RADAR SEMANAL · IA PARA ESCRITÓRIOS, GABINETES E ASSESSORIAS

IA em Cognição Nada Exauriente

LAVRADO EM 30 DE MAIO DE 2026, COM EFEITOS RETROATIVOS A 23 DE MAIO

Resumo

A semana girou em torno de um lançamento e de uma mudança de paradigma que vieram juntos. A **Anthropic publicou o Claude Opus 4.8** (28/mai), uma atualização que ficou quatro vezes menos propensa a deixar erros de código passarem sem aviso, e o acompanhou dos Dynamic Workflows — um recurso que faz um único pedido se desdobrar em centenas de agentes trabalhando em paralelo. A **NVIDIA abriu a família Nemotron Diffusion**, modelos que escrevem texto de um jeito diferente do usual e chegam a ser seis vezes mais rápidos. E a **Darrow estreou uma plataforma** que trata a carteira de litígios de um escritório como uma carteira de investimentos. No pano de fundo, papers de fronteira sobre alinhamento e segurança de agentes, uma leva de ferramentas de infraestrutura e, no Brasil, o governo federal abriu os números do projeto INSPIRE de IA no serviço público.

MODELO PROPRIETÁRIO

Claude Opus 4.8 e os Dynamic Workflows

A Anthropic lançou o Claude Opus 4.8 quarenta e um dias depois do 4.7 (anúncio oficial — Anthropic). A janela de contexto — a quantidade de texto que o modelo consegue ler de uma vez — é de 1 milhão de tokens, algo como milhares de páginas; a saída chega a 128 mil tokens (análise de Simon Willison). O preço de entrada seguiu em US$ 5 por milhão de tokens, e o modo rápido ficou bem mais barato — passou a custar o dobro da tarifa padrão, contra o múltiplo bem maior das gerações anteriores.

O ganho central não é velocidade, é honestidade. O modelo é **cerca de quatro vezes menos propenso** a deixar uma falha no código que ele mesmo escreveu passar sem comentar — e, em vez de chutar, prefere admitir quando não sabe. Para quem delega trabalho a uma máquina, a diferença entre um assistente que esconde a dúvida e um que a declara é a diferença entre revisar por cima e revisar no susto.

A Anthropic não detalhou esses números no anúncio oficial; os resultados abaixo vêm da cobertura do Latent Space, que tabulou os benchmarks da semana.

BENCHMARK	SCORE	O QUE MEDE
SWE-Bench Pro	69,2%	Resolução autônoma de tarefas reais de engenharia de software; +10 pontos sobre o GPT-5.5.
APEX-SWE (Pass@1)	45,3%	Acerto na primeira tentativa em tarefas de programação de alta dificuldade.
GDPval-AA	1890 Elo	Pontuação de capacidade econômica geral; +137 sobre o Opus 4.7.
AA Intelligence Index	61,4	Índice agregado de inteligência; +4,1 sobre o Opus 4.7.

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

O salto sobre o Opus 4.7 é incremental em capacidade bruta, mas a Anthropic priorizou comportamento: menos afirmações sem lastro, mais sinalização espontânea de problemas nos dados de entrada e de saída. Willison classificou como *melhoria modesta, porém tangível* — e ninguém entende melhor a diferença entre modesto e tangível do que quem revisa contrato alheio para viver. O que de fato muda a régua é o que veio junto, no **Claude Code** (release no GitHub).

O QUE ISSO DESBLOQUEIA NA PRÁTICA

Os **Dynamic Workflows** são o ponto de virada. Em vez de um agente — um programa de IA que executa uma tarefa de várias etapas sozinho — resolver tudo em sequência, o Claude escreve na hora um roteiro de orquestração e dispara dezenas a centenas de subagentes em paralelo, cada um cuidando de um pedaço, com uma etapa de conferência antes de devolver o resultado. Na demonstração, a base de código do projeto Bun foi migrada de uma linguagem (Zig) para outra (Rust) — 750 mil linhas, do primeiro commit ao merge, em 11 dias, com 99,8% dos testes passando (cobertura — Latent Space). Pense num escritório que precisa revisar dez mil contratos para checar uma única cláusula de foro: o trabalho que antes era enfileirado vira um enxame que ataca tudo ao mesmo tempo, com um revisor de segunda leitura embutido.

> *Opus 4.8 é aproximadamente quatro vezes menos propenso que seu antecessor a deixar falhas em código passarem sem comentário.*
>
> ANTHROPIC, CITADO POR SIMON WILLISON (28/MAI/2026)

IMPACTO PRÁTICO

Há uma ironia útil aqui. O recurso que mais impressiona — soltar um exército de agentes — só vale alguma coisa porque o recurso mais discreto — o modelo aprendeu a dizer "não sei" — veio junto. Cem agentes confiantes e errados produzem cem vezes mais estrago; cem agentes que sinalizam a própria dúvida produzem cem oportunidades de corrigir antes do dano. Para a operação jurídica, a lição é velha: a máquina rápida só é segura quando é honesta. Velocidade sem revisão é petição protocolada com o nome da parte errada — só que multiplicada por cem.

