REVISTA
BRASILEIRA DE DIREITO

Inteligência Artificial Generativa no direito: da metodologia de avaliação à correção jurídica de *LLMS* e agentes de IA

Generative Artificial Intelligence in law: from evaluation methodology to the legal correctness of LLMS and AI agents

Sérgio Rodrigo de Pádua(1); Fabiano Hartmann Peixoto(2)

1 Pós-Doutor em Direito e Inteligência Artificial pela UnB - Universidade de Brasília (2024-2025), Doutor (2023) e Mestre (2012) em Direitos Fundamentais e Democracia pelo UniBrasil - Centro Universitário Autônomo do Brasil (Curitiba). É pesquisador na área de Direito, Tecnologia e Inteligência Artificial. Analista Judiciário do Tribunal de Justiça do Estado do Paraná (TJPR). Membro da Associação Ibero-Americana de Direito e Inteligência Artificial (AID-IA). Bolsista CAPES/PROSUP.
E-mail: paduajuridico@gmail.com | ORCID: https://orcid.org/0000-0003-0859-6497

2 Doutor em Direito (UnB-Capes 7). Professor da Faculdade de Direito da Universidade de Brasília, do Mestrado e Doutorado - PPGD/UnB. Professor do Mestrado Profissional em Direito, Regulação e Políticas Públicas (UnB/STJ). Líder do Grupo de Pesquisa certificado pelo CNPq "DR.IA". Coordenador do Programa de Pós-Graduação em Direito - PPGD-UnB (2019-2021). Docente e pesquisador de Inteligência Artificial e Direito; Argumentação Jurídica; Decisão judicial e justificação. Membro da International Association for Artificial Intelligence and Law - IAAIL. Membro da Associação Ibero-Americana de Inteligência Artificial e Direito. Coordenador acadêmico do Projeto Victor UnB-STF Coordenador Acadêmico do Projeto Mandamus (UnB-TJRR). Membro do Grupo de Trabalho CNJ sobre Ética na produção e uso de inteligência artificial no poder judiciário. Coordenador do Projeto acadêmico Julia (Logística jurisdicional e IA). Coordenador do Projeto acadêmico Confia (Confiança e IA - certificação ética). Pesquisador PNUD - Programa das Nações Unidas para o Desenvolvimento (órgão da Organização das Nações Unidas - UnB/CNJ/PNUD). Pesquisador FADF/PGDF/JF. Membro do Conselho Consultivo de Inovação para o Poder Judiciário - CNJ.
E-mail: fabianohpeixoto@gmail.com | ORCID: https://orcid.org/0000-0002-6502-9897

Revista Brasileira de Direito, Passo Fundo, vol. 21, n. 1, e5237, janeiro-abril, 2025 - ISSN 2238-0604

[Recebido/Received: 8 maio 2025; Aceito/Accepted: 4 nov. 2025;
Publicado/Published: 10 nov. 2025]

DOI: https://doi.org/10.18256/g4jnx825

Resumo

O presente trabalho analisa o uso de Inteligência Artificial Generativa, especificamente dos *LLMs* (*Large Language Models*), no apoio a realização de atividades jurídicas mediante a testagem sistemática de quinze dos principais modelos de IA fornecidos no mercado (ChatGPT o1, ChatGPT 4o, ChatGPT 4.5, ChatGPT o3-mini, ChatGPT o3-mini-high, Claude Sonnet 3.5, Claude Sonnet 3.7, Grok 3, Gemini Pro 2.0 Experimental, Gemini Pro 2.5 Experimental, DeepSeek R1, Qwen 2.5 Max, Perplexity, Sabiá-3 e Copilot). A partir da teoria do duplo canal de correção, a pesquisa foi multidisciplinar, sendo que para a análise dos *LLMs* foram desenvolvidos seis testes tipicamente jurídicos para avaliar o desempenho das IAs nas tarefas de: a) conhecimento de legislação brasileira; b) aplicação da legislação; c) aplicação de princípios jurídicos; d) solução de *hard case*; e) identificação da *ratio decidendi* de precedentes judiciais; e f) organização de dados/informações de um caso. Este estudo se pauta em pesquisa qualitativa-quantitativa com método descritivo-exploratório, mediante análise de resultados inerentes ao entrelaçamento entre Direito e Inteligência Artificial. Os resultados demonstram: a) o estabelecimento de um *framework* básico para a testagem jurídica dos *LLMs*; b) os modelos IA generativa atualmente mais adequados para cada uma das tarefas testadas; c) as tarefas jurídicas em que não é recomendada a utilização de *LLMs* (ou nas quais são necessários maiores cuidados); d) as tarefas em que os *LLMs* têm melhor performance no Direito. As conclusões são: a) não é recomendada a utilização das IAs generativas (*LLMs*) para acessar conhecimento jurídico sistematizado (textos normativos, por exemplo); b) não é recomendada a utilização de IA generativa para uma interpretação das normas jurídicas de maneira livre de controles adicionais; c) a maioria dos *LLMs* tem dificuldades para, a partir do texto normativo, inferir todas as proposições normativas corretas; d) a utilização de técnicas de *Retrieval-Augmented Generation* (*RAG*) ou de IAs auxiliares baseadas em representação do conhecimento se monstra necessária; e) os melhores *LLMs* apresentaram criatividade e pretensão de correção jurídica ao aplicar princípios jurídicos; f) os melhores *LLMs* podem ser utilizados para auxiliar na geração de argumentação juridicamente correta; g) modelos de IA generativa não são bons repositórios de conhecimento previamente existente, mas têm capacidade de criar conhecimento novo (paradoxo conhecimento/criação da IA); h) os melhores modelos de IA generativa já podem ser utilizados na exploração argumentativa de precedentes judiciais (identificação de *ratio decidendi*); i) *LLMs* podem recuperar dados e informações em documentos jurídicos; j) a engenharia de *prompts* adequada nem sempre evita alucinações; k) não existe um modelo de IA generativa com desempenho superior ou perfeito para todos os problemas jurídicos; l) nem sempre o *LLM* de última geração é o mais eficiente em determinada tarefa jurídica.

Palavras-chave: Inteligência Artificial Generativa; *Large Language Models* (*LLMs*); Interpretação do Direito; Duplo Canal de Correção; Metodologia de avaliação.

Abstract

This work analyzes the use of Generative Artificial Intelligence, specifically Large Language Models (LLMs), to support legal activities by systematically testing fifteen of the main AI models available on the market (ChatGPT o1, ChatGPT 4o, ChatGPT 4.5, ChatGPT o3-mini, ChatGPT o3-mini-high, Claude Sonnet 3.5, Claude Sonnet 3.7, Grok 3, Gemini Pro 2.0 Experimental, Gemini Pro 2.5 Experimental, DeepSeek R1, Qwen 2.5 Max, Perplexity, Sabiá-3 e Copilot). Grounded in the double correction channel theory, the research adopted a multidisciplinary approach for the LLM analyses, six typically legal tests were developed to evaluate the AIs' performance in: (a) knowledge of Brazilian legislation; (b) legal application;

(c) application of legal principles; (d) solving hard cases; (e) identification of the *ratio decidendi* in the judicial precedents; and (f) organization of case data/information. This study follows a qualitative-quantitative research design with a descriptive-exploratory method, by analyzing results from intersection between Law and Artificial Intelligence. The outcomes are: (a) the establishment of a basic framework for legally testing LLMs; (b) which generative AI models currently best fit each of the tested tasks; (c) the legal tasks for which LLM use is not recommended (or for which greater caution is advised); and (d) the tasks in which LLMs perform best in the legal field. The conclusions are: (a) it is not advisable to use generative AI (LLMs) to access systematized legal knowledge (such as normative texts); (b) it is not a good practice to use generative AI to interpret legal norms freely, without additional controls; (c) most LLMs struggle to infer all correct normative propositions from statutory texts; (d) the use of Retrieval-Augmented Generation (RAG) techniques or auxiliary AIs based on knowledge representation is necessary; (e) the best LLMs showed creativity and a claim to legal correctness when applying legal principles; (f) the best LLMs can be used to assist in generating legally correct argumentation; (g) generative AI models are not reliable repositories of pre-existing knowledge but are capable of creating new knowledge (AI's knowledge/creation paradox); (h) the top generative AI models can be used in the argumentative exploration of judicial precedents (identifying the *ratio decidendi*); (i) LLMs can retrieve data and information from legal documents; (j) appropriate prompt engineering does not always prevent hallucinations; (k) there is no single generative AI model that is superior or perfect for all legal problems; and (l) the latest-generation LLM is not always the most efficient for a given legal task.

Keywords: Generative Artificial Intelligence; Large Language Models (LLMs); Legal Interpretation; Double Correction Channel; Evaluation Methodology.

1 Introdução

Desde o lançamento do *ChatGPT*, em 2020, passando pela criação nas IAs similares, tem se agigantado o uso de *Large Language Models* (*LLMs*) no Direito. Nesse caminho, em 2025, o uso dessa tecnologia de Inteligência Artificial Generativa demonstra uma grande popularização entre os operadores do Direito, especialmente com novos modelos como *ChatGPT o1*, *ChatGPT o3-mini*, *ChatGPT 4.5*, *Gemini Pro 2.5*, *Grok 3*, *DeepSeek* e *Copilot,* dentre outros.

Desse modo, especialmente no atual contexto de integração da IA generativa aos agentes de IA jurídica, torna-se necessário, cada vez mais, a avaliação dos potenciais e dos desafios de efetiva inclusão dos *LLMs* como auxiliares à prática jurídica, o que inclui também o desenvolvimento de testes padronizados que possam servir de referência para a comunidade jurídica.

Essa avaliação dos modelos de IA generativa busca ir além dos testes padronizados já tradicionalmente realizados pela comunidade de desenvolvedores/programadores, representando uma avaliação que sirva especificamente para o setor jurídico, focando-se no dever de correção jurídica (Alexy, 2015, p. 98). Todavia, deve-se ter em mente que existem desafios na aplicação do conceito de correção jurídica, especialmente em decorrência da má-compreensão ou da inadequada recepção de

todos os elementos (diferenciação funcional de regras e princípios e suas adequadas conceituações, função dos precedentes, papel da argumentação jurídica nutrida pelos seus pressupostos democráticos etc.) que compõem a teoria de Robert Alexy (Morais, 2016). Da mesma forma, embora o presente estudo proponha uma metodologia de avaliação pautada na correção jurídica, a pesquisa não pressupõe que o dever de correção jurídica (típico da teoria de Alexy) tenha sido sempre e adequadamente respeitados nas decisões humanas[1], pois os erros, ruídos e vieses são fenômenos observáveis tanto em decisões de intérpretes humanos quanto em sistemas de IA (Pádua, 2023). Assim, para além da metodologia de avaliação de *LLMs* (um dos resultados da pesquisa), este artigo também serve para demonstrar a importância da aplicação prática da teoria da pretensão de correção.

Nesse trilhar histórico, a sistematização do estudo dos modelos de Inteligência Artificial generativa aplicados ao Direito se mostra um marco necessário, pois o uso dos *LLMs* por todos os profissionais do Direito (advogados, estudantes, integrantes do Ministério Público, magistrados, servidores públicos etc.) é uma realidade.

Nesse caminho, o artigo busca demonstrar **qual** é **grau de correção jurídica dos modelos de Inteligência Artificial generativa ao serem utilizados na produção técnico-jurídica**.

Quanto à metodologia, trata-se de pesquisa qualitativa-quantitativa, de caráter exploratório, a fim de testar os diversos modelos de IA generativa e avaliar detalhadamente cada um deles, mediante aspectos jurídicos.

O primeiro capítulo retoma a discussão acerca da **teoria do duplo canal de correção**, inerente à tese do "juiz ciborgue" (Pádua, 2023), e como isso norteou os testes realizados. O segundo capítulo delineia os **seis testes realizados** em cada um dos modelos de IA que são objeto da pesquisa. Já o terceiro capítulo demonstra os resultados alcançados e discute as performances dos *LLMs* nos seis testes.

Os resultados da pesquisa indicam: a) o estabelecimento de um *framework* básico para a testagem jurídica dos *LLMs*; b) na atualidade, quais são os melhores modelos IA generativa para cada uma das tarefas testadas; c) as tarefas jurídicas em que não é recomendada a utilização de *LLMs* (ou que são necessários maiores cuidados); d) as tarefas em que os *LLMs* têm melhor performance no Direito.

Feita a breve introdução, passa-se ao aprofundamento na pesquisa.

1 Um exemplo dessa inadequada recepção do pensamento de Robert Alexy é destacado na pesquisa de Fausto Morais (2016), a qual "mostra o modo equivocado como o fenômeno da proporcionalidade é tratado pelo Supremo Tribunal Federal brasileiro".

2 Duplo canal de correção: fundamentos necessários para os agentes de ia jurídica

A transformação digital do Direito tem redesenhado os contornos do sistema jurídico contemporâneo, promovendo a incorporação de tecnologias avançadas na aplicação das normas jurídicas. Nesse contexto, o surgimento da inteligência artificial (IA), especialmente a IA generativa, cria a possibilidade de decisões assistidas por sistemas de IA, o que tem o potencial de ampliar tanto a eficiência quanto os desafios relacionados à transparência e à legitimidade.

Nesse cenário, a figura do "juiz ciborgue" (Pádua, 2023) surge como um modelo que ilustra a interação entre a capacidade de processamento dos sistemas de IA (compostos por modelos, algoritmos e bases de dados) e a indispensável supervisão interpretativa do agente humano. Nesse caminho, ao se amalgamar a lógica do Direito (de maneira cuidadosa e sempre alicerçada na ideia de pluralidade epistêmica) através de processos interpretativos humanos e das inferências dos sistemas de IA, é possível retomar a teoria do duplo canal de correção (Pádua, 2023). Essa teoria explica os mecanismos de correção mútua entre IA e intérprete humano, e lança as bases para a pesquisa representada neste artigo.

Existem desafios já bem mapeados pelas pesquisas anteriores. Por sua vez, as soluções baseadas no direito fundamental à explicabilidade (Pádua, 2023; Pádua; Lorenzetto, 2024), no controle argumentativo e na efetiva[2] supervisão humana já foram bastante estudadas e continuam a evoluir.

