, pretende-se oferecer subsídios para a incorporação responsável e eficiente dessas tecnologias em práticas pedagógicas, alinhando-se às

diretrizes globais para o uso de IA na educação (Holmes; Miao et al., 2023), e reforçando seu papel como ferramentas de apoio ao processo de aprendizagem e avaliação.

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 as 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 nos últimos 20 anos, 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 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 LLM's.

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 - GRANDE MODELO DE LINGUAGEM

Grandes Modelos de Linguagem (*Large Language Models*, LLM's) representam uma quebra de paradigma no uso da Inteligência Artificial (IA)(RAPOSO et al., 2024). Os LLM's 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 LLM's 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 LLM's 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 LLM's 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 LLM'S

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 décadas 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 décadas 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 (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 MODELOS

2.4.1 GPT

O GPT (Generative Pre-trained Transformer) constitui uma família de modelos de linguagem desenvolvida pela OpenAI, baseada na arquitetura Transformer proposta inicialmente por (Vaswani et al.). Sua implementação segue especificamente a variante decoder-only, na qual todas as camadas do modelo atuam como blocos de decodificação com mecanismos de self-attention unidirecional (OpenAI et al., 2024). Esse tipo de arquitetura permite que o modelo opere de forma auto-regressiva, gerando cada token com base na probabilidade condicional do próximo elemento dada a sequência anterior, conforme amplamente adotado em modelos autoregressivos de geração de linguagem (Brown; Mann; Ryder et al., 2020). O processo de treinamento do GPT é tradicionalmente dividido em duas fases. A primeira é o pré-treinamento, no qual o modelo é exposto a grandes volumes de dados textuais não estruturados com o objetivo de aprender regularidades linguísticas e padrões estatísticos gerais por meio de aprendizado auto-supervisionado (Brown; Mann; Ryder et al., 2020). A segunda etapa consiste no ajuste fino (fine-tuning), que pode ser realizado de maneira supervisionada ou mediante técnicas de alinhamento com feedback humano, como o RLHF (Reinforcement Learning from Human Feedback), visando adaptar o comportamento do modelo para tarefas específicas e aumentar sua segurança e confiabilidade (OpenAI et al., 2024). Ao longo de suas versões como GPT-3, GPT-4 e GPT-5 esses modelos demonstraram avanços substanciais em raciocínio, coerência discursiva e capacidade de seguir instruções complexas, estabelecendo um novo paradigma na área de LLM's. Pesquisas e relatórios técnicos da OpenAI destacam a evolução contínua em desempenho, robustez e alinhamento ético, definindo o GPT como uma das arquiteturas de referência no desenvolvimento de sistemas de IA generativa (OpenAI et al., 2024).

2.4.2 Gemini

O Gemini é um modelo multimodal de grande escala desenvolvido pela Google DeepMind, concebido desde o início para processar de forma integrada múltiplas modalidades, incluindo texto, imagens, áudio, vídeo e código (Team et al., 2025). Em contraste com modelos estritamente auto-regressivos baseados apenas em arquiteturas decoder-only, como o GPT, o Gemini adota uma arquitetura multimodal unificada, estruturada por componentes de encoders e decoders, permitindo que diferentes modalidades sejam representadas e integradas de maneira coerente antes da geração de saídas (Team et al., 2025). Essa arquitetura híbrida possibilita ao modelo realizar raciocínio multimodal de maneira mais eficiente e com maior profundidade contextual, uma vez que o sistema é capaz de analisar relações entre sinais perceptuais e linguísticos simultaneamente, evitando a necessidade de modelos auxiliares ou módulos externos de visão computacional. De acordo com o relatório técnico do Gemini, essa integração arquitetural melhora substancialmente

o desempenho em tarefas que envolvem interpretação de imagens, análise de vídeos, reconhecimento de áudio e compreensão combinada de múltiplos formatos de dados (Team et al., 2025). O treinamento do Gemini também se distingue pelo uso de um conjunto de dados multimodal de larga escala, composto por textos, imagens, documentos, áudio e conteúdos de código, o que permite ao modelo desenvolver representações alinhadas entre modalidades distintas. Essa diversidade de dados promove melhor capacidade de generalização e torna o modelo particularmente adequado para aplicações que requerem integração entre informação verbal e perceptual, como sistemas de análise visual, interpretação automática de vídeos, resolução de problemas multimodais e suporte a tarefas complexas que envolvem diferentes formas de representação simbólica (Team et al., 2025).

2.4.3 DeepSeek

O DeepSeek é um modelo de linguagem de código aberto desenvolvido pela DeepSeek AI que se destacou por apresentar alto desempenho aliado a um custo computacional significativamente reduzido, tornando-se uma alternativa competitiva em relação a modelos proprietários de larga escala (DeepSeek-AI; Bi; Chen et al., 2024). Assim como outros modelos autoregressivos modernos, sua arquitetura segue o paradigma decoder-only Transformer, originalmente baseado na formulação de (Vaswani et al.), porém incorpora um conjunto de otimizações estruturais e de treinamento voltadas para a eficiência (DeepSeek-AI; Bi; Chen et al., 2024). Entre as principais técnicas adotadas está o DeepSeekMoE, uma arquitetura inovadora de Mixture-of-Experts (MoE) que utiliza segmentação de especialistas em grão fino (fine-grained expert segmentation) e isolamento de especialistas compartilhados. Essa abordagem permite reduzir drasticamente a quantidade de parâmetros ativados durante a inferência, preservando a expressividade do modelo e diminuindo o custo computacional sem comprometer o desempenho geral (Dai; Deng; Zhao et al., 2024). Além disso, o DeepSeek utiliza otimizações avançadas de paralelismo, permitindo aproveitar de forma mais eficiente múltiplas GPUs ou hardwares equivalentes durante o treinamento (DeepSeek-AI; Bi; Chen et al., 2024). Esse tipo de estratégia reduz significativamente o tempo de convergência e o custo operacional em comparação com arquiteturas tradicionais densas. O modelo também foi projetado para execução eficiente e democratização do acesso, ampliando sua adoção pela comunidade acadêmica e por organizações sem acesso a infraestrutura computacional de grande porte (Dai; Deng; Zhao et al., 2024). Dessa forma, o DeepSeek se consolida como um exemplo relevante de como arquiteturas abertas e otimizadas podem viabilizar o uso de modelos de linguagem de grande escala.