MODELO OPEN-SOURCE

NVIDIA Nemotron Diffusion (3B / 8B / 14B)

A NVIDIA liberou a família Nemotron Diffusion, modelos de linguagem abertos que geram texto por difusão (HF blog — NVIDIA). A maioria dos modelos atuais escreve como quem fala: uma palavra de cada vez, da esquerda para a direita. Os modelos de difusão fazem diferente — preenchem blocos de 32 palavras de uma vez e vão refinando o rascunho em passadas sucessivas, como um escultor que parte de um bloco bruto e vai revelando a forma. A família tem versões de 3, 8 e 14 bilhões de parâmetros, mais um modelo que também enxerga imagens.

O mesmo modelo opera em três modos, e um deles entrega o ganho de velocidade sem perder qualidade na geração mais determinística. O atrativo é prático: gerar texto mais rápido no mesmo hardware significa custo menor por documento produzido.

BENCHMARK	SCORE	O QUE MEDE
Velocidade (Self-Speculation)	**~6×**	Aceleração sobre o modelo autoregressivo equivalente, sem perda de qualidade em temperatura zero.
Vazão em hardware B200	**~865 tok/s**	Tokens por segundo no modo mais agressivo.
Acurácia média vs. Qwen3 8B	**+1,2%**	Ganho de qualidade médio frente a um modelo aberto de referência da mesma faixa.

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

Modelos de difusão para texto não são novidade acadêmica, mas costumavam pagar a velocidade com qualidade pior. A aposta da NVIDIA é treinar um único modelo que faz as duas coisas — escreve à moda antiga quando convém e por difusão quando dá para acelerar — convertendo o método tradicional em modelo de difusão sem começar do zero. O resultado fica acima de um par aberto da mesma faixa de tamanho, em vez de abaixo.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

Para quem roda IA dentro de casa — caso de estruturas jurídicas com dado sensível que não pode sair do servidor próprio —, modelo aberto e rápido na faixa de 3 a 14 bilhões de parâmetros é o que cabe em hardware modesto sem virar fatura impagável de nuvem. Não é o modelo que vence o benchmark de manchete; é o que roda no computador que a estrutura já tem, sob a licença que permite uso interno sem pedir licença a ninguém.

> **IMPACTO PRÁTICO**
>
> A corrida de IA tem dois trilhos, e eles raramente aparecem na mesma manchete. Um é o do modelo gigante que quebra recordes; o outro, mais silencioso, é o do modelo pequeno que fica bom o bastante para rodar onde os dados precisam ficar. Para o sigilo profissional, o segundo trilho importa mais que o primeiro. De que adianta o modelo mais inteligente do mundo se o preço de usá-lo é mandar o processo do cliente para o servidor de outro?

FERRAMENTA

Darrow: litígio como carteira de investimentos

A Darrow estreou uma plataforma que usa agentes de IA para vasculhar dados públicos continuamente e mapear onde há exposição a litígio, para que escritórios de autores administrem seus casos como quem administra uma carteira de investimentos, com análise de risco (LawNext). O sistema trabalha sobre uma taxonomia própria de 164 "fraquezas legais" em campos como privacidade, concorrência, consumidor e ambiental, e compara cada caso novo contra um banco de mais de 220 precedentes parecidos.

Os números que a empresa divulga dão a dimensão da aposta: no ano anterior, a plataforma apontou **US$ 10,3 bilhões em exposição**, dos quais 77% viraram ações reais em até um ano. A camada de conversa permite ao advogado interrogar o próprio acervo — "quais casos de privacidade têm precedente favorável neste circuito?" — em linguagem natural.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

A novidade não é a IA achar um caso; é a IA transformar a seleção de causas numa decisão de portfólio, com risco estimado e precedente na mão antes do protocolo. Para a banca de contencioso de massa, isso muda a pergunta inicial: deixa de ser "este caso tem mérito?" e passa a ser "este caso, dado o que já temos, melhora ou piora a carteira?". É a lógica do gestor de fundos aplicada à mesa do litigante.

> **IMPACTO PRÁTICO**
>
> Vale o aviso de sempre, porque ele nunca envelhece: a máquina mede exposição, não justiça. Um banco de 220 precedentes diz onde a vitória é provável, não onde a causa é digna — e confundir as duas coisas é como escolher cliente por CEP. A ferramenta é poderosa para quem já sabe o que está fazendo e perigosa para quem terceiriza o juízo. O critério continua sendo humano; a IA só chega mais rápido à porta. Quem abre, e por quê, ainda é o advogado.

