Universidade de Brasília

Faculdade de Direito

Programa de Mestrado em Direito, Regulação e Políticas Públicas

A REVOLUÇÃO DO DEEPSEEK PARA OS GRANDES MODELOS DE LINGUAGEM E O QUE ISSO PODE PROPORCIONAR AO PODER JUDICIÁRIO BRASILEIRO

Edimara Alexandrino de Souza Macedo

Dissertação apresentada ao Programa de Mestrado em Direito, Regulação e Políticas Públicas da Faculdade de Direito da Universidade de Brasília, como requisito parcial para obtenção do grau de Mestre em Direito.

Orientador: Prof. Dr. Henrique Araújo Costa.

2025

A REVOLUÇÃO DO DEEPSEEK PARA OS GRANDES MODELOS DE LINGUAGEM E O QUE ISSO PODE PROPORCIONAR AO PODER JUDICIÁRIO BRASILEIRO

Edimara Alexandrino de Souza Macedo

Dissertação apresentada ao Programa de Mestrado em Direito, Regulação e Políticas Públicas da Faculdade de Direito da Universidade de Brasília, como requisito parcial para obtenção do grau de Mestre em Direito.

Orientador: Prof. Dr. Henrique Araújo Costa.

2025

Alexandrino de Souza Macedo, Edimara

AA382r

A REVOLUÇÃO DO DEEPSEEK PARA OS GRANDES MODELOS DE LINGUAGEM E O QUE ISSO PODE PROPORCIONAR AO PODER JUDICIÁRIO BRASILEIRO / Edimara Alexandrino de Souza Macedo; orientador Henrique Araújo Costa. Brasília, 2025.

63 p.

Dissertação (Mestrado Profissional em Direito, Regulação e Políticas Públicas) Universidade de Brasília, 2025.

1. Inteligência Artificial. 2. Poder Judiário. 3. DeepSeek. 4. STF. I. Araújo Costa, Henrique , orient. II. Título

EDIMARA ALEXANDRINO DE SOUZA MACEDO

A REVOLUÇÃO DO DEEPSEEK PARA OS GRANDES MODELOS DE LINGUAGEM E O QUE ISSO PODE PROPORCIONAR AO PODER JUDICIÁRIO BRASILEIRO

Dissertação apresentada ao Programa de Mestrado em Direito, Regulação e Políticas Públicas da Faculdade de Direito da Universidade de Brasília, como requisito parcial para obtenção do grau de Mestre em Direito.

Orientador: Prof. Dr. Henrique Araújo Costa.

Brasília, 7 de novembro de 2025.

Banca Examinadora

Dra. Amanda Nunes Lopes Espiñeira Lemos, MS

Examinadora Externa à Instituição

Dra. Maria Cristine Branco Lindoso, IDP

Examinadora Externa à Instituição

Dr. Wilson Roberto Theodoro Filho, UnB

Examinador Interno

Dr. Henrique Araujo Ccosta, UnB

Orientador

RESUMO

A pesquisa analisa a viabilidade técnico-jurídica e institucional de o Poder Judiciário brasileiro adotar um modelo próprio de linguagem (LLM), inspirado na abertura e eficiência computacional do DeepSeek. O avanço dos LLMs tem redefinido a produção, a interpretação e a gestão de informações públicas. Embora os tribunais brasileiros utilizem IA desde 2017, ainda prevalece o uso de ferramentas restritas a classificação, sumarização e triagem processual, com forte dependência de soluções privadas como GPT-4 e Copilot. A metodologia adotada foi documental e analítico-descritiva, com base em relatórios técnicos (Stanford, CNJ e DeepSeek AI), documentos normativos — especialmente a Resolução CNJ n. 615/2025 e o Projeto de Lei 2338/2023 — além da análise das iniciativas de IA no STF entre 2017 e 2025. Verificou-se uma evolução progressiva no STF: de modelos preditivos (Victor, VitórIA e RAFA 2030) até a ferramenta generativa MARIA, voltada à redação de atos judiciais. Essa trajetória revela maturidade institucional, mas também desafios críticos, como a dependência tecnológica estrangeira, o risco de vieses e a ausência de uma governança de dados plenamente transparente. O lançamento do modelo DeepSeek, com arquitetura Mixture of Experts, custo reduzido de treinamento e capacidade de operação offline, rompe o paradigma de que apenas grandes corporações podem criar modelos de IA em larga escala. Técnicas como LoRA, QLoRA e RAG permitem adaptar modelos abertos com menos recursos computacionais, o que indica que um caminho próprio para o domínio jurídico é tecnicamente concebível, desde que acompanhado de planejamento e investimento estratégico. Esse contexto, somado à Resolução CNJ n. 615/2025 e o Plano Brasileiro de Inteligência Artificial (PBIA) revelam uma oportunidade institucional concreta para o desenvolvimento de um LLM jurídico nacional. A viabilidade depende, sobretudo, de decisão institucional estável e de integração entre tribunais, universidades e centros públicos de pesquisa. Conclui-se que um LLM jurídico nacional representa inovação tecnológica, soberania digital e fortalecimento democrático do sistema de justiça brasileiro.

Palavras-chave: Inteligência Artificial; DeepSeek; Poder Judiciário; STF.

ABSTRACT

The research analyzes the technical, legal, and institutional feasibility of the Brazilian Judiciary adopting its own language model (LLM), inspired by the openness and computational efficiency of DeepSeek. The advancement of LLMs has redefined the production, interpretation, and management of public information. Although Brazilian courts have been using AI since 2017, the use of tools restricted to classification, summarization, and procedural screening still prevails, with a strong dependence on private solutions such as GPT-4 and Copilot. The methodology adopted was documentary and analytical-descriptive, based on technical reports (Stanford, CNJ, and DeepSeek AI), normative documents—especially CNJ Resolution No. 615/2025 and Bill 2338/2023—in addition to the analysis of AI initiatives in the STF between 2017 and 2025. A progressive evolution was observed in the STF: from predictive models (Victor, VitórIA, and RAFA 2030) to the generative tool MARIA, aimed at drafting judicial acts. This trajectory reveals institutional maturity, but also critical challenges, such as foreign technological dependence, the risk of bias, and the absence of fully transparent data governance. The launch of the DeepSeek model, with Mixture of Experts architecture, reduced training costs, and offline operating capability, breaks the paradigm that only large corporations can create large-scale AI models. Techniques such as LoRA, QLoRA, and RAG allow open models to be adapted with fewer computational resources, indicating that a path specific to the legal domain is technically conceivable, provided it is accompanied by strategic planning and investment. This context, coupled with CNJ Resolution No. 615/2025 and the Brazilian Artificial Intelligence Plan (PBIA), reveals a concrete institutional opportunity for the development of a national legal LLM. Feasibility depends, above all, on stable institutional decision-making and integration between courts, universities, and public research centers. It can be concluded that a national legal LLM represents technological innovation, digital sovereignty, and democratic strengthening of the Brazilian justice system.

Keywords: Artificial Intelligence; DeepSeek; Judiciary; STF.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

ARE	Agravo em Recurso Extraordinário
CENIA	Centro Nacional de Inteligência Artificial do Chile
CNJ	Conselho Nacional de Justiça
GPU	Unidade de Processamento Gráfico
IA	Inteligência Artificial
IAG	Inteligência Artificial Generativa
LGPD	Lei Geral de Proteção de Dados
LLM	Large Language Models
LoRA	Low-Rank Adaptation
MARIA	Módulo de Apoio para Redação com Inteligência Artificial
MCTI	Ministério da Ciência, Tecnologia e Inovação
MoE	Mistura de Especialistas
NIAC	Núcleo de Inteligência Artificial do Supremo Tribunal Federal
OCR	Optical Caracter Recognition
ODS	Objetivos de Desenvolvimento Sustentável
ONU	Organização das Nações Unidas
RE	Recurso Extraordinário
STF	Supremo Tribunal Federal
STJ	Superior Tribunal de Justiça
TED	Termo de Execução Descentralizada
TJGO	Tribunal de Justiça de Goiás
UNESCO	Organização das Nações Unidas para a Educação, a Ciência e a Cultura

SUMÁRIO

1 INTRODUÇÃO

O objetivo deste trabalho é analisar a viabilidade técnico-jurídica e institucional da adoção de um LLM próprio pelo Poder Judiciário brasileiro, identificando os fundamentos tecnológicos, regulatórios e éticos necessários à sua implementação.

O avanço acelerado da Inteligência Artificial (IA) transformou radicalmente a forma como governos, empresas e cidadãos produzem, acessam e interpretam informações. Entre as tecnologias mais disruptivas estão os grandes modelos de linguagem (*Large Language Model*s – LLMs), capazes de compreender e gerar texto de forma contextual, dinâmica e interativa. Esses modelos, originalmente desenvolvidos em centros de pesquisa e em grandes corporações de tecnologia, tornaram-se instrumentos centrais de inovação, automação e formulação de políticas públicas. No entanto, seu uso no setor público, especialmente no âmbito jurídico, ainda apresenta desafios éticos, técnicos e institucionais de grande complexidade.

O trabalho inicia com a explicação dos conceitos fundamentais de *machine learning*, *deep learning* e LLMs. Na sequência, examina-se a governança de IA no sistema de justiça brasileiro à luz do Conselho Nacional de Justiça (CNJ). Parte-se do diagnóstico da Resolução CNJ n. 332/2020 e de seu aperfeiçoamento pela Resolução CNJ n. 615/2025, que passa a tratar, de forma mais explícita, de princípios, gestão de riscos, transparência e uso responsável de IA tradicional e generativa no âmbito dos tribunais. Analisa-se, ainda, a plataforma SINAPSES como trilha de auditoria e repositório de projetos de IA, peça central para registrar, monitorar e auditar soluções em produção.

O trabalho, então, focaliza a experiência do Supremo Tribunal Federal (STF), que investe em IA desde 2017. São discutidas as soluções Victor (classificação por temas de repercussão geral), VitórIA (agrupamento de processos por similaridade), RAFA 2030 (classificação por ODS) e, mais recentemente, a MARIA, primeira iniciativa de IA generativa do Tribunal voltada à produção de minutas e resumos. Essa trajetória demonstra ganhos reais em triagem, organização e apoio à redação, ao mesmo tempo em que evidencia a necessidade de supervisão humana e de controles de qualidade.

Nesse contexto, a criação de modelos abertos e eficientes, como o DeepSeek, desenvolvido pela DeepSeek AI, inaugura uma nova fase na disputa pela soberania tecnológica global. O DeepSeek demonstrou que é possível alcançar desempenho comparável ao de modelos proprietários, como o GPT-4, com custo significativamente menor e uma arquitetura

eficiente baseada em Mixture of Experts (MoE). Essa inovação abre caminho para que países em desenvolvimento — como o Brasil — adotem estratégias de IA próprias, éticas e economicamente sustentáveis, especialmente em domínios sensíveis, como o sistema de justiça.

A relevância do estudo reside na necessidade de equilibrar a inovação tecnológica e a segurança jurídica. O uso de IA em decisões judiciais não pode comprometer direitos fundamentais nem gerar desigualdades no acesso à justiça. Assim, compreender como modelos abertos, auditáveis e economicamente viáveis — como o DeepSeek — podem servir de base para o desenvolvimento de soluções soberanas é um passo essencial para consolidar uma política pública de IA no Judiciário brasileiro, pautada em autonomia, transparência e responsabilidade algorítmica.

Por fim, o estudo posiciona o DeepSeek como um marco recente no ecossistema de modelos abertos. O lançamento do modelo elevou a discussão sobre custo, eficiência e possibilidade de adoção de LLMs, inspirando, inclusive, o Plano Brasileiro de Inteligência Artificial (PBIA) do MCTI (2025) a considerar viável o treinamento de modelos em língua portuguesa com recursos mais contidos. Esse contexto abre uma janela concreta para que, no médio prazo, o Poder Judiciário avalie a criação de um LLM jurídico em português, sob coordenação pública e em conformidade com as salvaguardas regulatórias nacionais. É essa viabilidade que será explorada no trabalho.

A pesquisa é essencialmente documental e analítico-descritiva, com base em relatórios do CNJ, publicações técnicas (Stanford, manuais oficiais das ferramentas, artigos técnicos), documentos normativos e na análise de casos de uso no STF e no STJ entre 2017 e 2025.

2 INTELIGÊNCIA ARTIFICIAL

De acordo com Shabbir e Anwer (2015), o termo inteligência refere-se à capacidade de adquirir e aplicar diferentes competências e conhecimentos para resolver um determinado problema. A inteligência está integrada a várias funções cognitivas, tais como a linguagem, a atenção, o planeamento, a memória e a percepção.

Já o conceito de Inteligência Artificial (IA) não tem um consenso entre os diversos cientistas da área, mas segundo Gabriel Filho (2023, p.24), é a:

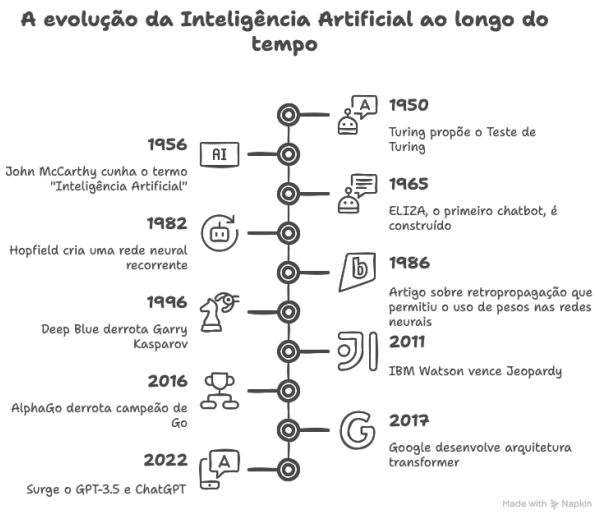
> "ciência multidisciplinar que tem o objetivo de prover as máquinas com a capacidade de executar tarefas que exigem concorrência de alguma habilidade característica do ser humano como inteligência, criatividade, atenção, perseverança etc. No campo da Psicologia, a inteligência é caracterizada pela manifestação de uma ou mais das seguintes qualidades: aprendizagem, adaptação e capacidade de resolver problemas".

A IA tenta copiar características humanas, realizando várias tarefas que exigem raciocínio e aprendizado, resolvendo problemas e tomando decisões, utilizando um conjunto de tecnologias que vão do machine learning às redes neurais (Hartmann Peixoto, 2020a).

Ao longo do tempo, a tecnologia evoluiu a ponto de hoje existir a chamada IA generativa, que não apenas rotula dados, mas também cria textos, imagens e vídeos.

Nesse contexto, é importante trazer os principais marcos da evolução da Inteligência Artificial (IA) no tempo (figura 1):

Figura 1 - Evolução da Inteligência Artificial ao longo do tempo



Fonte: autoria própria, baseada em Taulli (2020) e elaborada com o Napkin.

Nota-se, pela linha do tempo, que, desde os anos 50, o matemático Alan Turing buscava ensinar aos computadores comportamentos humanos, o que seria o aprendizado de máquina (*Machine Learning*) (Kneusel, 2024).

Com a evolução do aprendizado de máquina e do aprendizado profundo, a Inteligência Artificial evoluiu significativamente a partir da segunda metade da década de 90, pois os recursos computacionais se tornaram mais robustos e de custo menor. Além disso, com o advento da rede mundial de computadores, havia uma grande quantidade de dados disponíveis gratuitamente para testar modelos computacionais (Taulli, 2020).

Para falar de inteligência artificial, é necessário também apresentar os conceitos de *Machine Learning* e *Deep Learning*.

2.1 Machine Learning e Deep Learning

Machine Learning, ou aprendizado de máquina, trabalha com o treinamento de modelos que possuem parâmetros de entrada (vetores com as características do que se analisa) e de saída, para que a máquina, em determinado momento, possa criar resultados a partir de dados em que não foi treinada, fazendo associações com o que foi aprendido (Kneusel, 2024).

O aprendizado de máquina pode usar algoritmos de aprendizado supervisionado, não supervisionado ou de reforço (Taulli, 2020). No supervisionado, o ser humano rotula previamente os dados; no não assistido, a própria máquina os rotula; e, no de reforço, com a saída dos dados, o ser humano confirma se os resultados foram positivos ou negativos (Hartmann Peixoto, 2020a).

O GPT, um dos modelos mais conhecidos, utiliza o treinamento supervisionado. A IA usa modelos treinados com uma grande quantidade de dados. Existem vários tipos de modelos de IA; um dos mais utilizados para a IA generativa é o de redes neurais, mas também existem outros, como árvores de decisão, florestas aleatórias e máquinas de vetores de suporte (Kneusel, 2024).

Os modelos de rede neural simulam redes de neurônios humanas, operam com pesos e utilizam exemplos para que a máquina possa estabelecer correlações e gerar resultados para diversos problemas (Hartmann Peixoto, 2020a). São utilizados em grandes conjuntos de dados e conseguem capturar padrões mais complexos (Gabriel Filho, 2023).

Segundo Gabriel Filho (2023, p. 281), a árvore de decisão (*Decision Tree* – DT) é um método baseado no princípio da divisão e conquista. É amplamente utilizado para resolver

problemas de predição, tanto de regressão (com variáveis numéricas ou quantitativas) quanto de classificação (com variáveis não numéricas, qualitativas ou categóricas), geralmente relacionados à tomada de decisão.