2.5 TRABALHOS CORRELATOS

Nunes et al. (2023) em seu trabalho 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* valiosos 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 LLM's 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 (LLM's) em responder a questões objetivas do Exame Nacional do Ensino Médio (ENEM). Os LLM's 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 LLM's se limita ao contexto da língua inglesa, sem testes eficazes em cenários globalizados como o brasileiro. Nessa pesquisa utilizou-se as LLM's: 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 LLM's e também fazer váriações na temperatura das respostas entregues pela as mesmas. No geral dos acertos o Gemini obteve um percen-

tual maior nas questões com a temperatura definida como a padrão das LLM's, 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 quantidade de acertos, no entando, o Gemini ainda se saiu superior aos 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 LLM's mostraram maior dificuldade em Matemática e Ciências da Natureza.

Rodrigues et al. (2025) investigaram a capacidade das LLM's em sistemas de resposta a perguntas educacionais, que facilitam o aprendizado adaptativo e respondem às dúvidas dos alunos. O estudo investiga como as LLM's 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 LLM's 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 LLM's, 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 detalha 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 LLM's 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 (LLM's) 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 LLM's, 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 LLM's 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 LLM's 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.

4 RESULTADOS

O presente capítulo tem como propósito apresentar, contextualizar e analisar de maneira sistemática os resultados obtidos a partir do envio das questões aos modelos de linguagem (LLM's) e da subsequente coleta de suas respostas. Busca-se, assim, oferecer uma interpretação rigorosa e fundamentada dos dados produzidos pelos experimentos, evidenciando padrões, discrepâncias e insights relevantes para a compreensão do desempenho dos modelos avaliados. Além disso, este capítulo contribui para o aprofundamento da discussão sobre a eficácia das LLM's em cenários distintos de complexidade, parametrização e áreas do conhecimento, estabelecendo conexões diretas com os objetivos específicos delineados na etapa metodológica deste trabalho.

Para garantir a clareza da exposição e atender aos objetivos propostos, a análise dos resultados foi organizada em quatro etapas principais. Inicialmente, realiza-se a avaliação da taxa de acertos dos modelos ao longo dos anos e em diferentes configurações de temperatura. Em seguida, examina-se a acurácia dos modelos considerando exclusivamente a variação de temperatura. Posteriormente, é conduzida a análise do *recall*, também por temperatura, a fim de observar o comportamento dos modelos na recuperação correta de respostas. Por fim, apresenta-se um heatmap referente à taxa média de erro por área e por modelo, seguido de outros três heatmaps específicos para cada modelo, destinados a ilustrar a acurácia por temperatura e por área do conhecimento.

4.0.1 Taxa de Acerto Por Ano

A seguir, inicia-se a discussão pelos resultados relacionados à taxa de acertos dos modelos ao longo dos anos e das diferentes configurações de temperatura.

Tabela 1 – Taxa de acertos dos modelos por ano e temperatura

		Temperatura					
Ano	**Modelo**	**0.0**	**0.4**	**0.8**	**1.2**	**1.6**	**2.0**
2014	**GPT**	0.76%	0.76%	0.76%	0.76%	0.76%	0.69%
	GMN	0.72%	0.72%	0.76%	0.72%	0.72%	0.76%
	DS	*0.79%*	*0.79%*	*0.79%*	*0.79%*	*0.76%*	*0.83%*
2017	**GPT**	0.85%	0.85%	0.85%	0.85%	0.85%	0.90%
	GMN	*0.95%*	*0.95%*	*0.95%*	*1.00%*	*0.95%*	*1.00%*
	DS	0.90%	0.90%	0.90%	0.90%	0.85%	1.00%
2021	**GPT**	*0.91%*	*0.91%*	*0.91%*	*0.91%*	*0.91%*	*0.87%*
	GMN	0.74%	0.78%	0.83%	0.83%	0.74%	0.78%
	DS	0.78%	0.78%	0.83%	0.78%	0.74%	0.74%

A Tabela 1 apresenta as taxas de acerto dos modelos GPT, Gemini (GMN) e DeepSeek (DS) em diferentes anos 2014, 2017 e 2021 sob variações de temperatura de 0.0 a 2.0. Os valores mostrados estão entre 0 e 1, representando proporções de acerto, e refletem o desempenho de cada modelo sob diferentes condições de geração.

Em 2014, os três modelos exibem desempenhos muito próximos, com taxas situadas entre 0.69 e 0.83. O DeepSeek (DS) apresenta uma leve vantagem em relação aos demais, alcançando até 0.83 em temperatura 2.0, enquanto o GPT e o Gemini (GMN) mantêm-se próximos de 0.76 e 0.72, respectivamente. Nessa fase inicial, o DS se destaca pela consistência, enquanto o Gemini apresenta pequenas oscilações entre as temperaturas, sugerindo maior sensibilidade ao parâmetro.

No ano de 2017, observa-se uma melhora significativa no desempenho de todos os modelos. O Gemini (GMN) passa a liderar, atingindo valores de até 1.00, o que o coloca à frente do GPT e do DS em praticamente todas as temperaturas. O GPT, por sua vez, mantém valores em torno de 0.85, demonstrando boa estabilidade e confiabilidade nos resultados. O DeepSeek (DS) também mostra progresso em relação a 2014, chegando a 1.00 em temperatura 2.0, mas sem a consistência observada no Gemini, que apresenta desempenho sólido em todo o intervalo.