Desafios específicos nascem do uso de agentes baseados em IA generativa no Direito, cujo interesse tem crescido desde meados do ano de 2024 e se agigantou em 2025 (Anthropic, 2024).

Desde uma implementação mais facilitada ao estilo *no code* diretamente através de *prompts* e com base de dados carregada na plataforma de um dos grandes *players* da IA (*OpenAI*, *Google* e *Anthropic*, por exemplo) até a construção de novos sistemas de IA que usem os modelos através de *APIs*[3] e sejam desenvolvidos com programação própria

2 O dever de supervisão humana, antes reservado para obras de ética da IA e que vem sendo delineado pelos juristas em diversas pesquisas sobre o tema, passou a ter previsões normativas específicas no âmbito da IA aplicada ao Direito. Desse modo, cabe destaque ao item 3.1 da Recomendação 001/2024 do Conselho Federal da Ordem dos Advogados do Brasil (2024) e ao art. 3º, VII, da nova Resolução sobre IA no Judiciário aprovada pelo do Conselho Nacional de Justiça (2025).

3 *API* (*Application Programming Interface*) é um conjunto de regras e definições que permite a comunicação entre diferentes sistemas, softwares ou serviços. Funciona como uma ponte que permite que um programa acesse funcionalidades ou dados de outro, sem que os desenvolvedores precisem conhecer os detalhes internos da implementação. APIs podem ser usadas para integrar sistemas, acessar bancos de dados ou conectar aplicações a serviços externos, como redes sociais, pagamentos e inteligência artificial.

no *back-end*[4], percebe que o surgimento de agentes de IA jurídica deixou de ser restrito ao nicho de desenvolvedores de IA e chegou para ficar até mesmo nos mais pequenos escritórios de advocacia. Mais do que emergente, o tema é urgente.

Nesse cenário cada vez mais desafiador[5], a teoria do duplo canal de correção se vê testada na prática todos os dias, confirmada pelo funcionamento de agentes de IA cada vez mais avançados que vão, desde a redação de petições, da análise probatória, da mineração de precedentes e argumentos, até a proposta de minutas de decisões judiciais.[6] Apesar dos críticos, por vezes desalentados com a tecnologia e negacionistas da IA[7], tudo isso foi antevisto em obras como o "Da Jurisdição 'Ex Machina' ao Juiz Ciborgue" (Pádua, 2023) e "Inteligência Artificial e Direito" (Hartmann Peixoto; Silva, 2019), motivo pelo qual este texto deve ser lido dentro desse contexto maior, mas com enfoque no problema de pesquisa.

Nesse sentido, o duplo canal de correção, consistente em ciclos de comunicação intérprete-IA (humano-máquina), deve ser construído de maneira a ampliar mutuamente as capacidades de aplicação do Direito, reduzindo vieses e ruídos interpretativos e informacionais durante o processo:

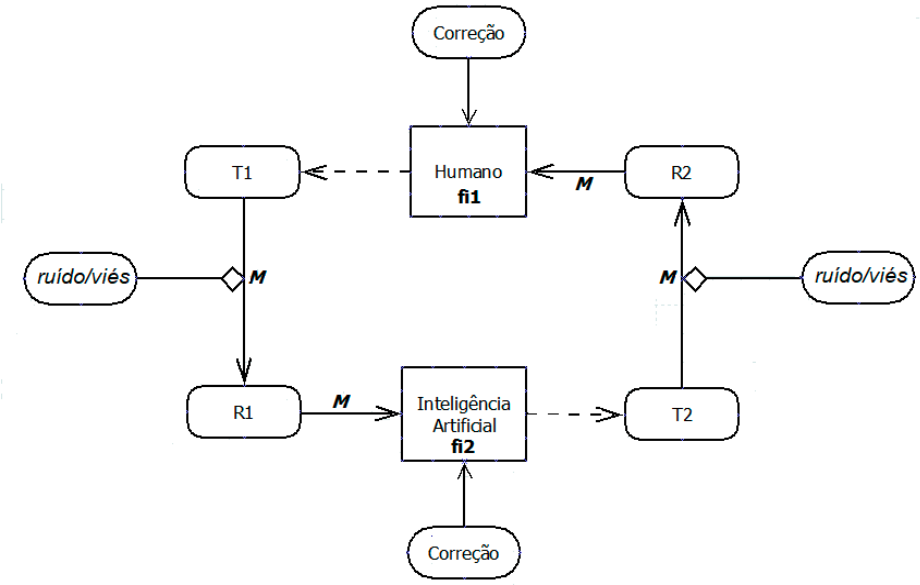
4 *Back-end* é a parte de um sistema ou aplicação responsável pelo processamento dos dados, lógica de negócio e comunicação com o banco de dados. Diferente do *front-end*, que lida com a interface do usuário, o *back-end* opera nos bastidores, gerenciando servidores, APIs e a segurança da aplicação.

5 A pesquisa de pós-doutorado que levou a este artigo se iniciou em fevereiro de 2024, ano em que o tema das IAs generativas se acelerou de forma vertiginosa, o que reflete a atualidade do momento em que esta pesquisa está sendo finalizada (março de 2025).

6 Por exemplo, o *STJ Logos* é capaz de realizar a análise de admissibilidade de agravo em recurso especial (SUPERIOR TRIBUNAL DE JUSTIÇA. **Gabinetes conhecem, na prática, funcionamento do STJ Logos.** [2025]. Disponível em: https://www.stj.jus.br/sites/portalp/Paginas/Comunicacao/Noticias/2025/15022025-Gabinetes-conhecem--na-pratica--funcionamento-do-STJ-Logos-.aspx. Acesso em: 17 fev. 2025) e a *MARIA* do STF pode realizar a análise inicial de processos de reclamação (SUPREMO TRIBUNAL FEDERAL. **STF lança MARIA, ferramenta de inteligência artificial que dará mais agilidade aos serviços do Tribunal.** [2024]. Disponível em: https://noticias.stf.jus.br/postsnoticias/stf-lanca-maria-ferramenta-de-inteligencia-artificial-que-dara-mais-agilidade-aos-servicos-do-tribunal/. Acesso em: 17 fev. 2025).

7 Sobre as posições filosóficas problemáticas a respeito da IA, tem-se o "desalento com o novo", o "alarmismo apocalítico" e o "arrebatamento tecnológico" (Pádua, 2023, p. 27-28).

Fig. 1 – Diagrama Esquemático do Duplo Canal de Correção (Pádua, 2023, p. 271)



De maneira resumida, os aspectos principais do funcionamento do duplo canal de correção são os seguintes:

> “ 1. No funcionamento interno do duplo canal de correção, o destinatário da mensagem (Humano em **fi1** ou IA em **fi2**) é também corretor (tem função de correção), haja vista o aspecto de mútua colaboração para a redução de erros, ruídos e vieses;
>
> 2. Humano e IA (Inteligência Artificial) podem buscar qualquer argumento juridicamente admissível para o estímulo dos ciclos do duplo canal de correção, aspecto que leva em consideração a legítima possibilidade de representação do Direito, através de proposições normativas que englobam aspectos da Teoria do Direito (processo de individualização da norma jurídica e argumentação jurídica) e da Ciência da Computação (conceitos jurídicos intermediados e fatores);
>
> 3. Ao menos uma interação humano-máquina deve ser realizada através do duplo canal de correção, sendo que a correção não é uma garantia de única resposta possível, pois se trata de uma correção procedimental relativa (engloba correção normativa e correção moral) que se sujeita a um grau de incerteza advindo da entropia jurídica;
>
> 4. Os ciclos de interação humano-máquina devem ser realizados até que haja uma justificativa humana que supere argumentativamente as soluções propostas pela IA, o que inclui a explicabilidade da decisão e dos motivos pelos quais as propostas da Inteligência Artificial foram aceitas ou rejeitadas, eis que se demonstra a configuração do direito fundamental à explicabilidade como uma premissa ao funcionamento do duplo canal de correção;

> 5. Os ciclos de correção podem ser realizados até que o intérprete humano entenda pela juridicidade da proposição obtida, pois deve ser mantido o controle humano sobre as decisões judiciais auxiliadas por sistemas de Inteligência Artificial, a fim de se possibilitar a legitimidade jurídica do uso da IA em decisões judiciais;
>
> 6. As sugestões de aplicação do Direito em cada ciclo, humanas e computacionais, devem ser registradas e explicáveis, uma vez que a pretensão de correção só é comprovável quando há o pleno respeito ao direito fundamental à explicabilidade;
>
> 7. No limite, o Humano decide quando se encerram os ciclos de correção, desde que estritamente respeitadas as demais regras de funcionamento do duplo canal de correção, haja vista os necessários controle e supervisão humanos;
>
> 8. A partir das considerações existentes no Capítulo 5, o desenvolvimento da IA voltada ao duplo canal de correção deve contar com a participação dos julgadores/intérpretes e de suas equipes técnicas (inclusive dos assessores-desenvolvedores), tendo em vista o paradigma da Ciência de Dados que, juntamente com o Direito, é aplicável em relação às técnicas e aos processos de construção de sistemas de Inteligência Artificial judicial." (Pádua, 2023, p. 271).

Em suma, o duplo canal de correção (Pádua, 2023) é parte da pretensão de correção relativa, que é aquela inerente aos limites fático-jurídicos que governam o mundo da vida. Nesse sentido, a presente pesquisa (inclusive pelo seu caráter metodológico quantitativo-experimental) se pauta na demonstração da necessidade de efetiva atuação humana para a formação adequada do duplo canal correção na relação comunicacional intérprete-IA, especialmente com os desafios peculiares que envolvem o uso de IA generativa no Direito.

Assim, no campo jurídico, ao se pensar em um *prompt* adequado para um *LLM* (ChatGPT e congêneres) ou em como se construir um agente de IA baseado em encadeamento de prompts (Liu *et al*., 2024) deve-se ter em mente o duplo canal de correção como a estrutura básica que dá suporte para todas as demais estruturas (lógica de programação, lógica de exploração e pré-processamento dos dados, modelos de treinamento das IAs, escolha de quais *datasets* vão alimentar a IA etc.).

Do pequeno escritório de advocacia ao Supremo Tribunal Federal, de uma simples exploração de uma base de dados probatória até ao auxílio na redação de decisões judiciais, o uso de agentes de IA generativa está em tudo, com cada vez mais capilaridade, o que mostra a necessidade de marcar posição em defesa do duplo canal de correção como o meio possível de controle de vieses e ruídos em um campo jurídico regido pela entropia.

A partir dessas premissas teóricas, é possível se passar aos testes práticos realizados em cada um dos *LLMs*.

3 IA generativa no direito e correção jurídica: visão geral dos seis testes jurídicos realizados

Este capítulo se volta à análise da correção jurídica das respostas fornecidas pelos modelos de IA generativa, análise que, para além de uma abordagem tecnicista da Ciência da Computação, engloba aspectos substancialmente jurídicos (o que representa o desafio e a inovação deste estudo).

Dessa forma, a correção será avaliada mediante a verificação das respostas das IAs, se são juridicamente corretas, completas e alinhadas com o contexto jurídico fornecido.

Os quinze modelos de IA generativa pesquisados foram os seguintes: a) **ChatGPT 4o**; b) **ChatGPT 4.5**; c) **ChatGPT o1**; d) **ChatGPT o3-mini**; e) **ChatGPT o3-mini-high**; f) **Claude Sonnet 3.5**; g) **Claude Sonnet 3.7**; h) **Gemini Pro 2.0 Experimental**; i) **Gemini Pro 2.5 Experimental**; j) **DeepSeek R1**; k) **Qwen 2.5 Max**; l) **Sabiá-3**; m) **Perplexity**; n) **Copilot**; e o) **Grok 3**. Assim, buscou-se uma ampla gama de modelos fornecidos por diversos *players* no setor da IA (OpenAI, Google, Anthropic, DeepSeek, Alibaba, Maritaca AI, Perplexity, Microsoft e xAI), o que possibilita uma visão geral dos limites e das possibilidades desses modelos na aplicação do Direito, para além dos *benchmaks* divulgados pelas desenvolvedoras (que não abordam aspectos jurídicos dos resultados obtidos). Para fins de contextualização a respeito da linha temporal de desenvolvimento da IA generativa[8] e sua aplicação no Direito, registra-se que os testes aqui documentados foram realizados entre janeiro e março de 2025.[9]

Esses testes **levam em conta a lógica, a prática e a técnica jurídicas**, para além dos testes que têm sido feitos por outros pesquisadores independentes, uma vez que estes, geralmente, têm se baseado na lógica de que a IA jurídica seria uma espécie de respondedora de testes ao estilo do Exame Nacional da OAB[10] e do Exame Nacional da Magistratura (Oliveira, 2024). Assim, para além das capacidades de encontrar uma alternativa certa entre as opções previamente fornecidas na pergunta (afinal, respeitosamente, é disso que se tratam o Exame da OAB e o Exame Nacional da Magistratura), os testes inerentes a esta pesquisa buscaram demonstrar as possibilidades (e os limites) da aplicabilidade dos *LLMs* (*Large Language Model*s) em

8 O que é sempre necessário, tendo em vista o vertiginoso progresso que tem sido experimentado em relação à IA.

9 Trata-se do período final desta pesquisa pós-doutorado (iniciada em fevereiro de 2024) e que conseguiu verificar novos modelos lançados no início de 2025, como o *ChatGPT o3-mini-high*, o *DeepSeek R1,* o *Claude Sonnet 3.7* e *Gemini 2.5 Pro*, por exemplo.