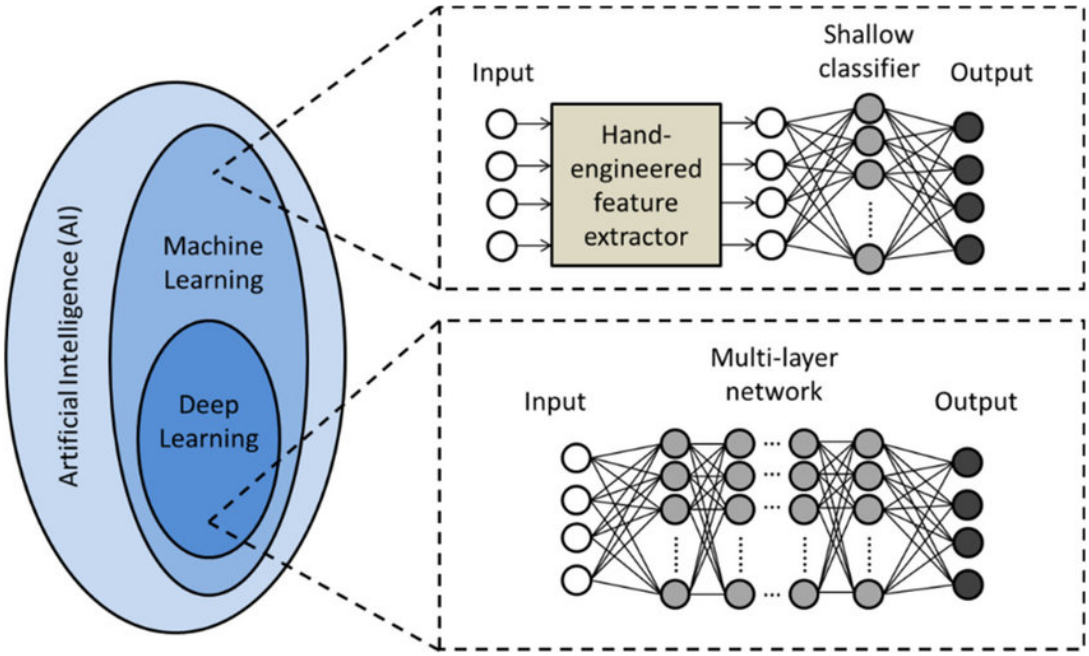
Como explica Taulli (2020), o nó raiz de uma árvore de decisão está no topo do fluxograma. A partir dele, a árvore se expande por meio de caminhos de decisão, chamados de divisões. Nessas etapas, um algoritmo será utilizado para tomar decisões e calcular uma probabilidade. No final da árvore, encontra-se a folha que representa o resultado. As árvores de decisão oferecem vantagens, como facilidade de compreensão, bom desempenho em grandes conjuntos de dados e transparência do modelo. Contudo, uma desvantagem é a propagação de erros: se uma divisão estiver incorreta, esse erro pode se propagar por toda a árvore. Além disso, à medida que a árvore cresce, sua complexidade aumenta, o que pode reduzir o desempenho do modelo.

As florestas aleatórias (*Random Forests*) são um método que emprega um conjunto de várias árvores de decisão, treinadas em diferentes subconjuntos, com amostragem aleatória das características. Essas árvores combinam suas decisões para aumentar a precisão e reduzir o *overfitting* (quando a rede memoriza os dados de treino e não consegue generalizar). Esse método é usado para processar um volume maior de dados, mas pode ser demorado, exigir mais recursos e ser mais complexo. (IBM).

Já a máquina de vetores de suporte (*Support Vector Machine* – SVM), segundo Gabriel Filho (2023), é uma técnica supervisionada usada para resolver problemas de regressão e de reconhecimento de padrões (classificação) binária, envolvendo duas categorias, por meio da aprendizagem baseada na vetorização dos dados de entrada. Seu objetivo principal é classificar dados de entrada, operando na separação destes em dois subconjuntos, que podem ser linearmente ou não linearmente separáveis. É aplicada especialmente quando há um número limitado de dados definidos.

O *Deep Learning*, ou aprendizado profundo, é uma subárea de *Machine Learning* que utiliza redes neurais em várias camadas, conhecidas como redes neurais profundas, aproximando-se cada vez mais do funcionamento dos neurônios humanos. Esses modelos revolucionaram o uso da IA, pois não apenas classificam dados, mas também aprendem com eles. A diferença entre *Machine Learning* e *Deep Learning* é que o aprendizado profundo utiliza muito mais camadas e muito mais dados para encontrar relacionamentos e padrões (Kneusel, 2024), como se observa na figura 2:

Figura 2 - A diferença entre Inteligência Artificial, Machine Learning, e Deep Learning



Fonte: BOON et al., 2025.

O Deep Learning também depende de uma grande quantidade de dados para “aprender” e, assim, de uma grande capacidade computacional para analisar tudo isso (Kneusel, 2024).

A evolução da IA está relacionada à quantidade de dados disponíveis para o treinamento de suas redes neurais. A evolução da internet e o aumento da quantidade de dados disponíveis na rede mundial de internet facilitaram muito o treinamento de modelos de inteligência artificial (KneuseL, 2024).

2.2 Grandes Modelos de Linguagem

O LLM é um modelo de aprendizado de máquina treinado a partir de uma grande quantidade de textos. Os LLMs são sistemas de IA Generativa usados para modelar e processar a linguagem humana (Gabriel Filho, 2023).

Os LLMs utilizam como entrada um *prompt* de texto fornecido pelo usuário, que consiste em instruções escritas por ele para definir o contexto da resposta da IA. Depois, os modelos geram saídas palavra a palavra, ou seja, *tokens*. Com base no seu treinamento, o modelo tenta predizer qual será a próxima palavra. Além disso, os modelos conseguem responder a perguntas, criar algoritmos de programação de computadores e realizar tarefas que demandam raciocínio lógico (Kneusel, 2024).

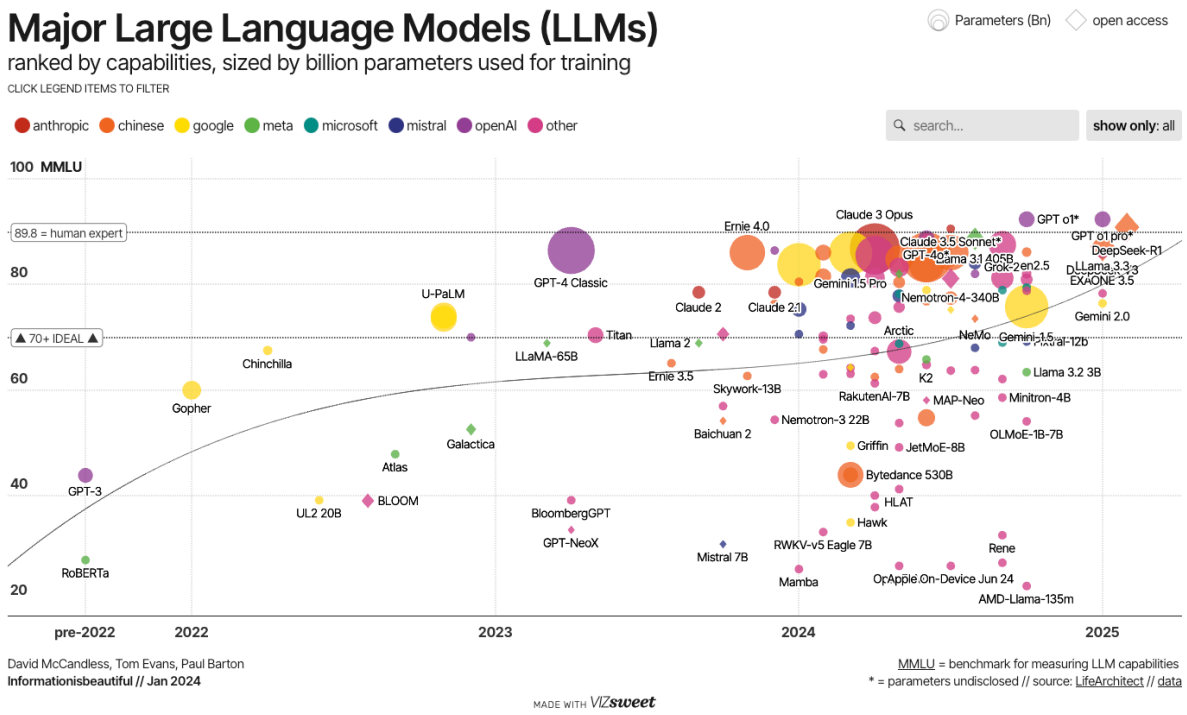
Segundo Gabriel Filho (2023), o termo “grandes” refere-se ao número de parâmetros que precisam ser ajustados durante um processo de aprendizado (treinamento) que utiliza alguma técnica de IA.

O uso da arquitetura *transformer* (um tipo de rede neural), desenvolvida pelo Google em 2017 e divulgada no artigo *Attention is All You Need*, revolucionou o uso de IA generativa, tornando o processamento muito mais rápido. A arquitetura utiliza a técnica da “atenção”, que é assim chamada, segundo Gabriel Filho (2023, p. 385):

> [...] devido à sua capacidade de conseguir captar e processar o ambiente ao seu redor de um determinado objeto de interesse [...], ou seja, as características relevantes do contexto em que o objeto se insere [...], podendo ser o reconhecimento de padrões (classificação) ou a tradução do texto de um idioma para outro”

Após a criação da arquitetura *transformer*, é possível verificar como os modelos evoluíram mais rapidamente de acordo com o gráfico abaixo (figura 3):

Figura 3 - A evolução dos grandes modelos de linguagem



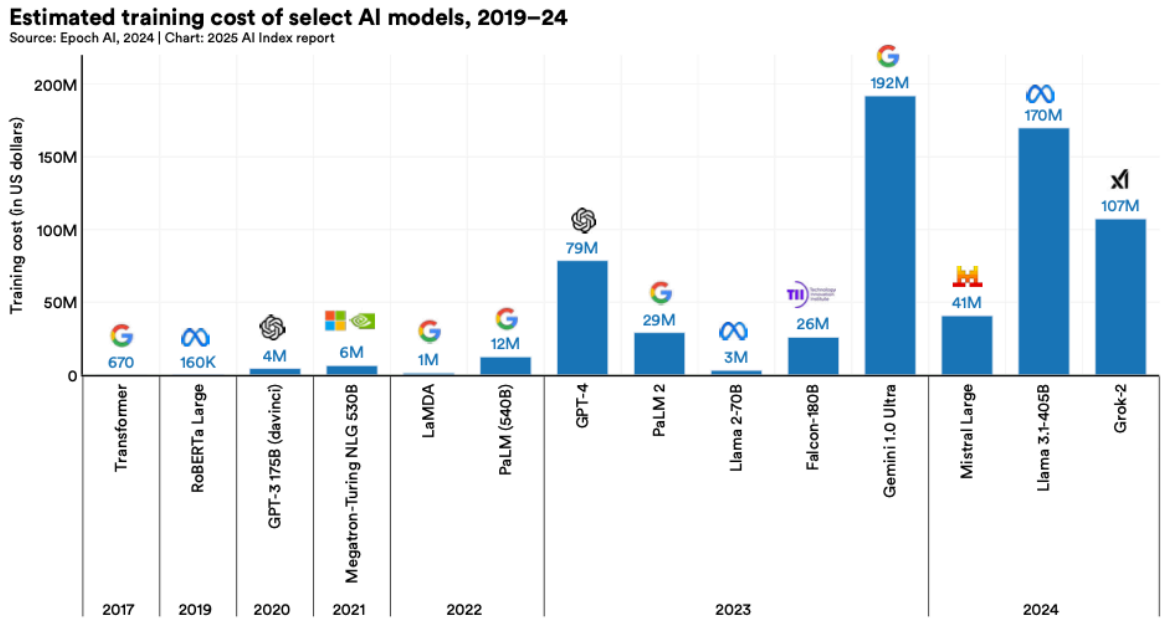
Fonte: MCCANDLESS; EVANS; BARTON, 2025.

A utilização de LLMs revolucionou o uso da IA para gerar textos. Antes, a IA chamada “preditiva” apenas “rotulava” as informações. Com esses modelos, veio a era da IA Generativa (IAG), que consegue criar textos por meio do Processamento de Linguagem Natural (PLN) e de redes neurais profundas.

A revolução no campo da IAG ocorreu principalmente com o lançamento do modelo GPT-3, da OpenAI, que, em 2022, adotou a arquitetura de *transformer* da Google (Kneusel, 2024). O grande diferencial desse modelo foi aproximar a IAG dos usuários comuns com o lançamento do ChatGPT. O ChatGPT é uma interface que permite que qualquer usuário, mesmo sem o conhecimento de programação, utilize a IAG (Kneusel, 2024). Assim, o usuário pode usar a IAG de maneira simples e intuitiva, criando apenas *prompts* (Porto; Araújo; Gabriel, 2024).

Para treinar os LLMs, é necessária uma base de dados muito grande de textos, muitos parâmetros para ajustar os modelos e algoritmos complexos que demandam meses ou até anos de treinamento. Isso exige uma grande capacidade de processamento computacional, elevando os custos a centenas de milhões de dólares. Os custos de treinamento do GPT-4, por exemplo, estimam-se em mais de 79 milhões de dólares (figura 4) (Stanford, 2025).

Figura 4 - O custo estimado do treinamento de modelos de IA



Fonte: Stanford, 2025.

Nota-se, pelo *Artificial Intelligence Index Report 2025* da Universidade de Stanford (Stanford, 2025), que os custos de treinamento dos grandes modelos de linguagem vêm aumentando ao longo do tempo, pois utilizam cada vez mais parâmetros e, com isso, necessitam de recursos computacionais cada vez maiores para a análise dos dados.

Contudo, em 2025, um modelo de uma startup chinesa, o DeepSeek-V3, quebrou esse paradigma ao apresentar desempenho comparável ao dos modelos mais caros do mercado, a um custo menor. Estima-se que os seus custos de treinamento tenham sido bem inferiores (cerca de 6 milhões de dólares), mesmo utilizando 671 bilhões de parâmetros (figura 5), o que é maior do que o número de parâmetros do modelo aberto LlaMa-3.1, da Meta, que usa 405 bilhões de parâmetros e custou em torno de 170 milhões de dólares para ser treinado.

Figura 5 - Quantidade de parâmetros por modelo

	Benchmark (Metric)	DeepSeek V2-0506	DeepSeek V2.5-0905	Qwen2.5 72B-Inst.	LLaMA-3.1 405B-Inst.	Claude-3.5-Sonnet-1022	GPT-4o 0513	DeepSeek V3
	Architecture	MoE	MoE	Dense	Dense	-	-	MoE
	# Activated Params	21B	21B	72B	405B	-	-	37B
	# Total Params	236B	236B	72B	405B	-	-	671B
English	MMLU (EM)	78.2	80.6	85.3	**88.6**	**88.3**	87.2	**88.5**
	MMLU-Redux (EM)	77.9	80.3	85.6	86.2	**88.9**	88.0	**89.1**
	MMLU-Pro (EM)	58.5	66.2	71.6	73.3	**78.0**	72.6	75.9
	DROP (3-shot F1)	83.0	87.8	76.7	88.7	88.3	83.7	**91.6**
	IF-Eval (Prompt Strict)	57.7	80.6	84.1	86.0	**86.5**	84.3	86.1
	GPQA-Diamond (Pass@1)	35.3	41.3	49.0	51.1	**65.0**	49.9	59.1
	SimpleQA (Correct)	9.0	10.2	9.1	17.1	28.4	**38.2**	24.9
	FRAMES (Acc.)	66.9	65.4	69.8	70.0	72.5	**80.5**	73.3
	LongBench v2 (Acc.)	31.6	35.4	39.4	36.1	41.0	48.1	**48.7**
Code	HumanEval-Mul (Pass@1)	69.3	77.4	77.3	77.2	81.7	80.5	**82.6**
	LiveCodeBench (Pass@1-COT)	18.8	29.2	31.1	28.4	36.3	33.4	**40.5**
	LiveCodeBench (Pass@1)	20.3	28.4	28.7	30.1	32.8	34.2	**37.6**
	Codeforces (Percentile)	17.5	35.6	24.8	25.3	20.3	23.6	**51.6**
	SWE Verified (Resolved)	-	22.6	23.8	24.5	**50.8**	38.8	42.0
	Aider-Edit (Acc.)	60.3	71.6	65.4	63.9	**84.2**	72.9	79.7
	Aider-Polyglot (Acc.)	-	18.2	7.6	5.8	45.3	16.0	**49.6**
Math	AIME 2024 (Pass@1)	4.6	16.7	23.3	23.3	16.0	9.3	**39.2**
	MATH-500 (EM)	56.3	74.7	80.0	73.8	78.3	74.6	**90.2**
	CNMO 2024 (Pass@1)	2.8	10.8	15.9	6.8	13.1	10.8	**43.2**
Chinese	CLUEWSC (EM)	89.9	90.4	**91.4**	84.7	85.4	87.9	90.9
	C-Eval (EM)	78.6	79.5	86.1	61.5	76.7	76.0	**86.5**
	C-SimpleQA (Correct)	48.5	54.1	48.4	50.4	51.3	59.3	**64.8**

Fonte: Deepseek, 2025.

Outra vantagem desse modelo é ser *open source* (código aberto), o que oferece maior transparência sobre como trata os dados, ou seja, “como ele pensa”. Assim, é possível entender como o modelo funciona e quais dados utiliza na análise.

Ainda existe a possibilidade de baixar o modelo DeepSeek-R1 e executá-lo *offline* em máquinas que não necessitam de tanto processamento quanto o que seria necessário para um GPT-4 (Matos, 2025). Assim, é possível utilizar o DeepSeek-R1 no contexto das organizações sem o risco de compartilhar os dados com a empresa que criou o modelo. Isso possibilita a soberania sobre os dados nacionais.

Destaca-se que as empresas chinesas enfrentam limitações no uso de recursos computacionais, tanto no hardware quanto no software. A empresa NVIDIA, que produz placas de vídeo e processadores amplamente utilizados em IA, não pode vender seus lançamentos mais recentes a empresas chinesas (Vargas, 2025).

O lançamento do DeepSeek representou uma revolução na área de IAG, pois a empresa conseguiu otimizar recursos computacionais e financeiros e, ainda assim, criar um modelo que não perde em qualidade nem em desempenho em relação aos grandes modelos do mercado.

Uma das desvantagens na utilização dos grandes modelos de linguagem é que eles são treinados em uma grande quantidade de textos, na maioria das vezes, em inglês, isso faz com que quando sejam utilizados em português possam não compreender bem o contexto das palavras e gerar as "alucinações" – especialmente no português jurídico –, embora os modelos mais recentes sejam treinados em múltiplos idiomas a maioria das informações não são em português.

Após o lançamento do DeepSeek, o Ministério da Ciência, Tecnologia e Inovação (MCTI) passou a considerar viável um modelo de IA treinado em português, pois o custo de treinamento do DeepSeek demonstrou que é possível criar um modelo de LLM com menos recursos computacionais e financeiros. Para isso, consta a proposta do Plano Brasileiro de Inteligência Artificial (PBIA), do MCTI, com previsão de investimentos de R$ 23 bilhões até 2028. O plano deve abranger áreas de saúde, educação, meio ambiente, segurança pública, agricultura e gestão governamental, com foco em modernizar serviços, combater desigualdades e promover a inclusão social (Agência Gov, 2025).

Dessa forma, abre-se a possibilidade de que não apenas as empresas bilionárias do setor de tecnologia possam construir seus modelos, mas até mesmo alguns países – e, quem sabe, o Poder Judiciário brasileiro – possam ter seus próprios LLMs em breve.

Compreendidos os fundamentos da IA e dos LLMs, passa-se à análise de como essas tecnologias vêm sendo aplicadas no Poder Judiciário brasileiro.

3 USO DE INTELIGÊNCIA ARTIFICIAL NO PODER JUDICIÁRIO

Nesse contexto, o uso de IA no Poder Judiciário vem crescendo nos últimos anos. Diversos tribunais no Brasil têm projetos de IA com o intuito de promover maior celeridade no julgamento.

