CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
CAMPUS TIMÓTEO

Marcio Gabriel Gonçalves Soares

DESEMPENHO DE GRANDES MODELOS DE LINGUAGEM (GPT-5 E GEMINI 2.5 PRO) NA RESOLUÇÃO DE QUESTÕES DO EXAME DA OAB: UM ESTUDO COMPARATIVO

Timóteo

2026

MARCIO GABRIEL GONÇALVES SOARES

DESEMPENHO DE GRANDES MODELOS DE LINGUAGEM (GPT-5 E GEMINI 2.5 PRO) NA RESOLUÇÃO DE QUESTÕES DO EXAME DA OAB: UM ESTUDO COMPARATIVO

Trabalho de Conclusão de Curso apresentado ao Curso de Engenharia de Computação do Centro Federal de Educação Tecnológica de MinasGerais, campus Timóteo, como requisito parcial para obtenção do título de Engenheiro de Computação.

Trabalho aprovado. Timóteo, 19 de junho de 2026:

Prof. Dr. Lucas Pantuza Amorim
Orientador

Prof. Dr. Bruno Rodrigues Silva
Professor Convidado

Prof. Dr. Marcelo de Sousa Balbino
Professor Convidado

Timóteo, 2026

Soares, Marcio Gabriel Gonçalves
S676d Desempenho de grandes modelos de linguagem (GPT-5 e Gemini 2.5 Pro) na resolução de questões do exame da OAB: um estudo comparativo / Marcio Gabriel Gonçalves Soares . – 2026.
51 f. : il.
Orientador: Lucas Pantuza Amorim
Trabalho de conclusão de curso (Graduação) – Centro Federal de Educação Tecnológica de Minas Gerais. Engenharia de Computação, Departamento de Computação e Construção Civil, Timóteo, 2026.
Bibliografia.

1. Inteligência artifcial . 2. GPT-5 3. Gemini 4. Engenharia de Computação. . I. Amorim, Lucas Pantuza II. Título.

CDU: 004.8

Ficha elaborada pela Biblioteca - *campus* Timóteo - CEFET-MG
Bibliotecário: Alexandre Augusto Gamberini - CRB6-2140

CÓPIA DE FOLHA DE ASSINATURAS Nº 1/2026 - DECOMTM (11.63.11)

(Nº do Protocolo: NÃO PROTOCOLADO)

(Assinado digitalmente em 23/06/2026 17:13)
BRUNO RODRIGUES SILVA
PROFESSOR ENS BASICO TECN TECNOLOGICO
DECOMTM (11.63.11)
Matrícula: ###759#5

(Assinado digitalmente em 23/06/2026 16:47)
LUCAS PANTUZA AMORIM
PROFESSOR ENS BASICO TECN TECNOLOGICO
DECOMTM (11.63.11)
Matrícula: ###974#1

(Assinado digitalmente em 23/06/2026 15:47)
MARCELO DE SOUSA BALBINO
PROFESSOR ENS BASICO TECN TECNOLOGICO
DECOMTM (11.63.11)
Matrícula: ###093#2

*Visualize o documento original em https://sig.cefetmg.br/documentos/ informando seu número: **1**, ano: **2026**, tipo: **CÓPIA DE FOLHA DE ASSINATURAS**, data de emissão: **23/06/2026** e o código de verificação: **975d9cdef3***

Dedico este trabalho aos meus pais e à minha família,
pelo amor incondicional, apoio paciente e
por serem os pilares que tornaram possível
a realização deste sonho acadêmico.

Agradecimentos

Agradeço primeiramente a Deus, pela sabedoria, proteção e forças concedidas ao longo desta jornada acadêmica. Aos meus pais e à minha família, pilares essenciais da minha formação, expresso minha profunda gratidão pelo apoio incondicional, pelas palavras de incentivo e por me ensinarem, desde cedo, o valor da disciplina, da responsabilidade e da perseverança.

Por fim, a todos que contribuíram direta ou indiretamente para esta pesquisa, meu sincero agradecimento.

“Nós só podemos ver um pouco do futuro,
mas o suficiente para perceber que há muito a fazer.”
Alan Turing

Resumo

Grandes Modelos de Linguagem (LLMs) têm demonstrado avanços significativos em tarefas que exigem compreensão textual e raciocínio. No entanto, a literatura brasileira carece de estudos que avaliem modelos de fronteira em contextos jurídicos nacionais. Este trabalho investigou o desempenho comparativo do GPT-5 e do Gemini 2.5 Pro na resolução das questões objetivas da 1ª fase do Exame da OAB, nos cenários zero-shot e zero-shot chain-of-thought (CoT). A pesquisa adotou um delineamento experimental controlado, com execução automatizada via APIs oficiais em ambiente de computação em nuvem, utilizando questões dos 43º e 44º Exames de Ordem. Os resultados demonstraram que ambos os modelos superam o limiar de aprovação de 50%: o GPT-5 atingiu acurácia global de 94,94% e o Gemini 2.5 Pro de 98,10% no cenário Standard. De forma contraintuitiva, o cenário Zero-shot CoT provocou redução de desempenho em ambas as arquiteturas, com quedas de 1,27 pontos percentuais para o GPT-5 e de 5,06 pontos percentuais para o Gemini 2.5 Pro. A estratificação por disciplina jurídica revelou que as maiores fragilidades relativas se concentraram em Ética Profissional, Direito Tributário e Direito Previdenciário. O estudo contribui para o entendimento das capacidades atuais dos LLMs em tarefas jurídicas em língua portuguesa e para o desenvolvimento de metodologias de avaliação reprodutíveis no contexto da Engenharia de Computação.

Palavras-chave: Inteligência Artificial; Grandes Modelos de Linguagem; GPT-5; Gemini PRO; Exame da OAB.

Abstract

Large Language Models (LLMs) have shown substantial progress in tasks involving textual understanding and reasoning. However, Brazilian literature lacks studies evaluating frontier models in national legal contexts. This work investigated the comparative performance of GPT-5 and Gemini 2.5 Pro on the multiple-choice questions of the Brazilian Bar Examination (OAB), under zero-shot and zero-shot chain-of-thought (CoT) prompting conditions. The study adopted a controlled experimental design, with automated execution via official APIs in a cloud computing environment, using questions collected from the 43rd and 44th Bar Exams. Results demonstrated that both models substantially surpass the 50% passing threshold: GPT-5 achieved a global accuracy of 94.94% and Gemini 2.5 Pro reached 98.10% under the Standard prompting scenario. Counter-intuitively, the Zero-shot CoT scenario produced accuracy decreases in both architectures, with reductions of 1.27 percentage points for GPT-5 and 5.06 percentage points for Gemini 2.5 Pro. Stratification by legal subject revealed that the greatest relative weaknesses were concentrated in Legal Ethics, Tax Law, and Social Security Law. The study advances the understanding of LLM capabilities in Portuguese-language legal tasks and provides a reproducible evaluation methodology relevant to the field of Computer Engineering.

Keywords: Artificial Intelligence; Large Language Models; GPT-5; Gemini PRO; Brazilian Bar Exam.

Lista de ilustrações

Lista de tabelas

Lista de quadros

Lista de abreviaturas e siglas

API	*Application Programming Interface*
CNN	*Convolutional Neural Networks* (Redes Neurais Convolucionais)
CoT	*Chain-of-Thought*
CSV	*Comma-Separated Values*
DL	*Deep Learning* (Aprendizado Profundo)
ENEM	Exame Nacional do Ensino Médio
FGV	Fundação Getulio Vargas
GPT	*Generative Pre-trained Transformer*
HPC	*High-Performance Computing* (Computação de Alto Desempenho)
IA	Inteligência Artificial
IME	Instituto Militar de Engenharia
ITA	Instituto Tecnológico de Aeronáutica
JSON	*JavaScript Object Notation*
LLM	*Large Language Model* (Grande Modelo de Linguagem)
LSTM	*Long Short-Term Memory*
ML	*Machine Learning* (Aprendizado de Máquina)
OAB	Ordem dos Advogados do Brasil
PLN	Processamento de Linguagem Natural
RNN	*Recurrent Neural Networks* (Redes Neurais Recorrentes)
UBE	*Uniform Bar Exam*

Sumário

1 Introdução

"O coração do entendido adquire o conhecimento,
e o ouvido dos sábios busca a sabedoria".
Provérbios 18:15

A ascensão da Inteligência Artificial (IA), impulsionada pelos avanços em deep learning (aprendizagem profunda), redefiniu as fronteiras da capacidade computacional em domínios antes restritos à cognição humana. Conforme elucidado por LeCun, Bengio e Hinton (2015), esse paradigma permitiu o uso de arquiteturas hierárquicas capazes de aprender representações cada vez mais abstratas a partir de grandes volumes de dados, ampliando consideravelmente a capacidade dos sistemas computacionais em tarefas complexas de percepção e linguagem. Um marco decisivo desse progresso foi a arquitetura Transformer, proposta por Vaswani et al. (2017), cuja adoção consolidou o desenvolvimento dos Grandes Modelos de Linguagem (Large Language Models — LLMs), hoje responsáveis por resultados de estado da arte em tarefas de compreensão, raciocínio e geração textual.

O desempenho desses modelos em domínios especializados, particularmente no campo jurídico, tornou-se objeto crescente de investigação acadêmica. Estudos internacionais, como o de Katz et al. (2024) demonstraram que modelos avançados foram capazes de atingir ou superar o desempenho médio de candidatos humanos em exames profissionais, como o Uniform Bar Exam (UBE). Esses resultados suscitam questionamentos relevantes sobre até que ponto LLMs modernos conseguem interpretar normas, identificar fundamentos jurídicos aplicáveis e raciocinar em contextos avaliativos de alta complexidade.

No Brasil, o Exame da Ordem dos Advogados do Brasil (OAB) desempenha papel equivalente ao Bar Exam norte-americano, sendo requisito indispensável para o exercício da advocacia. A primeira fase do exame, composta por 80 questões objetivas que abrangem diferentes disciplinas do Direito, é considerada uma das mais desafiadoras etapas de avaliação profissional do país. Por suas características, o exame configura um cenário ideal para investigar o desempenho dos LLMs em português jurídico.

Embora existam avaliações preliminares sobre modelos anteriores, como GPT-3.5 e GPT-4, observa-se uma notável escassez, até o momento, de estudos nacionais sistemáticos que avaliem a performance dos modelos de fronteira mais recentes, especificamente o GPT-5 e o Gemini 2.5 PRO, na resolução de questões da OAB. Ademais, apesar dos avanços internacionais, carecem-se de benchmarks específicos que considerem o ordenamento jurídico brasileiro. Essa lacuna é particularmente relevante, dado que a língua portuguesa apresenta estruturas sintáticas próprias e que os padrões jurisprudenciais e doutrinários do Brasil diferem substancialmente daqueles observados nos exames estrangeiros utilizados em pesquisas anteriores.

Diante desse cenário, este trabalho propõe um estudo comparativo focado no contexto

brasileiro. O objetivo é avaliar o desempenho dos modelos GPT-5 e Gemini 2.5 Pro na resolução de questões objetivas da 1ª fase da OAB, sob diferentes configurações de prompting, zero-shot Standard e zero-shot Chain-of-Thought, analisando a acurácia global de cada modelo, o impacto da técnica de indução de raciocínio passo a passo sobre o desempenho, e os padrões de acerto estratificados por disciplina jurídica.

1.1 Justificativa

A relevância deste estudo se sustenta em três pilares: tecnológico, prático-social e acadêmico.

Do ponto de vista tecnológico, a evolução dos LLMs é exponencial. Avaliar continuamente os modelos mais avançados, como o GPT-5 e o Gemini 2.5 PRO, em tarefas de alta complexidade como o Exame da OAB, é fundamental para mapear o estado da arte e compreender os limites e as potencialidades dessas tecnologias. O Exame de Ordem, pela diversidade temática e pelo nível de exigência cognitiva de suas questões, configura-se como um ambiente de teste robusto para examinar a capacidade desses modelos em interpretar normas, argumentos jurídicos e estruturas lógico-textuais em língua portuguesa.

No âmbito acadêmico, ainda são escassos os estudos nacionais que avaliam modelos de última geração em português jurídico, especificamente no contexto do Exame da OAB. Embora existam investigações internacionais envolvendo o Uniform Bar Exam (KATZ et al., 2024) e estudos nacionais sobre exames técnico-científicos e de ensino médio (Peres, 2023; Rodrigues, 2025), não foram identificadas, na revisão de literatura conduzida para este trabalho, evidências sistemáticas sobre o desempenho de modelos como GPT-5 e Gemini 2.5 Pro quando confrontados com as especificidades do ordenamento jurídico, da linguagem técnica e das particularidades normativas brasileiras. Este estudo oferece uma contribuição inédita ao fornecer uma avaliação comparativa estruturada, com metodologia reprodutível via APIs oficiais e delineamento fatorial controlado, ampliando a literatura sobre Inteligência Artificial aplicada ao Direito no contexto nacional.