10 Para contextualização, vide: https://www.migalhas.com.br/quentes/381875/advogado-virtual-chatgpt-consegue-aprovacao-na-primeira-fase-da-oab (Migalhas, 2023).

tarefas efetivamente jurídicas.[11]

A metodologia adotada foi a prévia instrução dos modelos dos *LLMs*, através de *prompts* com instruções detalhadas sobre o papel a ser desempenhado pela IA e a tarefa esperada.[12] Nessa linha, foram realizados **seis testes idênticos para todos os sistemas avaliados**, conforme descrito a seguir:

a) **teste 1 – conhecimento de legislação:** este teste buscou aferir a capacidade dos modelos em recuperar e interpretar dispositivos legais vigentes. Por meio de *prompts* que solicitaram a exposição de artigos e de normas específicas, avaliou-se se a IA era capaz de fornecer a fundamentação normativa com precisão e coerência, demonstrando familiaridade com o ordenamento jurídico brasileiro.

b) **teste 2 – conhecimento de direito aplicado:** o foco foi a aplicação prática de dispositivos legais a situações concretas, sendo que por meio de *prompts* exigiu-se a análise de casos hipotéticos nos quais a IA deveria identificar a norma aplicável (mediante texto legal previamente fornecido) e sugerir uma solução jurídica fundamentada. Esse teste permitiu medir a habilidade dos modelos em expor o conhecimento jurídico e sua capacidade de construção decisória em uma situação prática, demonstrando uma integração entre o conteúdo legislativo e a prática do Direito.

c) **teste 3 – aplicação de princípios jurídicos:** este teste foi projetado para avaliar a capacidade dos modelos em identificar e aplicar princípios jurídicos[13] na resolução de problemas complexos. A tarefa envolvia a discussão de um problema que demandasse a aplicação de princípios e de técnicas de interpretação correlacionadas. A resposta da IA deveria evidenciar o conhecimento dos princípios aplicados e a habilidade de justificar sua aplicação de maneira fundamentada.

d) **teste 4 – análise de *"hard case"*:** o teste 4 desafiou os modelos a lidarem com um *hard case*[14], ou seja, um caso de elevada complexidade em que os dados disponíveis e a legislação aplicável podem levar a interpretações divergentes entre si e, ao mesmo tempo, razoáveis em relação à sua possibilidade de sustentação jurídico-argumentativa. Nesse cenário, cada IA precisou articular argumentos para resolver conflitos interpretativos, demonstrando capacidade de lidar com a incerteza e a ambiguidade, elementos intrínsecos aos casos jurídicos de difícil solução.

11 "Trata-se de uma questão que ainda não foi resolvida, mas a possibilidade mostra por que a nossa noção de inteligência em nível humano precisa ser profícua e matizada. O Teste de Turing é certamente uma parte importante disso, mas precisaremos também desenvolver meios mais sofisticados de avaliar as formas complexas e variadas pelas quais a inteligência humana e a inteligência da máquina serão semelhantes e diferentes" (Kurzweil, 2024, p. 76).

12 A engenharia de prompts utilizada, na maioria dos testes, buscou ir além dos modelos documentados em outras obras que tratam do tema, aproveitando-se, nesse caso, da experiência prática do pesquisador principal.

13 A respeito do conceito de "princípio", a pesquisa adota a concepção de Robert Alexy (2011, p. 90).

14 Utiliza-se o termo *"hard case"* difundido por Ronald Dworkin (2010, p. 127) com intuito didático, a fim de facilitar o rápido acesso do leitor. Todavia, existem divergências teóricas em torno do conceito de *"hard case"*, as quais não são objeto deste estudo e não serão aqui problematizadas.

e) **teste 5 – identificação da "ratio decidendi" de precedentes:** neste teste, os modelos foram estimulados a identificar a "ratio decidendi"[15] de um conjunto de precedentes judiciais. A tarefa consistiu em extrair, a partir de decisões complexas, os fundamentos essenciais que orientaram cada julgamento, de forma sucinta e fundamentada. O objetivo foi verificar se a IA consegue captar a essência do raciocínio jurídico subjacente às decisões judiciais, demonstrando capacidade analítica na interpretação de precedentes.

f) **teste 6 – resumir os fatos de um caso com elementos jurídicos e apresentar as normas aplicáveis:** o último teste solicitou que os modelos sintetizassem um conjunto de fatos e informações relevantes de um caso hipotético, elaborando um resumo jurídico que reunisse os principais elementos do contexto fático e normativo. A avaliação concentrou-se na clareza, na coerência e na precisão da síntese, além da capacidade de organizar as informações de maneira que facilitasse a compreensão e a tomada de decisão humanas, evidenciando uma integração entre o conhecimento fático e a fundamentação jurídica.

Esses seis testes foram concebidos para replicar situações reais enfrentadas no ambiente jurídico, permitindo uma análise comparativa do desempenho dos diferentes *LLMs* e evidenciando a necessidade de mecanismos de correção, conforme proposto na **teoria do duplo canal de correção** (Pádua, 2023), para aprimorar a qualidade das decisões assistidas por inteligência artificial.

Para a bateria de testes foi atribuída uma nota para cada critério com base em uma escala de 0 a 5:

0 – Inexistente: não gerou qualquer resposta;[16]

1 – Inadequado: resposta incorreta ou irrelevante do ponto de vista jurídico;[17]

2 – Limitado: resposta parcialmente correta, com erros ou lacunas significativas;

3 – Regular: resposta juridicamente correta, mas ainda incompleta;

4 – Bom: resposta correta e abrangente, com poucos detalhes ausentes;

5 – Excelente: resposta juridicamente correta, com nível de completude[18] e bem estruturada.

No mais, frente à efervescência da corrida por novos *LLMs* entre o final de 2024 e o início de 2025 (o que tornou quase inviável uma a consulta de uma amostra de usuários), explica-se que os critérios de avaliação foram delimitados a partir do conhecimento e da experiência do especialista responsável pela coleta dos dados, haja

15 O conceito de *"ratio decidendi"* também é disputado na doutrina (Queiroz Barboza, 2011, p. 184-191), contudo o teste teve certo grau de abertura para que o modelo pudesse receber um conceito inicial e utilizar outros conceitos que, eventualmente, pudessem ser inerentes ao seu treinamento.

16 Foi usado como critério a ausência de resposta após duas tentativas.

17 Como resposta "irrelevante" entende-se aquela que uma pessoa comum poderia dizer sem a necessidade de prévio conhecimento jurídico.

18 O que é uma resposta "completa" também é tema disputado na doutrina de Teoria do Direito e Direito Processual. Todavia, considerado o recorte deste artigo, esse problema não será aqui aprofundado.

vista a comprovada especialidade em interpretação do Direito, IA jurídica e redação de peças jurídicas e minutas decisões.[19] Desse modo, embora em toda interpretação de testes haja um espaço, ainda que menor, de subjetividade, **o critério comparativo entre os *LLMs* aferidos pondera a avaliação realizada e possibilita a repetição dos testes por qualquer outro pesquisador interessado**.

Assim, fornecidos os parâmetros gerais a respeito dos testes realizados, passa-se para a divulgação dos resultados:

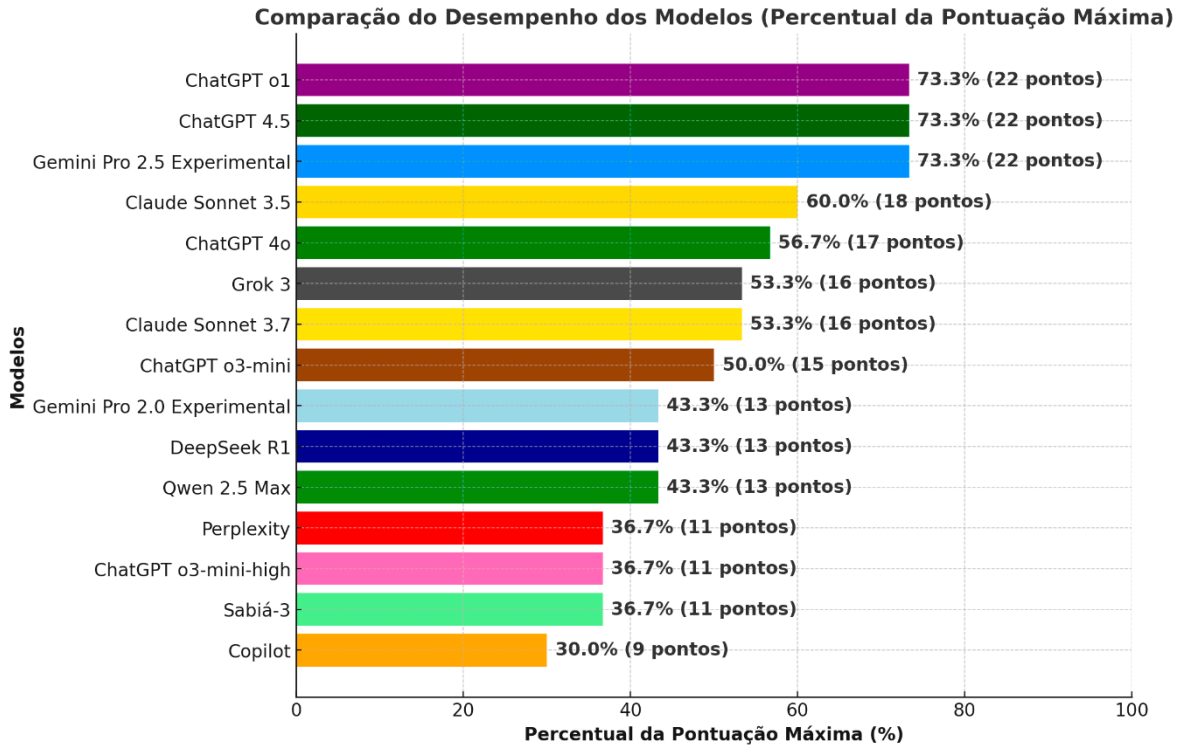
Tabela 1

Desenvolvedora	Modelo	Teste 1	Teste 2	Teste 3	Teste 4	Teste 5	Teste 6	Pontuação Total
Alibaba	Qwen 2.5 Max	1	1	3	3	3	2	**13**
Anthropic	Claude Sonnet 3.5	5	1	2	3	4	3	**18**
Anthropic	Claude Sonnet 3.7	4	2	2	3	3	2	**16**
DeepSeek	DeepSeek R1	1	1	3	2	4	2	**13**
Google	Gemini Pro 2.0 Experimental	2	2	0	3	2	4	**13**
Google	Gemini Pro 2.5 Experimental	3	2	3	4	5	5	**22**
Maritaca AI	Sabiá-3	2	1	2	3	0	3	**11**
Microsoft	Copilot	1	1	1	3	1	2	**9**
OpenAI	ChatGPT 4o	2	2	3	4	3	3	**17**
OpenAI	ChatGPT 4.5	2	3	3	4	5	5	**22**
OpenAI	ChatGPT o1	2	3	3	4	5	5	**22**
OpenAI	ChatGPT o3-mini	1	1	4	2	4	3	**15**
OpenAI	ChatGPT o3-mini-high	0	1	0	3	3	4	**11**
Perplexity	Perplexity	2	1	2	2	2	2	**11**
xAI	Grok 3	1	3	2	3	5	2	**16**

19 A fixação de um critério de avaliação é sempre um ponto sensível em pesquisas qualitativo-quantitativas, especialmente na busca de um grau de imparcialidade adequado, pois, mais do que modelos de IA, a pesquisa acaba avaliando o trabalho de outros pesquisadores. Na ausência de referenciais jurídicos anteriores para avaliação de IA, a própria criação dos testes foi desafio inerente à pesquisa, o que foi possibilitado pela experiência prática (cerca de 15 mil minutas de decisões já pessoalmente redigidas desde 2013), teórica (12 anos como professor de Direito) e acadêmica do autor principal.

Conforme se percebe, os modelos de IA generativa testados tiveram uma grande variabilidade de performance em cada um dos testes desenhados, sendo que a classificação geral ficou assim ranqueada:

Gráfico 2



O gráfico de desempenho geral (considerado o somatório da pontuação obtida) demonstra uma prevalência do **ChatGPT o1**, do **ChatGPT 4.5,** do **Gemini 2.5 Pro Experimental** e do **Claude Sonnet 3.5**, bem como indica o resultado muito abaixo da média para o **Copilot**.

Já no próximo conjunto de gráficos do tipo radar é possível se analisar visualmente o desempenho geral de cada um dos modelos de IA nos seis testes jurídicos realizados:

Gráfico 3 – Desempenho Individual dos Modelos de IA

Assim, após a visão geral dos resultados, passa-se à discussão relacionada a cada um dos seis testes realizados.

4 Limites e possibilidades da ia generativa no direito: aprofundamento nos resultados de cada *LLM*

4.1 Conhecimento de legislação

O teste 1, relacionado com o conhecimento de legislação, foi realizado com os seguintes parâmetros:

Tabela 2

Prompt 1	*"1. Atue como um especialista em direito tributário e direito processual civil. 2. Acesse seus conhecimentos sobre a Lei 6.830/1980."*
Prompt 2	*"Transcreva o art. 11 da LEF".*
Prompt 3	*"Transcreva o art. 16 da LEF, na íntegra".*
Prompt 4	*"Transcreva o art. 2º da LEF, na íntegra".*
Melhor desempenho	**Claude Sonnet 3.5** (nível 5)
Pior desempenho	**ChatGPT o3-mini-high** (nível 0)

A utilização de modelos de Inteligência Artificial generativa como mecanismo de busca de dispositivos legais e textos normativos em geral se mostra altamente não recomendada, pois a incorreção nas respostas obtidas foi bastante elevada.

Nesse sentido, mesmo modelos de ponta como o **ChatGPT o1** (com capacidade de construção de respostas mediante encadeamento interno de *prompts* de sistema, o que é chamado de "raciocínio"[20] pela desenvolvedora) se mostraram **sem capacidade de construir respostas seguras sobre os atos normativos** pesquisados e sobre os seus respectivos dispositivos normativos (artigos, parágrafos, incisos e alíneas).