Em 2021, ocorre uma inversão de cenário. O GPT mostra evolução contínua ao longo dos anos e se torna o modelo com o melhor desempenho geral, alcançando cerca de 0.91 em praticamente todas as temperaturas. Em contraste, o Gemini (GMN) e o DeepSeek (DS) apresentam queda de desempenho, atingindo no máximo 0.83. Essa mudança indica que o GPT conseguiu aprimorar sua arquitetura e manter estabilidade, enquanto os outros dois modelos mostraram sinais de retrocesso ou falta de adaptação a novas condições.

Quando analisamos o efeito da temperatura, percebe-se que, para temperaturas mais baixas (entre 0.0 e 0.8), todos os modelos mantêm resultados estáveis, o que sugere que a aleatoriedade controlada não afeta significativamente o desempenho. Entretanto, em temperaturas mais altas (acima de 1.6), o Gemini e o DeepSeek apresentam pequenas quedas, enquanto o GPT mantém quase a mesma taxa de acerto, demonstrando maior robustez a variações de temperatura.

De modo geral, o GPT evidencia uma trajetória de crescimento constante entre 2014 e 2021, consolidando-se como o modelo mais estável e preciso. O Gemini (GMN) mostra um pico de desempenho em 2017, mas perde eficiência posteriormente, enquanto o DeepSeek (DS) mantém consistência, porém sem avanço expressivo. Assim o GPT demonstra melhor adaptação, estabilidade e evolução ao longo do tempo, destacando-se como o modelo mais equilibrado.

Com base na taxa de acertos dos modelos ao longo dos anos e nas diferentes configurações de temperatura, não é possível determinar qual deles apresentou desempenho consistentemente superior, uma vez que, em cada ano analisado, um modelo específico mostrou-se levemente acima dos demais. Observa-se que, em 2014, o DeepSeek apresentou

maior precisão; em 2017, o destaque foi o Gemini; e, em 2021, o GPT obteve o melhor desempenho. Esse comportamento também se confirma ao se analisar os resultados com temperatura 0,8, reconhecida na literatura como um parâmetro mais equilibrado para comparação entre modelos.

Entretanto, o GPT demonstra maior consistência ao longo da variação das temperaturas, apresentando oscilações menores em comparação aos demais modelos. Tal característica sugere que sua arquitetura baseada no paradigma *decoder-only* pode exercer influência na estabilidade das respostas, mantendo níveis de coerência mais elevados mesmo diante do aumento da criatividade induzido pela temperatura.

4.0.2 Acurácia por Temperatura

A Tabela 2 apresenta a acurácia média dos modelos GPT, Gemini (GMN) e DeepSeek (DS) em diferentes níveis de temperatura, variando de 0.0 a 2.0. Os valores estão entre 0 e 1, representando proporções de acerto e não porcentagens. Essa análise permite observar o comportamento de cada modelo diante da variação da temperatura, que influencia o grau de aleatoriedade na geração de respostas.

Tabela 2 – Acurácia dos modelos por temperatura

	Temperatura					
Modelo	**0.0**	**0.4**	**0.8**	**1.2**	**1.6**	**2.0**
GPT	*0.83%*	*0.83%*	*0.83%*	*0.83%*	*0.83%*	*0.81%*
GMN	0.79%	0.81%	0.83%	0.83%	0.79%	0.83%
DS	0.82%	0.82%	0.83%	0.82%	0.78%	0.85%

De forma geral, o GPT apresenta o desempenho mais consistente e elevado entre os três modelos. Sua acurácia permanece praticamente estável em torno de 0.83 em todas as temperaturas, com uma ligeira redução para 0.81 em 2.0. Essa estabilidade indica que o modelo é pouco sensível à variação de temperatura, mantendo boa performance mesmo em condições de maior aleatoriedade.

O Gemini (GMN) apresenta resultados também estáveis, variando entre 0.79 e 0.83. Ele atinge seu melhor desempenho entre as temperaturas 0.8 e 1.2, mas apresenta pequenas quedas em 0.0 e 1.6, o que sugere leve sensibilidade à variação do parâmetro. Apesar disso, o comportamento geral é equilibrado, sem flutuações bruscas.

O DeepSeek (DS) mostra desempenho semelhante ao Gemini, mas com pequenas oscilações mais perceptíveis. Seu valor mínimo ocorre em 1.6 (0.78), e o máximo em 2.0 (0.85), indicando que o modelo tende a reagir melhor em temperaturas mais altas. Essa variação pode refletir uma arquitetura mais sensível à aleatoriedade, o que, em alguns casos, contribui para ganhos marginais de acurácia.

Os resultados ressaltam que os três modelos mantêm um nível de desempenho bastante próximo e estável em todas as temperaturas. O GPT continua se destacando como o modelo mais robusto e consistente, apresentando pequenas variações e a melhor acurácia média geral. O Gemini e o DeepSeek seguem próximos, com leve vantagem do DeepSeek em temperatura mais alta, enquanto o Gemini se mantém mais estável nas intermediárias. Esses dados sugerem que todos os modelos lidam bem com variações de temperatura, mas o GPT demonstra o equilíbrio mais sólido entre estabilidade e desempenho.

Ao analisar a acurácia exclusivamente em função da temperatura aplicada aos modelos, observa-se uma convergência entre eles quando a temperatura é configurada em 0,8. Esse padrão indica que, ao submeter qualquer questão nessa configuração, os modelos tendem a apresentar probabilidades de acerto semelhantes. No entanto, ao se considerar a variação completa das temperaturas, nota-se que, assim como verificado na análise da taxa de acertos por ano e temperatura apresentada anteriormente, o GPT demonstrou maior consistência, exibindo apenas uma leve redução de desempenho quando submetido ao limite máximo de criatividade (temperatura 2,0).