Sob a ótica prático-social, o campo do Direito é ligado à linguagem e ao raciocínio interpretativo. O desempenho expressivo de LLMs em um exame que certifica a aptidão para o exercício da advocacia sinaliza um potencial transformador para o setor jurídico. Conforme sugerido por Katz et al. (2024), a proficiência desses modelos pode impulsionar o desenvolvimento de ferramentas que ampliem a eficiência dos profissionais do Direito e contribuam para a democratização do acesso à informação jurídica. Os resultados deste trabalho fornecem evidências empíricas que embasam essa discussão no contexto brasileiro, identificando tanto as áreas de alta proficiência quanto as disciplinas em que os modelos ainda apresentam fragilidades relativas, informação relevante para pesquisadores, desenvolvedores e operadores do Direito que considerem a adoção dessas tecnologias.

1.2 Problema

O Exame da Ordem dos Advogados do Brasil (OAB) constitui uma avaliação rigorosa que busca aferir a proficiência mínima necessária para o exercício da advocacia. Diante do avanço recente dos Grandes Modelos de Linguagem e de seu desempenho expressivo em exames jurídicos estrangeiros, surge a necessidade de investigar como modelos de última geração se comportam quando submetidos às particularidades linguísticas, normativas e doutrinárias brasileiras.

Dessa forma, o problema de pesquisa que norteia este trabalho é: Qual é o desempenho comparativo dos modelos GPT-5 e Gemini 2.5 Pro na resolução das questões objetivas da primeira fase do Exame da OAB, e em que medida as taxas de acurácia alcançadas por esses sistemas atingem ou superam o limiar de acertos exigido para a aprovação?

1.3 Objetivos

Para responder ao problema de pesquisa, o objetivo geral deste trabalho é avaliar e comparar o desempenho dos modelos GPT-5 e Gemini 2.5 Pro na resolução de questões de múltipla escolha da 1ª fase do Exame da OAB, utilizando os cenários zero-shot e zero-shot chain-of-thought, analisando suas taxas de acerto e consistência por disciplina jurídica.

Também, objetivam-se mais especificamente:

1. Mensurar a acurácia global do GPT-5 e Gemini 2.5 Pro nas questões válidas da OAB, verificando objetivamente se os modelos superam a nota de corte exigida para aprovação na primeira fase.
2. Comparar o desempenho e as taxas de acerto entre os modelos e entre os cenários de prompt (Padrão vs. Chain-of-Thought), avaliando o impacto da indução de raciocínio passo a passo nas respostas.
3. Mapear o desempenho por disciplina jurídica (ex: Direito Civil, Penal, Constitucional), identificando as áreas de maior proficiência e as principais fragilidades de cada modelo.

Para alcançar os objetivos propostos e investigar o problema de pesquisa, será conduzida uma avaliação experimental e comparativa. Serão avaliadas as capacidades dos modelos GPT-5 e Gemini 2.5 PRO na resolução das questões do Exame da OAB, permitindo uma análise direta de seus desempenhos e a identificação de padrões relevantes. Os procedimentos detalhados para a condução desta avaliação são apresentados no capítulo metodologia.

2 FUNDAMENTAÇÃO TEÓRICA

“Não tenha pressa... Dê um passo de cada vez.
Talvez, quando vocês descobrirem a verdade por si mesmos,
cheguem a uma conclusão diferente da nossa”.
Eiichiro Oda

O principal objetivo deste capítulo é apresentar os fundamentos históricos, teóricos e metodológicos sobre os quais esta pesquisa é executada, além de mapear os principais trabalhos encontrados na literatura sobre a avaliação de Grandes Modelos de Linguagem (LLMs). As seções seguintes reúnem os principais e mais importantes trabalhos que versam sobre os conceitos essenciais que sustentam o desenvolvimento das tecnologias de Deep Learning e do processamento de linguagem natural (PNL). Além disso, serão discutidos os trabalhos utilizados como material referencial para o desenvolvimento desta pesquisa.

2.1 Evolução Histórica da Inteligência Artificial e do Aprendizado Profundo

Os fundamentos da Inteligência Artificial (IA) remontam à década de 1950, período em que pesquisadores pioneiros como Alan Turing (1950) e John McCarthy (1956) começaram a formular as primeiras abordagens computacionais com o objetivo de simular aspectos do raciocínio humano (TURING, 2009; MCCARTHY et al., 2006). Ao longo das décadas seguintes, o desenvolvimento da IA foi marcado por ciclos de grande otimismo e investimento, intercalados com períodos de estagnação e redução de financiamento, que ficaram conhecidos como "invernos da IA".

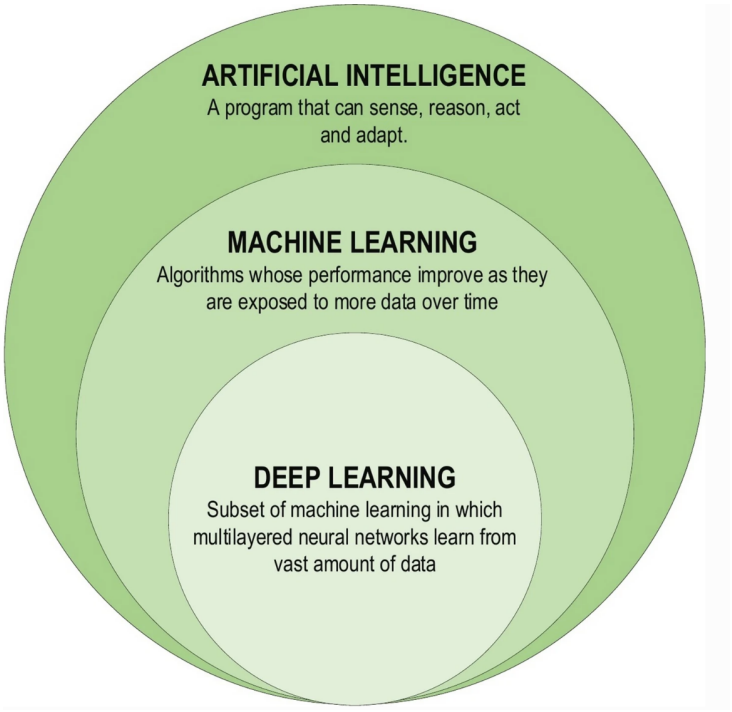
O ressurgimento da área ocorreu a partir dos anos 2010, impulsionado pelo aumento do poder computacional, pela disponibilidade de grandes volumes de dados e pelo amadurecimento das redes neurais artificiais profundas. É neste contexto que emerge o Aprendizado Profundo (Deep Learning - DL), um subconjunto específico do Aprendizado de Máquina (Machine Learning - ML). Como ilustrado na Figura 1, o DL se distingue por utilizar redes neurais multicamadas para aprender representações diretamente de vastas quantidades de dados. O DL permitiu a construção de sistemas capazes de aprender representações complexas e hierárquicas dos dados, superando limitações anteriores de generalização e desempenho.

Conforme LeCun, Bengio e Hinton (2015) e reforçado por Alzubaidi et al. (2021), o aprendizado profundo introduziu métodos eficazes de aprendizado de representações (também chamado de representation learning). Diferentemente das técnicas de ML tradicionais, que frequentemente exigiam uma etapa sequencial e trabalhosa de pré-processamento, extração manual de características (feature engineering) e seleção de características , o DL tem a capacidade de aprender automaticamente hierarquias de características diretamente dos da-

dos brutos. As camadas iniciais da rede aprendem características de baixo nível, enquanto camadas mais profundas combinam essas informações para extrair características de alto nível, mais abstratas. Essa capacidade de extração automática de características é uma das principais vantagens do DL e simula, em certo nível, o processamento sensorial do cérebro humano.

A combinação de modelos DL mais capazes (como as Redes Neurais Convolucionais - CNNs, que se tornaram extremamente populares), o aumento exponencial no poder computacional (especialmente com o uso de GPUs e computação de alto desempenho - HPC) e a disponibilidade de grandes volumes de dados consolidaram o DL como o "Padrão Ouro"na comunidade de ML (ALZUBAIDI et al., 2021). Ele passou a superar o desempenho humano em diversas tarefas complexas, como classificação de imagens e tem sido aplicado com sucesso em domínios variados, desde processamento de linguagem natural até diagnóstico médico. A capacidade atual dos modelos de linguagem, foco deste trabalho, decorre diretamente desses avanços fundamentais no campo do Deep Learning.

Figura 1 – Relação hierárquica entre Inteligência Artificial (AI), Aprendizado de Máquina (ML) e Aprendizado Profundo (DL)



Fonte: Alzubaidi et al. (2021)

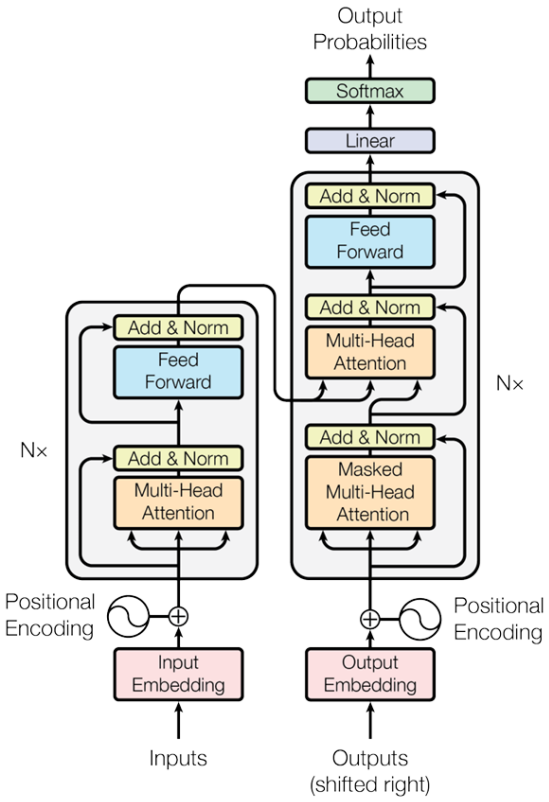
Entre as limitações mais significativas das arquiteturas tradicionais de Deep Learning está a dificuldade em lidar com dependências de longo alcance em dados sequenciais, especialmente em tarefas de processamento de linguagem natural. Modelos baseados em redes recorrentes (RNNs) e convolucionais (CNNs) apresentavam restrições quanto à paralelização e à retenção de contexto em sequências extensas, o que comprometia o desempenho em aplicações mais complexas. Nesse cenário, surge a arquitetura Transformer, proposta por Vaswani

et al. (2017), como uma solução inovadora para esse problema específico, ao substituir os mecanismos recorrentes pelo princípio de self-attention, capaz de modelar relações contextuais entre elementos de uma sequência de forma mais eficiente e escalável. Assim, a próxima seção abordará em detalhe os princípios e inovações que consolidaram os Transformers como a base da atual geração de Modelos de Linguagem de Grande Escala (LLMs).

2.2 A Revolução da Arquitetura Transformer

Um avanço fundamental para o Processamento de Linguagem Natural (PLN) foi a introdução da arquitetura Transformer por Vaswani et al. (2017) no artigo Attention is All You Need. Modelos anteriores, como Redes Neurais Recorrentes (RNNs) e Long Short-Term Memory (LSTM), processavam texto sequencialmente, o que dificultava a captura de dependências de longo alcance e limitava a paralelização. O Transformer abandonou a recorrência, baseando-se inteiramente em mecanismos de atenção (self-attention) para ponderar a importância de diferentes partes da sequência de entrada ao processar cada elemento. Isso permitiu um treinamento mais eficiente em larga escala e uma melhor modelagem do contexto linguístico. A arquitetura Transformer tornou-se a base para a maioria dos LLMs subsequentes.

Figura 2 – Arquitetura Transformers



Fonte: Vaswani et al. (2017).

A Figura 2 apresenta a estrutura geral do modelo Transformer proposta por Vaswani et al. (2017). O modelo opera através de dois módulos distintos: uma etapa de codificação, que

mapeia a entrada em representações vetoriais ricas em contexto, e uma etapa de decodificação, que utiliza essas informações para gerar a sequência de saída de forma autoregressiva. Cada um desses blocos é formado por múltiplas camadas (N vezes repetidas), contendo submódulos de multi-head attention e redes feed-forward posicionais, intercalados por camadas de normalização e adição residual. O uso de multi-head attention permite que o modelo aprenda relações contextuais sob diferentes perspectivas, enquanto o positional encoding fornece informação sobre a ordem sequencial dos tokens, algo essencial, já que a arquitetura não depende de recorrência.