Segundo o CNJ, o Brasil possui quase 84 milhões de processos em tramitação, distribuídos por 91 tribunais – mais de 80% na Justiça Estadual –, que contam com cerca de 18 mil juízes e 275 mil servidores responsáveis por sua solução. O índice de judicialização continua a crescer e, em 2023, atingiu 35 milhões de novos casos, um aumento de quase 9,5% em relação ao ano anterior (CNJ, 2024b).

O Poder Judiciário brasileiro possui um grande volume de processos que, sem o uso de novas tecnologias, é impossível julgar de forma mais célere e eficiente. Nesse cenário, a pesquisa realizada em junho de 2024 pelo Conselho Nacional de Justiça (CNJ) demonstrou que o uso de IA no Poder Judiciário cresceu 26% em relação a 2022 (CNJ, 2022). Contudo, com o crescente uso de Inteligência Artificial nas decisões públicas – sejam judiciais ou administrativas –, surge uma maior preocupação quanto ao uso adequado dessa nova tecnologia.

Vale destacar que o relatório da UNESCO de 2022 apresentou diretrizes sobre os modelos de IA destinados ao apoio às decisões públicas – judiciais e da Administração Pública –, que deveriam seguir os princípios da proporcionalidade, justiça, não discriminação, supervisão e determinação humana, transparência, explicabilidade e responsabilidade (Pádua; Lorenzetto, 2024). Assim, o crescente uso da IA no sistema de justiça também levanta questões fundamentais sobre transparência, imparcialidade e respeito aos direitos fundamentais (Carneiro; Araújo; França, 2025).

Diante desse contexto de aumento do uso da IA no Brasil, consta em tramitação o Projeto de Lei n. 2338/2023 no Senado Federal (Senado, 2025) que traz em seu art. 1º que essa lei estabelece “normas gerais de caráter nacional para o desenvolvimento, implementação e uso responsável de sistemas de Inteligência Artificial no Brasil, com o objetivo de proteger os direitos fundamentais e garantir a implementação de sistemas seguros e confiáveis, em benefício da pessoa humana, do regime democrático e do desenvolvimento científico e tecnológico”.

O Projeto de Lei n. 2338/2023 parte do princípio de que a inovação tecnológica deve avançar em paralelo à proteção da dignidade humana, dos direitos fundamentais e da democracia. Ele se fundamenta em princípios como a não discriminação, a transparência, a supervisão humana e a proteção de dados pessoais. A proposta classifica os sistemas de IA de

acordo com o nível de risco: os de baixo risco terão obrigações mínimas; os de alto risco deverão cumprir requisitos de governança, mitigação de vieses, auditoria e documentação técnica; e os de risco excessivo podem ser proibidos. Além disso, assegura aos cidadãos direitos importantes, como a informação prévia sobre o uso de IA, a explicabilidade das decisões automatizadas e o direito de contestá-las, especialmente quando afetarem interesses jurídicos ou pessoais.

Nessa linha do Projeto de Lei 2338/2023, Toledo e Pessoa (2023) ressaltaram a preocupação com o uso dessa nova tecnologia no Direito e a necessidade de transparência nos sistemas de IA, devido à possibilidade de vieses cognitivos se replicarem na decisão judicial por meio de IA. Com isso, verifica-se que a utilização da IAG deve ser transparente quanto ao uso de seus algoritmos, até para que as partes possam conhecer como a IAG foi utilizada para auxiliar em alguma decisão judicial e saber se foi utilizada alguma base que contenha discriminação ou outro viés ideológico (Bonat; Vale; Pereira, 2023).

Nos treinamentos de grandes modelos de linguagem, os parâmetros são ajustados a cada vez que a máquina erra na classificação dos resultados. Assim, quanto maiores forem as bases de dados de treinamento do modelo e o ajuste dos parâmetros, maior será a acurácia do modelo de *Machine Learning* (Kneusel, 2024). Dessa forma, é importante que os dados de treinamento dos modelos sejam abundantes e de boa qualidade para evitar o chamado "viés do algoritmo", que pode gerar resultados distorcidos da realidade (Kneusel, 2024).

Contudo, é importante destacar que não apenas a máquina, mas também o próprio ser humano pode apresentar distorções da realidade. Nesse caso, ocorre o que a psicologia moderna denomina "vieses da cognição". Nesse ponto, é importante destacar que não apenas a máquina, mas também o próprio ser humano pode sofrer distorções da realidade. Nesse caso, ocorre o que a psicologia moderna chama de "vieses da cognição humana", que decorrem do funcionamento do cérebro humano e são distorções ou ilusões cognitivas, erros sistêmicos de avaliação que acabam por influenciar o entendimento e as crenças da pessoa em relação ao mundo que está à sua volta (Salomão; Tauk, 2023b).

Tanto os seres humanos quanto as máquinas estão sujeitos a distorções na interpretação da realidade. Nos indivíduos, tais distorções configuram vieses cognitivos, isto é, erros sistemáticos de avaliação que afetam a percepção e a tomada de decisão (Salomão; Tauk, 2023b). De modo análogo, os algoritmos podem reproduzir vieses presentes nos dados de treinamento, gerando o chamado viés algorítmico. Por isso, a supervisão humana continua indispensável.

Apesar disso, algumas pessoas ainda tendem a confiar nas respostas fornecidas pelos sistemas, acreditando que a máquina será mais precisa por recorrer a métodos matemáticos, mas isso nem sempre ocorre (Salomão; Tauk, 2023b). Além disso, o algoritmo pode "alucinar" na resposta quando não consegue classificar corretamente e, com isso, cria informações inexistentes apenas para oferecer uma resposta, gerando doutrinas e jurisprudências que não existem.

Demonstrando a mesma preocupação que levou à proposição do Projeto de Lei 2338/2023, o CNJ publicou, em 2024, um relatório com informações sobre o uso de IAG no Poder Judiciário (CNJ, 2024d). O relatório, justamente, evidenciou a preocupação com a governança no uso da IAG, destacando a necessidade de revisão da Resolução n. 332/2020 do CNJ, que estabeleceu diretrizes sobre ética, transparência e governança na produção e no uso de inteligência artificial no Poder Judiciário. A atualização dessa resolução ocorreu com a publicação da Resolução CNJ n. 615, de 14 de março de 2025.

Um dos pontos de maior preocupação apresentados pelo relatório do CNJ foi a falta de transparência no uso desses modelos de IA. Com essa preocupação, o CNJ criou uma base de dados em 2020 (Resolução CNJ n. 332/2020) para o armazenamento das informações sobre os projetos de IA utilizados pelos tribunais no Brasil; a base de dados é chamada de "SINAPSES", que tem como objetivo "armazenar, testar, treinar, distribuir e auditar modelos de inteligência artificial, disponível na Plataforma Digital do Poder Judiciário (PDPJ-Br)" (Resolução CNJ n. 615/2025, p. 7).

A Resolução CNJ n. 615/2025 reforça a preocupação do CNJ com o uso responsável da Inteligência Artificial (IA) no Poder Judiciário, estabelecendo diretrizes para a governança de sistemas de IA tradicionais e generativos. O texto normativo enfatiza o respeito aos direitos fundamentais, à proteção de dados pessoais e à transparência algorítmica, alinhando-se aos princípios da Lei Geral de Proteção de Dados (LGPD) e às boas práticas internacionais.

A nova norma amplia o escopo da regulação, abrangendo expressamente as IAs generativas e os grandes modelos de linguagem (LLMs), incluindo *chatbots* e ferramentas de automação judicial.

Entre os principais avanços, destaca-se a instituição, por tribunal, de uma política de governança de IA, que prevê registro, *logs*, documentação técnica e rastreabilidade completa dos sistemas. Como toda decisão automatizada passa a exigir supervisão humana efetiva, são vedadas decisões autônomas de mérito tomadas por sistemas de IA.

A resolução incorpora os princípios de *privacy by design* (privacidade dos dados é uma prioridade desde a concepção do projeto) e *privacy by default* (utilização, por padrão, de alto nível de confidencialidade dos dados) (Resolução CNJ n. 615/2025, p. 8), além de prever mecanismos de anonimização de dados, relatórios de impacto algorítmico e de mitigação de vieses. Diferentemente da Resolução CNJ n. 332/2020 — que previa apenas o registro das soluções na base do SINAPSES — a nova norma exige auditoria técnica e o registro formal de todos os sistemas, com documentação pública e possibilidade de verificação externa. Também introduz programas de capacitação voltados a magistrados, servidores e jurisdicionados, a fim de promover o uso consciente e responsável da tecnologia.

A Resolução CNJ n. 615/2025 reforça a autonomia dos tribunais, mas condiciona-a à observância de padrões mínimos de segurança, transparência e auditoria, fortalecendo o controle social e o direito à informação. Determina a publicação de relatórios de impacto e desempenho, acessíveis ao público e às partes interessadas, ampliando a *accountability* institucional.

Entre suas inovações éticas, restringe práticas invasivas, proibindo o uso de IA para prever comportamento humano, reconhecer emoções ou realizar discriminações sem base legal. Estabelece, ainda, padrões técnicos de interoperabilidade e de segurança cibernética entre as soluções de IA do Judiciário, promovendo a cooperação entre os tribunais e o CNJ.

A norma reforça o dever de cuidado e a responsabilização por danos decorrentes do uso indevido de IA, definindo responsabilidades institucionais e pessoais na implantação, no monitoramento e na auditoria das soluções tecnológicas. A Resolução CNJ n. 615/2025 ainda determina que os tribunais realizem análises de risco quanto ao uso de IA, a fim de evitar que algoritmos apresentem vieses capazes de comprometer a equidade e a justiça decisória.

Por exemplo, o art. 19, § 6º da Resolução CNJ n. 615/2025 dispõe que os magistrados e servidores podem utilizar LLMs, SLMs e outros meios de IAG desde que sigam os padrões de segurança da informação e as regras impostas por essa resolução:

> Art. 19. Os modelos de linguagem de larga escala (LLMs), de pequena escala (SLMs) e outros sistemas de inteligência artificial generativa (IAGen) disponíveis na rede mundial de computadores poderão ser utilizados pelos magistrados e pelos servidores do Poder Judiciário em suas respectivas atividades como ferramentas de auxílio à gestão ou de apoio à decisão, em obediência aos padrões de segurança da informação e às normas desta Resolução.
>
> § 6º Quando houver emprego de IA generativa para auxílio à redação de ato judicial, tal situação poderá ser mencionada no corpo da decisão, a critério do magistrado, sendo, porém, devido o registro automático no sistema interno do tribunal, para fins de produção de estatísticas, monitoramento e eventual auditoria.

Contudo, apesar da exigência normativa de que o magistrado marque no sistema se utilizou de IA, não há a obrigação de mencionar isso no corpo da decisão. Assim, nota-se que o magistrado não é obrigado a deixar explícito, nos atos judiciais, que utilizou alguma ferramenta de IA. O ideal seria que, pelo princípio da transparência e da ampla defesa, esse uso fosse devidamente indicado. Trata-se de um ponto que precisa de melhor regulamentação pelos tribunais.

Quanto à utilização de IA pelos tribunais em 2023, consta na base do SINAPSES que 94 órgãos (tribunais e conselhos) responderam à pesquisa, dos quais 62 utilizaram IA, totalizando 140 projetos (CNJ, 2023). Já quanto à pesquisa referente aos dados de 2024, participaram 92, dos quais 58 utilizaram IA, totalizando 98 projetos cadastrados. Isso corresponde a 96,8% do total de órgãos. (CNJ, 2024c). Verifica-se que a pesquisa de 2024 criou uma seção específica sobre o uso de IAG pelos tribunais e conselhos. Nesse ponto, é importante ressaltar que o Supremo Tribunal Federal, por não estar subordinado ao CNJ, não participou da pesquisa. Assim, os projetos do STF não estão mapeados nessa base do CNJ.

Nos dados de 2023 constantes no SINAPSES, é possível verificar os principais usos de IA no Poder Judiciário, sendo que a grande maioria dos projetos trabalha com classificação, análise de similaridade de texto e busca semântica.

Outro dado interessante é que a maioria dos dados utilizados nos projetos de IA é proveniente dos próprios tribunais e continua a ser observada na pesquisa de 2024, o que demonstra preocupação com a utilização de dados próprios. Isso é importante para que a IA entenda o contexto de cada órgão e tenha maior controle sobre a qualidade dos dados utilizados, a fim de evitar também vieses.

Também é importante destacar que a pesquisa do SINAPSES mostra quais modelos de LLM os tribunais estão utilizando, com maior uso do GPT-4, da OpenAI, e do BERT, da Google. Os tribunais ainda possuem poucos projetos que utilizam LLMs, mas há interesse em investir nesses modelos, conforme pesquisa do SINAPSES (CNJ, 2023).

Apesar do uso de LLMs nos tribunais em 2023 ter sido pequeno, provavelmente ocorreu devido aos altos custos de investimento para utilizar essa tecnologia e à falta de pessoal capacitado para isso. Já na pesquisa sobre os dados de 2024, observou-se um aumento de 81,25% no interesse pela IAG , apesar de ainda aparecer que as principais dificuldades para a elaboração dos projetos de IAG sejam: de encontrar profissionais treinados em IA; questões relacionadas a privacidade e a segurança dos dados utilizados; dificuldade na obtenção de dados (quantidade, qualidade ou diversidade); necessidade de adaptar os processos e rotinas já

estabelecidos; complexidade de adaptar a IA com os sistemas existentes; dificuldades em obter os recursos financeiros; questões de ética e de transparência no uso de IA na tomada de decisões judiciais; e a resistência de servidores e magistrados na adoção de IA (CNJ, 2024c).

Além disso, as principais aplicações de algoritmos de máquina utilizados nos projetos de IA foram redes neurais, árvores de decisão, florestas aleatórias e máquinas de vetores. O mais utilizado ainda é o de redes neurais, o modelo mais utilizado em IA, conforme dados de 2023.

A dependência de plataformas corporativas, como Microsoft e Google, predomina, refletindo a escassez de infraestrutura pública e a falta de LLMs nacionais. Essa situação aumenta a necessidade de modelos abertos e auditáveis sob controle público, garantindo que o Brasil possa exercer controle efetivo sobre os dados utilizados pelas grandes empresas de tecnologia. Para manter a soberania sobre seus dados, o país precisa de uma gestão clara desse controle. Isso só será possível quando o armazenamento for responsabilidade pública e o uso for transparente. Atualmente, os modelos de IA nos tribunais geralmente estão hospedados em nuvens dessas grandes corporações, sem transparência quanto ao uso dos dados armazenados nesses ambientes.

Já quanto ao código-fonte, consta que a maioria é de propriedade do próprio tribunal. Isso é importante porque eles têm acesso ao código para alterá-lo e atualizá-lo, o que gera transparência na utilização do código e soberania sobre os dados. Contudo, em muitos projetos, o código-fonte não está disponível publicamente para reutilização e ainda é de terceiros, como demonstra a pesquisa do SINAPSES (CNJ, 2023). Isso prejudica o compartilhamento de tecnologia entre os tribunais e a dependência de terceiros.

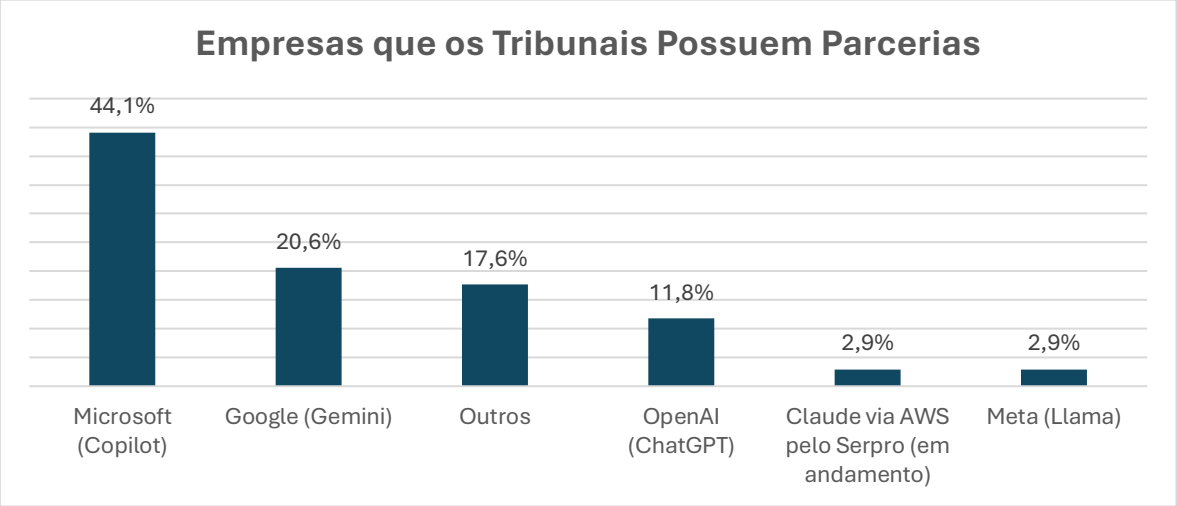
Apesar de desenvolver um algoritmo próprio para resumos e análises semânticas, a maior parte dos tribunais ainda depende de grandes modelos de linguagem de empresas de tecnologia, cujos processos de uso de dados e de geração de resultados não são transparentes.

Outro ponto de alerta é que 48,6% dos projetos de IA no Poder Judiciário em 2023 não possuíam documentação (CNJ, 2023), o que pode contribuir para a falta de transparência sobre como essas aplicações funcionam e para a necessidade de cumprir todas as diretrizes da Resolução CNJ n. 615/2025 que traz várias recomendações sobre o uso de IAG pelos tribunais.

Já a pesquisa de 2024 do SINAPES confirma que a maioria das ferramentas de IAG é utilizada por meio da interface oficial, como o ChatGPT ou o Gemini, diretamente ou por meio de APIs, o que evidencia a dependência dos órgãos de grandes empresas de tecnologia (CNJ, 2024c).

A maioria dos tribunais mantém parcerias com grandes empresas de tecnologia, como a Microsoft e a Google (figura 6) (CNJ, 2024c). Muitos órgãos públicos possuem contratos com a Microsoft para o sistema operacional e o pacote Office, o que explica um grande uso do *Copilot* que em geral é usado pelos servidores e magistrados como apoio nas atividades administrativas (Silva et al., 2025).

Figura 6 - Empresas que os tribunais possuem parcerias para IAG



Fonte: SINAPSES (CNJ, 2023).