Em contraste, os demais modelos apresentaram oscilações significativamente maiores na acurácia ao longo das diferentes configurações de temperatura. Esses resultados reforçam a percepção de que o GPT mantém maior estabilidade diante de cenários de maior aleatoriedade, enquanto DeepSeek e Gemini sofrem variações mais acentuadas em suas respostas.

4.0.3 *Recall* Por Temperatura

A Tabela 3 apresenta o *recall* dos modelos GPT, GMN e DS em diferentes temperaturas, variando de 0,0 a 2,0. Assim como observado na análise de precisão, os valores de *recall* também se mantêm bastante estáveis, indicando que a variação da temperatura exerce pouca influência sobre o desempenho dos modelos nesse aspecto.

Tabela 3 – *Recall* dos modelos por temperatura

	Temperatura					
Modelo	**0.0**	**0.4**	**0.8**	**1.2**	**1.6**	**2.0**
GPT	*0.82%*	*0.82%*	*0.82%*	*0.82%*	*0.82%*	*0.80%*
GMN	0.78%	0.80%	0.83%	0.82%	0.78%	0.83%
DS	0.81%	0.81%	0.82%	0.81%	0.76%	0.83%

O modelo GPT demonstrou o comportamento mais consistente entre os três, mantendo um *recall* de 0,82% em todas as temperaturas, com exceção da temperatura 2,0, em que houve uma leve queda para 0,80%. Esse resultado reforça a estabilidade do GPT,

mostrando que ele é pouco afetado por mudanças no parâmetro de temperatura e consegue manter um desempenho previsível e uniforme.

O modelo GMN, por sua vez, apresentou uma leve oscilação. Ele começou com 0,78% em temperatura 0,0, alcançou 0,83% em 0,8 e finalizou com o mesmo valor em 2,0. Apesar das pequenas variações, o GMN demonstrou uma tendência positiva em temperaturas mais altas, o que sugere que este modelo pode apresentar um ligeiro ganho de sensibilidade e capacidade de recuperação de informações conforme a temperatura aumenta.

Já o modelo DS apresentou um comportamento um pouco mais variável. Iniciando com 0,81%, o *recall* do DS se manteve estável até a temperatura 1,2, quando apresentou uma pequena queda para 0,76% em 1,6 e, depois, se recuperou para 0,83% na temperatura 2,0. Esse padrão indica que o modelo DS, assim como na análise de precisão, é o mais sensível às variações de temperatura, mas pode alcançar melhores resultados em níveis mais elevados.

O *recall* dos três modelos manteve-se em níveis semelhantes e estáveis, com pequenas flutuações. O GPT se destaca pela constância, o GMN mostra leve melhora em temperaturas mais altas, e o DS apresenta maior variabilidade, mas também potencial de melhor desempenho quando a temperatura é aumentada.

O GPT demonstra ser a opção mais robusta e previsível, mantendo um *recall* constante de 0,82% em praticamente todas as faixas de temperatura, apresentando apenas uma leve redução para 0,80% no limite superior (temperatura 2,0). Essa estabilidade pode ser atribuída à sua arquitetura *decoder-only*, amplamente otimizada para respostas coerentes mesmo sob maior aleatoriedade, tornando-o especialmente adequado para aplicações que demandam consistência e baixo nível de variabilidade.

No entanto, os modelos GMN e DS mostram-se mais voláteis e sensíveis à variação dos parâmetros, registrando quedas expressivas de desempenho na temperatura 1,6 ponto em que o DS atinge seu valor mínimo de *recall* 0,76%. Essa maior oscilação pode estar relacionada às arquiteturas mais complexas e sensíveis desses modelos, que incluem mecanismos de atenção esparsa ou estratégias internas de paralelização que ampliam a influência da temperatura no processo de geração. Contudo, essa instabilidade é compensada pelo fato de ambos alcançarem o maior índice de eficácia da tabela 0,83% nas temperaturas 0,8 e 2,0, refletindo o potencial dessas arquiteturas para superar o GPT em condições ideais. Dessa forma, enquanto o GPT se destaca como a escolha mais segura para resultados uniformes, os modelos GMN e DS revelam maior potencial de recuperação de dados quando suas temperaturas são ajustadas adequadamente para seus pontos ótimos.

4.0.4 Análise por área do conhecimento

A seguir, apresentamos os resultados estratificados por áreas de conteúdo, sem a segregação por ano de prova. Essa abordagem agregada busca isolar a variável temática,

possibilitando avaliar se determinados modelos possuem maior aptidão para componentes específicos de Ciência da Computação.