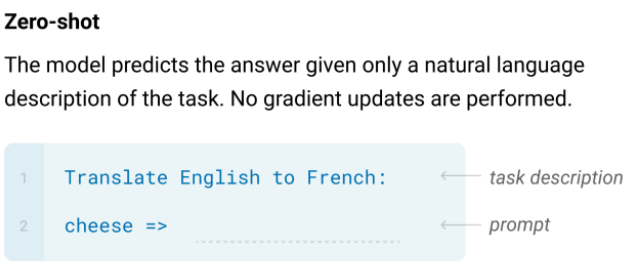
Os Transformers tornaram-se a base estrutural dos Grandes Modelos de Linguagem devido, primordialmente, à sua capacidade de paralelização eficiente, o que viabilizou o treinamento em largas escalas de dados e parâmetros. Foi nesse contexto de escalabilidade que o trabalho seminal de Brown et al. (2020), com o GPT-3 (175 bilhões de parâmetros), revelou que o aumento massivo da capacidade do modelo faz emergir a habilidade de aprendizado em contexto (in-context learning). Esse fenômeno permite que o modelo adapte seu comportamento para executar tarefas complexas, como a resolução de questões jurídicas, baseando-se exclusivamente em instruções em linguagem natural (zero-shot) ou em poucos exemplos demonstrativos, dispensando a necessidade de atualizações nos pesos da rede (fine-tuning). Portanto, é essa flexibilidade arquitetural que fundamenta tecnicamente as abordagens experimentais investigadas neste estudo.

2.3 Engenharia de Prompt e Paradigmas de Aprendizado em Contexto

Como introduzido na seção anterior, um dos avanços mais significativos trazidos pelos LLMs de grande escala, é a capacidade de aprendizado em contexto (in-context learning) . Essa habilidade permite que o modelo adapte seu comportamento para realizar novas tarefas com base apenas nas informações fornecidas no prompt de entrada, sem a necessidade de atualizações nos seus pesos (gradientes) através de fine-tuning . Brown et al. (2020) exploraram sistematicamente essa capacidade, definindo e avaliando três paradigmas principais que se distinguem pela quantidade de informação contextual fornecida ao modelo no momento da inferência.

Começando pelo Zero-shot, neste paradigma, o modelo recebe apenas uma descrição da tarefa em linguagem natural, sem qualquer exemplo de entrada e saída. A Figura 3 ilustra isso com um exemplo de tradução. Essa configuração avalia a capacidade do modelo de compreender e generalizar instruções puramente textuais, dependendo exclusivamente do conhecimento internalizado durante o pré-treinamento. Brown et al. (2020) destacam que o desempenho nesse regime reflete a competência do modelo em interpretar instruções explícitas e aplicar o raciocínio aprendido implicitamente. Os resultados mostraram que, embora o GPT-3 apresente desempenho modesto em tarefas complexas nesse cenário, ele supera modelos anteriores mesmo sem exemplos adicionais, indicando que o pré-treinamento em larga escala confere uma forma rudimentar de compreensão semântica generalista.

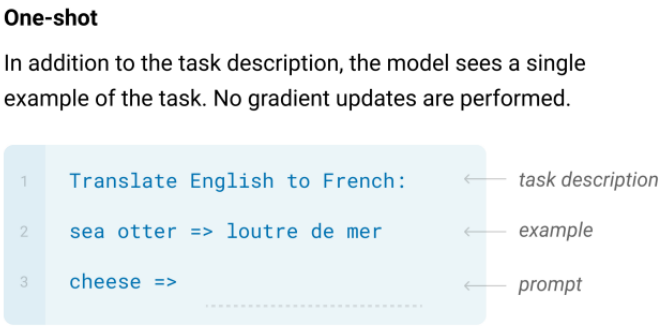
Figura 3 – Exemplo Zero-shot



Fonte: Brown et al. (2020).

Já o One-shot, além da descrição da tarefa, o modelo adiciona um único exemplo de demonstração antes da instrução principal. Esse exemplo serve como um guia de formato e estilo de resposta, reduzindo ambiguidades interpretativas que ocorrem no zero-shot. Brown et al. (2020) demonstraram que essa pequena adição já resulta em ganhos consistentes de desempenho, sobretudo em tarefas estruturadas (como tradução, completamento de sentenças e correção gramatical). Os autores sugerem que o one-shot atua como uma forma mínima de condicionamento contextual, em que o modelo ajusta seu comportamento à estrutura exibida, mesmo sem treinamento supervisionado adicional.

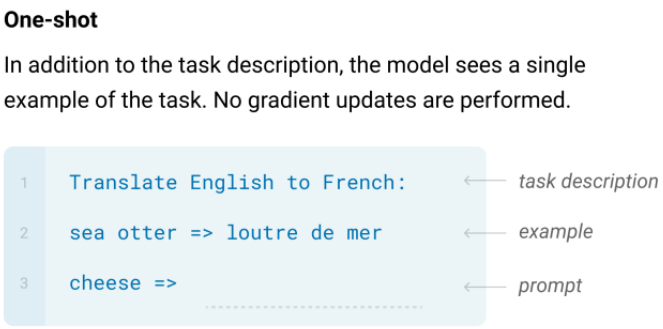
Figura 4 – Exemplo One-shot



Fonte: Brown et al. (2020).

Por fim, o paradigma Few-shot fornece ao modelo diversos exemplos de entrada e saída da tarefa antes da solicitação final. Esses exemplos não servem para ajuste de pesos, mas sim como contexto condicional, permitindo que o modelo identifique padrões e aprenda a tarefa de forma implícita durante a inferência. Brown et al. (2020) demonstraram que, conforme o número de exemplos aumenta, o desempenho do modelo melhora de maneira substancial, especialmente em tarefas que exigem raciocínio, compreensão semântica e inferência contextual.

Figura 5 – Exemplo Few-shot



Fonte: Brown et al. (2020).

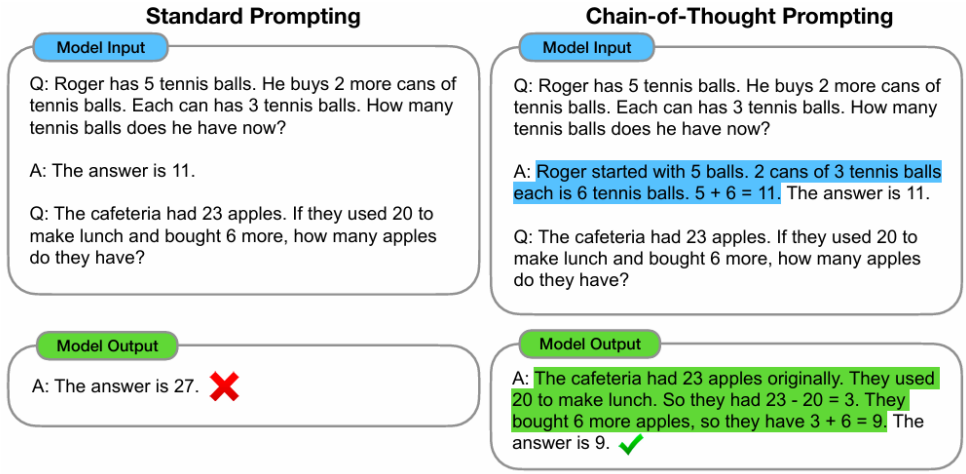
Brown et al. (2020) concluíram que o desempenho few-shot do GPT-3 se aproxima ou supera o de modelos supervisionados treinados especificamente para certas tarefas, evidenciando que o aumento de escala confere ao modelo capacidades emergentes de aprendizado e raciocínio adaptativo. Essa descoberta redefiniu a forma como os modelos de linguagem são avaliados e utilizados, inaugurando uma nova linha de pesquisa voltada à engenharia de prompts, ou seja, à formulação estratégica de instruções e exemplos para orientar o comportamento de modelos sem necessidade de re-treinamento.

No entanto, também observaram que abordagens de prompting padrão, que estabelecem um mapeamento direto entre a entrada (pergunta) e a saída (resposta), apresentavam limitações significativas em tarefas que exigem raciocínio complexo, especialmente aquelas compostas por múltiplas etapas lógicas ou inferências intermediárias. Essa limitação evidenciou a necessidade de métodos capazes de induzir o modelo a explicitar seu processo de raciocínio, o que motivou o surgimento de abordagens como o Chain-of-Thought Prompting (WEI et al., 2022).

Para mitigar essa limitação, Wei et al. (2022) propuseram a técnica de prompting denominada Cadeia de Pensamento (Chain-of-Thought - CoT). Diferente do prompting tradicional, que foca apenas no resultado, o CoT induz o modelo a externalizar seu processo cognitivo. Ao fornecer ao LLM, durante a fase de prompting, exemplos que não contêm apenas a pergunta e a resposta final, mas também uma série de etapas de raciocínio intermediárias, expressas em linguagem natural, que detalham o processo para alcançar essa resposta. A ideia é que, ao observar esses exemplos detalhados, o modelo aprenda a gerar sua própria cadeia de pensamento para novas perguntas, melhorando sua capacidade de raciocínio sequencial.

A Figura 6 ilustra a diferença fundamental: o prompting padrão (esquerda) fornece Pergunta -> Resposta, enquanto o CoT (direita) fornece Pergunta -> Passos de Raciocínio -> Resposta nos exemplos. Ao ser apresentado a uma nova pergunta, o modelo instruído via CoT tende a gerar primeiro os passos lógicos e, em seguida, a resposta final.

Figura 6 – Prompting padrão x Prompting CoT



Fonte: Wei et al. (2022)

Empiricamente, Wei et al. (2022) demonstraram que o CoT melhora significativamente o desempenho de LLMs de grande escala (acima de 100 bilhões de parâmetros) em diversos benchmarks de raciocínio, superando o prompting padrão, por vezes de forma expressiva. Por exemplo, no benchmark de problemas matemáticos GSM8K, o modelo PaLM de 540B parâmetros, com apenas oito exemplos CoT, atingiu um desempenho estado da arte na época, superando modelos como o GPT-3 que haviam passado por fine-tuning específico para a tarefa.

No entanto, a abordagem CoT também apresenta limitações. Primeiramente, sua eficácia parece ser uma habilidade emergente da escala do modelo; modelos menores geralmente não se beneficiam, podendo até ter seu desempenho prejudicado ao gerar cadeias de pensamento ilógicas. Em segundo lugar, a geração de uma cadeia de pensamento não garante a correção lógica dos passos intermediários ou da resposta final; o modelo pode cometer erros de cálculo ou de raciocínio ao longo do processo. Por fim, a necessidade de criar manualmente exemplos de alta qualidade com cadeias de pensamento detalhadas para o prompting few-shot pode ser custosa e exigir conhecimento especializado. Essa última limitação motivou o desenvolvimento de variações que buscam os benefícios do raciocínio passo a passo sem a necessidade de fornecer exemplos explícitos, como será detalhado a seguir.

2.3.1 Zero-shot Chain-of-Thought

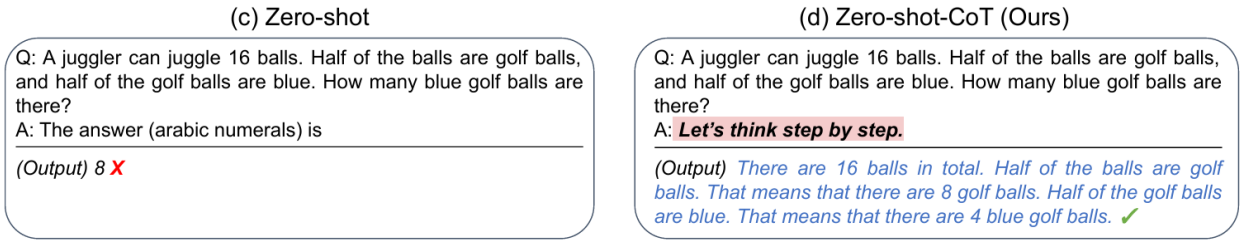
Apesar da eficácia demonstrada pelo prompting CoT no paradigma few-shot em tarefas complexas de raciocínio, a necessidade de elaborar manualmente exemplos de alta qualidade com cadeias de pensamento detalhadas representa um desafio prático. A criação desses exemplos pode ser custosa, demorada e exigir conhecimento especializado no domínio da tarefa.

Buscando superar essa limitação e explorar as capacidades intrínsecas de raciocínio

dos LLMs sem depender de demonstrações explícitas, Kojima et al. (2022) investigaram a possibilidade de elicitar o comportamento de Cadeia de Pensamento em um cenário zero-shot. Os autores descobriram que era possível induzir o modelo a gerar etapas de raciocínio intermediárias simplesmente anexando uma instrução simples e direta ao prompt original da tarefa.

A técnica, denominada Zero-shot Chain-of-Thought (Zero-shot CoT), consiste em adicionar uma frase gatilho, como "Let's think step by step"(em português, "Vamos pensar passo a passo"), imediatamente após a pergunta da tarefa, antes de onde o modelo começaria a gerar a resposta. Como ilustrado na Figura 7, essa instrução minimalista orienta o LLM a verbalizar seu processo de raciocínio antes de fornecer a resposta final, mimetizando o comportamento do CoT few-shot sem requerer nenhum exemplo

Figura 7 – Exemplo Zero-shot CoT



Fonte: Kojima et al. (2022).