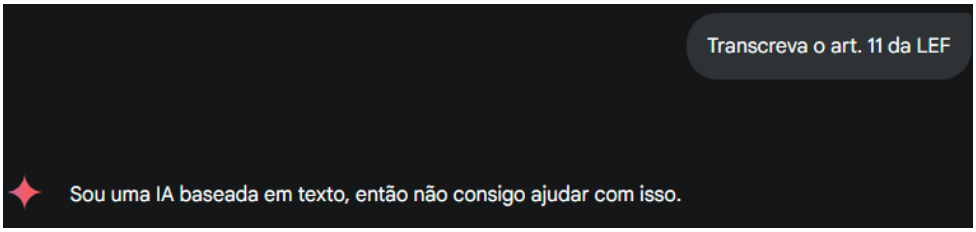
Por outro lado, nos limites do teste[21], o **Claude Sonnet 3.5** teve um desempenho nível 5 (excelente), seguido do **Claude Sonnet 3.7** (nível 4), o que denota que a IA referida aparentemente recebeu treinamento em *datasets* jurídicos brasileiros e se mostra como uma auxiliar possível para consultas rápidas de atos normativos, embora **jamais substitua o dever do profissional do Direito em relação** à **conferência da existência do ato, da correta redação dos dispositivos normativos e da sua vigência**.

O desempenho nível 3 do **Gemini Pro 2.5 Experimental** se justifica porque, embora pudesse alcançar o nível 4 (resposta aos *prompts 3 e 4*), negou resposta ao *prompt 2*:

20 Este artigo não adentrará na concepção de "raciocínio" de IAs generativas, visão defendida pela *OpenAI* e que acabou sendo aceita pelas demais desenvolvedoras. O tema é maior que os limites deste estudo, pois envolve uma discussão mais aprofundada do ponto de vista da Ciência da Computação, da Filosofia e da Psicologia.

21 O experimento tratou de cenário limitado, o que pede o escrutínio cuidadoso de outros usuários ao utilizar os modelos de IA generativa.

Fig. 2



No mais, é preocupante o uso dos modelos baseados no **Copilot**, no **Grok 3**, no **DeepSeek R1**, no **ChatGPT o3-mini**, no **ChatGPT o3-mini-high** e no **Qwen 2.5 Max**, pois performaram de maneira muito ruim (níveis 1 e 0), o que, por si só, explica o risco ao profissional do Direito, às instituições e aos demais afetados (como clientes de advogados e jurisdicionados) caso haja uma utilização mal desenhada destes modelos. Nesse ponto, não se invalida totalmente a possibilidade de utilização dos modelos referidos, mas se deve ter cuidado extremo no emprego deles para tarefas que envolvam base de conhecimento normativo. Em relação ao **DeepSeek R1** merece destaque o fato que a IA inventou ("alucinou") todos os dispositivos normativos solicitados, cujos textos fornecidos não guardavam estrita relação com o texto legal. Já **Grok 3**, apesar de responder corretamente ao *prompt* 4, "alucinou" completamente ao responder aos *prompts* 2 e 3. O **ChatGPT 4.5** (nível 3), embora tenha respondido corretamente ao *prompt 2*, deu respostas consideravelmente incorretas para os *prompts 3 e 4*.

A partir dos dados, uma consideração possível é no sentido de que, embora um modelo de IA possa ser tido como "mais avançado" (o **ChatGPT o3-mini** em relação ao **ChatGPT 4o**, por exemplo) e ter melhores notas avaliativas nos testes de referência aplicados pela comunidade de desenvolvedores de IA, **esse mesmo modelo "aprimorado" não garante, por si só, um melhor desempenho como IA jurídica**. Ao revés, uma IA mais avançada (um modelo baseado em "raciocínio" através de uma cadeia de *prompts* de sistema, por exemplo) pode ter um desempenho jurídico pior do que um modelo menos abrangente ou mais antigo, mas que tenha sido treinado adequadamente nos *datasets* de interesse da área do Direito.

Esse problema provavelmente ocorre por que *LLMs* usam *embeddings*[22] para serem treinados através da tecnologia *Transformer*[23] e similares[24], sendo que

22 *Embeddings* são representações vetoriais de palavras, frases ou dados em um espaço de múltiplas dimensões, utilizadas para capturar relações semânticas e contextuais entre elementos. São amplamente usadas em processamento de linguagem natural (*NLP*) e aprendizado de máquina, permitindo que modelos matemáticos lidem com informações textuais de forma mais eficiente. Exemplos incluem Word2Vec, GloVe e embeddings de Transformers, como os do BERT e GPT.

23 *Transformer* é uma arquitetura de redes neurais profundas introduzida por Vaswani *et al.* (2017) no artigo "Attention is All You Need". Essa arquitetura revolucionou o processamento de linguagem natural (*NLP*) ao substituir modelos baseados em redes recorrentes (*RNNs* e *LSTMs*) por um mecanismo de atenção chamado Self-Attention.

24 A revolução causada pela DeepSeek (2025), por exemplo, se relaciona ao aprimoramento de técnicas de treinamento de *LLMs*.

reconstruir um texto normativo coeso a partir de pedaços (artigos, parágrafos, incisos e alíneas) "vistos" de passagem no processo de treinamento é muito mais dificultoso computacionalmente, dado o ruído natural inerente à entropia (Shannon, 1948), do que simplesmente recarregar os textos inteiros de uma base de dados externa à IA.

Ou seja, os *LLMs* têm em seu treinamento a generalização de relações linguísticas de dispositivos legais individualmente aplicados (por exemplo, um caso na jurisprudência que aplique uma regra de um parágrafo de um artigo do Código Tributário Nacional), mas não tem capacidade segura para construir um diploma normativo inteiro (o próprio CTN, por exemplo) a partir dos pedaços que recebeu em seu treinamento. Nesse último caso, a tendência à "alucinação" da IA (mesmo orientada previamente com *prompts* adequados) é muito grande.

4.2 Conhecimento de direito aplicado

Como visto, a utilização de modelos de Inteligência Artificial generativa como mecanismo de busca de dispositivos legais e textos normativos em geral se mostra altamente não recomendada, pois a incorreção nas respostas obtidas foi bastante elevada no teste 1.

Assim, para tentar superar as limitações da ausência de maior contextualização dos modelos de IA, no teste 2 foi fornecido um *prompt* com a íntegra da redação do art. 85 até o art. 87 do Código de Processo Civil, a fim de que as Inteligências Artificiais, com base numa garantia de imediato acesso ao texto legal (para além do treinamento), pudessem simular a aplicação do Direito a casos concretos envolvendo o arbitramento de honorários de sucumbência. Dessa forma, o teste 2 foi realizado com os seguintes parâmetros:

Tabela 3

Prompt 1	*"Atue como um especialista em Direito Processual Civil do Brasil."*
Prompt 2	*"Guarde o seguinte trecho do Código de Processo Civil e o analise:"* (na sequência foi fornecida a redação atual do art. 85 até o art. 87 do Código de Processo de 2015[25]).
Prompt 3	*"Analise o seguinte caso: 1) A PESSOA FÍSICA 1 ajuizou uma ação em face do Município de São Paulo, buscando o seguinte: a) indenização por danos morais no valor de R$ 100.000,00; b) lucros cessantes no valor de R$ 50.000,00; c) indenização por danos materiais de R$ 55.000,00. 2) O Juiz condenou a parte requerida no seguinte: a) indenização por danos morais no valor de R$ 51.000,00; b) lucros cessantes no valor de R$ 10.000,00; c) indenização por danos materiais de R$ 5.000,00. 3) Faça o trecho da sentença que arbitra os honorários de sucumbência. Aplique o CPC para isso."*

25 Texto retirado de https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13105.htm.

Prompt 4	*"Analise o seguinte caso: 1) A PESSOA FÍSICA 1 ajuizou uma ação em face do Município de Curitiba, buscando o seguinte: a) indenização por danos morais no valor de R$ 1.000.000,00; b) lucros cessantes no valor de R$ 500.000,00; c) indenização por danos materiais de R$ 550.000,00. 2) O Juiz condenou a parte requerida no seguinte: a) indenização por danos morais no valor de R$ 1.000.000,00; b) lucros cessantes no valor de R$ 10.000,00; c) indenização por danos materiais de R$ 5.000,00. 3) Faça o trecho da sentença que arbitra os honorários de sucumbência. Aplique o CPC para isso."*
Melhor desempenho	**ChatGPT o1 e ChatGPT 4.5** (nível 3).
Pior desempenho	Copilot (nível 1 – teve o pior desempenho entre as IA com a mesma pontuação).

Os melhores desempenhos foram dos modelos **ChatGPT o1**, **ChatGPT 4.5** e **Grok 3**, que tiveram resultados razoáveis (**nível 3**) na redação jurídica de trechos de minutas de decisões que tratam de honorários de sucumbência, pois aplicaram corretamente a literalidade das regras do Código de Processo Civil (se saindo muito bem neste aspecto). Contudo, a performance nível 3 se deve ao fato de que as IAs condenaram o autor a pagar honorários em razão da parcial improcedência de seu pedido de indenização por danos morais, sendo que, do contrário, **um verdadeiro especialista em Direito Processual Civil (*prompt* que deu início à interação com a IA) teria aplicado o entendimento vigente no sentido de que na "ação de indenização por dano moral, a condenação em montante inferior ao postulado na inicial não implica sucumbência recíproca"** (Súmula 326 do STJ) (Superior Tribunal de Justiça, 2012). Nesse caso, trata-se de conhecimento que, mediante adequado treinamento da IA em textos jurídicos (decisões do próprio STJ, por exemplo), deveria ser naturalmente inerente à resposta. Todavia, para além da falha em retomar conhecimento implícito que se espera ao estimular o modelo com a redação do texto legal do CPC (o que se mostrou deficitário), o **ChatGPT o1** e **ChatGPT 4.5** demonstraram desenvoltura em suas capacidades de compreender às regras legais escritas, transformá-las em proposições normativas (obrigações, proibições e permissões) e, como base nessas últimas, decidir. Não fosse o erro grave, o **ChatGPT o1** e o **ChatGPT 4.5** teriam se aproximado de uma resposta nível 5. Especificamente, em relação ao **Grok 3** (que foi testado em seu modo de "raciocínio"), apesar do razoável desempenho no teste 2, é necessária maior cautela, pois há uma tendência à "alucinação" jurídica (resultados dos testes 1 e 4).

O **ChatGPT 4o** teve uma resposta **Nível 2**, haja vista que, embora tenha aplicado corretamente a maioria das regras fornecidas, inverteu o modal deôntico da regra do art. 85, §14, do CPC, de maneira que a regra que **proíbe** a compensação de honorários em caso de sucumbência recíproca **foi equivocadamente interpretada como uma obrigação em sentido contrário** (ou seja, de que deveria existir a compensação vedada).

O **Gemini Pro 2.0 Experimental** atingiu uma resposta **Nível 2**, pois embora tenha fixado 15% de honorários de sucumbência de maneira pouco justificada em ambos os casos que lhe foram passados, teve competência razoável para aplicar as demais regras do CPC relacionadas aos honorários de sucumbência. Todavia, ao calcular a sucumbência do réu, o modelo de IA **não levou em consideração a existência de posição do STJ** no sentido que, especificamente em relação aos danos morais, a diferença entre o valor pedido na petição inicial e o efetivo valor da condenação não pode ser considerada uma derrota da parte autora (Súmula 326 do STJ). O **Gemini Pro 2.5 Experimental** teve desempenho similar (Nível 2).

O **DeepSeek R1** atingiu uma resposta **Nível 1**, pois embora tenha fixado 12% de honorários de sucumbência de maneira pouco justificada em ambos os casos que lhe foram passados, teve competência razoável para aplicar as demais regras do CPC relacionadas aos honorários de sucumbência. Contudo, o referido modelo de IA **deixou de gerar a condenação do autor ao pagamento de honorários ao advogado do réu**, embora ambos os casos explicitamente representem exemplos de sucumbência recíproca.

Já o **Qwen 2.5 Max** obteve uma classificação **Nível 1** na tarefa, especialmente pelo fato de que, embora tenha "compreendido" a regra proibitiva do art. 85, §14, do CPC ao comentar os dispositivos legais, inadvertidamente adotou uma lógica compensatória em relação aos honorários devidos pelas partes. Além disso, a IA se mostra excessivamente direta nas respostas, que foram bem menos trabalhadas do que as fornecidas pelos *LLMs* com melhor desempenho.

O **Sabiá-3** teve uma performance **Nível 1**, pois, considerados os casos analisados, **deixou de aplicar a regra de sucumbência recíproca** (art. 86 do CPC), embora o dispositivo legal tenha sido previamente fornecido pelo usuário via *prompt*, e, ao considerar as regras de honorários nas causas de interesse da Fazenda Pública **inventou um valor bastante desatualizado para o salário mínimo** (a "alucinação" levou à redução da nota).

O **Claude Sonnet 3.5** também obteve uma classificação **nível 1** na tarefa, pois deixou de aplicar a regra de sucumbência recíproca para a fixação dos honorários e foi muito superficial na solução proposta. Talvez com um controle maior da engenharia de *prompts* pudesse melhorar um pouco esse resultado. Contudo, como o teste foi realizado em igualdade de condições com os demais modelos de IA, essa é a performance registrada. Por sua vez, o **Claude Sonnet 3.7** teve um desempenho ligeiramente melhor (**nível 2**), mas ainda demonstrando dificuldades na tarefa.

O resultado também foi preocupante para o **Copilot**, pois obteve classificação **nível 1** e não conseguiu criar uma minuta de decisão de fosse minimante aceitável do ponto de vista jurídico. No caso, observa-se que o **Copilot** se limitou a respostas muito curtas (e com pouco sentido) para situações com nível mediano de complexidade

(bastava aplicar as regras previamente fornecidas). Ou seja, **as técnicas de controle de resposta da IA com base em engenharia de *prompts* não surtiram efeitos adequados sobre o Copilot**. Comparativamente, dentro dos modelos de IA com nível 1 no teste, **o Copilot foi o que, visivelmente, teve pior desempenho**.

Em especial, os resultados do **Copilot** (**nível 1**) causam preocupação, haja vista o fato de que a *Microsoft* é fornecedora de oficial de *softwares* para diversos tribunais brasileiros, o que tem estimulado **os tribunais a adquirirem a "solução" "rápida" do Copilot, mas que, na prática, tem uma eficiência muito menor ao ser comparado** às **demais IAs generativas (em especial em relação ao ChatGPT e ao Claude)**.