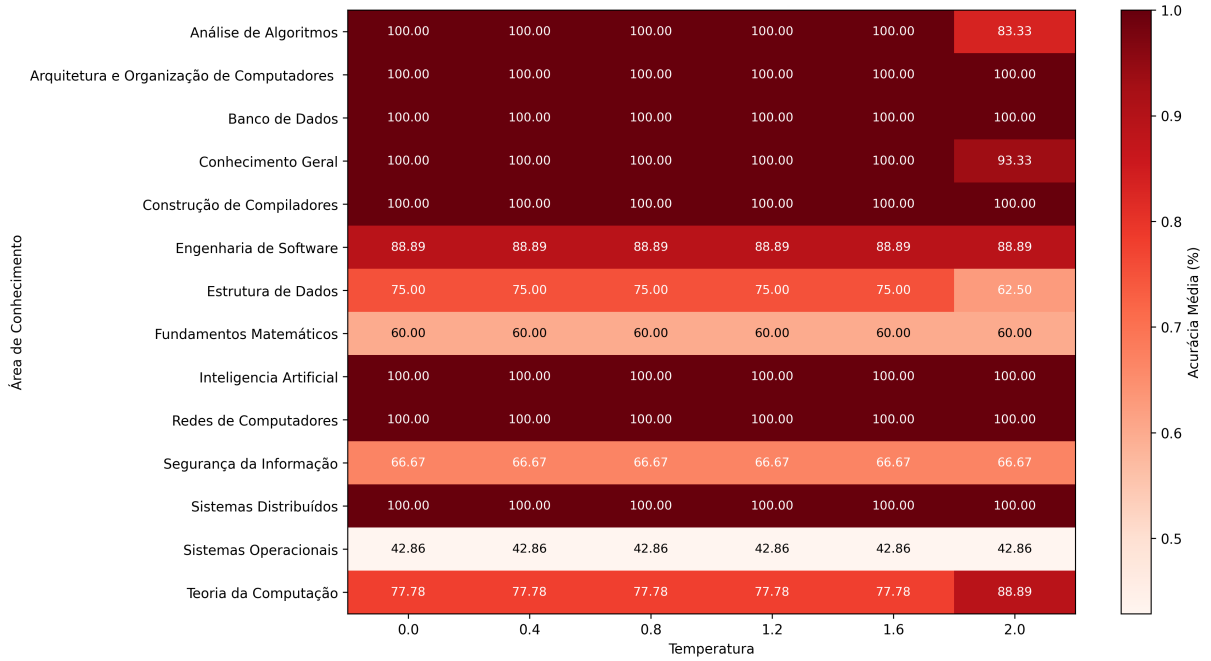

Figura 1 – Heatmap de acurácia por área de conhecimento GPT

Na figura 1, está presente o heatmap mostra que o modelo GPT apresenta um desempenho alto na maioria das áreas de conhecimento avaliadas. Campos como Análise de Algoritmos, Arquitetura e Organização de Computadores, Banco de Dados, Conhecimento Geral, Construção de Compiladores, Inteligência Artificial, Redes de Computadores e Sistemas Distribuídos mantêm acurácia de 100% em praticamente todas as temperaturas, isso indica que o modelo possui domínio consistente desses assuntos. Essa estabilidade sugere que são áreas amplamente cobertas nos dados de treinamento e com estruturas conceituais bem definidas.

No entanto, algumas áreas apresentam desempenho elevado, mas não perfeito. Engenharia de Software se mantém estável com cerca de 88,89% de acurácia, enquanto Segurança da Informação permanece em 66,67%. Esses valores mostram que o modelo compreende bem esses domínios, porém ainda há espaço para imprecisões, possivelmente devido à natureza interpretativa ou multidimensional dessas áreas. Porém Teoria da Computação apresenta: acurácia de 77,78% em temperaturas baixas e intermediárias, mas melhora para 88,89% quando a temperatura chega a 2.0, sugerindo que um pouco mais de variabilidade e criatividade nas respostas pode, de fato, auxiliar em perguntas mais conceituais.

Algumas áreas, no entanto, demonstram fragilidades claras. Fundamentos Matemáticos permanece constante em 60% de acurácia em todas as temperaturas, evidenciando dificuldades do modelo em lidar com conteúdo matemático mais formal ou rígido. Estru-

tura de Dados inicia com 75% e mantém esse valor até a temperatura 1.6, mas baixa para 62,5% na temperatura mais alta, indicando maior sensibilidade ao aumento de aleatoriedade. O pior desempenho é observado em Sistemas Operacionais, com 42,86% constantes em todas as temperaturas. Esse resultado sugere que o GPT tem dificuldade em gerar respostas precisas nessa área, independentemente do nível de variabilidade imposto pela temperatura.