Os resultados experimentais de Kojima et al. (2022) demonstraram que o Zero-shot CoT melhora significativamente o desempenho em comparação ao zero-shot padrão em uma ampla gama de tarefas de raciocínio, incluindo aritmética (MultiArith, GSM8K), simbólica (Last Letter Concatenation, Coin Flip) e lógica (Date Understanding, Tracking Shuffled Objects). Por exemplo, no dataset MultiArith, a acurácia saltou de 17.7% para 78.7%, e no GSM8K, de 10.4% para 40.7% usando o modelo InstructGPT (text-davinci-002). Destaca-se que essa melhoria foi alcançada usando o mesmo prompt gatilho ("Let's think step by step") em tarefas muito diversas, destacando a versatilidade e a natureza task-agnostic da técnica.

Os dois paradigmas (Zero-shot e Zero-shot Cot) serão diretamente explorados na metodologia deste estudo para avaliar o impacto da estrutura do prompt no desempenho jurídico.

2.4 Grandes Modelos de Linguagem (LLMs)

Os Grandes Modelos de Linguagem (LLMs) representam uma evolução do paradigma Transformer originalmente discutido em trabalhos de larga escala como GPT-3, no qual o aumento do tamanho do modelo e do corpus de treinamento ampliou significativamente a capacidade de aprendizado em contexto, raciocínio e generalização linguística. Modelos contemporâneos, como GPT-5 e Gemini 2.5 Pro, seguem essa mesma linha teórica, operando como arquiteturas de grande porte capazes de executar tarefas complexas de compreensão, síntese, argumentação e resolução de problemas textuais em um único sistema unificado. Ambos per-

tencem a uma geração de modelos que prioriza não apenas a potência computacional, mas especialmente mecanismos de alinhamento, segurança e controle de instruções, utilizando estratégias como RLHF e ajustamentos iterativos de preferências humanas descritos em seus documentos oficiais.

Apesar de serem sucessores diretos de avanços previamente consolidados na área, esses modelos permanecem parcialmente documentados, uma vez que tanto OpenAI quanto Google DeepMind divulgam apenas descrições de alto nível de suas capacidades, protocolos de segurança e datas de corte do treinamento. No caso do GPT-5, o System Card oficial (OpenAI, 2025b) indica um *training cutoff* em setembro de 2024, enfatizando melhorias em raciocínio estruturado, estabilidade e mitigação de alucinações, embora sem detalhar arquitetura, parâmetros ou composição exata dos dados. O Gemini 2.5 Pro, por sua vez, possui um relatório técnico mais abrangente, descrevendo seu caráter multimodal unificado e uma janela de contexto ampliada, com *training cutoff* em janeiro de 2025 (Gemini Team, Google, 2025), mas ainda sem revelar número de parâmetros ou informações completas sobre o pré-treinamento. Assim, mesmo com diferenças no nível de transparência entre as empresas, ambos se inserem no mesmo cenário metodológico: são modelos de larga escala avaliados principalmente por suas propriedades empíricas, e não por detalhes internos.

Embora detalhes completos sobre a arquitetura e os dados de treinamento de modelos recentes, como GPT-5 e Gemini 2.5 Pro, não sejam inteiramente públicos, é possível caracterizá-los como modelos de ampla escala, alinhados com técnicas modernas de ajuste fino, alinhamento com preferências humanas e mecanismos de segurança baseados em RLHF (*Reinforcement Learning from Human Feedback*). Nesta pesquisa, portanto, tais modelos são tratados como sistemas de caixa-preta, ou seja, avaliados exclusivamente por seu comportamento observável em inferência.

2.5 Métricas de Avaliação

A avaliação de desempenho de Grandes Modelos de Linguagem em benchmarks padronizados requer a definição de critérios objetivos que permitam a mensuração de acertos. Nesta seção, define-se a métrica primária adotada para a análise quantitativa dos resultados desta pesquisa.

2.5.1 Acurácia (Accuracy)

A acurácia constitui a métrica primária para tarefas de Processamento de Linguagem Natural (PLN) voltadas à resolução de questões de múltipla escolha (*Multiple Choice Question Answering*). Ela é definida como a razão entre o número de predições corretas fornecidas pelo modelo e o número total de instâncias do conjunto de dados (BROWN et al., 2020).

Matematicamente, para um conjunto de dados D contendo N questões, onde y_i é o gabarito oficial e $\hat{y}_i$ é a resposta predita pelo modelo para a questão i, a acurácia (Acc) é dada por:

$$Acc = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(y_i = \hat{y}_i) \tag{2.1}$$

Onde $\mathbb{I}(\cdot)$ é a função indicadora que retorna 1 se a condição for verdadeira e 0 caso contrário. No contexto do Exame da OAB, esta métrica é diretamente correlacionada ao critério de aprovação, que exige um limiar de acerto igual ou superior a 50% (40 questões) na primeira fase.

2.6 O Exame de Ordem Unificado como Benchmark

O Exame de Ordem Unificado, regulamentado pelo Provimento nº 144/2011 do Conselho Federal da Ordem dos Advogados do Brasil (OAB) (Conselho Federal da OAB, 2011), constitui o requisito legal indispensável para que bacharéis em Direito exerçam a advocacia no país. Devido ao seu alto grau de exigência técnica, abrangência temática e complexidade interpretativa, o exame transcendeu sua função original de certificação profissional.

Na literatura acadêmica recente, exames de proficiência jurídica têm se consolidado como *benchmarks* robustos para a avaliação de capacidades cognitivas em sistemas de Inteligência Artificial. Este movimento foi fundamentado pelo estudo seminal de Katz et al. (2024), que demonstrou a eficácia desse método ao aplicar o *Uniform Bar Exam* (UBE) dos Estados Unidos para testar o modelo GPT-4. Seguindo esse precedente metodológico, este trabalho adota o Exame da OAB como a métrica equivalente para avaliar a competência de LLMs no contexto do sistema jurídico brasileiro.

2.6.1 Estrutura e Critérios de Avaliação

O certame é composto por duas etapas eliminatórias distintas, aplicadas três vezes ao ano em território nacional. Embora ambas exijam profundo conhecimento jurídico, possuem formatos de avaliação diferentes:

- **1ª Fase (Prova Objetiva):** Foco deste estudo, consiste em 80 questões de múltipla escolha com quatro alternativas (A, B, C, D). Para a aprovação, o candidato deve obter um aproveitamento mínimo de 50% (40 acertos). O conteúdo programático é multidisciplinar, abrangendo áreas fundamentais como Direito Constitucional, Civil, Penal, Administrativo, além de Ética Profissional, que historicamente possui peso significativo na composição da nota.
- **2ª Fase (Prova Prático-Profissional):** Exige a redação de uma peça processual e a resolução de quatro questões discursivas em uma área de concentração escolhida pelo candidato (ex: Direito Penal ou Tributário). Nesta etapa, é permitida a consulta a legislação não comentada (*Vade Mecum*).

3 TRABALHOS RELACIONADOS

“Ninguém ignora tudo. Ninguém sabe tudo.
Todos nós sabemos alguma coisa.
Todos nós ignoramos alguma coisa.
Por isso aprendemos sempre”.
Paulo Freire

A avaliação de Grandes Modelos de Linguagem (LLMs) em exames padronizados tornou-se uma prática consolidada para mensurar capacidades de raciocínio e generalização. Esta revisão bibliográfica organiza o estado da arte em três eixos relevantes para esta pesquisa: (i) a avaliação de LLMs em exames de ordem jurídica no cenário internacional; (ii) o desempenho de modelos em exames vestibulares e de conhecimento geral no contexto brasileiro; e (iii) a aplicação de LLMs em exames de certificação profissional especializada no Brasil.

A seguir, discute-se como cada um desses eixos contribui para a fundamentação metodológica deste estudo e como a presente pesquisa preenche as lacunas identificadas.

3.1 LLMs em Exames Jurídicos: O Precedente Internacional

O marco inicial para a avaliação de IA no domínio jurídico foi estabelecido por Katz et al. (2024), em um estudo publicado na *Philosophical Transactions of the Royal Society A*. Os autores investigaram o desempenho do GPT-4 no *Uniform Bar Examination* (UBE), o exame de ordem dos Estados Unidos.

Ao submeter o modelo às três etapas do exame (MBE, MEE e MPT) em regime *zero-shot*, Katz et al. (2024) demonstraram que o GPT-4 não apenas superou seus predecessores, mas alcançou uma pontuação (297 pontos) que garantiria a aprovação em todas as jurisdições norte-americanas, superando a média dos candidatos humanos em cinco das sete áreas do Direito avaliadas. Este desempenho posicionou o modelo próximo ao percentil 90 dos examinandos humanos . Um diferencial crucial identificado pelos autores foi a capacidade superior do GPT-4 em tarefas de não-implicação (*non-entailment*), ou seja, a habilidade de descartar com precisão alternativas incorretas, simulando a estratégia de eliminação utilizada por especialistas humanos .

Embora a pesquisa de Katz seja um marco metodológico ao estabelecer que exames de ordem podem funcionar como benchmarks robustos para avaliar raciocínio jurídico, ela também apresenta limitações importantes. O estudo concentra-se exclusivamente na língua inglesa e no sistema jurídico da *Common Law*, cujo formato argumentativo e jurisprudencial difere substancialmente do modelo normativo brasileiro. Além disso, se observa que o prompting utilizado se restringe ao cenário *zero-shot*, o que impede uma análise mais refinada sobre a contribuição do raciocínio explícito induzido via *Chain-of-Thought*, técnica que tem se mostrado particularmente relevante em tarefas interpretativas. Assim, a utilidade do trabalho

de Katz para esta pesquisa é sobretudo metodológica: o estudo valida o uso de exames de ordem como instrumento de mensuração de raciocínio complexo, ao mesmo tempo em que evidencia lacunas que precisam ser supridas quando se busca transpor tal metodologia para o contexto da OAB e para o português jurídico.

3.2 Benchmarks em Língua Portuguesa e Contexto Educacional

A transição para o cenário nacional exige a compreensão de como os LLMs processam a língua portuguesa e a lógica de exames brasileiros. Neste eixo, destacam-se os trabalhos de Peres (2023) e Rodrigues (2025).

Peres (2023) conduziu uma investigação pioneira ao aplicar LLMs (da família GPT-3.5 e GPT-4) em vestibulares de alta complexidade do Instituto Militar de Engenharia (IME) e do Instituto Tecnológico de Aeronáutica (ITA). O estudo foi crucial para demonstrar que a técnica de engenharia de prompt *Chain-of-Thought* (CoT) originalmente proposta por Wei et al. (2022), eleva significativamente a acurácia dos modelos em questões que exigem raciocínio lógico, atingindo 55% de acerto com o GPT-4. Embora o foco de Peres tenha sido um domínio técnico-científico, sua pesquisa é relevante porque evidencia que o ganho de desempenho não decorre apenas do porte do modelo, mas também do método de prompting utilizado, o que fundamenta a decisão desta pesquisa de avaliar o cenário *zero-shot CoT*. Apesar disso, certos limites metodológicos do estudo são notáveis para o avanço da literatura: O estudo de Peres se limita a modelos disponíveis até 2023 (GPT-4) e, apesar de incluir a disciplina de Português, o foco do trabalho recai sobre vestibulares de engenharia (IME/ITA). Esses elementos reforçam a necessidade de estudos voltados especificamente para o discurso jurídico brasileiro, cujo raciocínio envolve categorias linguísticas e inferenciais significativamente distintas das exigidas por problemas matemáticos e físicos de IME/ITA.

Complementarmente, Rodrigues (2025) estabeleceu um *benchmark* focado no Exame Nacional do Ensino Médio (ENEM), avaliando modelos de código aberto e proprietários. O autor observou que, enquanto modelos como o Phi-4 e DeepSeek seguiram as instruções de formatação corretamente, outros modelos abertos (como Mistral) tiveram dificuldades significativas, exigindo pós-processamento intenso. Isso reforça a necessidade de rigor na engenharia de prompt para o português.

Sua contribuição central para este trabalho reside na identificação de desafios de formatação: modelos robustos em inglês muitas vezes falham em seguir instruções de saída em português. Rodrigues sugere o uso de instruções em inglês para tarefas em português e algoritmos de pós-processamento (Regex), técnicas que foram incorporadas à metodologia de desenvolvimento da ferramenta de avaliação deste TCC.

3.3 LLMs em Exames Profissionais Especializados no Brasil

Enquanto os trabalhos anteriores focaram em exames de admissão acadêmica, Filho et al. (2025) ampliaram a discussão para o âmbito da certificação profissional médica, explorando como LLMs se comportam em avaliações que exigem raciocínio técnico especializado,

análogo ao que ocorre no Exame da OAB. No estudo piloto conduzido pelos autores, 411 questões de residência médica em Nefrologia foram submetidas ao GPT-3.5 e ao GPT-4, revelando um salto de desempenho estatisticamente significativo entre gerações, variando de 56,29% para 79,80% de acurácia. Além de confirmar a superioridade cognitiva do GPT-4 em múltiplos subdomínios da Nefrologia, o estudo introduz aspectos metodológicos particularmente relevantes para esta pesquisa, como a estratificação temática detalhada, o uso rigoroso de testes estatísticos para comparar modelos e a distinção entre questões exclusivamente textuais e aquelas que exigiam interpretação visual. Tais escolhas metodológicas reforçam a importância de avaliar LLMs não apenas em termos globais, mas também por subáreas de conhecimento, abordagem que inspira a segmentação por grandes áreas jurídicas (Direito Civil, Penal, Constitucional) realizada neste trabalho.