Já segundo a pesquisa do CNJ de 2024 (2024c), os principais usos de IAG no Poder Judiciário continuam a ser a geração, a melhoria, a sumarização e a verificação ortográfica de textos e documentos. O que demonstra que os tribunais ainda utilizam a IA preditiva e generativa em atividades de baixo risco, considerando a classificação de riscos da Resolução CNJ n. 615/2015.

Com esse tipo de utilização da IAG, os principais benefícios foram: melhoria da produtividade na elaboração de documentos, aumento da velocidade e da eficiência dos processos judiciais, redução do tempo gasto em tarefas administrativas repetidas, auxílio na detecção de inconsistências e de possíveis erros em documentos produzidos (CNJ, 2024c).

Outra informação de alerta é que a maioria dos tribunais não elaborou diretrizes próprias para o uso de IAG (CNJ, 2024c). O que é essencial para uma boa governança da IAG.

A nova Resolução CNJ n. O 615/2025 trata da autonomia dos tribunais para criar e utilizar suas próprias inteligências artificiais no seu contexto, o que gera certa insegurança nos órgãos quanto ao uso de IA, na ausência de normatização específica.

Cabe ainda destacar que o Superior Tribunal de Justiça (STJ) tem papel de destaque na elaboração de projetos de IA, como o Athos, o Sócrates e o STJ Logos (FGV, 2025).

O Athos realiza a gestão de precedentes e a automação do exame de admissibilidade, além de agregar e monitorar processos repetitivos. Ele utiliza análise semântica, modelos de *machine learning* e planos vetoriais (FGV, 2025).

O Sócrates realiza a triagem de Recursos Especiais, aponta a ausência de requisitos e pesquisa automaticamente por precedentes. Ele utilizou a programação neurolinguística semântica (PNL) e vetores (FGV, 2025).

Já o STJ Logos é a ferramenta mais recente do STJ, a primeira IAG que apoia os gabinetes na produção de minutas e na tomada de decisões por meio de comandos em um chat. Utilizou LLM integrado à base do Tribunal (FGV, 2025).

Apesar de tantas iniciativas, ainda se nota a falta de maior cooperação entre os órgãos do Poder Judiciário para desenvolver algoritmos em conjunto e compartilhar tecnologias – inclusive por meio de compras públicas centralizadas de computadores de alta performance. Embora cada tribunal tenha suas particularidades, seria possível compartilhar mais recursos humanos e computacionais, já que as maiores dificuldades identificadas na pesquisa do SINPAPSES foram justamente de pessoal treinado e de recursos financeiros para o desenvolvimento de projetos de IA.

Além disso, há poucas iniciativas de chamamento público para que empresas privadas ofereçam soluções ao Poder Judiciário, apesar dos dados do SINAPSES indicarem vários contratos com empresas privadas, como a Microsoft e a Google. Verifica-se que, por se tratar de pesquisa mais recente do SINAPSES, referente a 2024, ainda não há registro de uso do DeepSeek pelos tribunais.

A pesquisa do SINAPES de 2024 mostra que ainda há receio, no âmbito jurídico, de que empresas privadas ofereçam soluções que influenciem decisões judiciais, já que a maioria dos modelos de LLM não é transparente quanto ao seu funcionamento, a chamada opacidade dos modelos. Além disso, há o risco de órgãos públicos fornecerem dados do Poder Judiciário para o treinamento de modelos privados de IA. Contudo, é importante destacar que, entre as diretrizes da Resolução CNJ n. 615/2025 consta a exigência de que os tribunais analisem os riscos associados à utilização de aplicações de IA. Contudo, ainda é muito difícil não depender de grandes empresas de tecnologia, devido à falta de pessoal treinado e de recursos computacionais para desenvolver seus próprios modelos e demais tecnologias necessárias.

Após essa visão geral do Poder Judiciário, prepara-se o caminho para uma análise mais detalhada das experiências do Supremo Tribunal Federal, que se destaca no contexto das inovações tecnológicas. Como a mais alta instância judicial, o STF lidera e influencia os demais órgãos do sistema de justiça, sendo uma referência tanto nacional quanto internacional na implementação de soluções de Inteligência Artificial e na prática de governança digital.

4 USO DE INTELIGÊNCIA ARTIFICIAL NO SUPREMO TRIBUNAL FEDERAL

A primeira inteligência artificial do STF, Victor, foi criada em parceria com a UnB a partir de 2017, por meio do Termo de Execução Descentralizada 1/2018. Essa inteligência artificial classifica os recursos extraordinários de acordo com temas de repercussão geral. Uma das preocupações do STF, à época, era a redução do acervo de processos do Tribunal. (Hartmann Peixoto, 2020b).

Segundo relatório de pesquisa de Salomão e Tauk foram utilizadas 22.000 petições de Recursos Extraordinários e selecionados 27 temas mais frequentes (2023, p. 29 e 30):

> Foram utilizadas 22.000 petições de RE (3 TB de dados) do período entre 2014 e 2017 para o treinamento do modelo atualmente implantado, as quais foram disponibilizadas ao Grupo de Aprendizado de Máquina (GPAM) da Universidade de Brasília para processamento. Para o treinamento, foram selecionados os 27 temas mais frequentes. Durante a fase de treinamento, identificou-se que deveriam ser priorizadas cinco peças na construção do sistema: o acórdão, o recurso extraordinário, o agravo de recurso extraordinário, o despacho e a sentença.
>
> Os autos processuais dos feitos recursais remetidos ao STF são submetidos ao modelo que identifica a presença de um ou mais temas de repercussão geral.
>
> O Victor utiliza a inteligência artificial aplicada em linguagem natural (texto) para executar a seguinte sequência de atividades: (i) conversão de imagens no processo digital ou eletrônico em textos: os recursos chegam ao STF, como regra, como peças digitalizadas em formatos que nem sempre permitem a leitura pela máquina. Por isso, precisam ser submetidas a uma fase de reconhecimento ótico de caracteres (OCR – Optical Caracter Recognition), que converte as imagens das peças em texto, viabilizando o uso de técnicas de processamento de linguagem natural (NLP); (ii) após, há a separação do começo e do fim de um docu- mento (peça processual, decisão etc.) no arquivo pdf; (iii) em seguida, há a classificação das peças processuais mais utilizadas nas atividades do STF (o acórdão, o recurso extraordinário, o agravo de recurso extraordinário, o despacho e a sentença); (iv) por fim, o sistema faz a identificação se o recurso protocolado se encaixa em um dos temas de repercussão geral de maior incidência para os quais foi treinado, sem elaboração de minuta. Todos os itens acima são realizados pelo sistema Victor em cerca de 5 (cinco) segundos. Trata-se, portanto, de um sistema que apoia a atividade de análise de admissibilidade recursal por meio da sugestão de um ou mais temas de repercussão geral, posteriormente sujeita à validação pelos servidores e pelos ministros.

A extração do texto, por meio de OCR, nas peças que vinham em vários formatos, foi uma das etapas mais complexas e demoradas, exigindo ajustes significativos nos algoritmos de processamento. Quando o projeto do Victor começou, ainda havia processos físicos; por isso, houve um grande trabalho de classificação das peças. A equipe da UnB teve que rotular os dados enviados pelo STF para a classificação das peças (Hartmann Peixoto, 2020b).

Segundo informações no Portal do STF, quando o Recurso Extraordinário (RE) era encaminhado ao STF, era necessário que um servidor separasse e identificasse suas peças, tarefa que demandava, em média, 30 minutos de serviço. O Victor realizava essa tarefa em apenas cinco segundos (STF, 2018).

Em abril de 2025, o Núcleo de Inteligência Artificial do STF informou que a acurácia do Victor se encontrava em 90% de acertos nas inferências de temas realizadas no mês. Cerca de 775 temas de Repercussão Geral (RG) estão atualmente classificados pela ferramenta. A ferramenta utiliza tecnologia de similaridade com o uso de metadados (Núcleo de IA do STF, 2025). Ressalta-se que a ferramenta realiza uma categorização ainda em revisão por um núcleo do STF; ou seja, não realiza uma análise jurídica (Nunes, 2025). Assim, não substitui a análise jurídica dos servidores e dos magistrados

Segundo o relatório de Salomão e Tauk (2023), não há evidências de um aumento significativo na produtividade com o uso do Victor, já que os servidores continuam revisando suas sugestões da mesma maneira que analisam recursos sem a ferramenta de IA.

Já a RAFA 2030, criada em 2022, classifica os processos de acordo com os Objetivos de Desenvolvimento Sustentável (ODS) da Agenda 2030 da Organização das Nações Unidas (ONU). Antes da utilização da IA, a atividade era realizada manualmente pelos servidores. A necessidade dessa classificação dos ODS surgiu em 2020, quando foi iniciado o projeto da Agenda 2030 na Corte. A importância dessa classificação reside em permitir ao Tribunal dar prioridade aos processos que atendam aos objetivos da Agenda 2030 (Nunes, 2025). A ferramenta, em abril de 2025, apresentava acurácia superior a 90% na classificação dos processos nos ODS 3, 8, 10 e 16 – que correspondiam a mais de 80% dos processos classificados no STF. Essa ferramenta foi desenvolvida pela equipe interna do STF em linguagem R (Núcleo de IA do STF, 2025).

A classificação dos processos é facilmente visível na consulta no site do STF; os classificadores indicam o número do ODS. Um exemplo, na ADPF 1107 aparecem os objetivos 10 – Redução das desigualdades e 16 – Paz, Justiça e Instituições Eficazes.

A VitórIA, lançada em 2023 (STF, 2023), agrupa recursos e reclamações por meio de textos e metadados. A VitórIA é parecida com o Athos do Superior Tribunal de Justiça (STJ). Até abril de 2024, a ferramenta do STF permitia a criação de 6 temas de repercussão geral (Núcleo de IA do STF, 2025).

Conforme apresentação no Youtube (STF, 2023), a VitórIA cria grupos de processos com base na similaridade com um processo utilizado como paradigma. A VitórIA está

configurada para trabalhar com classes recursais a partir de 2022 (STF, 2023). Os servidores dos gabinetes podem utilizar a ferramenta no STF Digital para agrupar seus processos (Núcleo de IA do STF, 2025). A ferramenta utilizou um agrupador próprio, baseado em similaridade, e foi desenvolvida na linguagem Python (Núcleo de IA do STF, 2025).

A área de IA do STF desenvolveu a RAFA 2030 (Redes Artificiais Focadas na Agenda 2030) e a VitórIA com recursos próprios utilizando componentes desenvolvidos internamente como: OCR, limpeza de textos, *embedding* que são representações numéricas que transformam tokens em vetores que os LLMs podem interpretar e utilizar para detectar padrões complexos em textos (Núcleo de IA do STF, 2025)

Encerrada a fase de projetos de IA que realizavam apenas a classificação e a rotulação dos dados (Victor, VitórIA, RAFA 2030), o STF ingressou em uma nova etapa: o uso de IAG para auxiliar na redação de atos judiciais, iniciada com a MARIA.

Em dezembro de 2023, o STF publicou o Edital de Chamamento Público n. 001/2023 com o "objetivo de permitir a participação de interessados no desenvolvimento de protótipos de soluções de IAG para a criação de sumários automatizados de processos judiciais no Tribunal" (STF, 2024a, p. 2).

Consta no Relatório do Chamamento que 39 pessoas jurídicas foram habilitadas a apresentar suas soluções, de um total de 60 interessados, mas apenas 23 concluíram o desenvolvimento e apresentaram seus protótipos em evento realizado em 18 de dezembro de 2023. Por fim, 22 empresas encaminharam os sumários gerados por suas ferramentas para apreciação jurídica (STF, 2024a, p. 2).

O Tribunal tinha o objetivo de verificar a possibilidade de utilizar modelos de linguagem, principalmente LLMs, nas atividades do STF. Embora os modelos de LLM possam gerar resumos textuais, as principais soluções existentes apresentavam a limitação de não ter seus modelos treinados com base em textos jurídicos em língua portuguesa.

As empresas teriam que desenvolver aplicações com LLMs capazes de analisar peças processuais e extrair, com precisão, informações relevantes de cada documento, resultando na elaboração de relatórios no formato predefinido no edital do chamamento.

É importante ressaltar que a iniciativa não gerou nenhum ônus para o STF, nem havia expectativa de contratação de qualquer participante do chamamento.

Consta no item 7 do Edital do Chamamento que os participantes deveriam atuar com "transparência, responsabilidade e visando à mitigação de riscos e vieses, seguindo as boas práticas globais e a Estratégia Brasileira de Inteligência Artificial" (STF, 2024a, p. 5)

A MARIA (Módulo de Apoio para Redação com Inteligência Artificial) é resultado desse chamamento público e foi lançada em dezembro de 2024, tornando-se a primeira IAG do STF. A EloGroup, em parceria com a Microsoft, cedeu o código-fonte ao STF (STF, 2024a).

Conforme notícia publicada no site do STF, em dezembro de 2024, a ferramenta será utilizada para resumos de votos, relatórios em processos recursais e análise inicial de processo de reclamação (STF, 2024b):

> Resumos de votos: a MARIA pode gerar automaticamente minutas de ementas, com o resumo do entendimento do ministro sobre a matéria em questão.
>
> Relatórios em processos recursais: a ferramenta pode resumir relatórios de ministros em Recursos Extraordinários (REs) e em Recursos Extraordinários com Agravo (AREs).
>
> Análise inicial de processos de reclamação: a MARIA analisa a petição inicial e apresenta respostas aos questionamentos que orientam o estudo inicial desse tipo de processo.

Consta no site do STF que, em setembro de 2025, a MARIA incorporou novas funcionalidades de IA, como a revisão gramatical e textual, que permite selecionar trechos para correção diretamente no sistema, e a consulta unificada de precedentes, recurso que apresenta decisões relacionadas ao caso em análise, sem necessidade de pesquisas externas. As ferramentas estão disponíveis no STF Digital, ambiente eletrônico que reúne os sistemas judiciais do Tribunal, e a IA não atua em processos sigilosos ou de segredo de justiça (FGV, 2025).

A MARIA também ampliou as classes de análise. Inicialmente restrita à geração de ementas no padrão do Conselho Nacional de Justiça (CNJ), relatórios em recursos extraordinários (RE) e recursos extraordinários com agravo (ARE) e questionários apenas para petições iniciais de reclamação, a plataforma hoje abrange (STF, 2025):

- Questionários para agravos regimentais
- Embargos de declaração
- Decisões monocráticas finais e o inteiro teor de RE e ARE
- Relatórios para reclamações, petições iniciais e agravos.

Dessa forma, é possível verificar a evolução das IAs do STF conforme a figura abaixo (figura 7):

Figura 7- Linha do Tempo da Evolução das Inteligências Artificiais do STF

Ano	Inteligência Artificial	Descrição e Função Principal	Base Tecnológica e Observações
2017–2018	VICTOR	Primeira IA do STF, criada em parceria com a UnB (TED nº 1/2018). Classifica Recursos Extraordinários segundo os temas de repercussão geral. Reduz o tempo de análise de 30 minutos para 5 segundos. Começou com 27 temas e hoje já analisa mais de 700.	Utiliza NLP e OCR para converter documentos digitalizados em texto. Acurácia atual: 90%.
2022	RAFA 2030	Classifica os processos de acordo com os ODS da Agenda 2030 da ONU, permitindo a priorização de ações que atendam aos ODS 3, 8, 10 e 16.	Desenvolvida internamente pelo STF em linguagem R. Acurácia superior a 90%. Representa avanço da IA voltada à gestão e governança judicial.
2023	VITÓRIA	Agrupa recursos e reclamações por similaridade textual e metadados, permitindo análise temática e automatizada de processos semelhantes.	Desenvolvida em Python com embeddings e vetorização semântica.
2024	MARIA	Primeira IA generativa efetiva do STF. Gera minutas de ementas, relatórios de votos e análises iniciais de processos de reclamação. Em 2025 expandiu as funções da IA para revisão gramatical e textual, consulta unificada de precedentes e ampliação das classes de análise.	Desenvolvida pela EloGroup e pela Microsoft e cedida ao STF. Integrada ao STF Digital, atua com controle humano e não é utilizada em processos sigilosos ou restritos.

Fonte: autoria própria, com base nas informações do Núcleo de IA do STF (2025) e da FGV (2025).

Essa trajetória evidencia a maturidade institucional do STF no uso da IA, criando uma base concreta para o desenvolvimento de um LLM jurídico nacional.

Assim, encerrada a análise das experiências institucionais, o estudo passa a examinar o potencial dos LLMs, com destaque para o DeepSeek, na transformação do uso de IA pelo Judiciário.

5 COMO GRANDES MODELOS DE LINGUAGEM, COMO O DEEPSEEK, PODEM IMPACTAR NO USO DE IA NO PODER JUDICIÁRIO

O surgimento de modelos de linguagem de grande porte (LLMs), como o DeepSeek, com custo de treinamento significativamente inferior ao do GPT da OpenAI abre caminho para mais inovações no Poder Judiciário.

Inicialmente, é importante diferenciar o custo de treinamento de um modelo de IA e o custo de sua inferência. O treinamento de aprendizado profundo e a inferência de IA são duas etapas de um mesmo processo voltado à obtenção de resultados úteis de um modelo de inteligência artificial. O treinamento ocorre primeiro: à medida que o modelo é treinado, ele adquire a capacidade de reconhecer níveis mais complexos de informação a partir dos dados. Já a inferência de IA ocorre após o treinamento, quando o modelo recebe uma solicitação para identificar esses elementos em novos conjuntos de dados (Erickson, 2024).

O treinamento de um algoritmo de inteligência artificial envolve um processo no qual se parte de um modelo base, ensinando-o a tomar decisões corretas com base em exemplos. Esse processo requer grandes volumes de dados e pode envolver diferentes níveis de supervisão humana. A quantidade de dados necessária está diretamente relacionada ao número de parâmetros definidos no algoritmo, bem como à complexidade do problema a ser resolvido. Tanto o número de parâmetros quanto o tamanho do conjunto de dados influenciam significativamente os recursos de processamento necessários e, consequentemente, os custos de treinamento. Por outro lado, a etapa de inferência exige muito menos recursos computacionais, sendo, portanto, mais barata e mais simples de integrar algoritmos de IA já treinados aos sistemas existentes das organizações (Doyle, 2023).

O treinamento de LLMs demanda um investimento financeiro considerável, devido à quantidade de parâmetros e ao alto consumo de energia computacional. Normalmente, esse processo utiliza GPUs de alta performance ou aceleradores de IA especializados, o que implica custos elevados tanto na compra quanto na manutenção (Ohiri; Poole, 2025).