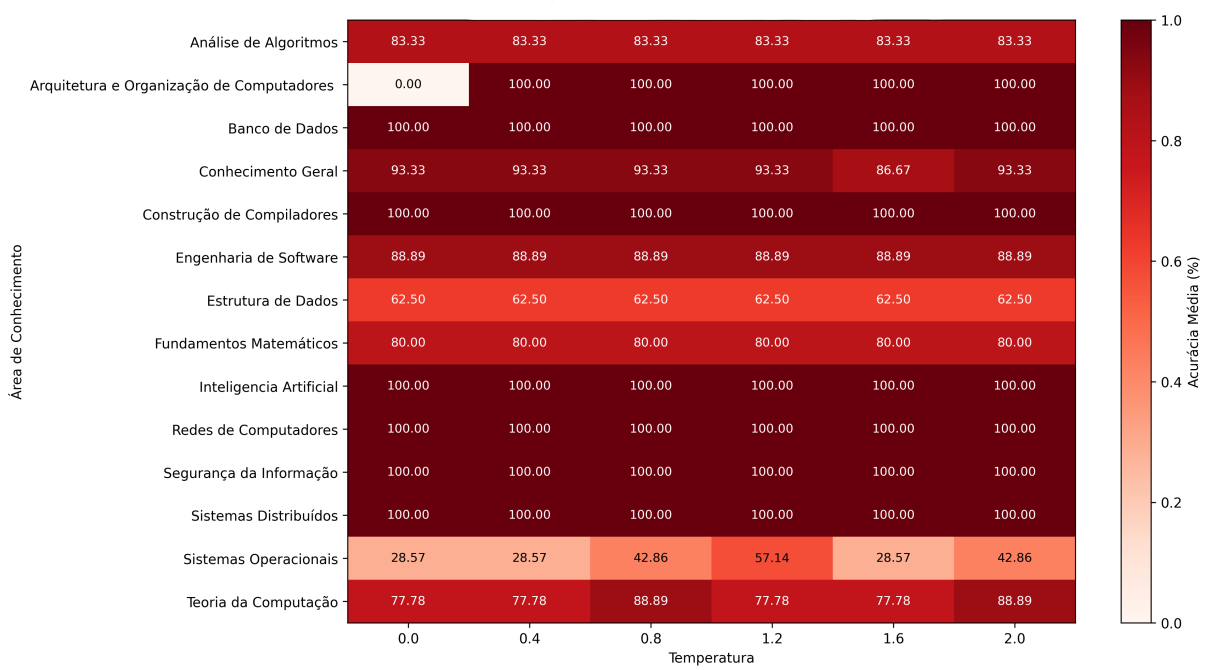

Figura 2 – Heatmap de acurácia por área de conhecimento Gemini

Na figura 2, está presente o heatmap de acurácia do modelo Gemini revela um comportamento bastante consistente em várias áreas de conhecimento, mas também apresenta pontos de instabilidade marcantes, especialmente em baixas temperaturas. A maioria das disciplinas apresenta desempenho excelente, com acurácia de 100% em praticamente todas as faixas de temperatura incluindo Banco de Dados, Construção de Compiladores, Inteligência Artificial, Redes de Computadores, Segurança da Informação e Sistemas Distribuídos. Esse padrão mostra que o Gemini domina bem esses temas e mantém respostas corretas mesmo quando o nível de aleatoriedade é aumentado. Esses resultados sugerem que o modelo possui uma base sólida e confiável nas áreas mais estruturadas e amplamente abordadas em ciência da computação.

Algumas áreas apresentam acurácia alta, mas não perfeita, mantendo valores estáveis e confiáveis ao longo das temperaturas. É o caso de Engenharia de Software, que permanece em 88,89%, e de Teoria da Computação, que varia entre 77,78% e 88,89%, mostrando leve sensibilidade à variação da temperatura, mas ainda demonstrando boa performance geral. O Conhecimento Geral também se destaca com uma acurácia próxima de 93%, com uma pequena queda na temperatura 1.6. Esses comportamentos sugerem que o

modelo compreende bem esses domínios, mas eles ainda possuem elementos interpretativos, em que pequenas variações de temperatura podem influenciar o resultado.

No entanto, algumas áreas se destacam pela baixa acurácia ou por variações abruptas. Um caso bastante evidente é o de Arquitetura e Organização de Computadores, que apresenta 0% de acurácia na temperatura 0.0 e 100% em todas as demais temperaturas. Essa discrepância indica que o modelo pode ter dificuldades em fornecer respostas determinísticas nessa área, mas melhora drasticamente mesmo com pequenas variações de aleatoriedade. Sistemas Operacionais também merece atenção: embora a acurácia suba para 57,14% na temperatura 1.2, ela permanece baixa nas demais faixas, entre 28,57% e 42,86%. Esse comportamento revela instabilidade e falta de domínio pleno do conteúdo, semelhante ao observado no heatmap do GPT. Estrutura de Dados, por sua vez, tem-se acurácia constante de 62,5%, o que indica um desempenho moderado e estável, mas inferior ao ideal.

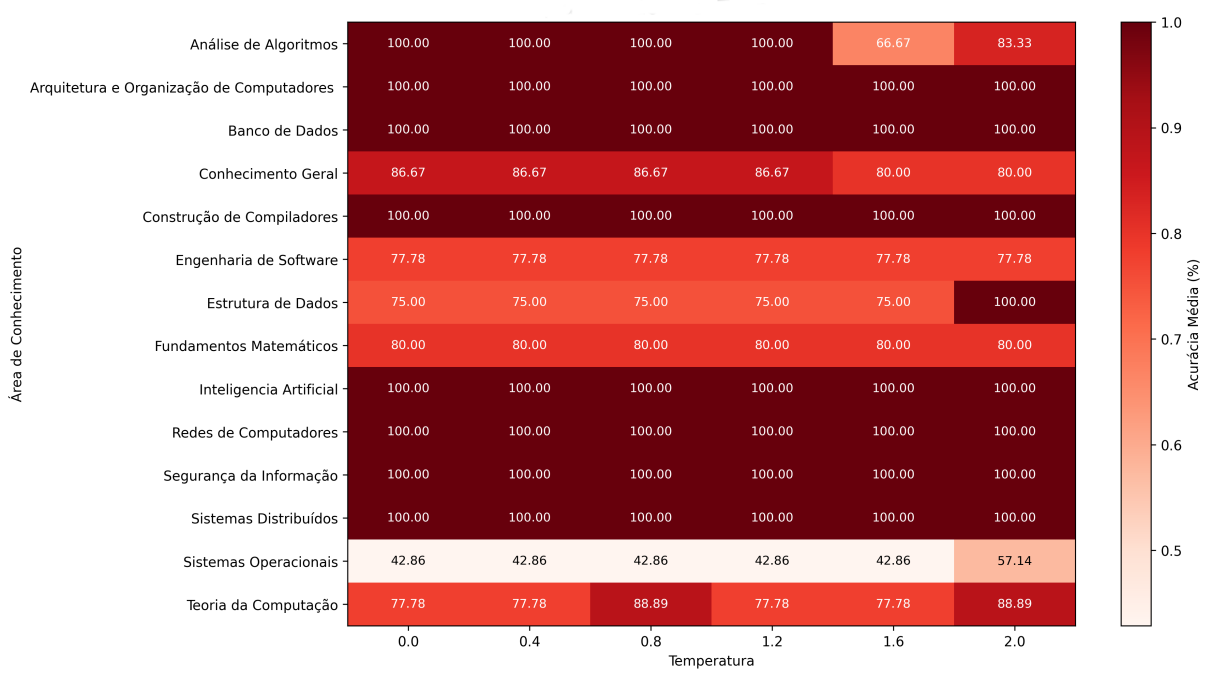

Figura 3 – Heatmap de acurácia por área de conhecimento DeepSeek

Por fim na figura 3, está presente o heatmap de acurácia do DeepSeek revela um desempenho relevante na maioria das áreas de conhecimento, mas também apresenta pontos de oscilação relevantes conforme a temperatura aumenta. Em grande parte das disciplinas como Banco de Dados, Construção de Compiladores, Inteligência Artificial, Redes de Computadores, Segurança da Informação e Sistemas Distribuídos; o modelo mantém uma acurácia de 100% em todas as temperaturas, demonstrando domínio sólido nessas áreas, mesmo com variações na aleatoriedade das respostas.