3.4 Síntese e Posicionamento da Pesquisa

A análise dos trabalhos relacionados revela uma evolução clara na capacidade dos LLMs, porém evidencia a carência de pesquisas nacionais com os modelos mais recentes. A tabela 1 sumariza o posicionamento deste TCC frente à literatura consultada.

Tabela 1 – Comparativo entre trabalhos relacionados e a presente pesquisa

Trabalho	**Domínio**	**Idioma**	**Modelos Principais**	**Foco da Análise**
Katz et al. (2024)	Direito (Bar Exam)	Inglês	GPT-4	Aprovação no exame (EUA)
Peres (2023)	Exatas (IME/ITA)	Português	GPT-3.5 / GPT-4	Impacto do Chain-of-Thought
Rodrigues (2025)	Geral (ENEM)	Português	Llama / Phi-4 / Deepseek	Benchmark de modelos abertos
Filho et al. (2025)	Medicina (Nefro)	Português	GPT-3.5 / GPT-4	Comparação entre gerações
Este Trabalho	**Direito (OAB)**	**Português**	**GPT-5 / Gemini 2.5 Pro**	**Frontier Models + Impacto do CoT**

Fonte: Elaborado pelo autor.

Observa-se que, embora existam estudos sobre Direito nos EUA (Katz) e exames gerais no Brasil (Peres, Rodrigues), não foi identificada, na literatura consultada, uma avaliação sistemática que combine o **domínio jurídico brasileiro** com os **modelos de fronteira de 2025** (GPT-5 e Gemini 2.5 Pro). Com o objetivo de ampliar o conhecimento sobre esse tema, esta pesquisa aplica o rigor metodológico e técnicas de *prompting* avançadas para determinar se a nova geração de IAs superou as barreiras de raciocínio jurídico no contexto nacional.

4 Metodologia

“Um bom começo é a metade”.
Aristóteles

Este trabalho caracterizou-se como uma pesquisa experimental de natureza quantitativa. A adoção dessa abordagem justificou-se pela necessidade de mensurar objetivamente a eficácia dos Grandes Modelos de Linguagem, permitindo uma análise comparativa rigorosa sobre os dados coletados. A estratégia adotada viabilizou a comparação direta entre as arquiteturas (GPT-5 e Gemini 2.5 PRO), os cenários de aplicação (zero-shot e chain-of-thought) e as diferentes disciplinas jurídicas avaliadas. A estrutura metodológica apoiou-se inicialmente na revisão bibliográfica para a fundamentação teórica e, na sequência, foi desdobrada em quatro etapas experimentais:

1. **Coleta e Tratamento dos Dados**: foram selecionadas questões objetivas da 1ª Fase do Exame da OAB com a exclusão de itens anulados ou dependentes de suporte visual.
2. **Seleção dos Modelos**: definição dos LLMs representativos do estado da arte (GPT-5 e Gemini 2.5 PRO) configurados os acessos via API para garantir a reprodutibilidade dos parâmetros de inferência.
3. **Definição da Engenharia de Prompt**: estruturação da comparação entre a abordagem direta (Zero-shot Standard) e a indução de raciocínio passo a passo (Zero-shot Chain-of-Thought).
4. **Avaliação e Análise**: extração das respostas, tabulação dos dados e cálculo das métricas de acurácia global e por disciplina.

4.1 Coleta de Dados e Tratamento de Dados

Como fonte primária para a constituição do *corpus* avaliativo desta pesquisa, foram utilizados os cadernos de prova e os gabaritos oficiais definitivos da 1ª Fase do Exame de Ordem Unificado (OAB), organizados pela Fundação Getulio Vargas (FGV).A 1ª Fase do certame é composta por 80 questões objetivas de múltipla escolha com quatro alternativas (A, B, C, D), exigindo do candidato (ou do modelo) um aproveitamento mínimo de 50% (40 acertos) para aprovação.

A escolha desse exame como benchmark justificou-se pela complexidade de seus enunciados. Diferente de testes de memorização simples, as questões da OAB frequentemente apresentam situações-problema hipotéticas ("casos concretos"), exigindo a identificação de fatos relevantes, a correlação com a legislação e a aplicação de raciocínio lógico-dedutivo para a seleção da única solução jurídica válida .

Para a composição da amostra final deste estudo, optou-se por utilizar exclusivamente as questões de duas edições contemporâneas: o 43º Exame de Ordem Unificado e o 44º Exame de Ordem Unificado.Essa escolha metodológica baseou-se na necessidade de mitigar o fenômeno de contaminação de dados (data leakage). Como os modelos selecionados possuem datas de corte de treinamento (training cutoff) anteriores à aplicação destas provas, garantiu-se que os sistemas exerceram raciocínio sobre questões estritamente inéditas, e não apenas resgataram respostas memorizadas de sua base de pré-treinamento.

O conjunto de dados inicial, totalizando 160 questões, passou por uma etapa de pré-processamento e curadoria. Foram aplicados critérios rigorosos de exclusão para garantir que a avaliação focasse exclusivamente na capacidade de processamento textual das IAs. Desse modo, foram removidas as questões anuladas pela banca organizadora (devido a inconsistências legais ou dubiedade) e aquelas que exigiam dependência visual (interpretação de imagens ou gráficos) para sua resolução. A tabela 2 detalha a composição da amostra, evidenciando o quantitativo de questões originais e o total de questões válidas após a aplicação dos critérios de exclusão.

Tabela 2 – Distribuição das questões coletadas por Exame

Estrato	Edição do Exame	Ano	Questões Originais	Questões Válidas
1	43º Exame	2025	80	78
2	44º Exame	2025	80	80
Total	**-**	**-**	**160**	**158**

Fonte: Elaborado pelo autor.

As questões válidas remanescentes foram classificadas de acordo com as disciplinas oficiais estipuladas no edital da OAB (como Direito Civil, Processo Civil, Direito Penal, Direito Constitucional, Ética Profissional, entre outras). Para viabilizar a submissão aos modelos, o conteúdo textual foi estruturado em formato JSON (JavaScript Object Notation), contemplando os campos de identificação, disciplina, enunciado integral, alternativas e o gabarito oficial .

Ao final, espera-se obter um conjunto de dados estruturado digitalmente em formato JSON (*JavaScript Object Notation*), contendo:

- **id:** Identificador único da questão no banco de dados;
- **exame:** Edição do Exame da OAB (ex: XXX);
- **ano:** Ano de aplicação da prova;
- **disciplina:** Área do Direito correspondente (ex: Direito Penal);
- **enunciado:** Texto integral da situação-problema apresentada;
- **alternativas:** Dicionário contendo as opções A, B, C e D;
- **gabarito:** Letra correspondente à resposta correta oficial.

4.2 Seleção dos Modelos de Linguagem

Para este estudo comparativo, foram selecionados dois modelos de linguagem de grande escala que representavam o estado da arte de suas respectivas desenvolvedoras no momento da realização dos experimentos (2025):

1. **GPT-5 (OpenAI):** trata-se do modelo *frontier* da família GPT, sucessor direto do GPT-4 e projetado com ênfase em capacidades avançadas de raciocínio, alinhamento e compreensão de instruções complexas. Modelos da linha GPT têm apresentado consistentemente desempenho superior em tarefas avaliativas de alto nível, como exames padronizados e problemas que exigem inferência jurídica, o que os torna referência em benchmarks de raciocínio (KATZ et al., 2024).
2. **Gemini 2.5 Pro (Google DeepMind):** representa a geração mais recente dos modelos Gemini, com avanços relevantes em análise textual, grounding factual e resolução de problemas em múltiplos domínios. Além de ser um competidor direto dos modelos GPT, o Gemini 2.5 Pro integra uma arquitetura distinta, permitindo comparar não apenas desempenho, mas também abordagens tecnológicas divergentes entre as principais famílias de LLMs do mercado (Gemini Team, Google, 2025).

A escolha desses modelos baseou-se nos seguintes critérios: (i) ambos eram considerados frontier models em 2025, encapsulando o estado da arte da inteligência artificial generativa; (ii) representavam famílias tecnológicas distintas, permitindo uma comparação interparadigmática; (iii) apresentavam resultados de ponta em benchmarks de raciocínio avançado, adequando-se ao rigor exigido pela OAB ; e (iv) estavam disponíveis via APIs oficiais durante o período experimental, garantindo padronização, reprodutibilidade, controle de versões e rastreabilidade das chamadas.

4.3 Engenharia de Prompt e Delineamento Experimental

A qualidade das respostas geradas por Grandes Modelos de Linguagem é diretamente atrelado à qualidade das instruções fornecidas. Neste estudo, a estratégia de *Prompt Engineering* foi desenhada para garantir a reprodutibilidade e minimizar alucinações. O experimento seguiu um delineamento fatorial 2×2, resultando em quatro cenários de teste distintos para cada questão do banco de dados, conforme sumarizado na Tabela 3.

Tabela 3 – Matriz do Delineamento Experimental Fatorial

Cenário	Modelo (Fator A)	Técnica de Prompt (Fator B)
1	GPT-5	Zero-shot Standard
2	GPT-5	Zero-shot CoT
3	Gemini 2.5 Pro	Zero-shot Standard
4	Gemini 2.5 Pro	Zero-shot CoT

Fonte: Elaborado pelo autor.

Para a formatação dos *prompts*, optou-se pela utilização de instruções em inglês (*English-centric prompting*). Embora o conteúdo das questões (enunciado e alternativas) seja mantido integralmente em português, as instruções de sistema ("*system instructions*") foram fornecidas em inglês. Essa decisão metodológica baseia-se em evidências de que LLMs, sendo treinados majoritariamente em corpora anglófonos, tendem a seguir instruções complexas com maior fidelidade quando estas são apresentadas em sua língua nativa de treinamento (RODRIGUES, 2025).

Foram desenvolvidos dois *templates* rígidos de *prompt*, descritos a seguir.

4.3.1 Cenário A: Zero-shot Standard

Neste cenário, o objetivo foi simular o comportamento padrão de um especialista respondendo diretamente a uma questão objetiva. O modelo não recebe exemplos prévios (*zero-shot*) e foi instruído a fornecer diretamente a resposta final. A estrutura do *prompt* foi desenhada para forçar uma saída estruturada, facilitando a extração automática posterior. A Figura 8 ilustra o modelo utilizado.

Figura 8 – Estrutura do Prompt para o cenário Zero-shot Standard

Template de Prompt (Standard)

System Instruction:
You are an expert lawyer in Brazilian Law. Answer the following multiple-choice question from the Brazilian Bar Exam (OAB).
Constraint:
Return ONLY the letter corresponding to the correct alternative (A, B, C, or D) inside braces. Do not explain. Example: {B}.
Input Data:
Question: {enunciado}
A) {alternativa_a}
B) {alternativa_b}
C) {alternativa_c}
D) {alternativa_d}
Output:

Fonte: Elaborado pelo autor.

4.3.2 Cenário B: Zero-shot Chain-of-Thought (CoT)

Para investigar se a explicitação do raciocínio melhoraria o desempenho em questões jurídicas complexas, utilizou-se a técnica *Zero-shot Chain-of-Thought*. Diferente do cenário padrão, este *prompt* inclui um gatilho cognitivo ("*Let's think step by step*") que induz o modelo a decompor o problema em etapas lógicas antes de concluir. A Figura 9 detalha esta estrutura.

A formatação das saídas foi projetada para atender ao rigor da avaliação. Enquanto no cenário *Standard*, a restrição de saída é severa (apenas a letra), no cenário *CoT*, permitiu-se a geração de texto livre ("*verbose*"), exigindo-se apenas uma *tag* final padronizada para viabilizar o processamento da resposta final.

Figura 9 – Estrutura do Prompt para o cenário Zero-shot Chain-of-Thought

Template de Prompt (CoT)

System Instruction:
You are an expert lawyer in Brazilian Law. Your task is to solve the following question from the Brazilian Bar Exam (OAB).
Reasoning Strategy:
Let's think step by step. Analyze the legal principles involved, evaluate each alternative, and justify why one is correct and the others are incorrect.
Constraint:
At the very end of your response, explicitly state the final answer inside braces, like this: FINAL ANSWER: {C}.
Input Data:
Question: {enunciado}
A) {alternativa_a}
B) {alternativa_b}
C) {alternativa_c}
D) {alternativa_d}
Output:

Fonte: Elaborado pelo autor, adaptado de Kojima et al. (2022).

4.4 Plataformas e Ambientes de Execução

Para a submissão das questões da OAB aos modelos avaliados, optou-se pela utilização da API em lote (Batch Processing) por meio de serviços de computação em nuvem das respectivas desenvolvedoras. Essa abordagem permite o envio assíncrono de múltiplas requisições estruturadas em formato JSONL (JSON Lines). Os Quadros de 1 a 4 apresentam a estrutura exata do *payload* enviado para as APIs, contendo o identificador da questão, os parâmetros e a construção do prompt.