Embora a maioria das empresas não divulgue os custos de treinamento de seus modelos, o *transformer* original, de 2017, que introduziu a arquitetura central dos LLMs, custou cerca de US$ 900 para ser treinado. Já o custo de computação para o treinamento do GPT-3, com 175 bilhões de parâmetros, foi estimado em 2020 entre US$ 500 mil e US$ 4,6 milhões, dependendo do hardware e das técnicas de otimização empregadas. Em comparação, os modelos mais recentes apresentam custos significativamente mais elevados. (Ohiri; Poole, 2025).

O treinamento do GPT-4, desenvolvido pela OpenAI, custou mais de US$ 100 milhões, com estimativas indicando cerca de US$ 78 milhões apenas em custos de recursos computacionais. Já o modelo Gemini Ultra, do Google, teve um custo estimado de US$ 191 milhões em recursos computacionais para treinamento. Esses valores elevados refletem, em parte, o aumento significativo do tamanho e da complexidade dos modelos ao longo do tempo (Ohiri; Poole, 2025).

As maiores empresas de tecnologia construíram supercomputadores na nuvem para viabilizar o treinamento de LLMs. A Microsoft, por exemplo, desenvolveu um supercomputador no Azure com mais de 10.000 GPUs e uma rede de altíssima velocidade, projetado especificamente para o treinamento de modelos da OpenAI (Ohiri; Poole, 2025).

Contudo, o custo de alugar essa infraestrutura é elevado. Na NVIDIA GTC 2024, Jensen Huang, CEO da NVIDIA, explicou que treinar o modelo GPT-MoE-1.8T levou de três a cinco meses usando 25.000 GPUs. Ele também estimou que treiná-lo na arquitetura Hopper (H100) com 8.000 GPUs levaria cerca de 90 dias, em um período semelhante. Na prática, a maioria dos usuários não treina LLMs do zero devido ao alto custo. Em vez disso, preferem usar modelos pré-treinados criados por grandes organizações e centros de pesquisa, como o ChatGPT ou o Llama (Ford, 2024).

Sobre a inferência de IA, vários tipos podem atender a diferentes casos de uso (Red Hat, 2025):

- **Inferência em lotes:** recebe esse nome porque processa dados em grandes lotes. Em vez de realizar inferências em tempo real, esse método processa as informações em blocos — às vezes de hora em hora ou até diariamente —, dependendo da quantidade de dados e da eficiência do modelo de IA. Essas inferências também podem ser denominadas *offline* ou estáticas (Red Hat, 2025).
- **Inferência online:** também conhecida como inferência dinâmica, essa abordagem é capaz de gerar respostas em tempo real. Essas inferências exigem hardware e software capazes de reduzir a latência e viabilizar previsões em alta velocidade. A inferência online é especialmente útil na *edge*, ou seja, quando a IA processa os dados diretamente no local onde são gerados. Isso pode ocorrer em dispositivos como celulares, automóveis ou estações remotas com conectividade limitada (Red Hat, 2025).

  O ChatGPT da OpenAI é um bom exemplo de inferência online. Ele exige uma infraestrutura operacional robusta para oferecer respostas rápidas e precisas (Red Hat, 2025).

- **Inferência em streaming:** descreve um sistema de IA que não é voltado à interação direta com pessoas. Em vez de receber prompts ou solicitações específicas, o modelo processa um fluxo contínuo de informações para realizar previsões e atualizar seu banco de dados interno. A inferência em streaming permite monitorar mudanças, manter a regularidade e até prever a ocorrência de um problema antes que ele se manifeste (Red Hat, 2025).

Segundo Erickson (2024), a inferência de IA requer a integração de muitas fontes de dados e uma arquitetura que permita ao modelo de IA funcionar com eficiência. Estas são as principais tecnologias que permitem que a IA funcione da melhor forma:

- **Unidade de processamento central (CPU)**
    - Uma CPU é o cérebro do computador. É um chip com circuitos complexos que reside na placa-mãe do computador e executa o sistema operacional e as aplicações. Uma CPU ajuda a gerenciar os recursos de computação necessários ao treinamento e à inferência de IA, como o armazenamento de dados e as placas gráficas (Erickson, 2024).
- **Unidade de processamento gráfico (GPU)**
    - As GPUs são componentes de hardware essenciais para a inferência em IA. Assim como as CPUs, as GPUs são chips com circuitos complexos; contudo, diferem por terem sido projetadas especificamente para realizar cálculos em alta velocidade, oferecendo suporte ao processamento de gráficos e imagens. Esse poder de computação é o que torna possível tanto o treinamento quanto a inferência de modelos de IA que demandam elevada capacidade de processamento (Erickson, 2024).
- **Matriz de portas programáveis em campo (FPGA)**
    - Um FPGA é um circuito integrado que pode ser programado pelo usuário final para executar funções específicas. Na inferência de IA, o FPGA pode ser configurado para oferecer a combinação ideal entre velocidade de hardware e paralelismo, dividindo o processamento de dados entre múltiplas unidades que operam simultaneamente. Essa característica permite que o modelo de IA realize previsões sobre diferentes tipos de dados, como textos, gráficos e vídeos (Erickson, 2024).
- **Circuito integrado específico de aplicação (ASIC)**

- Os ASICs são outra ferramenta utilizada por equipes de TI e cientistas de dados para realizar inferências de IA com a velocidade, o custo e a precisão necessários. Um ASIC é um chip de computador que integra diversos circuitos em um único componente, podendo ser otimizado para uma carga de trabalho específica — como reconhecimento de voz, processamento de imagens, detecção de anomalias ou outros processos orientados por IA (Erickson, 2024).

O avanço dos processadores gráficos (GPUs) e de técnicas como LoRA e QLoRA reduziu drasticamente os custos de treinamento e de adaptação de modelos. Esse contexto torna viável que instituições públicas, com infraestrutura moderada, executem ajustes finos (*fine-tuning*) em LLMs de código aberto, como o DeepSeek, adequando-os a domínios jurídicos específicos.

Existe ainda a técnica RAG (*Retrieval-Augmented Generation*), que complementa a inferência dos LLMs por meio do uso de bases de dados externas — como documentos, repositórios de informação ou sites —, fornecendo contexto adicional ao modelo. Essa abordagem contribui para reduzir as chamadas "alucinações" do modelo, tornando as respostas mais precisas e fundamentadas (Syal, 2024).

Também é possível realizar o chamado *fine-tuning*, que consiste em adaptar modelos de LLMs previamente treinados a uma tarefa, a um conjunto de dados ou a um domínio específico. Em vez de treinar o modelo do zero, o *fine-tuning* aproveita o conhecimento já existente e ajusta seus pesos com base em novos dados, tornando-o mais preciso e eficiente para casos de uso especializados — como na análise de peças jurídicas (Sujatha R, 2025).

A configuração adequada da GPU desempenha um papel fundamental no ajuste fino de LLMs e redes neurais, uma vez que esse processo demanda elevada capacidade de armazenamento e de processamento computacional. A escolha da GPU impacta diretamente o desempenho, a viabilidade, o custo e a escalabilidade do fluxo de trabalho de treinamento. Por exemplo, um modelo com 13 bilhões de parâmetros pode exigir mais de 200 GB de VRAM para um ajuste fino completo, o que requer o uso de múltiplas GPUs A100 ou H100. Com a técnica QLoRA, no entanto, o mesmo modelo pode ser executado em uma GPU RTX 4090 com apenas 24 GB de VRAM (Sujatha R, 2025).

O primeiro passo consiste em estimar a quantidade de VRAM necessária para o modelo-alvo. Uma diretriz comumente citada indica que o ajuste fino completo requer aproximadamente 16 GB de VRAM por bilhão de parâmetros, o que significa que um modelo

com 7 bilhões de parâmetros pode demandar mais de 100 GB de memória de GPU se treinado sem otimizações.

Com métodos de ajuste fino eficientes em termos de parâmetros, como LoRA e QLoRA, esses requisitos podem ser drasticamente reduzidos — para menos de 24 GB. Por exemplo, um modelo LLaMA-2 de 7 bilhões de parâmetros pode ser treinado em uma única NVIDIA RTX 4090 (24 GB) utilizando QLoRA (Sujatha R, 2025).

Após dimensionar a carga de trabalho, a etapa seguinte é mapeá-la para a classe de GPU adequada. As GPUs de consumo, como RTX 4000 Ada, RTX 6000 Ada e L40S, são apropriadas para métodos de ajuste fino com eficiência de parâmetros (como LoRA e QLoRA), para o treinamento de modelos menores — com até 7 bilhões de parâmetros — e para a execução de tarefas de inferência ou prototipagem (Sujatha R, 2025). Já modelos de classe corporativa, como o NVIDIA H100 (80 GB), são indispensáveis em projetos de larga escala, voltados ao treinamento completo de parâmetros ou ao aprendizado por reforço com *feedback* humano (RLHF) (SUJATHA R, 2025). O resultado é um modelo adaptado ao domínio, mais caro e demorado do que o RAG e menos flexível para atualização (precisa de novo treino quando surgem novos dados) (Sujatha R, 2025).

É importante ressaltar que nem todos os modelos permitem *fine-tuning*, pois isso depende da arquitetura técnica do modelo e da licença de uso. Os modelos de código aberto (*open source*) geralmente permitem ajuste fino, uma vez que seus pesos estão disponíveis publicamente — como é o caso do Llama 3, Mistral, Falcon e DeepSeek. Já nos modelos fechados (proprietários), como ChatGPT, Gemini e Claude, não é possível realizar o retreinamento direto, pois seus pesos não são públicos. Nesses casos, o *fine-tuning* só pode ser efetuado se as próprias empresas disponibilizarem essa funcionalidade oficialmente, por meio de interfaces ou APIs dedicadas (Magalhães, 2024).

Low-Rank Adaptation (LoRA), ou Adaptação de Baixo Posto, é uma técnica de ajuste fino voltada para modelos de inteligência artificial, em especial para grandes modelos de linguagem (Magalhães, 2024). Seu diferencial está em ser um método de ajuste fino eficiente em termos de parâmetros (*Parameter-Efficient Fine-Tuning* – PEFT), permitindo adaptar modelos robustos de forma mais econômica e com menor consumo de recursos computacionais. Em vez de atualizar todos os pesos originais, o LoRA aprende apenas matrizes adicionais de baixa dimensão, que contêm muito menos parâmetros. O modelo original permanece congelado (isto é, não é alterado), e apenas os parâmetros do LoRA são treinados. Essa abordagem permite reduzir o consumo de memória, treinar GPUs menos potentes e combinar múltiplos LoRAs —

por exemplo, um voltado à linguagem jurídica e outro ao atendimento ao público. Já na técnica *Adapters*, pequenas camadas adicionais são inseridas entre as camadas do modelo original. Assim como no LoRA, os pesos originais permanecem congelados, e apenas esses módulos extras são ajustados. A principal diferença é que os *Adapters* adicionam novos módulos à arquitetura, enquanto o LoRA modifica camadas já existentes por meio de fatores de baixa dimensionalidade (Magalhães, 2024).

Já em relação ao DeepSeek, de acordo com o *Artificial Intelligence Index Report 2025* da Universidade de Stanford (2025), o modelo alcançou um desempenho excepcionalmente alto, exigindo muito menos recursos computacionais do que muitos LLMs líderes.

Uma outra vantagem do DeepSeek é que ele possui código aberto e isso permite que os usuários tenham mais controle sobre o modelo e a capacidade de executá-lo de acordo com sua própria estrutura, com sua própria governança e privacidade de dados, além de permitir o *fine-tuning* (ajuste fino), a versão R1 pode ser baixada e rodada *offline* em um computador com no mínimo 16GB de RAM e 8GB de VRAM (Matos, 2025), o que não seria possível com um GPT-4 que não disponibiliza essa opção e exigiria uma quantidade enorme de recursos computacionais (Databricks, 2024).

Além do DeeSeek, existem outros modelos de código aberto muito utilizados, como Llama, Mistral e Phi-3.x/4, Qwen 2.5, Gemma e Falcon (Parsadanyan; Dmitrevna, 2025).

O Llama 3, da Meta, não utiliza o MoE, mas há previsão de utilizá-lo nas próximas versões. O modelo é utilizado principalmente para tarefas de geração de texto geral, de tarefas multilíngues, de geração de código e de conteúdos extensos e possibilita o ajuste fino para domínios específicos (Parsadanyan; Dmitrevna, 2025).

O Mistral, da Mistral IA, usa o MoE e é utilizado principalmente para tarefas de alta complexidade, processamento multilíngue, geração de código, compreensão de imagens, computação complexa, uso em dispositivos e chamada de funções (Parsadanyan; Dmitrevna, 2025).

O Phi-3.x/4, da Microsoft, utiliza o MoE e é empregado na geração de texto em geral, em tarefas multilíngues, na compreensão de código, no raciocínio matemático, na compreensão de imagens e na inferência no dispositivo (Parsadanyan; Dmitrevna, 2025).

O Qwen 2.5, da Alibaba, não utiliza o MoE e tem como principais usos a geração de texto geral, tarefas multilíngues, geração de código, raciocínio matemático, processamento de dados estruturados (Parsadanyan; Dmitrevna, 2025).

O Gemma, do Google, não utiliza o MoE e tem como principais usos a geração de texto geral, a resposta a perguntas, a elaboração de resumos, a geração de código e o ajuste fino para domínios específicos (Parsadanyan; Dmitrevna, 2025).

Por fim, o Falcon 3, da TII, não utiliza o MoE e tem como principais usos a geração de texto geral, a geração de código, tarefas matemáticas, conhecimento científico e aplicações multilíngues, além de permitir o ajuste fino para domínios específicos (Parsadanyan; Dmitrevna, 2025).

Destaca-se que o DeepSeek, mesmo sendo um modelo de linguagem aberto, é tão bom quanto modelos como o GPT da OpenAI que custou centenas de milhões de dólares (Stanford, 2025).

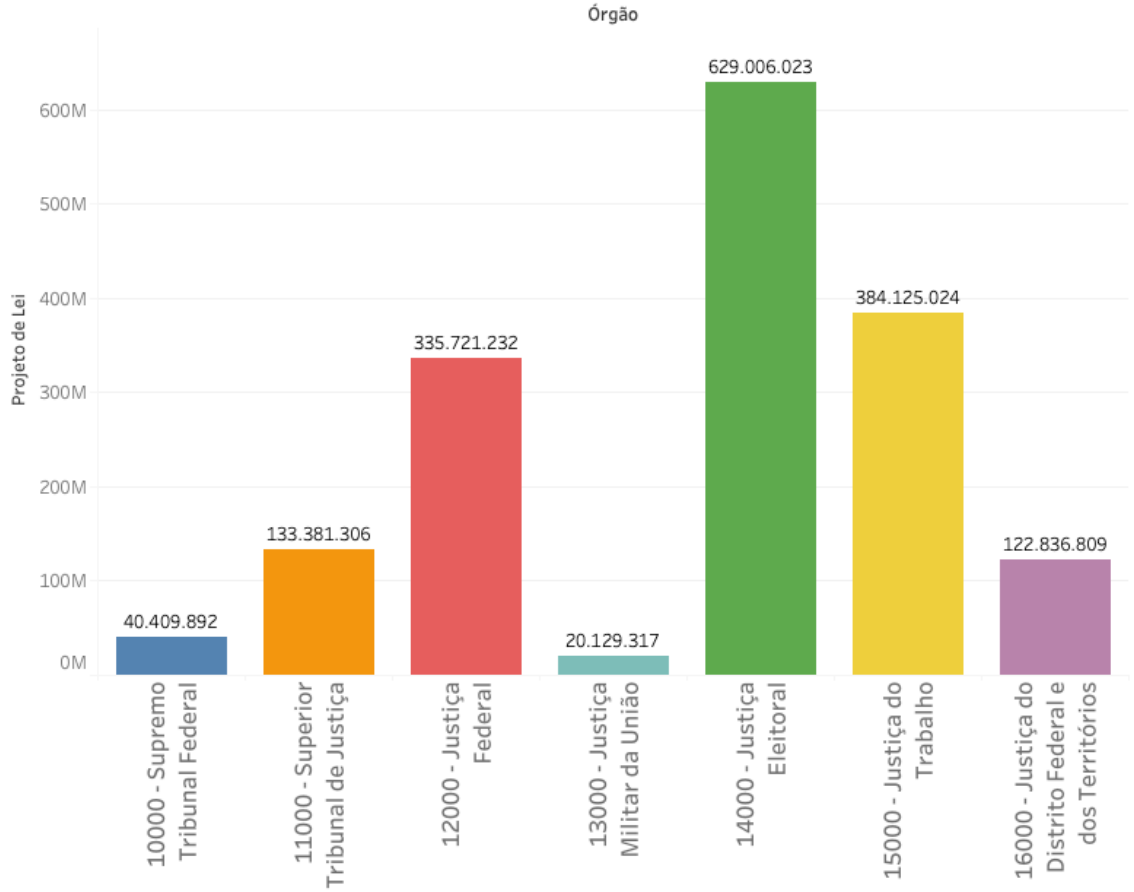
A vantagem de um modelo de código aberto é que qualquer pessoa pode auditar, modificar e reutilizar o modelo livremente, o que favorece a transparência e a adaptabilidade, como o uso de *fine-tuning* para adaptar os modelos aos contextos do Poder Judiciário.

Como foi dito, o Poder Executivo já vislumbra a possibilidade de que, diante de um modelo mais acessível, o Brasil em breve tenha um modelo próprio.

Os modelos atuais são treinados em multilíngues, mas um modelo treinado apenas em português seria muito mais assertivo nas respostas, pois toda variação semântica do português altera significativamente o significado das palavras. No Poder Judiciário, isso poderia ser viável no futuro, pois um modelo que custou cerca de 6 milhões de dólares em treinamento é mais próximo da realidade orçamentária do Poder Judiciário.

O orçamento de 2025 para despesas de TI do Poder Judiciário Federal (figura 8) está previsto em projetos de lei no valor de R$ 1.665.609.603,00. Demonstrando que, apesar de esse valor incluir o custeio de despesas já existentes, possivelmente, se os tribunais unissem esforços, seria possível investir em um LLM brasileiro para o Poder Judiciário ou, pelo menos, na maior utilização do *fine-tuning* de modelos existentes (MPO, 2025).

Figura 8 - Orçamento previsto em tecnologia da informação para 2025 para o Poder Judiciário



Fonte: MPO, 2025.

Ressalta-se que não foi possível identificar os valores investidos pelo Poder Judiciário em IA, informação que poderia ser obtida nas próximas pesquisas do CNJ.

Contudo, pelas pesquisas do SINPASES, é possível verificar que há custos com investimentos em treinamentos e parcerias com empresas de tecnologia. Apesar de a pesquisa não trazer os valores desses investimentos, é possível citar, por exemplo, que, em uma licitação recente, o STF e o STJ, por meio do Pregão Eletrônico n. 90016/2025 (Comprasnet, 2025) pretende adquirir, por meio de registro de preços, servidores de computação de alto desempenho (HPC), que são "supermáquinas" para o aprimoramento de IAG nesses órgãos; somente com essas compras está previsto um investimento de mais de 13 milhões de reais. Esses órgãos possuem núcleos de IA com servidores especializados nessas áreas e estão investindo cada vez mais em IAG.