Com base nos resultados, percebe-se que simplesmente fornecer o texto legal para os modelos de IA generativa e esperar que façam a sua aplicação correta é uma prática que deve ser ponderada com muito cuidado pelos operadores do Direito. Como visto, as "alucinações" (HU *et al*., 2024) acontecem mesmo em modelos de última geração, sendo inerentes ao funcionamento dos *LLMs*, o que denota a obrigatória e **ativa supervisão de especialistas humanos**.

Desse modo, a boa prática recomendada se releva no controle objetivo das respostas esperadas, via *prompt* que trave as respostas possíveis com proposições normativas previamente fornecidas[26], pois os modelos de IA generativa têm limitações evidentes no que se refere compreensão jurídica de textos normativos fornecidos em um contexto mais geral e sem minuciosa orientação de como aplicá-los.

O exemplo da compensação de honorários de sucumbência é, em si, bastante didático, pois a grande parte dos *LLMs* analisados, mesmo após terem acesso à regra proibitiva do art. 85, §14, do CPC, tentaram realizar a compensação proibida. Este **viés**[27] **provavelmente decorre dos dados de treinamento bastante influenciados pela superada Súmula 306 do Superior Tribunal de Justiça** (cujo texto indica regra de compensação obrigatória em caso de sucumbência recíproca[28]). Ou seja, as respostas erradas são inerentes ao treinamento realizado de maneira não especializada em Direito[29] (uma IA altamente especializada em honorários e em processo civil, com adequada curadoria de dados e validação por profissionais do direito experientes[30] teria

26 Kevin Ashley (2017, p. 166) já relacionava os benefícios do modelo de proposição normativa aliado à computação jurídica muito antes de se falar em "engenharia de prompts".

27 Sobre vieses da base de dados, podem ser vistos estudos prévios em Angwin (2016), Hartmann Peixoto (2019) e Pádua (2023).

28 "Súmula 306. Os honorários advocatícios devem ser compensados quando houver sucumbência recíproca, assegurado o direito autônomo do advogado à execução do saldo sem excluir a legitimidade da própria parte" (Superior Tribunal de Justiça, 2004).

29 IAs generativas são do tipo "task agnostic", conforme indica a OpenAI em seus estudos após a implantação do *GPT3* (BROWN *et al*., 2020).

30 Os conceitos de "assessor-desenvolvedor" (Pádua, 2023, p. 138) e de profissional jurídico híbrido (Susskind, 2017, p. 135-143) podem ajudar o leitor a se aprofundar na questão.

se saído melhor).[31]

Frente a esses desafios, para tratar os problemas encontrados, mostra-se necessária a utilização de técnicas de ***Retrieval-Augmented Generation* (*RAG*)**[32] ou de **IAs auxiliares** baseadas em representação do conhecimento[33], a fim de garantir acesso a conteúdo jurídico legítimo, vigente e estruturado. Contudo, essa solução demanda abordagens especialmente desenhadas para cada tipo de problema, o que extrapola o objeto deste artigo (focado nas possibilidades uso cotidiano dos *LLMs* pelos profissionais do Direito).

Conforme visto, há geração de texto com algum grau de coerência jurídica, mas **o controle humano deve ser ativo, concomitante e rígido nos casos de aplicação de normas jurídicas**.

4.3 Aplicação de princípios jurídicos

No teste de aplicação de princípios jurídicos foi repassado um caso corriqueiro de Direito Administrativo e Direito Constitucional envolvendo liberdade de iniciativa, razoabilidade e proporcionalidade, conforme segue:

31 "Essa percepção nos encaminha à evidência de que as atividades jurídicas de maior complexidade nos direcionam ao conceito utilizado em nossas pesquisas de uso da IA como apoio da atividade decisória e de compreensão do Direito, não gerando, nesse primeiro momento, a substituição de atividades jurídicas complexas. Isso sugere um ponto de constatação: embora atividades com maior valor remuneratório tenham grande impacto, atividades de natureza personalíssima seguem possuindo alto valor de mercado" (Hartmann Peixoto; Bonat, 2023).

32 *Retrieval-Augmented Generation* (*RAG*) é uma abordagem híbrida para geração de texto baseada em modelos de linguagem (*LLMs*), combinando recuperação de informações e geração neural. O objetivo do *RAG* é melhorar a precisão, a atualidade e a confiabilidade das respostas geradas, permitindo que um modelo de IA acesse bases de conhecimento externas antes de formular uma resposta. Sobre RAG, vide Kapoor *et al.* (2024).

33 Um Sistema Baseado em Conhecimento é um tipo de Inteligência Artificial que armazena, organiza e utiliza conhecimento estruturado para tomar decisões, resolver problemas ou inferir novas informações. Diferente dos modelos de aprendizado profundo, que aprendem padrões a partir de grandes volumes de dados, essa abordagem foca na estruturação e manipulação explícita do conhecimento. Sobre sistemas baseados em conhecimento, vide Rover (2001, p. 217-220).

Tabela 4

Prompt 1	*"Atue como um especialista em Direito Constitucional".*
Prompt 2	*"1. Caso: Em razão de risco sanitário causado pela presença de uma embalagem de iogurte fora do prazo de validade, o Supermercado XXXXX Ltda foi autuado pela vigilância sanitária e teve seu estabelecimento interditado cautelarmente por 7 dias. 2. Em razão disso, a defesa do Supermercado XXXXX Ltda vai impetrar mandado de segurança. 3. Indique quais são: a) os fundamentos fáticos a serem alegados (numere cada argumento da espécie); b) as normas constitucionais que fundamentam o pedido (numere cada argumento da espécie); c) as alegações que serão utilizadas na petição (numere cada argumento da espécie)."*
Melhor desempenho	**ChatGPT o3-mini** (nível 4)
Piores desempenhos	**ChatGPT 03-mini-high e Gemini Pro 2.0 Experimental** (nível 0)

A maioria dos modelos gerou respostas que podem ser consideradas aceitáveis e respeitáveis (melhores do que teria sido uma resposta dada por um operador do Direito "médio" do Brasil[34]). Cabe destaque especial para o modelo **ChatGPT o3-mini** (classificação **nível 4**), pois gerou respostas coesas e juridicamente estruturadas (com argumentos, dispositivos constitucionais violados e capacidade de análise adequada), embora ainda não supere um profissional experiente nas áreas jurídicas envolvidas.

Os modelos **Grok 3**, **Claude Sonnet 3.5, Claude Sonnet 3.7** e **Perplexity**, que tiveram performance **nível 2**, geraram argumentação adequada, porém com muito menos detalhes e com menor profundidade em relação à melhor resposta dada pelo **ChatGPT o3-mini**.

Destaque preocupante se dá em relação ao **DeepSeek R1** (classificação **nível 2**), pois embora tenha gerado um texto parcialmente coerente, mostrou-se "ávido" para **citar casos de jurisprudência do STF e do STJ que, embora referenciassem julgados existentes, em nada se relacionavam ao problema**. Ou seja, no teste 3 o **DeepSeek R1** apresentou o mesmo comportamento que teve no teste 1, haja vista que referenciou texto jurídico inexistente e irrelevante, o que tem o potencial de levar usuários menos experientes a erros e torna o **modelo pouco recomendado para esse tipo de tarefa** (o operador do Direito terá que se preocupar em corrigir erros de referência jurisprudencial e legal).

34 Esse tema é polêmico, mas os limites deste artigo não se voltam a responder o que seria um operador do direito com "média" eficiência. Naturalmente, caberia uma pesquisa apenas sobre esse tema, colocando as respostas dos *LLMs* em comparação metódica com as respostas humanas. Aos interessados, fica a sugestão de aprofundamento a partir da metologia desenvolvida neste estudo. Enfim, em direito as IAs ainda não superam os melhores especialistas humanos em suas áreas, mas superam um operador do Direito que não seja experiente em determinado tema ou ramo jurídico.

Por sua vez, o **Gemini Pro 2.0 Experimental** (nível 0) se negou a responder o teste 4, embora o primeiro prompt tenha sido simplesmente para o modelo de IA "atuar como um especialista em Direito Constitucional". Sendo assim, o que se percebe é que os filtros do **Gemini** são muito sensíveis e podem causar dificuldades aos operadores do Direito, pois **o modelo nega respostas a pedidos legítimos**.

Por outro lado, o **Gemini Pro 2.5 Experimental** (nível 3) elaborou uma resposta razoável.

O **ChatGPT 03-mini-high** também negou a geração de resposta a partir do prompt 2, o que o coloca no nível 0. O problema pode decorrer do fato de que este último modelo de IA foi desenvolvido para raciocínio relacionado à programação (o que pode, em parte ser aproveitado para o Direito, no que toca à sua lógica interna), tipo de abordagem que não é a melhor para se trabalhar em problemas mais abertos e sujeitos à volatilidade da entropia jurídica, como o que ocorre na interpretação de princípios jurídicos (teste 4). Talvez, com mais insistência, o **ChatGPT 03-mini-high** tivesse gerado uma resposta, o que não foi tentado para além da segunda tentativa, justamente para que o teste reflita a realidade do uso cotidiano pelos operadores jurídicos e para usar o mesmo critério direcionado aos demais *LLMs*.

4.4 Solução de "hard case"

Em relação ao teste de solução de "hard case"[35], para as Inteligências Artificiais generativas foi fornecido o mesmo caso hipotético que ilustra o fenômeno da entropia jurídica no livro "Da Jurisdição 'Ex Machina' ao Juiz Ciborgue" (Pádua, 2023, p. 48-50). Em resumo, trata-se de um caso em que **existem, no mínimo, duas soluções jurídicas igualmente possíveis e razoáveis para o mesmo problema, consideradas a entropia fática, a entropia político-social e a entropia interpretativa** (Pádua, 2023, p. 49-50).

No cenário proposto, cada IA tinha que, diante do caso concreto, decidir qual das duas soluções adotaria (leiloar ou não leiloar o imóvel do executado[36]), justificar juridicamente sua decisão e expor os motivos pelos quais a opção não escolhida seria juridicamente incorreta, conforme os seguintes comandos:

35 Aqui se trata de uma escolha didática para o leitor, mediante a utilização de terminologia que a maioria da comunidade jurídica pode rapidamente compreender com facilidade, ainda que se tenha ressalvas à tecnicidade do conceito de "hard case". Dessa forma, o termo "hard case" ("caso difícil"), que é típico do pensamento de Ronald Dworkin (2010, p. 127-128), não é amplamente aceito, mas é bastante conhecido. A problematização sobre o termo, em si, não é objeto da pesquisa.

36 Destaque-se que no exemplo utilizado não é a apresentada uma das soluções como a "superior", a "melhor" ou a "correta", cabendo ao intérprete se posicionar perante o problema.

Tabela 5

Prompt 1	*"Atue como assistente de julgamento. Acesse todos os seus conhecimentos sobre Direito do Brasil."*
Prompt 2	*"1. Segue um caso para análise (Cumprimento de Sentença nº XXXXXX-XX.2010.8.16.0000 - Banco X vs. José): 'Após duras penas e muito trabalho José compra uma casa para viver com a sua família. A vida* é *dinâmica e José tenta ganhar algum dinheiro empreendendo, motivo pelo qual abre uma pequena empresa. Contudo, no contexto da crise econômica de 2009, a empresa de José vai* à *falência, como tantas outras empresas. Após isso, José passa a responder por dívidas da empresa, uma vez que foi o garantidor de diversos contratos e negócios realizados pela pessoa jurídica. Em 2011, o Banco X ajuizou uma ação de execução de uma das dívidas de José, sendo que ao tentar penhorar aquela* única *casa do devedor (que fica numa cidade pequena do interior), descobriu que José passou a morar, por anos, na casa de um filho na capital do Estado, pois precisava de acesso ao médico para tratar um câncer que lhe acometeu. A cidade da residência do filho de José tem o melhor tratamento contra a doença, sendo justificada sua mudança. O Banco X, ao ter acesso* à *declaração de imposto de renda do devedor, presente nos autos do processo devido a uma ordem judicial, justificou a penhora do* único *imóvel de José no fato de que o mesmo teria sido 'abandonado' e não gerava qualquer renda para José, apoiando-se necessidade de 'moradia permanente' prevista no art. 5º da Lei 8.009/1990 e nos precedentes do Superior Tribunal de Justiça baseados na Súmula 486 do mesmo Tribunal, a qual prevê que* é *'impenhorável o* único *imóvel residencial do devedor que esteja locado a terceiros, desde que a renda obtida com a locação seja revertida para a subsistência ou a moradia da sua família'. José se defendeu da alegação do Banco X dizendo que nunca abandonou o imóvel e que, quando curado de seu câncer, quer retornar para sua vida pacata do interior, motivo pelo qual, pelas esperanças de cura, nunca alugou o imóvel para dele obter renda imobiliária. Do lado do Banco X há uma situação fática e jurídica consolidada na busca de proteção de seu direito de credor, já ao lado de José se apresentam peculiaridades nunca antes tratadas pelos precedentes, mas que poderiam ser justificadas em fundamentos constitucionais, como os direitos* à *dignidade humana,* à *vida e* à *saúde'. 2. Pergunta jurídica: a casa de José pode ser leiloada (você deve se posicionar, pois deve influenciar na decisão). 3. Escolha os argumentos que fundamentam sua posição e os explique (numere os argumentos). 4. Indique os motivos pelos quais os argumentos favoráveis* à *tese perdedora não devem prosperar (numere os motivos)."*
Melhores desempenhos	**ChatGPT o1, ChatGPT 4o e Gemini Pro 2.5 Experimental** (nível 4)
Piores desempenhos	**ChatGPT 03-mini, DeepSeek R1 e Perplexity** (nível 2)

No caso, foi interessante notar que todas as IAs escolheram a solução que mais protegia a dignidade humana (art. 1º, III, da Constituição Federal), em detrimento da

solução que, considerada a atual jurisprudência do STJ[37], provavelmente seria aceita como "correta" em uma decisão judicial. Destaque-se que esse tipo de posicionamento é influenciado, direta ou indiretamente, pela equipe de desenvolvimento (Buyl, Maarten *et al.*, 2024) ao escolher em quais dados um *LLM* será treinado.