Assim como o GPT e o Gemini, em algumas áreas mantêm desempenho elevado, mas não perfeito. Engenharia de Software e Teoria da Computação apresentam acurácia estável

de cerca de 77,78%, enquanto Estrutura de Dados se mantém em 75% até a temperatura 1.6, subindo para 100% na temperatura 2.0 um comportamento incomum, mas que sugere que respostas um pouco mais criativas ou exploratórias podem ajudar o modelo nessa área específica. Em Fundamentos Matemáticos, apresenta acurácia constante de 80%, mostrando um desempenho moderadamente forte. Conhecimento Geral sofre uma leve queda em temperaturas mais altas, indo de 86,67% para 80%, o que indica sensibilidade moderada à aleatoriedade.

Entretanto, duas áreas chamam atenção por comportamentos instáveis ou baixos. Análise de Algoritmos apresenta uma queda significativa na temperatura 1.6, despencando de 100% para 66,67%, embora recupere parcialmente para 83,33% em 2.0. Esse comportamento sugere que, nessa disciplina, maior aleatoriedade prejudica a precisão do modelo. O caso mais crítico continua sendo Sistemas Operacionais, que apresenta acurácia baixa e praticamente estável, variando entre 42,86% e 57,14% dependendo da temperatura. Assim como nos outros modelos comparados, DeepSeek também demonstra fragilidade nessa área, indicando possível lacuna no treinamento ou dificuldade intrínseca do modelo em lidar com perguntas específicas desse domínio.

4.0.4.1 Conclusão da avaliação por área do conhecimento

Ao analisar os heatmaps das Figuras 1, 2 e 3, observa-se que existem áreas em que ambos os modelos apresentam elevado desempenho e respondem corretamente de forma consistente. Entre essas áreas de conhecimento destacam-se Bancos de Dados, Construção de Compiladores, Inteligência Artificial, Redes de Computadores e Sistemas Distribuídos. Esse comportamento sugere que, durante o processo de treinamento, os modelos provavelmente foram expostos a um corpus amplo e proporcionalmente semelhante nesses domínios, o que pode explicar a elevada consistência de acertos observada em ambos os cenários.

A baixa acurácia geral observada na categoria de Sistemas Operacionais entre 28% e57% indica que essa área apresenta desafios particulares para os modelos, especialmente porque muitas questões dependem de diagramas, visualizações de estados de memória ou operações específicas de hardware, que são de difícil interpretação apenas por meio de texto. Além disso, a própria natureza dos Sistemas Operacionais contribui para essa dificuldade, pois o campo abrange múltiplas implementações e perguntas pouco específicas podem induzir o modelo a adotar um contexto incorreto, comprometendo a precisão das respostas.

Nota-se também que o GPT apresentou o pior desempenho entre os três modelos na área de Segurança da Informação. Esse resultado sugere que Gemini e DeepSeek provavelmente foram expostos a um volume maior de conteúdos específicos dessa área durante o treinamento, enquanto o GPT, por possuir um caráter mais generalista, não conseguiu se adaptar com a mesma eficácia. Outro fator que pode contribuir para esse

desempenho inferior relaciona-se às próprias limitações impostas pela OpenAI, que adota políticas mais restritivas para evitar que o modelo forneça informações potencialmente sensíveis ou prejudiciais nesse domínio. Essas restrições acabam reduzindo a abrangência e a assertividade das respostas do GPT em tópicos de segurança, o que se reflete diretamente nos resultados obtidos.

O GPT se torna um modelo ideal em análise de algoritmos e situações em que a consistência das respostas é essencial. Ele demonstrou ser o modelo mais estável e confiável no uso generalista, apresentando excelente desempenho na interpretação de algoritmos, especialmente em temperaturas mais acertivas e menos criativas, nas quais manteve médias superiores às do DeepSeek. Além disso, foi o único modelo que atingiu desempenho máximo de forma consistente em conhecimentos gerais, mostrando-se particularmente adequado para perguntas amplas, históricas ou conceituais. Sua baixa variação de comportamento com mudanças de temperatura também o torna a melhor escolha quando é importante evitar erros ou alucinações.

O Gemini, destaca-se em áreas que exigem precisão lógica rigorosa, domínio técnico especializado ou conhecimento de nicho. Ele superou amplamente o GPT em Segurança da Informação, apresentando desempenho de 100%. Da mesma forma, demonstrou melhor performance em Fundamentos Matemáticos e em Construção de Compiladores, onde obteve acerto absoluto. Contudo, é necessário ter cautela no uso de temperatura 0.0 para questões de Arquitetura de Computadores, pois o modelo apresentou uma falha crítica nesse cenário.

O DeepSeek é particularmente recomendado para tarefas técnicas complexas e para o estudo de Estruturas de Dados. Ele foi o único modelo capaz de alcançar 100% de acurácia nessa categoria, especialmente em temperaturas mais altas (como 2.0), enquanto os demais variaram entre 60% e 75%. Assim, é o mais adequado para implementações de árvores, grafos ou heaps em ambientes que toleram maior criatividade controlada. Além disso, o DeepSeek se mostra uma alternativa sólida ao Gemini em áreas como Segurança da Informação e Matemática, mantendo desempenho igualmente elevado. Entretanto, deve-se evitar o uso de temperaturas superiores a 1.2 para Análise de Algoritmos, pois o modelo apresenta queda significativa de performance nessas condições.

5 CONCLUSÕES

O presente trabalho teve por objetivo avaliar comparativamente o desempenho das LLM's, especificamente GPT, Gemini e DeepSeek na resolução de questões objetivas do ENADE aplicadas aos cursos da área de Computação. A análise, fundamentada nos gabaritos oficiais, permitiu traçar um panorama detalhado sobre a assertividade, a estabilidade e as competências específicas de cada modelo frente a desafios acadêmicos de nível superior.