Para comparação, o DeepSeek-V3 é treinado em um cluster com 2048 GPUs NVIDIA H800. Cada nó do cluster H800 contém 8 GPUs conectados entre si por NVLink e NVSwitch. (Vaswani, 2017) (Deepseek, 2025).

Assim, com relação às máquinas a serem adquiridas pelo STF e pelo STJ no Pregão Eletrônico n. 90016/2025, elas não são suficientes para treinar um modelo como o DeepSeek do zero, pois o DeepSeek-V3 precisou de 2.048 GPUs H800 interligadas por *InfiniBand*, algo viável apenas em supercomputadores de *big techs* ou de governos. O edital prevê máquinas com 2 a 4 GPUs cada, o que está muito abaixo dessa escala.

Contudo, é possível utilizá-lo para inferência, pois servidores com GPUs NVIDIA de alto desempenho (A100 ou H100, com 40–80 GB de memória) conseguem carregar o modelo ou suas versões otimizadas e responder em segundos.

Também é possível fazer o *fine-tuning* (ajuste fino) no domínio jurídico usando técnicas mais leves, como LoRA ou *adapters*. Com essas técnicas é possível treinar o modelo com decisões, jurisprudência e textos legais, ajustando-o ao contexto brasileiro. Além disso, é possível usar modelos com menos parâmetros.

Além disso, segundo o Relatório Técnico do DeepSeek-V3, o modelo adota uma arquitetura chamada *Mixture-Of-Experts* (MoE, mistura de especialistas). A versão do DeepSeek-V3 utiliza 671 bilhões de parâmetros, mas apenas 37 bilhões são ativados por token gerado, o que faz com que o modelo seja executado com menos recursos computacionais.

Aumentar a escala do modelo, ou seja, o número de parâmetros, tem sido fundamental para o sucesso da aprendizagem profunda, resultando em previsões mais precisas com conjuntos de dados maiores. No entanto, os custos de treinamento e de inferência aumentam proporcionalmente à capacidade, pois mais parâmetros são ativados para processar cada token. Para expandir o tamanho do modelo sem elevar esses custos, foi introduzida a abordagem MoE. (Zhang, 2025).

As principais técnicas que contribuem para a eficiência e o desempenho do DeepSeek incluem: *Mixture-of-Experts (MoE), Multi-Head Latent Attention (MLA), Multi-Token Prediction, DualPipe* e *FP8 Training* (Zhang, 2025).

A combinação de especialistas (*Mixture of Experts – MoE*) é uma técnica de aprendizado de máquina que organiza um modelo de inteligência artificial em subredes independentes (os “especialistas”), cada uma dedicada a um subconjunto específico dos dados de entrada. Esses especialistas atuam de forma coordenada, permitindo que o modelo execute tarefas de forma mais eficiente e direcionada (Bergmann, 2024).

Apesar de a maior parte das arquiteturas modernas de combinação de especialistas (MoE) ter sido desenvolvida na última década, a ideia central remonta ao artigo "*Adaptive Mixture of Local Experts"* de 1991. Nesse trabalho, os autores demonstraram que seu modelo experimental alcançava a mesma precisão de um modelo convencional, mas com metade do tempo de treinamento, evidenciando ganhos consideráveis de eficiência. (Bergmann, 2024).

Desde então, a arquitetura MoE passou a ser incorporada a alguns dos principais modelos de linguagem (LLMs), como o Mixtral 8x7B, da Mistral, e o GPT-4, da OpenAI (Bergmann, 2024).

Vários modelos de LLM utilizam a arquitetura MoE (Wolfe, 2025):

• **Mixtral (8x7B e 8x22B):** Uma versão de código aberto do Mistral-7B, que emprega oito experts por camada, com dois ativos por token (Wolfe, 2025).

• **Grok (da xAI):** O modelo Grok-1 é um MoE com 314 bilhões de parâmetros, dos quais aproximadamente 25% estão ativos por token (~70-80 bilhões de parâmetros ativos) (Wolfe, 2025).

• **DBRX (da Databricks):** Um MoE "granular" com 16 experts por camada (4 ativos por token), que melhorou a eficiência do pré-treinamento e apresentou desempenho destacado em tarefas de programação (Wolfe, 2025).

• **DeepSeek-v2/v3:** empregam experts granulares e a ideia de experts compartilhados (por onde passam todos os tokens), além de um treinamento chamado *Multi-Token Prediction* (MTP) para aumentar a eficiência (Wolfe, 2025).

Assim, em vez de depender de um único modelo massivo para abordar todos os aspectos de um problema, a arquitetura MoE divide a tarefa entre redes menores e especializadas, cada uma com foco em um domínio ou subtarefa específico (AlkameL, 2025).

Segundo Sayed Ali Alkamel, uma parte essencial do MoE é a rede de portas, que atua como um gerente ao determinar qual especialista é mais adequado para cada entrada. Ela avalia a entrada e a direciona de forma inteligente ao(s) especialista(s) mais relevante(s), assegurando um processamento eficiente e preciso.

Além disso, o MoE oferece uma vantagem significativa, pois, ao contrário dos modelos tradicionais que ativam todos os parâmetros para cada entrada, o MoE ativa apenas os especialistas necessários para uma determinada tarefa. Essa ativação seletiva reduz significativamente o custo computacional e melhora a eficiência, permitindo que os modelos MoE alcancem tamanhos massivos sem exigir um aumento proporcional no poder computacional (Alkamel, 2025).

Dessa forma, o uso de arquiteturas MoE reduz os custos de pré-treinamento de um LLM e ainda permite um desempenho melhor, pois são ativados apenas alguns especialistas específicos para cada tarefa, em vez de toda a rede neural (Alkamel, 2025).

Assim, o DeepSeek tem um uso eficiente de recursos, maior precisão nas tarefas e maior escalabilidade, pois pode incluir mais especialistas sem impactar muito os requisitos computacionais (Alkamel, 2025).

Por outro lado, Sayed Ali Alkamel traz que o MoE também possui alguns desafios como:

- **Instabilidade de Treinamento**: Modelos MoE tendem a sofrer colapso de roteamento (*routing collapse*) quando apenas alguns especialistas são escolhidos repetidamente. Isso impede que os demais aprendam e gera mau uso da rede.
  O DeepSeek reduz esse problema por meio do balanceamento de carga sem perdas auxiliares e de outras otimizações de treinamento. Implementa balanceamento de carga sem perdas auxiliares (*auxiliary-loss-free load balancing*). Ajusta dinamicamente o *bias* dos especialistas: se um especialista é sobrecarregado, seu *bias* cai; se é pouco usado, aumenta. Assim, todos os especialistas são utilizados de forma mais uniforme (Deepseek, 2025).
- **Desequilíbrio de carga**: Em MoE, alguns especialistas recebem muito mais tokens do que outros, criando gargalos e subutilização. Solução DeepSeek: Adota o algoritmo de roteamento *Expert Choice*, implementa técnicas de balanceamento para garantir a distribuição uniforme de tokens, evita sobrecarga e melhora a eficiência (Deepseek, 2025).
- **Altos Requisitos de Memória:** Mesmo os especialistas inativos precisam estar carregados na memória, o que aumenta o consumo de RAM e de GPU. Solução DeepSeek: Criação de versões destiladas (*distilled*) de seus modelos, com menos especialistas e com compressão de parâmetros, reduz o custo de memória em ambientes com recursos limitados (Deepseek, 2025).
- **Generalização durante o ajuste fino**: No *fine-tuning*, o MoE pode sofrer sobreajuste (overfitting) em especialistas específicos, prejudicando a capacidade de generalizar. Solução DeepSeek: Uso de técnicas de regularização (*dropout, weight decay, data augmentation*) e de estratégias de treinamento específicas para MoE, mantendo a generalização equilibrada (Deepseek, 2025).

- O *dropout* é uma técnica em que, durante o treinamento, alguns neurônios da rede são “desligados” aleatoriamente. Em cada iteração, uma porcentagem (ex.: 20–50%) das conexões é zerada. Isso evita que a rede dependa demais de um subconjunto específico de neurônios. Isso reduz *overfitting* (quando a rede decora os dados de treino e não generaliza) e faz a rede aprender representações mais robustas (Srivastava et al., 2014).
- *Weight Decay*: adiciona uma penalização aos valores dos pesos elevados da rede. A função de perda é ajustada para incluir a soma dos quadrados dos pesos. Isso força os pesos a permanecerem menores durante o treinamento. Isso evita que a rede atribua importância excessiva a conexões específicas e melhora a generalização em novos dados (Krogh; Hertz, 1991).
- *Data Augmentation*: consiste em aumentar artificialmente a diversidade dos dados de treino por meio de transformações controladas. Em imagens: rotações, espelhamentos, recortes e alterações de cores. Em texto: substituição de sinônimos, traduções reversas (*back-translation*), mascaramento de tokens. Em áudio: distorções leves, adição de ruído. Isso ajuda o modelo a generalizar melhor, pois aprende a lidar com variações e aumenta o conjunto de treino, sem precisar coletar novos dados reais (Shorten; Khoshgoftaar, 2019).

- **Limitações da inferência MoE**: A inferência MoE apresenta: alto consumo de memória e risco de token *dropping* (quando um token não encontra especialista disponível). Solução DeepSeek: otimizações de arquitetura e inferência, balanceamento eficaz que evita perda de tokens e melhor aproveitamento de cache e paralelismo (Deepseek, 2025).

O DeepSeek aprimorou a arquitetura MoE padrão com duas modificações (Dickson, 2025):

• **Segmentação de Especialistas de Granulação Fina (*Fine-grained expert segmentation*):** Os especialistas padrão foram divididos em **subespecialistas, cada qual** ainda mais especializado. Isto aumenta a flexibilidade e a capacidade do modelo de lidar com nuances, mantendo o cálculo total por token (Dickson, 2025).

• **Isolamento de Especialista Compartilhado (*Shared expert isolation*):** Um especialista foi designado como **"compartilhado"**. Cada token passa por este

especialista, que adquire conhecimento comum aplicável a diversos contextos (por exemplo, regras gramaticais ou raciocínio básico). Isto reduz a redundância e permite que os especialistas direcionados se tornem ainda mais especializados (Dickson, 2025).

Mesmo com todos os desafios, como verificado no relatório de Stanford (2025), o DeepSeek-V3 supera outros modelos de código aberto e atinge desempenho comparável ao dos melhores modelos fechados do mercado, especialmente em tarefas que exigem grande raciocínio matemático. Em diversos testes padronizados de conhecimento e raciocínio, ele atingiu resultados no nível de modelos como o GPT-4 da OpenAI. O modelo apresenta desempenho excepcional, sendo treinado de forma eficiente com menos recursos computacionais do que os modelos fechados (Deepseek, 2025).

A metodologia de desenvolvimento do DeepSeek focou na eficiência computacional. Os pesquisadores empregaram *mixed precision* com FP8 (8-bit) – validando, pela primeira vez, o treinamento em escala extrema com precisão reduzida – e técnicas de sobreposição de comunicação/cálculo para treinar o MoE em múltiplos nós, quase sem gargalos de rede. Além disso, foi introduzido um objetivo de treinamento de múltiplos tokens (MTP), no qual o modelo prevê mais de um token por vez. Esse objetivo *multi-token* demonstrou melhorar a performance e pode ser explorado para acelerar a geração de texto (via *speculative decoding*) (Deepseek, 2025)

O DeepSeek-V3 inovou ao eliminar a necessidade de perda auxiliar para balancear os especialistas do MoE, adotando uma estratégia de balanceamento de carga *auxiliar-free* que evita degradar o desempenho enquanto distribui melhor as tarefas entre os especialistas (Deepseek, 2025).

Diferentemente de muitos LLMs clássicos, com limite de contexto de 2k a 4k tokens (ou, em alguns modelos recentes, de 32k), o DeepSeek suporta entradas de até 128 mil tokens. Essa janela de contexto extremamente ampla permite ao modelo ler documentos extensos por completo, o que é particularmente relevante para o domínio jurídico – em que peças processuais, decisões e legislações podem ser muito longas –, sendo um excelente modelo *open-source* para ser utilizado no Poder Judiciário para *fine-tuning* (Deepseek, 2025).

O modelo foi pré-treinado em um conjunto massivo de dados – 14,8 trilhões de tokens de texto, de alta qualidade e de diversos tipos. Embora os detalhes não estejam todos publicados, esse volume sugere a inclusão de múltiplos domínios de conhecimento (literatura, web, código etc.) possivelmente em vários idiomas. Após o pré-treinamento, o DeepSeek passou por *fine-*

*tuni*ng supervisionado e por etapas de *Reinforcement Learning* (provavelmente com feedback humano, RLHF) para alinhamento, explorando ao máximo suas capacidades. Assim, o modelo otimizou o uso de recursos computacionais e reduziu significativamente os custos de treinamento (Deepseek, 2025).

Além disso, por ser aberto e por adotar uma arquitetura modular (especialistas que poderiam ser direcionados a domínios específicos), o DeepSeek é altamente adaptável (DeepSeek, 2025). A comunidade já produz adaptações publicadas no *Hugging Face*, facilitando experimentos e usos específicos (Hugging Face, 2025).

Essa adaptabilidade indica que o modelo pode ser ajustado para novas línguas ou campos de conhecimento, sem precisar recomeçar do zero. Por exemplo, módulos de especialistas poderiam ser treinados especificamente com textos jurídicos em português, integrando-se ao modelo principal.

O DeepSeek combina escala massiva, eficiência de treinamento, código aberto e alto desempenho, sendo considerado o modelo de MoE mais eficiente. Além disso, o modelo pode ser baixado para rodar *offline*. Essas características o tornam um forte candidato como ponto de partida para um LLM jurídico brasileiro, inicialmente para o *fine-tuning*, pois oferecem uma base técnica robusta sobre a qual se pode especializar o conhecimento do modelo. Também, dependendo da aplicação, podem ser utilizados outros modelos abertos. A adoção de um modelo inspirado no DeepSeek, adaptado ao contexto brasileiro, tem potencial para transformar significativamente a operação do Judiciário e do sistema jurídico do país. Um modelo de linguagem jurídica em português poderia atuar como assistente virtual para magistrados, capaz de analisar petições iniciais ou recursos, resumir os pontos principais, listar os pedidos das partes e até sugerir estruturas para decisões.

Além disso, com a capacidade de linguagem natural, poderia indicar jurisprudência relevante ao elencar precedentes similares dos tribunais superiores, economizando muitas horas de pesquisa. Também auxiliaria na redação de minutas de decisão, gerando rascunhos que seriam posteriormente revisados pelos juízes. Isso está bem próximo da realidade, como a MARIA no STF, mas, ao utilizar um modelo treinado em língua portuguesa na doutrina e na jurisprudência, a acurácia do modelo seria muito maior.

Esse LLM jurídico brasileiro poderia ainda gerar outros documentos padronizados, como despachos de citação, relatórios processuais resumidos e minutas de contratos ou de pareceres jurídicos para uso interno. Em áreas com processos repetitivos, como cobrança e

execuções fiscais, o modelo poderia preencher documentos automaticamente, liberando servidores para tarefas estratégicas ou que exigem juízo humano.

O grande volume da jurisprudência brasileira serviria de base para que o modelo respondesse a perguntas complexas, apoiando-se na doutrina e na jurisprudência nacionais, beneficiando juízes, advogados, promotores e defensores públicos com pesquisas mais ágeis.

Além disso, o LLM poderia integrar sistemas judiciais para classificar e encaminhar automaticamente novas ações, priorizar casos urgentes e ajudar na triagem de recursos, de forma semelhante ao Victor do STF. Com grandes acervos digitalizados, poderia extrair dados estatísticos e gerar relatórios gerenciais, auxiliando na gestão processual.

Para além do uso interno, o modelo poderia ser disponibilizado como um *chatbot* jurídico para advogados, defensores públicos e cidadãos, ampliando o acesso à informação jurídica, especialmente em regiões com poucos recursos.

Estrategicamente, um LLM jurídico nacional colocaria o Brasil como líder de inovação no Direito na América Latina, reduzindo a dependência tecnológica estrangeira, protegendo a soberania dos dados e gerando conhecimento especializado com potencial de exportação. A adoção interna poderia acelerar a prestação jurisdicional, enfrentando problemas estruturais, como a lentidão e o acúmulo processual, promovendo a eficiência e a democratização do conhecimento jurídico.

O uso da IA no sistema judicial reduz os custos de mão de obra, de tempo e de recursos financeiros, que estão cada vez mais escassos no setor público. Esta tecnologia pode automatizar tarefas repetitivas e demoradas, como a revisão de documentos e o registo de informações (Jadidi, 2025).

Experiências internacionais mostram o aumento do uso de IA no sistema judiciário e como ela pode ajudar (Jadidi, 2025).

A pesquisa de Vahid Jadidi (2025) é interessante porque detalha o uso de IA no Poder Judiciário de diversos países.

No Reino Unido, o software de IA tem sido usado para analisar casos, reduzindo a necessidade de contratar muitos funcionários para realizar essas tarefas. A JustisOne ajuda os advogados a gastar menos tempo em pesquisas jurídicas, fornecendo acesso rápido a leis e casos semelhantes (Jadidi, 2025).

No sistema judicial canadense, as ferramentas de IA são utilizadas na análise de documentos jurídicos e reduzem, em média, os custos judiciais em 30% (Jadidi, 2025).

Na Austrália, a IA já é utilizada para analisar argumentos jurídicos e determinar sentenças. Este sistema está sendo capaz de reduzir significativamente os erros jurídicos na revisão de casos complexos (Jadidi, 2025)

Na China, tribunais online equipados com inteligência artificial podem receber documentos relacionados a casos, analisá-los e fornecer resultados preliminares aos juízes. Esse processo reduziu significativamente o tempo necessário para a análise dos casos (Jadidi, 2025).

Ainda na China, Pequim tornou-se a primeira cidade do mundo a introduzir um centro de serviços judiciais baseado na Internet, com um juiz de inteligência artificial para alguns serviços. A chamada "Xinhua" é uma juíza robótica feminina com corpo, rosto, voz e maneirismos modelados a partir de uma juíza do Poder Judiciário de Pequim. A juíza virtual será utilizada principalmente em casos semelhantes e repetitivos. Em vez de emitir sentenças finais, ela lidará com petições e orientações online (Jadidi, 2025).