Para a avaliação do **GPT-5**, utilizou-se a *OpenAI Batch API* (OpenAI, 2025a), acessada por meio da plataforma *OpenAI API Dashboard* com interface programada em linguagem Python. O método empregado distingue-se das requisições síncronas tradicionais por operar por meio do *endpoint* `/v1/batches`, permitindo o processamento de grandes volumes de dados em lote. O conjunto de questões foi estruturado no formato `JSONL` (*JSON Lines*), no qual cada linha corresponde a uma requisição individual contendo um `custom-id` (identificador único da questão), a especificação do modelo e o *prompt* construído conforme os *templates* definidos nas seções 4.3.1 e 4.3.2. O processamento assíncrono viabilizou o envio integral do banco de questões em uma única operação, com retorno dos resultados, incluindo respostas e metadados de consumo de *tokens*, em prazo inferior a 24 horas e com custos operacionais reduzidos.

Para a avaliação do **Gemini 2.5 Pro**, o experimento foi conduzido por meio do *Google Cloud Console*, utilizando o serviço **Vertex AI**. O *dataset* contendo os *prompts* foi previamente armazenado em um *bucket* no *Google Cloud Storage*, e a predição em lote foi configurada e executada por meio do serviço *Vertex AI Batch Prediction* (Google Cloud, 2025). O processo

foi integralmente monitorado pelo *Vertex AI Dashboard*, que forneceu métricas detalhadas de consumo, incluindo *tokens* de entrada, *tokens* de saída e, de particular relevância para a análise do cenário *Zero-shot Chain-of-Thought* (CoT), os *Thinking Tokens*, marcadores do processamento de raciocínio interno gerado pelo modelo antes da emissão da resposta final.

4.4.1 Estruturação dos Payloads de Submissão em Lote

A coleta de dados foi operacionalizada por meio do processamento assíncrono em lote (*Batch Processing*), exigindo a formatação das requisições no padrão JSONL (*JSON Lines*), onde cada linha do arquivo estende um objeto independente de controle. Devido às divergências nas especificações das APIs fornecidas pela OpenAI e pela Google, os arquivos de submissão foram divididos em quatro matrizes estruturais distintas, mapeando os cenários *Zero-shot Standard* (STD) e *Zero-shot Chain-of-Thought* (CoT) para ambos os provedores.

Os Quadros 1 e 2 descrevem o padrão adotado nas requisições direcionadas ao modelo Gemini 2.5 Pro. Nota-se que o ecossistema da Google adota uma árvore de objetos centrada na chave contents, discriminando o papel da inferência em role e encapsulando as instruções em blocos de texto puros (*parts*).

Quadro 1: Payload de envio da API Gemini 2.5 Pro no cenário Zero-shot Standard

```
{
  "request": {
    "contents": [{
      "role": "user",
      "parts": [{
        "text": "You are an expert lawyer in Brazilian Law taking the Brazilian
            Bar Exam (OAB).\n\nAnswer the following multiple-choice question by
            selecting the correct alternative.\nReturn ONLY the letter
            corresponding to the correct alternative (A, B, C, or D) inside
            braces. Do not explain. Example: {B}.\n\nQuestion:\nAntonia,
            advogada atuante na area previdenciaria... [...] \n\nAlternatives:\
            nA) Antonia advogou contra... [...]"
      }]
    }],
    "generationConfig": { "temperature": 0.0 }
  }
}
```

Fonte: Elaborado pelo autor.

Quadro 2: Payload de envio da API Gemini 2.5 Pro no cenário Zero-shot Chain-of-Thought

```
{
  "request": {
    "contents": [{
      "role": "user",
      "parts": [{
        "text": "You are an expert lawyer in Brazilian Law taking the Brazilian
            Bar Exam (OAB).\n\nYour task is to solve the following question from
             the Brazilian Bar Exam (OAB).\nLet's think step by step. Analyze
            the legal principles involved, evaluate each alternative, and
            justify why one is correct and the others are incorrect based on
            Brazilian legislation, doctrine, or jurisprudence.\nAt the very end
            of your response, explicitly state the final answer inside braces,
            like this: FINAL ANSWER: {C}.\n\nQuestion:\nAntonia, advogada
            atuante na area previdenciaria... [...] \n\nAlternatives:\nA)
            Antonia advogou contra... [...]"
      }]
    }],
    "generationConfig": { "temperature": 0.0 }
  }
}
```

Fonte: Elaborado pelo autor.

Por outro lado, a arquitetura de submissão do ecossistema OpenAI para o modelo GPT-5 requer parâmetros explícitos de gerenciamento de requisição no escopo externo do JSON, tais como o identificador único da tarefa (custom_id), o método HTTP (method) e o *endpoint* alvo (url). Ademais, a engenharia de *prompts* difere ao segmentar linearmente o contexto de comportamento (system) da submissão da pergunta jurídica (user), conforme demonstrado nos Quadros 3 e 4.

Quadro 3: Payload de envio da API GPT-5 no cenário Zero-shot Standard

```
{
  "custom_id": "44_01-std",
  "method": "POST",
  "url": "/v1/chat/completions",
  "body": {
    "model": "gpt-5",
    "messages": [
      { "role": "system", "content": "You are an expert lawyer in Brazilian Law
          taking the Brazilian Bar Exam (OAB)." },
      { "role": "user", "content": "Answer the following multiple-choice
          question by selecting the correct alternative.\nReturn ONLY the letter
           corresponding to the correct alternative (A, B, C, or D) inside
          braces. Do not explain. Example: {B}.\n\nQuestion:\nAntonia, advogada
          atuante na area previdenciaria... [...] \n\nAlternatives:\nA) Antonia
          advogou contra... [...]" }
    ]
  }
}
```

Fonte: Elaborado pelo autor.

Quadro 4: Payload de envio da API GPT-5 no cenário Zero-shot Chain-of-Thought

```
{
  "custom_id": "44_01-cot",
  "method": "POST",
  "url": "/v1/chat/completions",
  "body": {
    "model": "gpt-5",
    "messages": [
      { "role": "system", "content": "You are an expert lawyer in Brazilian Law
          taking the Brazilian Bar Exam (OAB)." },
      { "role": "user", "content": "Your task is to solve the following question
           from the Brazilian Bar Exam (OAB).\nLet's think step by step. Analyze
           the legal principles involved, evaluate each alternative, and justify
           why one is correct and the others are incorrect based on Brazilian
          legislation, doctrine, or jurisprudence.\nAt the very end of your
          response, explicitly state the final answer inside braces, like this:
          FINAL ANSWER: {C}.\n\nQuestion:\nAntonia, advogada atuante na area
          previdenciaria... [...] \n\nAlternatives:\nA) Antonia advogou contra
          ... [...]" }
    ]
  }
}
```

Fonte: Elaborado pelo autor.

4.5 Tabulação e Análise Estatística dos Dados

Após a coleta dos arquivos de saída gerados pelas APIs, procedeu-se à extração automatizada das respostas, que foram consolidadas em planilhas eletrônicas para análise comparativa. O banco de dados final registrou cada questão individualmente, associando-a à edição do exame, à respectiva disciplina jurídica e ao gabarito oficial da banca organizadora, contrastando essas informações com as predições geradas pelos modelos nos cenários *Zero-shot Standard* (STD) e *Zero-shot Chain-of-Thought* (CoT).

A avaliação dos resultados foi conduzida por meio de estatística descritiva, parametrizada pelas seguintes dimensões analíticas:

- **Acurácia Global:** A taxa de acerto de cada configuração experimental foi calculada como a razão entre o número de predições corretas e o total de questões válidas processadas, conforme a fórmula definida na seção 2.5.1. Essa métrica constitui o indicador primário de desempenho e permite a comparação direta com o limiar de aprovação de 50% estabelecido pela OAB.

- **Impacto do Chain-of-Thought:** Para avaliar o efeito da indução de raciocínio passo a passo sobre o desempenho de cada modelo, calculou-se a diferença entre as acurácias obtidas nos dois cenários de *prompting*, expressa pela fórmula:

$$\Delta = Acurácia_{CoT} - Acurácia_{STD}$$

  Valores negativos nesse indicador evidenciam queda de desempenho associada ao uso da técnica CoT. Ressalta-se que essa comparação é de natureza descritiva; a significân-

cia estatística das diferenças observadas não foi verificada por testes inferenciais, o que é reconhecido como limitação metodológica e discutido na seção 5.5.

- **Estratificação por Disciplina Jurídica:** Para o mapeamento de proficiências e fragilidades temáticas, os dados foram agrupados por disciplina jurídica. Utilizou-se lógica de soma condicional para contabilizar os acertos em cada área específica (ex: Ética Profissional, Direito Civil, Direito Tributário). A acurácia por disciplina foi obtida dividindo-se o total de acertos da área pelo número de questões que compunham aquele extrato no certame, resultando nas tabelas e gráficos de radar apresentados no Capítulo 5.

5 Resultados e Discussão

5.1 Resultados

Neste capítulo, são apresentados e discutidos os resultados obtidos a partir da submissão das questões da 1ª Fase do Exame da OAB aos modelos de linguagem GPT-5 e Gemini 2.5 Pro. A análise está estruturada em quatro dimensões principais: a caracterização da amostra validada, o desempenho global dos modelos frente ao limiar de aprovação, o impacto da técnica de indução de raciocínio passo a passo e, por fim, a estratificação do desempenho por disciplina jurídica.

5.1.1 Curadoria do Banco de Questões

O primeiro resultado refere-se à consolidação do banco de dados utilizado nos experimentos. A partir das duas edições contemporâneas do certame (43º e 44º Exames), obteve-se um conjunto inicial de 160 questões. Após a aplicação dos critérios de exclusão metodológica, remoção de itens anulados pela banca, o conjunto final resultou em 158 questões válidas, sendo 78 provenientes do 43º Exame e 80 do 44º Exame, conforme detalhado na Tabela 4.

Tabela 4 – Distribuição final das questões coletadas por Exame e Ano

Edição do Exame	Ano	Data de Aplicação	Q. Originais	Q. Válidas
43º Exame de Ordem	2025	27 de abril	80	78
44º Exame de Ordem	2025	Agosto	80	80
Total	-	-	**160**	**158**

Fonte: Elaborado pelo autor.

A escolha de edições cujas datas de aplicação (abril e agosto de 2025) são posteriores às datas de corte de treinamento de ambos os modelos (setembro de 2024 para o GPT-5 e janeiro de 2025 para o Gemini 2.5 Pro) constitui uma medida para mitigar o fenômeno de contaminação de dados (data leakage), assegurando que os modelos exerceram raciocínio genuíno sobre questões estritamente inéditas.

5.1.2 Análise arquitetural das respostas brutas

Antes de proceder à análise estatística quantitativa de acertos e erros dos modelos avaliados, faz-se estritamente necessário compreender a "anatomia"dos dados brutos retornados pelas requisições em lote (*Batch Processing*). A análise das estruturas JSON (*JavaScript Object Notation*) geradas pelas APIs da Google e da OpenAI permite mapear como o paradigma moderno de processamento de linguagem lida com a execução de tarefas complexas de raciocínio jurídico.

Uma das principais contribuições técnicas desta investigação reside na diferenciação entre o texto visível gerado pelos modelos e o volume de computação interna gasta oculta-

mente. Conforme observado nos dados analíticos das requisições, as versões mais recentes das LLMs adotam uma arquitetura de computação que desacopla os *tokens* de pensamento interno da *string* final entregue ao usuário.

Inicialmente, avalia-se o comportamento do modelo Gemini 2.5 Pro sob a ótica dos cenários *Zero-shot Standard* (STD) e *Zero-shot Chain-of-Thought* (CoT). O Quadro 5 ilustra a resposta obtida no cenário STD, onde o *prompt* exigia exclusivamente o caractere da alternativa correta.

Quadro 5: Saída bruta da API Gemini 2.5 Pro no cenário Zero-shot Standard


```
{
  "request": { "contents": [{"parts": [{"text": "You are an expert lawyer...
      Return ONLY the letter... Question: O advogado Gomes representou Dênis
      ..."}]}] },
  "response": {
    "candidates": [{
      "content": {
        "parts": [{"text": "{C}"}],
        "role": "model"
      },
      "finishReason": "STOP"
    }],
    "modelVersion": "gemini-2.5-pro",
    "usageMetadata": {
      "promptTokenCount": 433,
      "candidatesTokenCount": 3,
      "thoughtsTokenCount": 1929,
      "totalTokenCount": 2365
    }
  }
}
```


Fonte: Elaborado pelo autor.

A análise do Quadro 5 revela que, embora o texto final gerado pelo Gemini tenha consumido apenas 3 tokens para exibir {C}, o modelo consumiu internamente um total de 1.929 *tokens* de pensamento (thoughtsTokenCount). Isso evidencia que a restrição imposta no *prompt* para omitir explicações não impediu o modelo de realizar uma busca semântica e analítica interna na sua cadeia de pesos.