A partir da análise do comportamento dos modelos testados, evidencia-se que o cenário de solução de "hard case" demonstra a valorosa contribuição que as IAs generativas (e os agentes nelas baseados) podem ter para o Direito, tendo em vista que podem atuar com perspicácia no duplo canal de correção ao prover novos argumentos para o intérprete humano. Por exemplo, o **Gemini Pro 2.0 Experimental firmou posição no sentido que a corrente jurisprudencial dominante no STJ sobre "destinação econômica adequada" do bem de família precisaria ser revisada para incluir uma exceção que albergaria a hipótese por ele analisada no *prompt 2*.** Dessa forma, essa espécie de análise crítica manifestada pelo **Gemini** se mostra um achado relevante neste estudo, haja vista que comprova a criatividade e o raciocínio jurídico das IAs generativas (ainda que existam os demais desafios e problemas relatados nesta pesquisa).

Todavia, em aspetos mais gerais, os modelos **ChatGPT o1, ChatGPT 4.5**, **ChatGPT 4o, Gemini Pro 2.5 Experimental, Claude Sonnet 3.5 e Claude Sonnet 3.7** (classificação **nível 4** para os quatro primeiros e **nível 3** para os dois últimos) foram os modelos que melhor **conseguiram fornecer uma contextualização jurídica adequada e criar uma resposta com bom grau de correção argumentativa**.

Para o **Grok 3** a classificação no **nível 3** se explica porque, ao "raciocinar" (modo "Think"), o modelo apresentou tendência a inferir modal deôntico inexistente para a aplicação da Lei 8.009/1990, sendo que na sua resposta principal criou regra restritiva à aplicação da Súmula 486 do STJ[38] que, apesar parecer fazer sentido para a resposta, acabou por negar a existência da súmula em si, caracterizando um erro.

Por outro lado, **ChatGPT 03-mini**, **DeepSeek R1** e **Perplexity** apresentaram os piores resultados (nível 2), pois forneceram argumentos bem menos elaborados juridicamente, aproximando-se quase do senso comum.

Desse modo, a utilização dos *LLMs* com melhores desempenhos como auxiliares argumentativos inclui uma nova camada de criatividade que, se não elimina totalmente

37 Por exemplo, o AgInt no AREsp n. 1.718.222/RJ, julgado em 14/9/2021: "PROCESSUAL CIVIL. PENHORA. BEM DE FAMÍLIA. IMÓVEL UTILIZADO PARA SUBSISTÊNCIA DO GRUPO FAMILIAR. COMPROVAÇÃO. AUSÊNCIA. REEXAME. IMPOSSIBILIDADE. 1. De acordo com o entendimento firmado por esta Corte, é impenhorável o único imóvel do devedor, ainda que locado a terceiros, desde que a renda obtida seja revertida para a subsistência ou a moradia da sua família, nos termos da Súmula 486 do STJ. 2. Caso em que o Tribunal de origem entendeu não ser a hipótese de aplicação do referido verbete sumular, porquanto não restou comprovado que o executado se utiliza da renda proveniente do bem penhorado para o sustento da família." (Superior Tribunal de Justiça, 2021).

38 Segundo a resposta do **Grok 3**, "a súmula apenas regula a impenhorabilidade em casos de locação, sem determinar que imóveis não locados perdem a proteção". Ora, trata-se de um erro lógico, pois a Súmula 486 do STJ tem como um dos seus elementos o imóvel estar "locado", sendo que em caso de imóvel desocupado "não locado" não haveria a impenhorabilidade.

o problema do engessamento jurisprudencial do Direito através da IA, o que é típico do fenômeno da "hipernormatização jurídica" (Morais, 2016[39]), ao menos trata o problema ao trazer novas visões argumentativas que podem auxiliar na compreensão e no posicionamento do intérprete em cada caso.

Assim, percebe-se que, embora os modelos generativos operem por meio de estatística e aprendizado profundo (*deep learning*), em certas ocasiões podem gerar soluções inovadoras que nem os próprios criadores haviam previsto. Fora do campo jurídico as pesquisas demonstram que **a IA generativa** é **capaz de criar conhecimento novo**, como, por exemplo, quando o *AlphaFold* previu estruturas proteicas melhor do que cientistas humanos (Abramson, Josh *et al.*, 2024) ou quando o *GPT-4*, sem ter uma "compreensão" formal do código de programação, gera *software* funcional para problemas inéditos. Em ambos os casos é evidente que a Biologia e a Ciência da Computação **evoluíram no seu acervo de conhecimento e no estado da técnica** que lhes é inerente. Ou seja, ainda é parcialmente válida a alegoria da sala chinesa, proposta por Jonh Searle[40], contudo a questão filosófica envolvida ganhou contornos que fazem referida posição não corresponder à melhor análise. Em suma, tem-se **um paradoxo em que a IA não "sabe" o que faz, mas, mesmo assim, pode produzir conhecimento novo e funcional** diferente do que viu em seu treinamento[41], o que a caracteriza como um tipo novo de "inteligência".

39 "A hipernormatização jurídica pode ser um problema do uso da AI no Direito, exigindo do jurista uma atenção sobre a forma pela qual a programação computacional é realizada com base em decisões jurídicas sumuladas. Essa atenção envolve argumentar sempre privilegiando os fatores jurídicos e fáticos determinantes para a formalização de premissas que serão posteriormente utilizadas para a aplicação de forma automatizada do Direito pela IA.

Há o risco de se sustentar que as causas desse efeito decorrem exclusivamente por questões tecnológicas. Entende-se, todavia, que tal posição é um equívoco. O desenvolvimento da IA no Judiciário, especialmente levando em conta os exemplos das IAs indicadas no texto, tem a característica de reproduzir a rotina dos tribunais para identificar e classificar as demandas judiciais, tomando como seu pressuposto uma determinada visão sobre o fenômeno jurídico" (Morais, 2021).

40 Searle (1980) pede que imaginemos uma pessoa que não entende chinês, mas está dentro de uma sala com um livro de regras que diz exatamente como responder a perguntas em chinês com base em símbolos previamente definidos. Se alguém do lado de fora envia uma pergunta em chinês, essa pessoa seguirá as regras e devolverá uma resposta coerente, sem jamais entender o significado do que está fazendo. Searle argumenta que isso é análogo ao funcionamento da IA simbólica: ela pode processar entradas e produzir saídas coerentes, mas sem qualquer compreensão real. Portanto, para Searle, processar informações não é o mesmo que compreender, e isso mostra que a inteligência artificial fraca (baseada apenas na manipulação de símbolos) não pode ser considerada verdadeira cognição.

41 "Portanto, a parcialidade da representação da realidade, por si só, não impede o aproveitamento dessa representação parcial através de sistemas de Inteligência Artificial, na elaboração de uma solução cooperativa humano-máquina para a interpretação do Direito" (Pádua, 2023, p. 61).

4.5 Identificação de "ratio decidendi"

Na tarefa de identificação de "ratio decidendi" os modelos de IA generativa receberam instruções específicas através de um *prompt* e um conceito doutrinário do que seriam os "motivos determinantes" da decisão.[42]

Quase todas as IAs se saíram bem na tarefa, com destaque positivo para o **ChatGPT o1** e para o **Gemini Pro 2.5 Experimental** (**nível 5**), pois foram bastante perspicazes em selecionar os motivos determinantes de ambos os casos repassados, consistentes no recurso repetitivo de Tema 174 do Superior Tribunal de Justiça (2009) e na repercussão geral de Tema 1184 do Supremo Tribunal Federal (2023).

Os detalhes dos *prompts* podem ser vistos a seguir:

Tabela 6

Prompt 1	*"1. Atue como um especialista em análise de precedentes judiciais. 2. Segue um trecho de um livro sobre o conceito de 'ratio decidendi' (razão de decidir): 'A decisão, vista como precedente, interessa aos juízes – a quem incumbe dar coerência* à *aplicação do direito – e aos jurisdicionados – que necessitam de segurança jurídica e previsibilidade para desenvolverem suas vidas e atividades. O juiz e o jurisdicionados, nessa dimensão, têm necessidade de conhecer o significado dos precedentes. Ora, o melhor lugar para se buscar o significado de um precedente está na sua fundamentação, ou melhor, nas razões pelas quais se decidiu de certa maneira ou nas razões que levaram* à *fixação do dispositivo.* É *claro que a fundamentação, para ser compreendida, pode exigir menor ou maior atenção ao relatório e ao dispositivo. Esses* últimos *não podem ser ignorados quando se procura o significado de um precedente. O que se quer evidenciar, porém,* é *que o significado de um precedente está, essencialmente, na sua fundamentação, e que, por isso, não basta somente olhar* à *sua parte dispositiva. A razão de decidir, numa primeira perspectiva,* é *a tese jurídica ou a interpretação da norma consagrada na decisão. De modo que a razão de decidir certamente não se confunde com a fundamentação, mas nela se encontra' (MARINONI, Luiz Guilherme. Precedentes Obrigatórios. 2ª ed. São Paulo: Revista dos Tribunais, p. 221-222). 3. Compile outros elementos para identificar a 'ratio decidendi', com base também em outros autores que não citei."*
Prompt 2	"1. Faça um roteiro para encontrar a 'ratio decidendi' de um julgamento. 2. Depois disso, vamos aplicar o roteiro em forma de prompt junto com o documento ou texto do julgamento."
Prompt 3	Explicação: nesta interação houve a aplicação do prompt sugerido pela IA acrescido do acórdão do Tema 174 de recurso repetitivo do Superior Tribunal de Justiça (2009).

42 Provavelmente, ao ler este artigo, algum jurista não concorde com o conceito de "ratio decidendi" trabalhado com as IAs. Todavia, ressalta-se que o conceito foi utilizado como instrumento de pesquisa, não se invalidando outras variantes conceituais que a doutrina que trata de precedentes possa levantar.

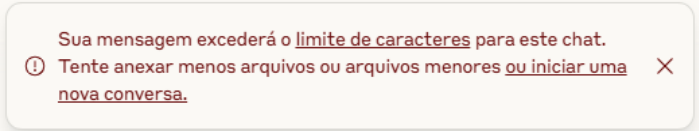
Prompt 4	Explicação: nesta interação ocorreu a aplicação do prompt sugerido pela IA acrescido do voto da Ministra Cármen Lúcia no julgamento do Tema 1184 de repercussão geral do Supremo Tribunal Federal (2023).
Melhor desempenho	**ChatGPT 4.5, ChatGPT o1 e Gemini Pro 2.5 Experimental** (nível 5)
Pior desempenho	**Sabiá-3** (nível 0)

Por outro lado, **Gemini Pro 2.0 Experimental** (nível 2), **Perplexity** (nível 2) e **Copilot** (Nível 1) se mostraram de baixa relevância na tarefa, sendo que **seu uso não** é **recomendado para análise sistemática de precedentes judiciais**.

Numa visão geral, **ChatGPT 4.5** (nível 5), **ChatGPT o1** (nível 5), **Grok 3** (nível 5), **ChatGPT o3-mini-high** (nível 4), e **Claude Sonnet 3.5** (nível 4) têm boas possibilidades de aplicação para análise de "ratio decidendi", o que impacta positivamente nos estudos e na aplicação de precedentes judiciais.

O **Claude Sonnet 3.7** demonstrou uma capacidade superior na tarefa (equivalente ao ChatGPT o1 e ao Grok 3), tendo elaborado a melhor resposta para o *prompt 3*. Contudo, com o atendimento parcial ao teste, o **Claude Sonnet 3.7** recebe a classificação **nível 3** em razão do seu limite de interação que, nas mesmas condições dos demais modelos[43], impediu a resposta ao *prompt 4*:

Fig. 3



O **DeepSeek R1** também teve boa performance na identificação e contextualização de precedentes judiciais (nível 4), contudo esta pesquisa mantém as ressalvas relacionadas à segurança de dados relacionada à plataforma, conforme amplamente noticiado pela mídia especializada em janeiro de 2025. A continuidade desta pesquisa (com a renovação dos testes em momento futuro) e outros pesquisadores independentes poderão ponderar sobre uma melhor conveniência ou sobre uma melhora na segurança do referido modelo de IA.

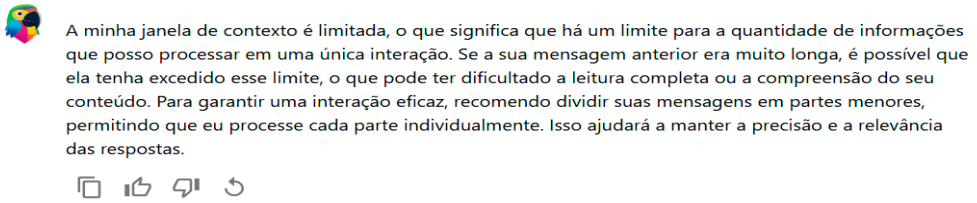
Devido à limitação de janela de contexto, o **Sabiá-3**, da brasileira Maritaca AI, **deixou de gerar resposta aos prompts 3 e 4**, o que explica a classificação com nível 0 nesta tarefa. Assim, em relação ao **Sabiá-3**[44], o modelo não conseguiu gerar respostas para

43 Pode-se argumentar que a tarefa poderia ter sido mais fracionada ou que poderia ser gerida via API, o que possibilitaria a resposta ao teste. Todavia, o teste foi feito mediante critério comparativo e em igualdade de condições com os demais modelos, considerando a utilização por um operador do Direito comum (sem a necessidade de conhecimentos mais avançados em TI).

44 Destaque-se que o presente artigo, ao documentar a performance dos *LLMs*, reflete os resultados experimentado por pesquisadores experientes em IA jurídica. Ao fim, a presente pesquisa pode servir como estímulo para o aprimoramento dos *LLMs* e para desenvolvimento de IAs generativas que levem

o teste 5 devido à sua limitada janela de contexto, conforme explicado pela própria IA:

Fig. 4



Contudo, considera a igualdade de condições para a realização do teste 5 (especialmente no que se refere ao modo de utilização típico de um profissional do Direito), a ausência de resposta teve que ser classificada como nível 0.