Nos Estados Unidos, o software Lex Machina fornece decisões jurídicas mais precisas em casos comerciais, com base em dados históricos (Jadidi, 2025). Os tribunais em alguns estados estão usando ferramentas de IA, como o ROSS Intelligence, para localizar documentos jurídicos e agilizar a emissão de sentenças (Jadidi, 2025). No estado de Nova Iorque, esta ferramenta ajudou os advogados a fornecer respostas mais precisas e rápidas em casos jurídicos complexos (Jadidi, 2025). Já a ferramenta COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*) utiliza dados de casos criminais anteriores para estimar a probabilidade de reincidência (Jadidi, 2025). O sistema COMPAS foi muito criticado por seu algoritmo ter demonstrado preconceito na análise das sanções, com base em determinados grupos étnicos. (Engel; Linhardt; Schubert, 2024).

Na Índia, um projeto chamado SUPACE utiliza IA para analisar rapidamente documentos jurídicos, reduzindo o tempo necessário para processar casos complexos (Jadidi, 2025).

Sobre a transparência e linguagem acessível, na Índia, sistemas baseados em IA explicam ao público as razões das decisões judiciais em linguagem simples (Jadidi, 2025).

Na França, ferramentas de previsão jurídica baseadas em IA são utilizadas para analisar dados de processos criminais. Essas ferramentas têm atingido até 80% de precisão na previsão de resultados (Jadidi, 2025).

Na Holanda, ferramentas de IA são usadas para gerir processos judiciais de forma digital. Esses sistemas podem categorizar informações sobre os processos e apresentá-las aos juízes (Jadidi, 2025).

Já em Singapura, os sistemas de gestão de processos baseados em IA reduziram a densidade de processos nos tribunais (Jadidi, 2025).

A ferramenta da Blue J Legal utiliza IA para prever os resultados de processos fiscais. A ferramenta tem conseguido aumentar a precisão das previsões em 90%. No Canadá, os advogados que utilizaram esta ferramenta conseguiram acelerar significativamente a tomada de decisão em processos fiscais (Jadidi, 2025).

A LawGeex gera rascunhos de contratos com base em regras e condições definidas e permite que os advogados os editem e os preparem com apenas alguns cliques. Em um grande escritório de advocacia na Alemanha, essa ferramenta reduziu o tempo necessário para redigir contratos de várias horas para apenas alguns minutos (Jadidi, 2025).

Vahid Jadidi (2025) ainda levanta as limitações e desafios com a implementação de IAs nas cortes judiciais:

- Falta de infraestrutura adequada: é necessária uma infraestrutura técnica e jurídica para a implementação bem-sucedida da IA nos tribunais. Em muitos países, as deficiências do sistema judicial digital impedem o pleno uso da tecnologia (Jadidi, 2025).
- Resistência tradicionalista: alguns juízes e advogados estão preocupados com o uso de IA; enxergam nisso uma ameaça à independência do sistema judicial e ao espírito humano do Poder Judiciário. Há necessidade de encontrar um equilíbrio entre a tecnologia e os princípios jurídicos tradicionais (Jadidi, 2025).

Outra iniciativa promissora que demonstra como é viável uma maior utilização do *fine-tuning* no contexto do Poder Judiciário brasileiro é a do Chile, um grande modelo de linguagem aberto da América Latina, adaptado por *fine-tuning* com dados característicos dessa região. O modelo deverá ser chamado de LatamGPT, e o Brasil está entre os países colaboradores do projeto. O modelo é baseado na tecnologia Llama 3 e é adaptado por meio de uma rede regional de computadores, incluindo instalações na Universidade de Tarapacá, no Chile, e sistemas em nuvem. O projeto de código aberto, coordenado pelo Centro Nacional de Inteligência Artificial (CENIA) do Chile, em parceria com mais de 30 instituições regionais, busca ampliar a adoção e a acessibilidade da IA em toda a América Latina, incluindo línguas indígenas no treinamento, além do espanhol e do português. Os pesquisadores destacaram que o modelo não pretende ser um concorrente do ChatGPT, mas sim ser utilizado como assistente virtual em serviços públicos e em sistemas de educação personalizados para a cultura da América Latina (Reuters, 2025).

Estes exemplos fornecem lições valiosas sobre governança, transparência, riscos e o potencial da IA jurídica, que o Brasil pode adaptar à sua realidade ao criar um LLM brasileiro.

No entanto, a implementação enfrenta desafios diversos:

- Infraestrutura computacional: Treinar e operar um modelo em larga escala exige alto investimento em hardware de ponta, como clusters de GPUs, energia e manutenção. Técnicas como o LoRA podem ajudar a reduzir os custos, mas a necessidade de centros de supercomputação ou de parcerias com nuvens confiáveis é essencial, o ideal seria uma nuvem governamental brasileira para manter o controle dos dados pelo poder público.
- Curadoria e disponibilidade de dados em português: embora exista um grande volume de dados jurídicos brasileiros, a dispersão, os formatos variados (PDF, HTML), os direitos autorais e a necessidade de limpeza e classificação exigem esforço técnico e jurídico para montar bases de dados adequadas e, principalmente, livres de vieses para treinamento.
- Adaptação linguística e cultural: partir de modelos pré-treinados, principalmente em inglês, exige *fine-tuning* focado no português e em termos jurídicos nacionais, para garantir o entendimento preciso da terminologia e dos conceitos legais próprios do Brasil. Isso já vem sendo realizado com modelos com uma quantidade menor de parâmetros.
- Privacidade e segurança: Devido à sensibilidade dos dados jurídicos, é essencial anonimizar informações pessoais, restringir o acesso, utilizar uma infraestrutura nacional segura e implementar mecanismos para evitar a exposição de dados confidenciais pelo modelo.
- Questões regulatórias e éticas: O uso deve seguir as normas do CNJ e a legislação vigente, como a LGPD, com princípios de transparência, supervisão humana e mitigação de vieses. A IA deve ser uma ferramenta auxiliar, nunca substituir a decisão humana.
- Resistência cultural e capacitação: Será necessário treinamento para magistrados e servidores, explicando as limitações e as vantagens, além de promover programas-piloto e criar cargos para a curadoria e a manutenção do sistema.

Assim, para avançar no desenvolvimento e implantação de um LLM jurídico brasileiro, será necessário:

- Implementação de políticas públicas voltadas ao Judiciário, com financiamento e fomento específicos.
- Construção de infraestrutura computacional robusta em parcerias públicas e privadas.
- Cooperação estreita entre tribunais e a academia para o desenvolvimento, a avaliação e a capacitação, como foi feito com o Victor no STF.
- Criação de grupos de trabalho e de pilotos em tribunais para testar e ajustar o modelo em ambiente controlado.
- Incentivos econômicos para atrair startups e empresas inovadoras.
- Estabelecimento de frameworks éticos e de sistemas de transparência, qualidade e auditoria.
- Escolha casos de uso estratégicos de alto impacto para gerar adesão e apoio.
- Plano de atualização contínua diante da dinâmica legislativa e tecnológica.

Verifica-se que, para treinar um modelo do zero, são necessários muitos recursos computacionais e humanos, o que acaba sendo muito caro. Se existir um modelo do governo brasileiro, em parceria com centros de pesquisa e com treinamento em português, como o governo espera, seria muito mais fácil adaptar-se ao Poder Judiciário brasileiro.

Hoje, de forma isolada, os tribunais não teriam capacidade de treinar um modelo do zero. Contudo, vários tribunais estão investindo isoladamente em máquinas com capacidade para realizar o *fine-tuning* de modelos existentes, nem que sejam com menos parâmetros, como o STF e o STJ planejam, que é uma forma de já treinar os modelos com a jurisprudência brasileira, capaz de atuar como “assistente” do operador do Direito, elevando a produtividade e consistência das decisões judiciais, pois existe um grande orçamento para a área de tecnologia informação, existem núcleos de IA sendo criados em vários tribunais com pessoal mais capacitado e estão cada dia mais investindo em compras de equipamentos para ter uma capacidade operacional maior para treinar os modelos de IA nos contextos de cada tribunal.

Assim, a experiência do DeepSeek demonstra que a soberania tecnológica e linguística é alcançável mesmo em contextos de restrição orçamentária, desde que acompanhada de políticas de governança, infraestrutura compartilhada e padrões éticos consolidados. O Poder Judiciário brasileiro, pela natureza pública de seus dados e pela capilaridade institucional, reúne as condições ideais para liderar esse movimento.

6 CONCLUSÃO

O trabalho demonstrou que a IAG, embora seja uma tecnologia nova no contexto jurídico, já se consolida como um instrumento estratégico para o Poder Judiciário brasileiro. As Resoluções n. 332/2020 e n. 615/2025 do CNJ estabeleceram um marco regulatório essencial, que impõe princípios de transparência, proteção de dados e gestão de riscos, garantindo que a adoção dessas ferramentas não comprometa a legitimidade e a segurança jurídica. Nesse cenário, a plataforma SINAPSES atua como um elo técnico e institucional fundamental para consolidar a governança, registrar as iniciativas e assegurar a possibilidade de auditorias, apesar de muitos tribunais ainda não constarem de sua base de dados, como o STF e o STJ, que possuem grandes iniciativas no campo da IA.

O exame das experiências do STF, como o Victor, VitórIA, RAFA 2030 e, sobretudo, a MARIA, confirma que o uso de IA já trouxe ganhos concretos em triagem, classificação e elaboração de minutas. Esses projetos revelam tanto o potencial da tecnologia quanto os limites que exigem supervisão humana qualificada, sobretudo diante dos riscos de vieses, alucinações e falta de rastreabilidade.

O estudo demonstrou que, no âmbito tecnológico, modelos abertos de alto desempenho, como o DeepSeek, oferecem ao Estado brasileiro uma oportunidade concreta de soberania sobre os dados. A sua arquitetura eficiente e de custo relativamente baixo aumenta a viabilidade de desenvolver, no futuro, um grande modelo de linguagem (LLM) jurídico, treinado em português e adaptado ao vocabulário e às particularidades normativas do Brasil. No entanto, o desafio vai além do acesso ao hardware, abrangendo a curadoria e a anonimização de dados, a criação de mecanismos de avaliação contínua, a governança de dados e a capacitação de equipes.

Uma solução para o uso de IAG no Poder Judiciário pode ser o uso de um modelo de código aberto como o Deepseek-R1, pois, como foi dito, pode ser baixado e utilizado *offline* sem a possibilidade de fornecer dados do Poder Judiciário para as empresas que criaram os modelos e dependendo da capacidade computacional dos tribunais já pode ser utilizado para *fine-tuning*. Não é necessário criar um LLM do zero; hoje, já podem ser utilizados modelos abertos, armazenados localmente e aplicados no contexto de cada órgão. Muitos tribunais já estão fazendo isso ao criar seus núcleos de IA e ao adquirir máquinas com maior capacidade de processamento de dados.

Contudo, o ideal seria o Poder Judiciário criar seu próprio LLM treinado em português jurídico, com base na jurisprudência brasileira. Em breve, isso será possível devido à disponibilidade de excelentes modelos, como o Deepseek-R1, com custo muito menor de

recursos financeiros e computacionais para o treinamento. Até porque a tecnologia está evoluindo muito rapidamente e os valores estão ficando mais acessíveis.

Entretanto, é necessário considerar que os custos não se limitam ao uso do modelo de LLM em si, mas também a toda a infraestrutura por trás dele. Antes do uso de um modelo de LLM, é necessária uma grande capacidade de processamento de dados e pessoas treinadas para tratar os dados a serem utilizados pelos modelos.

O Poder Executivo, após o DeepSeek, já fala em um modelo de IA brasileiro. Após a criação de um modelo treinado em português, o treinamento futuro em português jurídico e na jurisprudência brasileira seria mais fácil.

Cada tribunal investe em tecnologia de forma isolada, comprando máquinas e treinando pessoal próprio. Uma solução para o Poder Judiciário adotar um modelo de IA seria o Conselho Nacional de Justiça investir em um centro de tecnologia com maior capacidade de processamento de dados. A evolução das ferramentas de IAG tem se tornado cada vez mais rápida, e não é impossível que exista um LLM brasileiro no futuro próximo, desde que os tribunais unam esforços para esse fim e façam parcerias com centros de pesquisa. Inclusive, várias universidades estão adquirindo computadores com maior capacidade de processamento.

O surgimento do DeepSeek demonstrou que não é impossível um governo, em parceria com centros de pesquisa ou até mesmo com o Poder Judiciário, unir esforços e recursos humanos e computacionais para desenvolver seu próprio modelo de LLM, mesmo que seja um modelo com menos parâmetros. Antes, a visão era de que isso seria tão caro que apenas as grandes empresas de tecnologia teriam essa capacidade.

Os modelos ainda “alucinam” e geram jurisprudência e doutrinas. Contudo, quando testados em um ambiente mais controlado, eles tendem a “delirar” menos e a dar respostas mais assertivas. A qualidade das respostas dos modelos mais avançados e pagos é muito melhor, pois, como já foi dito, eles possuem mais parâmetros de validação e um custo de uso e manutenção maior. Assim, treinar modelos abertos pode ser uma vantagem, e o DeepSeek é um modelo aberto com excelentes resultados. Apesar de ser uma ferramenta com maior foco em raciocínio matemático, o DeepSeek-V3.2-Exp foi lançado recentemente e deve melhorar a produção de texto.

Hoje, a inteligência artificial ainda é mais uma ferramenta de apoio que precisa de supervisão humana por um especialista, pois somente os especialistas conseguem validar se a IA vai ou não “alucinando” nas respostas.

Ainda, para as decisões judiciais, é necessário um olhar humano que uma máquina ainda não possui condições de analisar. Até porque muitas respostas podem vir erradas ou apresentar vieses algorítmicos. Ademais, há uma grande preocupação com as ideologias por trás das ferramentas de IA e com como isso pode influenciar as respostas. Não há muita transparência quando se usam esses modelos de mercado.

Tanto no ambiente acadêmico quanto no de trabalho, as pessoas utilizam a IA o tempo todo para diversos fins. A Resolução CNJ n. 615/2025 permite a utilização de IA nas atividades de apoio no Poder Judiciário, inclusive com contas particulares em ferramentas de IA do mercado. Contudo, isso gera receio quanto ao uso dessas ferramentas, especialmente em relação às informações fornecidas, o que suscita preocupações éticas e de privacidade. As pessoas estão sendo treinadas pelos tribunais, pois também precisam entender que as respostas das ferramentas requerem supervisão humana.

Apesar disso, não tem como deixar de usar a IA no Poder Judiciário; é um caminho sem volta. O sistema judiciário brasileiro tem um enorme número de processos e servidores limitados; a IA pode tornar o trabalho mais rápido e eficiente, trazendo um grande ganho para a sociedade, que terá suas decisões judiciais resolvidas mais rapidamente.

O cuidado que se deve ter é que nenhuma IA, hoje, pode realizar uma análise sem revisão humana. Um LLM brasileiro treinado com base na jurisprudência brasileira pode reduzir significativamente os erros, mas os vieses que já existem nas bases de dados vão continuar e precisam ser analisados por um ser humano. Agora, milhares de atividades operacionais podem ser otimizadas com IA.

A IA erra, mas o ser humano também erra; comprovadamente, uma IA bem treinada pode reduzir significativamente os erros operacionais na análise de processos judiciais e ajudar na análise de milhares de dados, criando padrões entre eles e realizando avaliações mais rápidas e eficazes do que qualquer ser humano. Isso pode levar a uma tomada de decisão mais precisa e consistente. Além disso, o robô pode trabalhar 24 horas por dia em tarefas repetitivas. Assim, o ser humano pode ser utilizado em outras atividades mais relevantes e na revisão das respostas da IA.

A IA não vai substituir o julgador; será uma ferramenta para auxiliar nas tarefas repetitivas e proporcionar uma justiça mais rápida, que atenda melhor à sociedade, com um modelo brasileiro muito mais eficiente.

Conclui-se, portanto, que o DeepSeek não é apenas um sucesso técnico, mas também um símbolo de viabilidade e democratização tecnológica que pode inspirar a construção de um

LLM jurídico brasileiro — público, ético e soberano — alinhado às diretrizes do Plano Brasileiro de Inteligência Artificial (PBIA) e da Resolução CNJ n. 615/2025. O futuro da inteligência artificial no Judiciário dependerá da capacidade de converter potencial tecnológico em política pública, transformando a automação em instrumento de justiça.

REFERÊNCIAS

ALKAMEL, Sayed Ali. **DeepSeek e o poder da mistura de especialistas (MoE)**. Disponível em: https://dev-to.translate.goog/sayed_ali_alkamel/deepseek-and-the-power-of-mixture-of-experts-moe-ham?_x_tr_sl=en&_x_tr_tl=pt&_x_tr_hl=pt&_x_tr_pto=tc. Acesso em: 22 jun. 2025.

AGÊNCIA GOV. Ministério da Ciência, Tecnologia e Informação. **Brasil quer acelerar a revolução da inteligência artificial com uma estratégia**. Disponível em: https://agenciagov.ebc.com.br/noticias/202501/brasil-quer-acelerar-revolucao-da-inteligencia-artificial-com-estrategia-do-mcti. Acesso em: 10 abr. 2025.

ASHOORI, Maryam. **O Llama 4 Maverick e o Llama 4 Scout da Meta agora estão disponíveis no watsonx.ai**. 2025. Disponível em: https://www.ibm.com/br-pt/new/announcements/Meta-llama-4-maverick-and-llama-4-scout-now-available-in-watsonx-ai. Acesso em: 14 set. 2025.

BERGMANN, Dave. **O que é combinação de especialistas?** Disponível em: https://www.ibm.com/br-pt/think/topics/mixture-of-experts. 2024. Acesso em: 14 set. 2025.

BONAT, Débora; VALE, Luís Manoel Borges do; PEREIRA, João Sergio dos Santos Soares. **Inteligência Artificial Generativa e a Fundamentação da Decisão Judicial**. Revista de Processo, v. 346, 2023. Revista dos Tribunais Online.

BOON, Yi Di; JOSHI, Sunil C; BHUDOLIA, Somen Kumar Bhudoli; GOREL, Goram Gohel. **Recent Advances on the Design Automation for Performance-Optimized Fiber Reinforced Polymer Composite Components**. 2020. Disponível em: https://www.researchgate.net/figure/Relationship-between-AI-machine-learning-and-deep-learning_fig3_341844335. Acesso em 07 abril de 2025.

CARNEIRO, Mariana Telles de Oliveira; ARAÚJO, André Silva; FRANÇA, Mardoqueu Geraldo Lima. **Revista Foco**, v. 18, n. 6, e 8774, p. 01-30, 2025.

COELHO, Alexandre Zavaglia; BARBOSA, Maria Juliana do P. **Inteligência Artificial Aplicada aos Serviços Jurídicos**. São Paulo: Thomson Reuters Brasil, 2024.