Em contrapartida, quando o modelo é estimulado explicitamente a externalizar o seu raciocínio passo a passo através da técnica *Chain-of-Thought*, o comportamento textual altera-se significativamente. O Quadro 6 apresenta a estrutura mapeada no cenário CoT para o Gemini 2.5 Pro.

Quadro 6: Saída bruta da API Gemini 2.5 Pro no cenário Zero-shot Chain-of-Thought


```
{
  "request": { "contents": [{"parts": [{"text": "Let's think step by step...
     Question: Euclides é advogado..."}]}] },
  "response": {
    "candidates": [{
      "content": {
        "parts": [{"text": "Com certeza! [...] Análise do art. 70, 3 da Lei
           8.906/94... FINAL ANSWER: {C}"}],
        "role": "model"
      },
      "finishReason": "STOP"
    }],
    "modelVersion": "gemini-2.5-pro",
    "usageMetadata": {
      "promptTokenCount": 505,
      "candidatesTokenCount": 1497,
      "thoughtsTokenCount": 1953,
      "totalTokenCount": 3955
    }
  }
}
```


Fonte: Elaborado pelo autor.

Ao comparar os metadados do cenário CoT (Quadro 6), observa-se que o *thoughtsTokenCount* manteve-se estável (1.953 tokens), indicando uma base de computação interna similar à do cenário Standard. Contudo, o campo *candidatesTokenCount* expandiu para 1.497 tokens, correspondendo à fundamentação jurídica em língua portuguesa gerada no corpo da resposta antes da declaração da resposta final.

Para fins de comparação arquitetural, a pesquisa investigou os mesmos cenários de submissão utilizando o modelo GPT-5 da OpenAI. O Quadro 7 expõe o objeto JSON de retorno do GPT-5 no cenário *Standard*.

Quadro 7: Saída bruta da API GPT-5 no cenário Zero-shot Standard


```
{
  "id": "batch_req_69d68f5169748190b33933f486a70418",
  "custom_id": "44_02-std-B",
  "response": {
    "status_code": 200,
    "body": {
      "model": "gpt-5-2025-08-07",
      "choices": [{
        "index": 0,
        "message": {
          "role": "assistant",
          "content": "{A}"
        },
        "finish_reason": "stop"
      }],
      "usage": {
        "prompt_tokens": 262,
        "completion_tokens": 3020,
        "total_tokens": 3282,
        "completion_tokens_details": {
          "reasoning_tokens": 3008
        }
      }
    }
  }
}
```


Fonte: Elaborado pelo autor.

A assinatura de resposta da OpenAI difere estruturalmente do padrão da Google, encapsulando os dados de consumo dentro do objeto usage. Fato assemelhado ao Gemini ocorre no GPT-5: para responder apenas a *string* {A}, o modelo despendeu um total de 3.020 *completion_tokens*, dos quais 3.008 foram alocados estritamente como reasoning_tokens ocultos. Esse dado quantifica a alta complexidade computacional interna ativada para a validação das questões do exame da OAB.

Por fim, o Quadro 8 exibe o comportamento do GPT-5 quando parametrizado para o cenário *Chain-of-Thought*.

Quadro 8: Saída bruta da API GPT-5 no cenário Zero-shot Chain-of-Thought


```
{
  "id": "batch_req_69d68f5155508190b9dd4313dee86536",
  "custom_id": "44_02-cot-B",
  "response": {
    "status_code": 200,
    "body": {
      "model": "gpt-5-2025-08-07",
      "choices": [{
        "index": 0,
        "message": {
          "role": "assistant",
          "content": "Vamos por partes. Regra aplicável - Estatuto da Advocacia
              ... [...] FINAL ANSWER: {A}"
        }
      }],
      "usage": {
        "prompt_tokens": 300,
        "completion_tokens": 6136,
        "total_tokens": 6436,
        "completion_tokens_details": {
          "reasoning_tokens": 5696
        }
      }
    }
  }
}
```


Fonte: Elaborado pelo autor.

No cenário CoT (Quadro 8), o volume total de tokens expande para 6.436, impulsionado tanto pelo crescimento do raciocínio em camadas internas (reasoning_tokens saltando para 5.696) quanto pelo desdobramento do texto em português contido no campo content.

É possível observar, a partir do mapeamento dessas quatro estruturas, que a imposição de restrições rígidas no prompt, exigindo a geração de saídas textuais mínimas e a omissão de justificativas, não inibe a raciocínio complexo internamente. Sob a ótica da engenharia, isso evidencia um diferencial estrutural significativo das atuais arquiteturas de fronteira: a restrição imposta na camada de saída não impede a busca analítica profunda e o processamento latente internamente dos modelos. E isso estabelece o alicerce técnico necessário para compreender as variações de acurácia global e por disciplina que serão apresentadas nas seções subsequentes.

5.1.3 Desempenho Global e Limiar de Aprovação

O segundo resultado diz respeito à capacidade das arquiteturas em atingir o nível mínimo de acertos exigido pela OAB. Considerando-se o cenário Standard (Zero-shot), o GPT-5 obteve uma acurácia global de 94,94%, enquanto o Gemini 2.5 Pro atingiu 98,10%. Ambos os resultados superam com ampla margem o limiar de aprovação de 50% estabelecido pela FGV para a primeira fase do certame, demonstrando um nível de proficiência jurídica bastante elevado para sistemas baseados em Inteligência Artificial.

Esses índices são notavelmente superiores aos reportados por Katz et al. (2024) para

o GPT-4 no Uniform Bar Exam (UBE) norte-americano, em que o modelo obteve acurácia de 75,7% na seção de múltipla escolha (MBE), ainda que suficiente para aprovação naquele exame. A diferença entre os contextos não deve ser interpretada como simples superioridade do exame brasileiro, uma vez que os modelos avaliados neste trabalho pertencem a uma geração posterior (GPT-5 e Gemini 2.5 Pro versus GPT-4), e o formato da OAB, com quatro alternativas, difere estruturalmente do MBE, que apresenta a mesma quantidade de opções, mas com questões de maior extensão e complexidade integrada entre as seções. Ainda assim, a comparação evidencia a rápida evolução das capacidades dos LLMs em tarefas jurídicas avaliativas, corroborando a trajetória de progresso não linear descrita por Katz et al. (2024) ao analisar a progressão histórica dos modelos GPT no MBE.

Tabela 5 – Acurácia global dos modelos GPT-5 e Gemini 2.5 Pro por cenário de prompting

Modelo	**Cenário**	**Questões Válidas**	**Acertos**	**Acurácia (%)**
GPT-5	Zero-shot Standard	158	150	94,94
	Zero-shot CoT	158	148	93,67
Gemini 2.5 Pro	Zero-shot Standard	158	155	98,10
	Zero-shot CoT	158	147	93,04

Fonte: Elaborado pelo autor.

5.1.4 O Impacto do Raciocínio Passo a Passo (Chain-of-Thought)

O terceiro resultado diz respeito ao efeito do cenário Zero-shot Chain-of-Thought (CoT) sobre o desempenho dos modelos. Contrariando a hipótese inicial de que a indução de raciocínio passo a passo produziria ganhos de acurácia, os dados revelaram um impacto negativo consistente para ambas as arquiteturas. O GPT-5 apresentou redução de 94,94% para 93,67% (queda de 1,27 pontos percentuais), ao passo que o Gemini 2.5 Pro sofreu redução mais pronunciada, passando de 98,10% para 93,04% (queda de 5,06 pontos percentuais), conforme ilustrado na Figura 10.

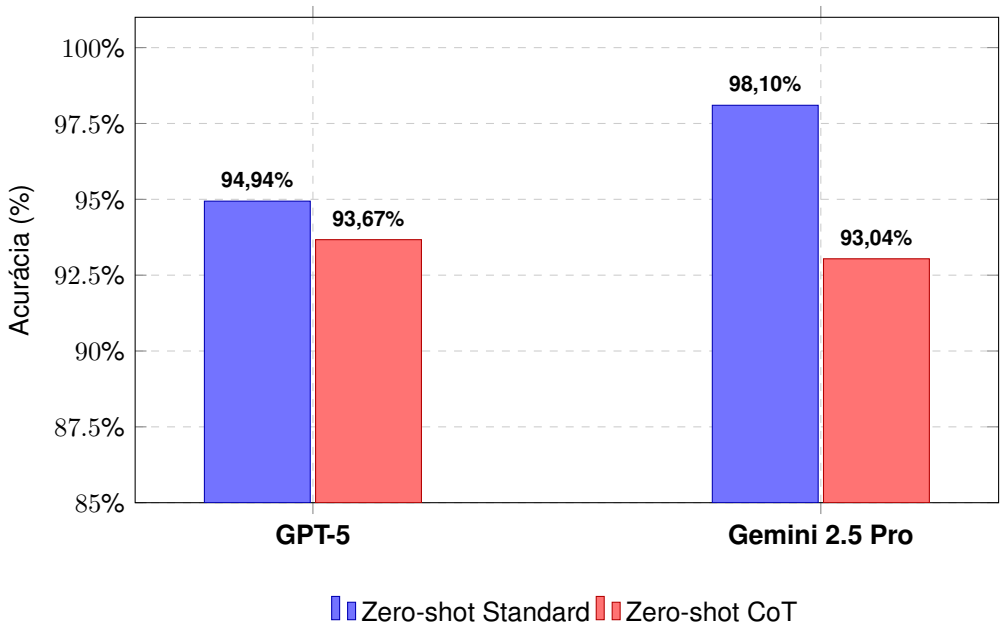
Este achado contrasta diretamente com os resultados obtidos por Peres (2023), que observou ganhos decorrentes do uso de CoT ao aplicar o GPT-4 em vestibulares de alta complexidade técnica do IME e do ITA, atingindo 55% de acerto, resultado superior ao desempenho obtido sem a técnica. A divergência entre os dois estudos pode ser explicada por uma diferença estrutural fundamental entre os domínios avaliados: questões de engenharia (física e matemática) exigem encadeamento lógico-dedutivo explícito e sequencial para chegar à resposta correta, contexto no qual o CoT atua como andaime cognitivo eficaz. Questões jurídicas da OAB, por outro lado, frequentemente demandam o reconhecimento direto e preciso de uma norma, dispositivo ou entendimento consolidado, um tipo de recuperação de conhecimento para o qual modelos de fronteira de alta capacidade já estão suficientemente calibrados sem necessidade de verbalização intermediária do raciocínio.

Uma segunda hipótese explicativa reside na natureza probabilística do processo de

geração de texto em LLMs. Ao elaborar uma cadeia de pensamento livre antes de concluir, o modelo introduz oportunidades adicionais para desvios e alucinações parciais: ao construir argumentos jurídicos detalhados, pode reforçar uma linha de raciocínio plausível, porém equivocada, comprometendo a resposta final. Esse fenômeno é reconhecido na literatura como uma limitação do CoT em modelos que já detêm alto desempenho em tarefas de recuperação de conhecimento estruturado (WEI et al., 2022). Ademais, o formato de saída "verbose"adotado no cenário CoT, em contraposição à saída restrita a uma única letra no cenário Standard, aumenta a complexidade do pós-processamento e pode introduzir falhas de formatação que impactam negativamente o cômputo final da acurácia.

A magnitude da queda foi assimétrica entre os modelos: o Gemini 2.5 Pro, que apresentou desempenho superior no cenário Standard, também exibiu maior sensibilidade negativa à adoção do CoT. O GPT-5, com desempenho de base ligeiramente inferior, mostrou-se mais robusto à variação de técnica de prompting, sustentando acurácias próximas nos dois cenários.

Figura 10 – Acurácia global dos modelos GPT-5 e Gemini 2.5 Pro nos cenários Zero-shot Standard e Zero-shot Chain-of-Thought.



Fonte: Elaborado pelo autor.

5.1.5 Desempenho por Disciplina Jurídica

O quarto resultado corresponde ao mapeamento estratificado do desempenho por disciplina jurídica, conforme detalhado na tabela 6 de acurácia por disciplina. De modo geral, ambos os modelos apresentaram desempenho consistentemente elevado na ampla maioria das áreas avaliadas. As disciplinas que revelaram maior fragilidade relativa foram Ética Profissional, Direito Tributário e Direito Previdenciário. Em Ética Profissional, o GPT-5 (Standard e CoT) e o Gemini CoT obtiveram 86,67%, enquanto o Gemini Standard atingiu 100%. Em Direito Tributário, os resultados variaram entre 80% (GPT-5 em ambos os cenários e Gemini

CoT) e 90% (Gemini Standard). O Direito Previdenciário apresentou a maior disparidade: enquanto o GPT-5 manteve 100% em ambos os cenários, o Gemini Standard caiu para 75% e o Gemini CoT atingiu apenas 50%, sugerindo que essa disciplina é especialmente sensível ao efeito negativo do CoT no modelo do Google.