Em resposta à primeira questão de pesquisa (QP01), sobre quão assertivas podem ser as LLM's na resolução de questões em computação, os resultados demonstraram que os modelos atuais atingiram um patamar elevado de competência. Observou-se uma convergência de desempenho, especialmente na temperatura 0,8, em que todos os modelos apresentaram probabilidades de acerto semelhantes e consistentes. Isso valida o uso dessas ferramentas como auxiliares no processo de aprendizado e revisão de conceitos fundamentais da área.

Em relação a segunda questão (QP02), referente a qual modelo obteve o melhor resultado, a análise revelou nuances importantes entre "pico de desempenho"e "consistência". O GPT consolidou-se como o modelo mais robusto e estável. Sua arquitetura demonstrou baixa sensibilidade à variação de temperatura, mantendo alta acurácia e recall constante (0,82%) mesmo em cenários de maior aleatoriedade. Por outro lado, o Gemini e o DeepSeek, embora apresentem oscilações maiores, revelaram-se capazes de superar o GPT em condições específicas de temperatura, sugerindo que, quando devidamente calibrados, possuem um teto de performance ligeiramente superior para recuperação de informações.

Quanto à terceira questão (QP03), sobre a superioridade em áreas específicas, o estudo identificou especializações claras:

- **GPT:** Mostrou-se superior em Análise de Algoritmos e Conhecimentos Gerais, sendo a escolha ideal para cenários que exigem consistência e baixa taxa de alucinação. No entanto, seu desempenho foi limitado em Segurança da Informação, provavelmente devido a filtros de segurança mais restritivos.

- **Gemini:** Destacou-se em Segurança da Informação, Fundamentos Matemáticos e Compiladores, atingindo acertos absolutos nessas categorias.

- **DeepSeek:** Revelou-se a melhor opção para Estruturas de Dados, sendo o único capaz de atingir 100% de acurácia nesta área em uma temperatura criativa (2.0), além de ser uma alternativa sólida ao Gemini em matemática e segurança.

Uma limitação transversal identificada nos três modelos foi o baixo desempenho nos conteúdos de Sistemas Operacionais (acurácia entre 28% e 57%). Conclui-se que essa dificuldade advém da natureza visual e diagramática de muitas questões dessa área

(estados de memória, hardware), que impõem barreiras à interpretação puramente textual das LLM's atuais em algumas áreas.

Por fim, este estudo evidencia que não há uma "melhor LLM"universal para a Computação, mas sim ferramentas com perfis complementares. Para tarefas que exigem rigor lógico e estabilidade, o GPT é o mais indicado; para tarefas que envolvem criatividade técnica, matemática formal ou domínios específicos como segurança, o Gemini e o DeepSeek apresentam vantagens competitivas.

Como sugestão para trabalhos futuros, recomenda-se a expansão da análise para modelos multimodais (capazes de interpretar as imagens) para analise de questões que fazem uso de recursos visuais como diagramas, gráficos ou estruturas voltadas a computação., e a investigação mais aprofundada sobre como o prompt engineering pode mitigar as alucinações observadas nas temperaturas mais elevadas do DeepSeek e Gemini.

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.

DAI, Damai; DENG, Chengqi; ZHAO, Chenggang et al. DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models. **arXiv preprint arXiv:2401.06066**, 2024.

DAS, Bidyut et al. Multiple-choice question generation with auto-generated distractors for computer-assisted educational assessment. **Multimedia Tools and Applications**, v. 80, p. 31907–31925, 2021. DOI: `10.1007/s11042-021-11222-2`.

DEEPSEEK-AI; BI, Xiao; CHEN, Deli et al. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. **arXiv preprint arXiv:2401.02954**, 2024.

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].

GRÉVISSE, Christian. LLM-based automatic short answer grading in undergraduate medical education. **BMC Medical Education**, Springer, v. 24, n. 1, p. 1060, 2024.

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.

HOLMES, Wayne; MIAO, Fengchun et al. **Guidance for generative AI in education and research**. [S.l.]: Unesco Publishing, 2023.

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.

K, Manish Kumar Jain; BHADANI, Charisma; PATHAK, Rishika. The Societal Impact of Artificial Intelligence. **International Journal on Science and Technology**, 2025. DOI: `10.71097/ijsat.v16.i1.3027`.

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.

KUNG, Tiffany H et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. **PLoS digital health**, Public Library of Science, v. 2, n. 2, e0000198, 2023.

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.

OPENAI et al. **GPT-4 Technical Report**. [S.l.*:* s.n.], 2024. arXiv: `2303.08774 [cs.CL]`. Disponível em: `https://arxiv.org/abs/2303.08774`.

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.

SUPERBI, Joao et al. Enhancing Large Language Model Performance on ENEM Math Questions Using Retrieval-Augmented Generation. In: SBC. BRAZILIAN e-Science Workshop (BreSci). [S.l.*:* s.n.], 2024. p. 56–63.

TEAM, Gemini et al. **Gemini: A Family of Highly Capable Multimodal Models**. [S.l.*:* s.n.], 2025. arXiv: `2312.11805 [cs.CL]`. Disponível em: `https://arxiv.org/abs/2312.11805`.

TILTON, L.; ARNOLD, T. Introduction to Natural Language Processing, p. 947–948, 2016.

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

VIEGAS, Cayo; GHEYI, Rohit; RIBEIRO, Márcio. Assessing the Capability of LLMs in Solving POSCOMP Questions. **arXiv preprint arXiv:2505.20338**, 2025.

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].

ZHAO, Wayne Xin et al. A Survey of Large Language Models. **arXiv preprint arXiv:2303.18223**, 2023. Disponível em: `https://arxiv.org/abs/2303.18223`.