COMPRASNET. **Portal de Compras do Governo Federal. Pregão Eletrônico 90016/2025**. Disponível em: https://cnetmobile.estaleiro.serpro.gov.br/comprasnet-web/public/compras/acompanhamento-compra?compra=04000105900162025. Acesso em: 02 set. 2025.

CONSELHO NACIONAL DE JUSTIÇA (CNJ). **Uso de IA no Judiciário cresceu 26% em relação a 2022, aponta pesquisa**. 2022. Disponível em: https://www.cnj.jus.br/uso-de-ia-no-judiciario-cresceu-26-em-relacao-a-2022-aponta-pesquisa/. Acesso em: 07 abr. 2025.

CONSELHO NACIONAL DE JUSTIÇA (CNJ). **Painel de Pesquisa sobre Inteligência Artificial 2023. Plataforma Sinapses**. 2023. Disponível em:

https://paineisanalytics.cnj.jus.br/single/?appid=43bd4f8a-3c8f-49e7-931f-52b789b933c4&sheet=e4072450-982c-48ff-9e2d-361658b99233. Acesso em: 13 abr. 2025.

CONSELHO NACIONAL DE JUSTIÇA (CNJ). **Ferramenta de IA desenvolvida pela Justiça goiana reduz o tempo de tramitação processual**. 2024a. Disponível em: https://www.cnj.jus.br/ferramenta-de-ia-desenvolvida-pela-justica-goiana-reduz-o-tempo-de-tramitacao-processual/. Acesso em: 24 ago. 2025.

CONSELHO NACIONAL DE JUSTIÇA (CNJ). **Justiça em Números 2024: Barroso destaca aumento de 9,5% no número de novos processos**. 2024b. Disponível em: https://www.cnj.jus.br/justica-em-numeros-2024-barroso-destaca-aumento-de-95-em-novos-processos/. Acesso em: 24 ago. 2025.

CONSELHO NACIONAL DE JUSTIÇA (CNJ). **Painel que apresenta os resultados do levantamento de 2024 do Conselho Nacional de Justiça (CNJ) sobre projetos de Inteligência Artificial (IA) no Poder Judiciário**. 2024c. Disponível em: https://paineisanalytics.cnj.jus.br/single/?appid=51977be5-96d0-4362-98ff-ed3eb3337781. Acesso em: 26 set. 2025.

CONSELHO NACIONAL DE JUSTIÇA (CNJ). **Relatório de Pesquisa: O uso de Inteligência Artificial Generativa no Poder Judiciário Brasileiro**. 2024d. Disponível em: https://www.cnj.jus.br/wp-content/uploads/2024/09/cnj-relatorio-de-pesquisa-iag-pj.pdf. Acesso em: 28 set. 2025.

CONSELHO NACIONAL DE JUSTIÇA (CNJ). **Tribunais de todo o país já podem utilizar a primeira IA generativa integrada à PDPJ-Br**. 2025. Disponível em: https://www.cnj.jus.br/tribunais-de-todo-o-pais-ja-podem-utilizar-primeira-ia-generativa-integrada-a-pdpj-br/. Acesso em: 24 ago. 2025.

DATABRICKS. **O livro completo de IA Generativa**. 2024.

DEEPSEEK-V3. **Technical Report**. 2025. Disponível em: https://arxiv.org/abs/2412.19437. Acesso em: 25 jul. 2025.

DICKSON, Ben. **Under the hood: The Innovations powering DeepSeek's AI breakthrough**. 2025. Disponível em: https://bdtechtalks.com/2025/04/07/deepseek-innovations/. Acesso em: 22 set. 2025.

DOYLE, Stephanie. **IA 101: Treinamento vs. Inferência**. Disponível em: https://www-backblaze-com.translate.goog/blog/ai-101-training-vs-inference/. Acesso em: 07 set. 2025.

ENGEL, C.; LINHARDT, L. & SCHUBERT, M. **Code is law: how COMPAS affects the way the judiciary handles the risk of recidivism.** ***Artif Intell Law*** **33**, 383–404 (2025). https://doi.org/10.1007/s10506-024-09389-8

ERICKSON, Jeffrey. **O que é a inferência de IA?** Disponível em: https://www.oracle.com/br/artificial-intelligence/ai-inference/. Acesso em: 08 set. 2025.

FIGUEIREDO, Silva Guilherme. **Projeto Athos: Um Estudo de Caso sobre a inserção do Superior Tribunal de Justiça na Era da Inteligência Artificial**. 2022.

FORD, Leon. **The Rising Costs of Training Large Language Models (LLMs)**. 2024. Disponível em: https://www.layerstack.com/blog/the-rising-costs-of-training-large-language-models-llms/. Acesso em: 08 set. 2025.

GABRIEL FILHO, Oscar. **Inteligência Artificial e Aprendizado de Máquina: Aspectos Teóricos e Aplicações**. São Paulo: Edgar Blucher Ltda., 2023.

HARTMANN PEIXOTO, Fabiano. **Direito e Inteligência Artificial**. Coleção Inteligência Artificial e Jurisdição. v. 2. DR.IA. Brasília, 2020a. Disponível em: www.dria.unb.br. DOI: 10.29327/521174.

HATMANN PEIXOTO, Fabiano. **Projeto Victor: Relato do Desenvolvimento da Inteligência Artificial na Repercussão Geral do Supremo Tribunal Federal**. Revista Brasileira de Inteligência Artificial e Direito, v. 1, n. 1, 2020b.

HERO, Sami. **Open Source vs. Proprietary LLMs: Arms Race for AI Leadership**. Disponível em: https://www.ellie.ai/blogs/open-source-vs-proprietary-llms-arms-race-for-ai-leadership. 2025. Acesso em: 22 set. 2025.

HUGGING FACE. **DeepSeek Models – quantizations and finetunes**. 2025. Disponível em: https://huggingface.co/models?search=deepseek. Acesso em: 10 ago. 2025.

JADIDI, Vahid. **The Impact of Artificial Intelligence on Judicial Decision-Making Processes**. AJMHSS, 2025; 1(4): Article Press.

KNEUSEL, Ronald T. **Como a Inteligência Artificial Funciona: da magia à ciência**. São Paulo: Novatec Editora Ltda., 2024.

KROGH, Anders; HERTZ, John. **A Simple Weight Decay Can Improve Generalization**. 1991. Disponível em: https://proceedings.neurips.cc/paper/1991/file/8eefcfdf5990e441f0fb6f3fad709e21-Paper.pdf. Acesso em: 14 set. 2025.

MAGALHÃES, Dimmy. **Entendendo o processo de fine-tuning com LoRA**. Disponível em: https://dimmymagalhaes.medium.com/entendo-o-processo-de-fine-tuning-com-lora-82c4945aff76. Acesso em: 19 set. 2025.

MATOS, Rodrigo. **Como instalar o DeepSeek localmente usando o Ollama e o LM Studio**. 2025. Disponível em: https://bitandsolder.com/deepseek-offline-guia-completo-para-rodar-no-windows-e-linux-2025/. Acesso em: 13 out. 2025.

MCCANDLESS, David; EVANS, Tom; BARTON, Paul. **A visualisation of major large-language models (LLMs), ranked by performance, using MMLU (Massive Multitasks Language Understanding), a benchmark for evaluating the capabilities of large language models.** Disponível em: https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt/. Acesso em 18 abril 2025.

MINISTÉRIO DO PLANEJAMENTO E ORÇAMENTO (MPO). **Painel do Orçamento Federal. Despesas de TI. 2025.** Disponível em: https://www1.siop.planejamento.gov.br/QvAJAXZfc/opendoc.htm?document=IAS%2FExecucao_Orcamentaria.qvw&host=QVS%40pqlk04&anonymous=true?lang=en-US&opendocqs=. Acesso em: 24 ago. 2025.

NÚCLEO DE INTELIGÊNCIA ARTIFICIAL DO STF (NIAC). **Recebimento de e-mail com as informações sobre as inteligências artificiais do STF.** 20 abril 2025.

NUNES, Dierle. **Evento da Comissão de Inteligência Artificial no Direito da OAB/MG, presidida por Dierle Nunes, recebendo Júlio Luz Sisson de Castro - Supervisor do Núcleo de Gerenciamento de Precedentes do Supremo Tribunal Federal.** Inteligência artificial nos Tribunais Superiores - STF e os sistemas RAFA 2030 e VitorIA no STF. Disponível em: https://www.youtube.com/watch?v=6ZwaxelHgco&t=1404s. Acesso em: 17 abril 2025.

PÁDUA, Sérgio Rodrigo de; Lorenzetto, Bruno Meneses. **Direito Fundamental à Explicabilidade da Inteligência Artificial Utilizada em Decisões Estatais.** Revista da AGU - Brasília-DF - v. 23 - n. 02 - jun/2024.

PARSADANYAN, Eduard; DMITRIEVNA, Yulia. **The 11 best open-source LLMs for 2025.** 2025. Disponível em: https://blog.n8n.io/open-source-llm/. Acesso em: 28 set. 2025.

POOLE, Richard; OHIRI, Emmanuel. **What is the cost of training large language models?** Disponível em: https://www.cudocompute.com/blog/what-is-the-cost-of-training-large-language-models?utm_source=chatgpt.com. 2025. Acesso em: 7 set. 2025.

PORTO, Fábio Ribeiro; ARAÚJO, Valter Shyunquener; Gabriel, Anderson de Paiva. **Inteligência Artificial Generativa no Direito: um guia de como usar os sistemas (ChatGPT, Google Gemini, Claude, Mistral e Bing) na prática jurídica.** São Paulo, SP: Thomson Reuters Brasil; 2024.

IBM. O que é Random Forest? Disponível em: https://www.ibm.com/br-pt/think/topics/random-forest. Acesso em: 14 out. 2025.

INFORMATION IS BEAUTIFUL. **Major Large Language Models (LLMs) ranked by capabilities, sized by billion parameters used for training.** https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt/. Acesso em: 7 abril 2025.

REDHAT. **O que é inferência de IA?** Disponível em: https://www.redhat.com/pt-br/topics/ai/what-is-ai-inference. 2025. Acesso em: 7 set. 2025.

REUTERS. **Latin American countries to launch their own AI model in September.** 2025. Disponível em: https://www.reuters.com/world/americas/latin-american-countries-launch-own-ai-model-september-2025-06-17/. Acesso em: 24 ago. 2024.

RHYMESAI. **Aria: First Open Multimodal Native MoE Model.** 2024. Disponível em: https://huggingface.co/blog/RhymesAI/aria?utm_source=chatgpt.com. Acesso em: 14 set. 2025.

SALOMÃO, Luis Felipe; TAUK, Caroline Somesom et al. **Inteligência Artificial: tecnologia aplicada à gestão de conflitos no âmbito do Poder Judiciário brasileiro.** 3ª ed. Rio de Janeiro: FGV, 2023a.

FGV. **Inteligência artificial [livro eletrônico]: tecnologia aplicada à gestão dos conflitos no âmbito do Poder Judiciário brasileiro** / coordenação Luis Felipe Salomão, Elton Leme, Dierle Nunes. -- 4. ed. -- Rio de Janeiro : Fundação Getulio Vargas, 2025. PDF

SALOMÃO, Luis Felipe; TAUK, Caroline Somesom. **Inteligência Artificial no Judiciário Brasileiro: Estudo empírico sobre algoritmos e discriminação.** Revista Jurídica. https://doi.org/10.36113/dike.23.2023.3819. 2023b.

SENADO. **Projeto de Lei n. 2338, de 2023**. Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/157233/ . Acesso em: 14 set. 2025.

SHABBIR, Jahanzaib; ANWER, Tarique. **Artificial Intelligence and its Role in Near Future.** Journal of Latex Class Files, Vol. 14, n. 8, august 2015. Disponível em https://arxiv.org/pdf/1804.01396.pdf. Acesso em: 7 abril 2025.

SHORTEN, C.; KHOSHGOFTAAR, T.M. **A survey on Image Data Augmentation for Deep Learning.** 2019. Disponível em: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0197-0#citeas. Acesso em: 14 set. 2025.

SILVA, E. C. M.; ROCHA, I.; VAZ, J. C.; VENEZIANI, J. R. A.; MODANEZ, C. C.. **Contratos, Códigos e Controle: A Influência das Big Techs no Estado Brasileiro.** São Paulo - SP, Brasil, jul. 2025. Disponível em: https://bit.ly/contratos-big-techs. Acesso em: 13 out. 2025.

SRIVASTAVA, Nitish; HINTON, Geoffrey; KRIZHEVSKY, Alex; SUTKESVER, Ilya; SALAKHUTDINOV, Ruslan. **Dropout: A Simple Way to Prevent Neural Networks from Overfitting.** 2014. Disponível em: https://jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf. Acesso em: 14 set. 2025.

STANFORD. **Artificial Inteligence Index Report 2025.** https://hai-production.s3.amazonaws.com/files/hai_ai_index_report_2025.pdf. Acesso em: 10 abril 2025.

SUPERIOR TRIBUNAL DE JUSTIÇA (STJ). **Revolução tecnológica e desafios da pandemia marcaram gestão do ministro Noronha na presidência do STJ.** 2020. Disponível em: https://www.stj.jus.br/sites/portalp/Paginas/Comunicacao/Noticias/23082020-Revolucao-tecnologica-e-desafios-da-pandemia-marcaram-gestao-do-ministro-Noronha-na-presidencia-do-STJ.aspx. Acesso em: 24 ago. 2025.

SUPERIOR TRIBUNAL DE JUSTIÇA (STJ). **STJ lança novo motor de inteligência artificial generativa para aumentar a eficiência na produção de decisões.** 2025. Disponível em: https://www.stj.jus.br/sites/portalp/Paginas/Comunicacao/Noticias/2025/11022025-STJ-lanca-novo-motor-de-inteligencia-artificial-generativa-para-aumentar-eficiencia-na-producao-de-decisoes.aspx. Acesso em: 24 ago. 2025.

SUPREMO TRIBUNAL FEDERAL (STF). **Projeto VICTOR do STF é apresentado em congresso internacional sobre tecnologia.** 2018. Disponível em STF: https://portal.stf.jus.br/noticias/verNoticiaDetalhe.asp?idConteudo=390818. Acesso em: 15 abril 2025.

SUPREMO TRIBUNAL FEDERAL (STF). **STF Digital: nova plataforma integra sistemas e aprimora a prestação jurisdicional.** 31 de agosto de 2020. Disponível: https://portal.stf.jus.br/noticias/verNoticiaDetalhe.asp?idConteudo=450698&ori=1. Acesso em: 19 abril 2025.

SUPREMO TRIBUNAL FEDERAL (STF). **Inteligência artificial permitirá a classificação dos processos do STF sob a ótica dos direitos humanos.** 2022. Disponível em: https://portal.stf.jus.br/noticias/verNoticiaDetalhe.asp?idConteudo=487134&ori=1. Acesso em: 19 abril 2025.

SUPREMO TRIBUNAL FEDERAL (STF). **STF apresenta como funciona o aplicativo da nova ferramenta de Inteligência Artificial – VitórIA**. 17 de maio de 2023. Disponível em: https://www.youtube.com/watch?v=xHoi0FMOvK8. Acesso em: 19 abril 2025.

SUPREMO TRIBUNAL FEDERAL (STF). **STF apresenta a nova ferramenta de Inteligência Artificial – VitórIA.** 17 de maio de 2023. Disponível em: https://www.youtube.com/watch?v=xuw1U1OredQ. Acesso em: 19 abril 2025.

SUPREMO TRIBUNAL FEDERAL (STF). **Ministra Rosa Weber lança robô VitórIA para agrupamento e classificação de processos.** 17 de maio de 2023. Disponível em: https://noticias.stf.jus.br/postsnoticias/ministra-rosa-weber-lanca-robo-vitoria-para-agrupamento-e-classificacao-de-processos/. Acesso em: 19 abril 2025.

SUPREMO TRIBUNAL FEDERAL. **Inteligência Artificial e Justiça. Relatório Geral Chamamento Público** 001/2023. 2024a. https://www.stf.jus.br/arquivo/cms/noticiaNoticiaStf/anexo/RELATORIOCHAMAMENTO.INTELIGENCIA.ARTIFICIAL.pdf. Acesso em: 16 de abril de 2025.

SUPREMO TRIBUNAL FEDERAL (STF). **STF lança MARIA, ferramenta de inteligência artificial que dará mais agilidade aos serviços do Tribunal.** 16 de dezembro de 2024. 2024b. 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: 18/04/2025.

SUPREMO TRIBUNAL FEDERAL. **STF amplia o uso de inteligência artificial em apoio à atividade jurisdicional.** 2025. Disponível em:

https://noticias.stf.jus.br/postsnoticias/stf-amplia-uso-de-inteligencia-artificial-em-apoio-a-atividade-jurisdicional/. Acesso em: 28 set. 2025.

SUJATHA R. **GPU Options for Finetuning Large Models: Choose the Right Setup. 2025.** Disponível em: https://www.digitalocean.com/resources/articles/gpu-options-finetuning. Acesso em 28 set. 2025.

SYAL, Anirudh. **RAG vs. LLM: Understanding the Difference and Synergy.** 2024. Disponível em: https://medium.com/%40anirudhsyal/rag-vs-llm-understanding-the-difference-and-synergyintroduction-the-magic-of-retrieval-augmented-3e4bd33f2465. Acesso em: 7 set. 2025.

TAULLI, Tom. **Introdução à Inteligência Artificial: uma abordagem não técnica.** São Paulo, SP: Novatec Editora Ltda.; 2020.

TOLEDO, Cláudia; PESSOA, Daniel. **O uso de inteligência artificial na tomada de decisão judicial. Revista de Investigações Constitucionais.** ISSN 2359-5639. D DOI: 10.5380/rinc.v10i1.86319.

VARGAS, Eduardo. **Revista Isto é Dinheiro: 'ChatGPT da China' abre portas e IA não substituirá ser humano, diz diretor da Nvidia**. Disponível em: https://istoedinheiro.com.br/nvidia-deepseek-china-dinheiro-entrevista/. Acesso em: 10 abril 2025.

WOLFE, Cameron. R. **Mixture-of-Experts (MoE) LLMs.** 2025. Disponível em: https://cameronrwolfe.substack.com/p/moe-llms?utm_campaign=post&utm_medium=web&triedRedirect=true. Acesso em: 22 set. 2025.

ZHANG, Jinpeng. **DeepSeek Technical Analysis — (1) Mixture-of-Experts.** Disponível em: https://dataturbo.medium.com/key-techniques-behind-deepseek-models-10x-efficiency-1-moe-9bd2534987c8. 2025. Acesso em: 22 set. 2025.