O Direito do Trabalho também merece destaque analítico: o GPT-5 Standard obteve 88,89%, o GPT-5 CoT apresentou queda para 77,78%, enquanto o Gemini Standard atingiu 100% e o Gemini CoT retornou a 88,89%. Esse padrão confirma que o CoT tende a impactar negativamente ambos os modelos, mas de forma heterogênea entre as disciplinas, sendo mais pronunciado em áreas de alta especificidade normativa nas quais a precisão na aplicação de dispositivos legais específicos é determinante para o acerto.

Esse padrão de variação disciplinar encontra paralelo nos achados de Katz et al. (2024), que identificaram diferenças significativas de desempenho do GPT-4 por área jurídica no MBE: enquanto Contratos (88,1%) e Evidências (85,2%) foram os pontos fortes do modelo, Processo Civil (61,1%) e Ilícitos Civis (64,9%) representaram suas maiores fragilidades. Os autores atribuíram essa heterogeneidade à exposição diferencial do modelo a determinadas áreas durante o pré-treinamento. De modo análogo, as variações disciplinares observadas neste estudo sugerem que a composição e densidade dos dados de treinamento em matérias jurídicas específicas, particularmente aquelas de natureza mais técnica ou de menor circulação em corpora textuais públicos, como o Direito Previdenciário, continuam influenciando o desempenho diferencial dos LLMs, mesmo nas gerações mais recentes.

Tabela 6 – Acurácia por disciplina jurídica nos quatro cenários experimentais

Disciplina	**GPT-5 STD**	**GPT-5 CoT**	**Gemini STD**	**Gemini CoT**
Ética Profissional	86,67%	86,67%	100,00%	86,67%
Filosofia do Direito	100,00%	100,00%	100,00%	75,00%
Direito Constitucional	100,00%	100,00%	91,67%	91,67%
Direitos Humanos	100,00%	100,00%	100,00%	100,00%
Direito Eleitoral	100,00%	100,00%	100,00%	100,00%
Direito Internacional	100,00%	100,00%	100,00%	100,00%
Direito Financeiro	100,00%	100,00%	100,00%	100,00%
Direito Tributário	80,00%	80,00%	90,00%	80,00%
Direito Administrativo	90,00%	100,00%	100,00%	100,00%
Direito Ambiental	100,00%	100,00%	100,00%	100,00%
Direito Civil	100,00%	100,00%	100,00%	100,00%
ECA	100,00%	100,00%	100,00%	100,00%
Direito do Consumidor	100,00%	100,00%	100,00%	100,00%
Direito Empresarial	75,00%	75,00%	100,00%	87,50%
Direito Processual Civil	100,00%	100,00%	100,00%	100,00%
Direito Penal	100,00%	100,00%	100,00%	100,00%
Direito Processual Penal	100,00%	91,67%	100,00%	100,00%
Direito Previdenciário	100,00%	100,00%	75,00%	50,00%
Direito do Trabalho	88,89%	77,78%	100,00%	88,89%
Direito Processual do Trab.	100,00%	90,00%	100,00%	90,00%
Acurácia Geral	**94,94%**	**93,67%**	**98,10%**	**93,04%**

Fonte: Elaborado pelo autor.

5.2 Considerações e limitações

O delineamento desta pesquisa apresenta limitações metodológicas inerentes à avaliação de sistemas baseados em inteligência artificial. Em primeiro lugar, o escopo temporal da amostra foi restrito a apenas duas edições do Exame da OAB (43º e 44º Exames de 2025). Embora essa restrição tenha sido fundamental para garantir a não contaminação de dados, ela impede generalizações históricas mais amplas sobre a estabilidade do desempenho dos modelos ao longo de múltiplas bancas e ao longo do tempo.

Em segundo lugar, a natureza generativa e probabilística dos Grandes Modelos de Linguagem implica que os resultados podem sofrer leves variações em inferências repetidas, mesmo com parâmetros de temperatura controlados. As análises empreendidas neste estudo focaram-se na estatística descritiva (acurácia quantitativa), sem a aplicação de testes formais de significância inferencial, como o Teste de McNemar para amostras relacionadas dicotômicas, que permitiriam atestar matematicamente a significância estatística das diferenças observadas entre modelos e entre cenários de prompting.

6 Conclusão

Este trabalho investigou o desempenho comparativo dos modelos GPT-5 e Gemini 2.5 Pro na resolução das questões objetivas da 1ª fase do Exame da OAB, considerando os cenários de prompting Zero-shot Standard e Zero-shot Chain-of-Thought. Os resultados demonstram que ambos os modelos superam o limiar de aprovação de 50%, com acurácias de 94,94% (GPT-5) e 98,10% (Gemini 2.5 Pro) no cenário Standard. Esses índices indicam que a geração atual de modelos de linguagem de fronteira alcança um nível de proficiência jurídica superior ao mínimo exigido de um bacharel em Direito para a obtenção do registro na OAB.

Um achado relevante deste estudo foi o efeito negativo do Chain-of-Thought sobre o desempenho de ambos os modelos. Diferentemente do observado por Peres (2023) em questões de vestibulares de exatas, nos quais o CoT produziu ganhos ao estruturar o raciocínio matemático-lógico, no domínio jurídico da OAB a técnica provocou redução de acurácia. Esse resultado sugere que, para tarefas de recuperação e aplicação de conhecimento normativo, as arquiteturas avaliadas operam com melhor precisão quando instruídas a responder de forma direta, uma vez que a verbalização do raciocínio pode introduzir ruído e desvios na geração de tokens.

Os dados obtidos dialogam com os achados de Katz et al. (2024), que documentaram a aprovação do GPT-4 no Uniform Bar Exam norte-americano. Ao comparar aquele cenário com os índices alcançados pelo GPT-5 e pelo Gemini 2.5 Pro, observa-se um avanço de desempenho: os modelos atuais superaram as margens de aprovação em um exame formulado em língua portuguesa, cuja estrutura jurídica é distinta da Common Law americana. Isso evidencia a capacidade de generalização dos LLMs avaliados em diferentes idiomas e sistemas jurídicos. Por fim, as ferramentas baseadas em inteligência artificial para o domínio jurídico, descritas por Katz et al. (2024) como em fase de consolidação, demonstram, com a geração atual, uma aplicabilidade prática madura. Os resultados apontam a consolidação dessas arquiteturas como alternativas tecnicamente viáveis para a interpretação e análise no contexto do Direito.

6.1 Trabalhos futuros

Em primeiro lugar, sugere-se que pesquisas futuras apliquem testes de hipóteses apropriados para amostras relacionadas de dados dicotômicos, como o Teste de McNemar. A adoção dessas ferramentas estatísticas de inferência permitirá validar matematicamente a significância das diferenças de desempenho observadas entre os modelos e entre os diferentes cenários de *prompting*, conferindo um rigor estatístico ainda maior à análise comparativa.

Como segunda frente de trabalho, dado o efeito negativo do *Chain-of-Thought* identificado neste estudo, investigações futuras poderiam explorar variações de *prompting* mais sofisticadas. Estratégias como o *few-shot* CoT, acompanhado de exemplos jurídicos cuidado-

samente curados, ou abordagens estruturadas baseadas em IRAC (*Issue, Rule, Application, Conclusion*), seriam essenciais para verificar se formatos de raciocínio explicitamente calibrados para o domínio jurídico seriam capazes de reverter o efeito negativo observado e produzir ganhos de desempenho consistentes.

No âmbito da engenharia de software, uma frente de trabalho de natureza mais aplicada consiste no desenvolvimento de um *pipeline* em Python. Estruturado no padrão de projeto *Pipeline*, esse sistema seria capaz de automatizar a ingestão do banco de questões, a construção dinâmica de *prompts*, a comunicação com as APIs dos modelos e a persistência estruturada dos resultados em formato auditável. A disponibilização do código-fonte em repositório público alinharia o estudo às práticas de Ciência Aberta e viabilizaria a replicabilidade por outros pesquisadores.

Sugere-se também a investigação da aplicabilidade desses modelos como ferramentas de apoio ao estudo para candidatos da OAB. O objetivo seria avaliar se o alto nível de acurácia demonstrado nesta pesquisa se traduz em valor prático para a preparação de bacharéis em Direito, ponderando, simultaneamente, os limites éticos e pedagógicos dessa utilização.

Por fim, considerando que este estudo adotou a estratégia de English-centric prompting (instruções de sistema em inglês e questão em português), um trabalho futuro relevante seria a execução e comparação dos mesmos cenários utilizando as instruções de prompt integralmente em língua portuguesa. Isso permitiria avaliar se o contínuo avanço e alinhamento dos modelos de fronteira já é capaz de superar totalmente a barreira linguística no seguimento de instruções complexas e restrições de formatação.

Referências

ALZUBAIDI, L. et al. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, SpringerOpen, [S.l.], v. 8, n. 53, March 2021. Disponível em: <https://doi.org/10.1186/s40537-021-00444-8>. Citado nas páginas 18 e 19.

BROWN, T. B. et al. Language models are few-shot learners. In: NEURAL INFORMATION PROCESSING SYSTEMS, 33., 2020, Virtual Conference. *Advances in Neural Information Processing Systems (NeurIPS 2020)*. [S.l.]: Curran Associates, Inc., 2020. p. 1877–1901. Citado nas páginas 21, 22, 23 e 26.

Conselho Federal da OAB. *Provimento nº 144, de 13 de junho de 2011*: Dispõe sobre o exame de ordem. 2011. Acesso em: 28 nov. 2025. Disponível em: <https://www.oab.org.br/leisnormas/legislacao/provimentos/144-2011>. Citado na página 27.

FILHO, H. N. F. et al. Chatgpt performance in answering medical residency questions in nephrology: a pilot study in brazil. *Brazilian Journal of Nephrology*, Sociedade Brasileira de Nefrologia, v. 47, n. 4, p. e20240254, Oct 2025. ISSN 0101-2800. Disponível em: <https://doi.org/10.1590/2175-8239-JBN-2024-0254en>. Citado na página 29.

Gemini Team, Google. *Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities.* [S.l.], 2025. Citado nas páginas 26 e 33.

Google Cloud. *Vertex AI Batch Prediction Documentation*. 2025. Acesso em: maio 2025. Disponível em: <https://cloud.google.com/vertex-ai/docs/predictions/batch-predictions>. Citado na página 35.

KATZ, D. M. et al. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, The Royal Society, [S.l.], v. 382, n. 20230254, 2024. Disponível em: <https://doi.org/10.1098/rsta.2023.0254>. Citado nas páginas 15, 16, 27, 28 e 33.

KOJIMA, T. et al. *Large Language Models are Zero-Shot Reasoners*. 2022. Disponível em: <https://arxiv.org/abs/2205.11916>. Citado na página 25.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, Nature Publishing Group, [S.l.], v. 521, p. 436–444, May 2015. Citado nas páginas 15 e 18.

MCCARTHY, J. et al. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI Magazine*, v. 27, n. 4, p. 12, Dec. 2006. Disponível em: <https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1904>. Citado na página 18.

OpenAI. *Batch API – OpenAI Platform Documentation*. 2025. Acesso em: maio 2025. Disponível em: <https://platform.openai.com/docs/guides/batch>. Citado na página 35.

OpenAI. *GPT-5 System Card*. [S.l.], 2025. Citado na página 26.

PERES, R. S. *Grandes Modelos de Linguagem na Resolução de Questões de Vestibular: O Caso dos Institutos Militares Brasileiros*. Agosto 2023. Dissertação (Dissertação de Mestrado) — Universidade Federal do Estado do Rio de Janeiro (UNIRIO), Rio de Janeiro, Agosto 2023.

Orientador: Carlos Eduardo Ribeiro de Mello; Coorientadora: Laura de Oliveira Fernandes Moraes. Citado na página 29.

RODRIGUES, C. G. d. C. *Um benchmark para avaliação de grandes modelos de linguagem no português usando questões do exame nacional do ensino médio*. Quixadá: [s.n.], 2025. 59 p. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) – Universidade Federal do Ceará, Campus de Quixadá. Orientador: Regis Pires Magalhães. Coorientador: Luis Gustavo Coutinho do Rêgo. Disponível em: <http://repositorio.ufc.br/handle/riufc/81270>. Citado nas páginas 29 e 34.

TURING, A. M. Computing machinery and intelligence. In: ____. *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Dordrecht: Springer Netherlands, 2009. p. 23–65. ISBN 978-1-4020-6710-5. Disponível em: <https://doi.org/10.1007/978-1-4020-6710-5_3>. Citado na página 18.

VASWANI, A. et al. Attention is all you need. In: ANNUAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 31., 2017, Long Beach, CA, USA. *Proceedings...* Long Beach, CA, USA: NIPS, 2017. p. 5998–6008. Citado nas páginas 15 e 20.

WEI, J. et al. Chain-of-thought prompting elicits reasoning in large language models. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NEURIPS), 36., 2022, New Orleans, LA, USA. *Advances in Neural Information Processing Systems*. New Orleans, LA, USA: Curran Associates, Inc., 2022. p. 24824–24837. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html>. Citado nas páginas 23, 24 e 29.