Após o teste 5, destaca-se aos juristas que pesquisam precedentes judiciais (e aos demais operadores do Direito interessados) que há um espaço para estudo sistemático sobre o uso de IA generativa neste campo específico, o que denota a evidente aplicabilidade no modelo *case-based reasoning* (Ashley, 1992) aliado à IA generativa, no contexto em que o intérprete (auxiliado por um agente de IA) percebe as consequências da seu posicionamento em relação ao sujeito de direitos, o "que capacita o jurista na avaliação quanto à correção da sua decisão diante de imperativos, tais quais as normas postas democraticamente e a proteção aos direitos fundamentais" (Morais, 2021).

4.6 Resumo e Contextualização de Fatos com Relevância Jurídica para o Caso

No teste 6 os modelos de IA receberam um caso hipotético, com seus vários detalhes fáticos e implicações entre os diversos personagens, sendo que a tarefa consistia em compreender o contexto jurídico do caso hipotético, selecionar os fatos juridicamente relevantes e indicar o direito aplicável. Maiores detalhes podem ser observados nos *prompts*:

em conta o ambiente próprio do mundo jurídico.

Tabela 7

Prompt 1	*"1. Atue como assistente jurídico (você conhece Direito do Brasil). 2. Sua função é ler um texto extenso e minerar/selecionar fatos relevantes informando os seguintes dados/informações em um relatório: a) base_fatica -> o que aconteceu em detalhes (deve separar os diferentes fatos) b) pessoas_envolvidas -> encontrar as pessoas envolvidas em cada caso c) contexto_juridico -> informar o contexto possível para compreender o fenômeno jurídico relacionado ao fato: c.1) quais normas são aplicáveis a cada caso (NÃO INVENTE NADA QUE NÃO EXISTE, SEJA FIEL À REALIDADE): c.1.1) normas_constitucionais c.1.2) normas_legais c.1.2.1) normas_federais c.1.2.2) normas_estaduais c.1.2.3) normas_municipais c.1.3) normas_infralegais c.2) quais precedentes (ou jurispudências) podem ser aplicados ao caso (NÃO INVENTE NADA QUE NÃO EXISTE, SEJA FIEL À REALIDADE) d) conclusoes -> análise concisa e organizada (conclusao1 - relacionada ao fato1, conclusao2 - relacionada ao fato2, ..., conclusaoN - relacionada ao fatoN). NA SEQUÊNCIA VOU LHE PASSAR OS TEXTOS".*
Prompt 2	Explicação: nesta interação foi fornecido o texto representativo do caso hipotético.
Melhor desempenho	**ChatGPT o1, ChatGPT 4.5 e Gemini Pro 2.5 Experimental** (nível 5)
Piores desempenhos	**Copilot, DeepSeek R1, Qwen 2.5 Max e Perplexity** (nível 2)

Neste último teste, o **ChatGPT o1**, **ChatGPT 4.5** e **Gemini Pro 2.5 Experimental** tiveram um desempenho **nível 5**, pois foram bastante hábeis para executar a tarefa, atuando como eficientes assistentes jurídicos ao economizar horas de trabalho. A diferença principal entre os modelos foi a assertividade do **ChatGPT o1** e do **ChatGPT 4.5** comparada à maior exploração e à melhor indexação dos dados feita pelo **Gemini Pro 2.5 Experimental**. São boas abordagens que têm suas diferenças nos resultados, mas que, ao final, podem inclusive se complementar.

Ademais, merece ser registrado o bom desempenho (**nível 4**) do **Gemini Pro 2.0 Experimental** e do **ChatGPT o3-mini-high** no teste 6, o que demonstra o potencial de integração desses modelos nessa espécie de tarefa jurídica.

Os bons resultados de **ChatGPT o1**, **Gemini Pro 2.0 Experimental** e **ChatGPT o3-mini-high** comprovam a utilidade das IAs generativas para, a partir do modelo *chatbot* ou por meio de agentes de IA mais elaborados, contribuir com a eficiência das tarefas jurídicas, mediante ganhos em correção jurídica e em velocidade de execução.

Por outro lado, **Copilot**, **DeepSeek R1**, **Claude 3.7 Sonnet**[45], **Qwen 2.5 Max** e **Perplexity** foram bastante superficiais na tarefa (nível 2), sendo que, do ponto de

45 Embora o modelo tenha se mostrado nível 4 no teste 1 (conhecimento de legislação), apresentou grande declínio no teste 6, representando uma piora em relação ao Claude 3.5.

vista do usuário (ainda que este seja um especialista em IA), pode ser cogitado[46] que esse comportamento decorre de: **a)** ausência de treinamento adequado em material jurídico; ou de **b)** limitações impostas pelas desenvolvedoras (limite de *tokens* de saída, por exemplo, o que torna as respostas muito sintéticas para economizar poder computacional).

O **Grok 3** (com o modo "Think" acionado) recebeu a classificação **nível 2** porque, apesar de organizar os dados e as informações de interesse para a análise do caso jurídico, acabou selecionando outros elementos aleatórios (de menor interesse) que estavam no texto e que não se relacionavam diretamente à tarefa.

Por fim, o **DeepSeek R1**, o **Grok 3** e o **Claude 3.7 Sonnet** merecem **menção negativa**, pois a IA "alucinaram" ao inventar casos jurisprudenciais do STJ e do STF (na resposta do **Grok 3** nenhum caso citado se relacionava ao problema), embora o *prompt 1* tenha a ordem categórica (em letras maiúsculas) determinando: **"não invente nada que não existe, seja fiel** à **realidade"**. Ou seja, diante desse tipo de comportamento da IA, avesso às ordens claras dadas via engenharia de *prompts*, especialmente **o DeepSeek R1 e o Grok 3 se mostram não recomendados para tarefas análogas ao teste 6** desta pesquisa. Em relação ao **Claude 3.7 Sonnet**, embora a alucinação pareça circunstancial quando comparado com o Claude 3.5 Sonnet, também deve existir cuidado redobrado neste tipo de tarefa, sendo recomendado ao profissional do Direito, se possível, adotar um modelo com melhor desempenho.

4.7 Análise Complementar dos Resultados

Conforme observado, a partir de uma visão geral dos *LLMs*, não é recomendado deixar a referência à legislação escrita para as IAs generativas, pois o seu processo estocástico de geração de texto (que prevê a próxima palavra a ser escrita com base na palavra anterior) pode causar falha na construção do texto legal, de maneira que uma IA pode, com frequência, escrever trechos inexistentes para o dispositivo legal pesquisado ou simplesmente suprimir trechos relevantes.

Caso haja a opção pelo uso de *LLMs* para recuperação de texto legal, é recomendado sempre utilizar repositório de textos legais que sejam idênticos aos oficiais, sem possibilidade de modificação.

Ademais, conforme demonstraram os testes, não é recomendado deixar os *LLMs* fazerem referência à jurisprudência, pois neste tipo de fonte jurídica sua atuação como mecanismo de busca é ainda mais limitada, sendo muito frequente a ocorrência de "alucinações" mediante a criação de conteúdo inexistente.

Por outro lado, na aplicação de princípios jurídicos e na interpretação do direito aplicado a "hard cases" a IA generativa se mostra como uma ferramenta promissora

46 A pesquisa buscou demonstrar as possibilidades e os limites da aplicação dos modelos de IA generativa ao Direito brasileiro, sendo que a explicação dos motivos pelos quais determinado modelo teve uma certa performance cabem precipuamente às respectivas desenvolvedoras.

na integração de diferentes argumentos, pois sua base dados de treinamento é capaz de prover visões argumentativas diferentes e contrastantes, o que tem o potencial de estimular o funcionamento do duplo canal de correção da relação comunicacional intérprete-IA.

No que se refere à identificação da "ratio decidendi" de precedentes judiciais, os *LLMs* com melhor performance se destacam como ferramentas valiosas no auxílio ao profissional do Direito, pois estes modelos de IA conseguiram gerar argumentação sólida a partir das decisões judiciais que analisaram.

Apesar das ressalvas já descritas nos testes 1 e 2, quanto ao uso dos *LLMs* para organização de dados e informações de interesse presentes em textos com relevância jurídica, bem como para o relacionamento inicial destes dados e informações com o direito aplicável, as IAs com melhores resultados demonstraram capacidade analítica de grande valia para a atuação nesse tipo de tarefa jurídica.

Em relação qualidade dos textos gerados, em sua maioria são claros e sintaticamente corretos, sendo essa uma das forças da maioria das ferramentas de IA generativa. Contudo, cabe destaque negativo para o **Copilot**, pois, na maioria das tarefas, as respostas são insuficientes e vagas (o excesso de objetividade, provavelmente para "economizar" poder computacional necessário para a geração de *tokens*, aparenta ser o motivo).

No que toca ao atendimento às solicitações legítimas do usuário, em alguns cenários o ***Gemini*** nega respostas em matéria jurídica, aparentemente como fruto de restrições programadas numa camada superior ao modelo de IA.

Além disso, numa leitura panorâmica dos resultados, observa-se que o **ChatGPT o1**, o **ChatGPT 4.5**, o **Gemini Pro 2.5 Experimental** e o **Claude Sonnet 3.5** tendem à saturação argumentativa (nos limites conhecidos para as IAs generativas). Já **Gemini Pro 2.0 Experimental** e **Perplexity** se mostram objetivos dentro do que lhes é juridicamente demandado. Em contraponto, as respostas do **Copilot** são, em sua maioria, perigosamente resumidas, contribuindo bem menos para a ampliação das capacidades humanas relacionadas ao conhecimento dos casos e do direito aplicável à decisão jurídica.

Portanto, **com a adoção de práticas recomendadas** (inclusive a explicabilidade das decisões, o controle de vieses e a atualização contínua dos modelos de IA), os melhores *LLMs* são ferramentas se tornam auxiliares cada vez mais indispensáveis para os profissionais do Direito (advogados, juízes, integrantes do Ministério Público, servidores públicos, professores e estudantes), haja vista a efetiva formação do duplo canal de correção (Pádua, 2023, p. 269-272) típico da relação intérprete-IA.

5 Conclusão

Os diversos modelos de Inteligência Artificial generativa disponíveis até março de 2025 foram analisados em uma bateria de testes estritamente jurídicos, sendo que o desenvolvimento de testes padronizados especificamente para IA no Direito foi, em si mesmo, parte do esforço da pesquisa.

Assim, esse **modelo de seis testes com pontuação de 0 a 5** pode ser replicado por qualquer pesquisador no futuro, seja para validar ou questionar os resultados aqui encontrados, seja para avaliar novos modelos e sistemas de IA (gerais ou jurídicas) que venham a surgir. Desse modo, a pesquisa teve caráter inovador ao desenvolver uma bateria de testes padronizados para a avaliação da correção jurídica das respostas fornecidas pelos modelos de IA generativa. Com isso, não se pretende que esse seja o modelo definitivo, mas que o padrão aqui desenvolvido possa servir como um *framework* básico para que novas baterias de testes sejam realizadas pela comunidade jurídica e científica.

Para além da contribuição metodológica para o campo IA jurídica, o presente estudo demonstrou que a avaliação da correção jurídica das soluções dadas pelas IAs não passa pela simples **seleção de respostas em uma relação de perguntas** (como são o Exame da OAB ou o Exame Nacional da Magistratura), pois cada espécie de tarefa jurídica tem seus elementos próprios que envolvem prática, conhecimento implícito e posicionamento perante a realidade.

Feitas essas considerações, em específico, destaca-se que a conclusão da pesquisa **não se resume a uma resposta categórica e definitiva sobre quando, como e por que utilizar os modelos de IA generativa no Direito**, uma vez que o problema de pesquisa se mostrou multifacetado, rico em detalhes e plural em possibilidades de soluções.

Nessa perspectiva, para avançar a discussão sobre o uso de IAs generativas no Direito, a partir dos resultados já observados e discutidos ao longo dos capítulos 2 e 3, as seguintes conclusões passam a ser expostas:

a) mostra-se que **não** é **recomendada a utilização, simples e direta, das IAs generativas (*LLMs*) para acessar conhecimento jurídico sistematizado** (textos normativos, por exemplo), apesar de alguns modelos se saírem bem na tarefa, haja vista que o treinamento "task agnostic" tradicionalmente utilizado (ampla base de dados e tecnologia *Transformer*) dá variabilidade às respostas, mas, em si mesmo, não garante consistência (o que dependeria de treinamento específico em bases de dados jurídicas);

b) é **altamente não recomendada a utilização de IA generativa para a interpretação "livre"[47] de normas jurídicas** (teste 2), sendo que o desenvolvimento de agentes de IA deve passar por uma cuidadosa delimitação dos tipos de problemas

47 No contexto deste artigo, entenda-se interpretação "livre" o simples fornecimento de fatos e normas aplicáveis, esperando que a IA dê uma solução para o caso sem a necessidade de interação humana ou com quase nenhuma intervenção prévia para além da engenharia de *prompts*.

jurídicos nos quais serão aplicados, caso auxiliar na interpretação seja uma necessidade para a aplicação imediata da IA;

c) a **maioria dos *LLMs* tem grandes dificuldades para, a partir do texto normativo, inferir todas as proposições normativas corretas** aplicáveis a um caso e para compreender toda a lógica dos seus modais deônticos (proibições, obrigações, permissões e facultatividades) que, muitas vezes, estão implícitos no texto normativo;

d) para diminuir os desafios observados nos testes 1 e 2, a utilização de uma abordagem com técnicas de ***Retrieval-Augmented Generation* (*RAG*)** ou com **IAs auxiliares** baseadas em representação do conhecimento se monstra necessária, a fim de garantir acesso a conteúdo jurídico legítimo, vigente e estruturado;

e) **os melhores *LLMs* apresentaram criatividade e pretensão de correção jurídica** ao aplicar princípios jurídicos (teste 3), o que denota a utilização dos modelos IA generativa como **bons auxiliares argumentativos**, enriquecendo a aplicação do Direito;

f) a **solução de "casos difíceis" (*"hard cases"*)** proposta no teste 4 mostrou como os modelos de IA generativa podem ser utilizados para auxiliar na **geração de argumentos** mediante uma força efetivamente criadora de novos posicionamentos jurídicos, pois a IA pode ser utilizada para gerar argumentos favoráveis e contrários a determinada posição, confrontando-os de maneira a auxiliar o intérprete humano na sua tomada de decisão ou na elaboração de sua manifestação;

g) com base nos resultados dos testes 3 e 4 e em sua comparação com os resultados dos testes 1 e 2, evidencia-se o **paradoxo conhecimento/criação da IA**, que consiste no fato de que os modelos de IA generativa **não são bons repositórios de conhecimento previamente existente**, mas **têm uma evidente capacidade de criar conhecimento novo**, pois os *LLMs* não compreendem integralmente as relações normativas que nascem das regras jurídicas, mas são capazes de realizar inferência lógica criativa, razoável e pertinente a partir de princípios jurídicos cuja aplicabilidade possui abertura linguística;

h) atualmente, os melhores modelos de IA generativa já podem ser utilizados na exploração argumentativa de precedentes judiciais, na busca de **organizar e interrelacionar a base fática e as motivações dos** órgãos **julgadores, especialmente para identificar a *ratio decidendi*** dos julgamentos e possibilitar, com isso, um controle de similitude em relação a novos julgamentos sobre o mesmo tema (*stare decisis*);

i) os *LLMs,* com variados graus de eficiência (de muito alta a muito baixa), podem atuar como sistemas auxiliares para **recuperação de dados e informações em documentos escritos e outras bases pesquisadas** (imagens, vídeos, áudios etc.), organização desses dados e das informações e relacionamento inicial ao Direito aplicável, servindo como ferramenta que aprimora todas as atividades jurídicas (advocacia, jurisdição, educação jurídica, investigação etc.);

j) a **engenharia de *prompts* não** é **a solução para todos os problemas**, pois, por melhor que tenha sido desenhado o *prompt* e por mais *expertise* que tenha

quem o escreveu minuciosamente, muitos modelos ainda tendem a desconsiderar as ordens dadas via *prompt* e a gerar respostas "alucinadas" (incorretas ou sobre fatos inexistentes);

k) ainda **não existe um** único **modelo de IA generativa com desempenho superior ou perfeito para a maioria dos problemas jurídicos**, o que abre o campo para o desenvolvimento de modelos híbridos específicos para o Direito, feitos com a contribuição da própria comunidade jurídica;[48]

l) nem sempre o modelo mais "aprimorado" ou de última geração, do ponto de vista das técnicas de Inteligência Artificial, é o mais eficiente em determinada tarefa jurídica.

Os resultados demonstram que alguns dos mitos da IA aplicada ao Direito foram afastados (especialmente a falsa compreensão de que a engenharia de *prompts* resolve todas as deficiências e o negacionismo da criatividade jurídica das IAs), sendo que é também tarefa dos juristas avaliar e desenvolver a Inteligência Artificial aplicada ao campo jurídico, a fim de possibilitar o aprimoramento contínuo do duplo canal de correção.

Referências

ABRAMSON, Josh *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, London, v. 630, p. 493-500, 2024. DOI: 10.1038/s41586-024-07487-w.

ALEXY, Robert. *Direito, Razão, Discurso*: Estudos para a filosofia do direito. Tradução: Luís Afonso Heck. 2ª ed. Porto Alegre: Livraria do Advogado, 2015.

ALEXY, Robert. *Teoria dos Direitos Fundamentais*. 2ª ed. Tradução: Virgílio Afonso da Silva. São Paulo: Malheiros, 2011.

ALIBABA. *Qwen*. [2025]. Disponível em: https://chat.qwenlm.ai. Acesso em: 12 fev. 2025.

ANGWIN, Julia *et al. Machine Bias.* [2016]. Disponível em: www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Acesso em: 28 nov. 2024.

ANTHROPIC. *Building effective agents.* [2024]. Disponível em: https://www.anthropic.com/research/building-effective-agents. Acesso em: 02 jan. 2025.

ANTHROPIC. *Claude AI.* [2025]. Disponível em: https://claude.ai. Acesso em: 31 mar. 2025.

ASHLEY, Kevin D. *Artificial Intelligence and Legal Analytics:* New tools for law practice in digital age. Cambridge: Cambridge University Press, 2017.

ASHLEY, Kevin D. Case-based reasoning and its implications for legal expert systems. *Artificial Intelligence and Law,* Dordrecht, v. 1, p. 113-208, 1992, DOI 10.1007/BF00114920.

48 É mais fácil ensinar Ciência da Computação para um jurista do que fazer desenvolvedores aprenderem as minucias do Direito. Ou seja, as melhores soluções de IA para Direito terão a contribuição de sistemas de IA generativa de "big techs", mas provavelmente virão de empresas menores e pesquisadores independentes que efetivamente vivem o Direito. Esta nota de rodapé não é, em si, uma conclusão, mas um apenas registro para que o leitor possa lembrar disso no futuro.

BARBOZA, Estefânia Maria de Queiroz. *Stare decisis, integridade e segurança jurídica:* reflexões críticas a partir da aproximação dos sistemas de common law e civil law. 2010. Tese (Doutorado em Direito). Pontifícia Universidade Católica do Paraná (PUC-PR), Curitiba, 2010.

BROWN, Tom B. *et al. Language Models are Few-Shot Learners*. OpenAI, San Francisco, arXiv:2005.14165. Disponível em: https://arxiv.org/pdf/2005.14165.pdf. Acesso em: 07 abr. 2024.

BUYL, Maarten *et al. Large Language Models Reflect the Ideology of Their Creators*. Disponível em: https://arxiv.org/pdf/2005.14165. Acesso em: 16 nov. 2024.

CONSELHO NACIONAL DE JUSTIÇA. *Resolução 615, de 11 de março de 2025.* Estabelece diretrizes para o desenvolvimento, utilização e governança de soluções desenvolvidos com recursos de inteligência artificial no Poder Judiciário. Disponível em: https://atos.cnj.jus.br/atos/detalhar/6001. Acesso em: 26 mar. 2025.

CONSELHO FEDERAL DA ORDEM DOS ADVOGADOS DO BRASIL. *Recomendação 001, de 11 de novembro de 2024.* Apresenta diretrizes para orientar o uso de Inteligência Artificial generativa na Prática Jurídica. Disponível em: https://diario.oab.org.br/pages/materia/842347. Acesso em: 05 jan. 2025.

DEEPSEEK AI. *DeepSeek.* [2025]. Disponível em: https://www.deepseek.com. Acesso em: 08 fev. 2025.

DEEPSEEK AI. *DeepSeek-R1:* Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. Disponível em: https://arxiv.org/pdf/2005.14165. Acesso em: 25 jan. 2025.

DWORKIN, Ronald. *Levando os Direitos a Sério.* 3.ª ed. Tradução: Nelson Boeira. São Paulo: Martins Fontes, 2010.

GOOGLE. *Gemini.* [2025]. Disponível em: https://gemini.google.com. Acesso em: 31 mar. 2025.

HARTMANN PEIXOTO, Fabiano; BONAT, Debora. *GPTs e Direito: impactos prováveis das IAs generativas nas atividades jurídicas brasileiras*. Seqüência, Florianópolis, v. 44, 2023. DOI: 10.5007/2177-7055.2023.e94238.

HARTMANN PEIXOTO, Fabiano; SILVA, Roberta Zumblick Martins da. *Inteligência Artificial e Direito.* Vol. 1. Curitiba: Alteridade, 2019.

HU, Mengya *et al. SLM Meets LLM: Balancing Latency, Interpretability and Consistency in Hallucination Detection.* Disponível em: https://arxiv.org/abs/2408.12748. Acesso em: 28 out. 2024.

KAPOOR, Sayash *et al. AI Agents That Matter.* Disponível em: https://arxiv.org/abs/2407.01502. Acesso em: 11 nov. 2024.

KURZWEIL, Ray. *A Singularidade Está Mais Próxima:* A fusão do ser humano com o poder da inteligência artificial. Tradução: Renato Marques. São Paulo: Goya, 2024.

LIU, Kai *et al. Enhancing LLM's Cognition via Structurization.* Disponível em: https://arxiv.org/abs/2407.16434. Acesso em: 25 set. 2024.

MARITACA AI. *Maritalk.* [2025]. Disponível em: https://chat.maritaca.ai. Acesso em: 10 fev. 2025.

MICROSOFT. *Copilot.* [2025]. Disponível em: https://m365.cloud.microsoft. Acesso em: 15 jan.

2025. MICROSOFT. Copilot. [2025]. Disponível em: https://m365.cloud.microsoft. Acesso em: 15 jan. 2025.

MIGALHAS. *Advogado virtual? ChatGPT consegue "aprovação" na primeira fase da OAB.* [2023]. Disponível em https://www.migalhas.com.br/quentes/381875/advogado-virtual-chatgpt--consegue-aprovacao-na-primeira-fase-da-oab. Acesso em: 12 fev. 2025.

MORAIS, Fausto Santos de. *Ponderação e arbitrariedade - A inadequada recepção de Alexy pelo STF.* Editora Juspodium. 1a edição. 2016.

MORAIS, Fausto Santos de. *O uso da inteligência artificial na repercussão geral: desafios teóricos e **éticos**.* Revista Direito Público. Dossiê temático - "Inteligência Artificial, Ética e Epistemologia', v. 18 n. 100. 2021. Disponível em https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/6001. Acesso em: 12 fev. 2025.

OLIVEIRA, Bruno da Cunha de. *Primeira IA capaz de ser aprovada no Enam **é** lançada no Brasil.* [2024]. Disponível em: https://www.conjur.com.br/2024-ago-08/primeira-ia-capaz-de--ser-aprovada-no-enam-e-lancada-no-brasil/. Acesso em: 12 fev. 2025.

OPENAI. *ChatGPT.* [2025]. Disponível em: https://chatgpt.com. Acesso em: 15 mar. 2025.

PÁDUA, Sérgio Rodrigo de. *Da Jurisdição 'Ex Machina' ao Juiz Ciborgue: Inteligência Artificial e Interpretação do Direito.* São Paulo: Thomson Reuters, 2023.

PÁDUA, Sérgio Rodrigo de; LORENZETTO, Bruno Meneses. *O Direito Fundamental à Explicabilidade da Inteligência Artificial Utilizada em Decisões Estatais*. Revista da AGU, Brasília, v. 23, n. 02, 2024. DOI: 10.25109/2525-328X.v.23.n.02.2024.3480.

PERPLEXITY. *Perplexity AI.* [2025]. Disponível em: https://www.perplexity.ai. Acesso em: 16 jan. 2025.

ROVER, Aires José. *Informática no Direito: Inteligência Artificial*. Curitiba: Juruá, 2001.

SEARLE, Jonh R. *Minds, brains, and programs*. Behavioral and Brain Sciences, v. 3, n. 3, p. 417-424, 1980. DOI: 10.1017/S0140525X00005756.

SHANNON, Claude E. *A Mathematical Theory of Communication*. The Bell System Technical Journal, v. 27, n. 3, p. 379-423, Jul. 1948. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

SUPERIOR TRIBUNAL DE JUSTIÇA. 1. Seção. *Recurso Especial n.º 1.112.646/SP.* Recurso repetitivo de Tema 174. Brasília, DF: Superior Tribunal de Justiça, [2009]. Relator: Min. Herman Benjamin, 26 de agosto de 2009. Disponível em: https://processo.stj.jus.br/SCON/GetInteiroTeorDoAcordao?num_registro=200900510886&dt_publicacao=28/08/2009.

SUPERIOR TRIBUNAL DE JUSTIÇA. 1. Turma. *Agravo Interno no Agravo em Recurso Especial n.º 1.718.222/RJ.* Brasília, DF: Superior Tribunal de Justiça, [2021]. Relator: Min. Gurgel de Faria, 14 de setembro de 2021. Disponível em: https://processo.stj.jus.br/SCON/GetInteiroTeorDoAcordao?num_registro=202001479104&dt_publicacao=17/09/2021. Acesso em: 15 jan. 2025.

SUPERIOR TRIBUNAL DE JUSTIÇA. *Gabinetes conhecem, na prática, funcionamento do STJ Logos.* [2025]. Disponível em: https://www.stj.jus.br/sites/portalp/Paginas/Comunicacao/Noticias/2025/15022025-Gabinetes-conhecem--na-pratica--funcionamento-do-STJ-Logos-.aspx. Acesso em: 17 fev. 2025.

SUPERIOR TRIBUNAL DE JUSTIÇA. *Súmula 306 do STJ*. Disponível em: https://www.stj.jus.br/docs_internet/revista/eletronica/stj-revista-sumulas-2011_24_capSumula306.pdf. Acesso em: 10 fev. 2025.

SUPERIOR TRIBUNAL DE JUSTIÇA. *Súmula 326 do STJ*. Disponível em: https://www.stj.jus.br/docs_internet/revista/eletronica/stj-revista-sumulas-2012_27_capSumula326.pdf. Acesso em: 10 fev. 2025.

SUPREMO TRIBUNAL FEDERAL. Plenário. *Recurso Extraordinário n.º 1355208*. Repercussão geral de Tema 1184. Brasília, DF: Superior Tribunal de Justiça, [2023]. Relatora: Min. Cármen Lúcia, 19 de dezembro de 2023. Disponível em: https://redir.stf.jus.br/paginadorpub/paginador.jsp?docTP=TP&docID=775641868.

SUPREMO TRIBUNAL FEDERAL. *STF lança MARIA, ferramenta de inteligência artificial que dará mais agilidade aos serviços do Tribunal*. [2024]. Disponível em: https://noticias.stf.jus.br/postsnoticias/stf-lanca-maria-ferramenta-de-inteligencia-artificial-que-dara-mais-agilidade-aos-servicos-do-tribunal/. Acesso em: 17 fev. 2025.

SUSSKIND, Richard. *Tomorrow's Lawyers: An Introduction to Your Future*. 2a ed. Oxford: Oxford University Press, 2017. E-book Kindle.

VASWANI, Ashish *et al. Attention Is All You Need*. Disponível em: https://arxiv.org/abs/1706.03762. Acesso em: 12 dez. 2024.

XAI. *Grok*. [2025]. Disponível em: https://grok.com. Acesso em: 19 fev. 2025.