Pesquisa e métodos

Foi uma semana forte em segurança e alinhamento de agentes. O Google DeepMind publicou avaliações do tipo "honeypot" — armadilhas montadas em repositórios reais de pesquisa — e concluiu que os modelos Gemini não tramam por conta própria; só ensaiam sabotagem quando o prompt explicitamente os empurra para agir com autonomia ou lhes planta uma meta oculta (arXiv — DeepMind). Na direção oposta da euforia com enxames de agentes, pesquisadores de Berkeley formalizaram que o desempenho de um agente depende de seis peças do sistema, não só do modelo — o *harness*, o arcabouço que executa o modelo, pesa tanto quanto o cérebro (arXiv — Berkeley).

Dois resultados merecem o olho de quem trabalha com pesquisa assistida. Um paper mostrou como envenenar silenciosamente um sistema RAG — a técnica que combina o modelo com uma busca em base de documentos própria — inserindo textos que parecem legítimos mas sequestram a resposta, com apenas 0,016% da base contaminada (arXiv — SilentRetrieval); recado direto para quem monta repositório de jurisprudência para a IA consultar. E um método de converter modelos MoE — arquitetura que ativa só uma fração dos seus "especialistas" a cada palavra — de volta em modelos densos comuns abriu caminho para rodar em máquina com pouca memória, com ganho de qualidade sobre a poda tradicional (arXiv — Pruning MoE).

Ferramentas do ecossistema

- **vLLM 0.22.0:** o motor de inferência mais usado para servir modelos abertos ganhou cache de contexto em múltiplas camadas com descarga para disco, o que afrouxa o teto de memória ao rodar modelos grandes, e 28,9% menos latência num modo de cálculo específico (release).
- **langchain 1.3.2:** o framework de orquestração de LLM passou a redigir dados pessoais (PII) no meio do fluxo de resposta, sem esperar o texto completo — útil para quem precisa filtrar nome e CPF antes de a resposta sair (release).

- **llama.cpp (b9414):** o runtime que roda modelos abertos em hardware comum passou a suportar o DeepSeek-OCR 2, modelo de leitura de documentos com resolução adaptável — útil para extrair texto de digitalizações de qualidade irregular, o pão de cada dia de quem lida com autos escaneados (release).
- **Agente fiscal auto-corrigível (OpenAI + Thrive):** um sistema que processou 7.000 declarações de imposto usa as correções dos contadores como combustível para se reescrever — o acerto de preenchimento subiu de 25% para 86% em seis semanas, exemplo concreto de agente que melhora sozinho com supervisão humana (OpenAI).

Análise

Demis Hassabis, da DeepMind, antecipou seu palpite para a chegada da AGI — a IA de capacidade geral comparável à humana — de 2030 para a janela de 2029 a 2030, e descreveu o momento atual dos agentes como um "ensaio geral" para sistemas bem mais potentes, listando quatro lacunas ainda abertas: modelos de mundo, memória, consistência e aprendizado contínuo (Axios). Do lado de quem constrói, Boris Cherny, da Anthropic, contou que o Claude Code é escrito 100% por ele mesmo há mais de seis meses, e que hoje roda milhares de agentes por noite, contra os 30 segundos de autonomia de 18 meses atrás (Platformer).

Mas o texto mais útil para o profissional do direito veio de Ethan Mollick, revisando quatro estudos sobre IA e aprendizado (One Useful Thing). A conclusão é desconfortável: IA usada como dadora de respostas piora o desempenho de quem a usa; usada como tutora que puxa o raciocínio, melhora. Num dos estudos, programadores que delegaram tudo à máquina depois "não sabiam responder o que tinham feito". O recado para quem terceiriza a fundamentação da própria petição é direto — a muleta que carrega o peso também atrofia o músculo.

Brasil

O governo federal abriu os números do **INSPIRE**, projeto de R$ 390 milhões de IA no serviço público coordenado pelo CPQD, após sete meses (gov.br). O destaque é o Chat GOV.BR, atendimento por IA generativa que resolveu sozinho mais de 60% das demandas em teste, cobrindo 22 políticas públicas; nos bastidores, a qualificação automática de 77 milhões de registros de endereços e a auditoria do catálogo nacional de dados, que caiu de sete horas para uma.

Na mesma semana, a CGU anunciou o **Informa.BR**, plataforma que deixa o cidadão consultar em linguagem natural o Portal da Transparência e os pedidos já respondidos pela Lei de Acesso à Informação, com lançamento previsto para junho (Agência Brasil). E o Itaú, com a CI&T, abriu

sob licença MIT o **Business Complexity Points**, ferramenta de IA generativa que mede a complexidade de um software lendo as histórias de usuário e pontuando regras de negócio, interface e fronteiras de dados, com 83% de acerto na classificação (FIDC News).

> **IMPACTO PRÁTICO**
>
> Repare no padrão da semana brasileira: nenhum modelo de fronteira, mas três aplicações que efetivamente entregam número. Enquanto boa parte do mercado ainda discute se vai adotar IA, o setor público federal publicou métricas de resolutividade e um banco privado abriu o código de uma ferramenta que usa de verdade. É o avesso do palco de costume — e é exatamente onde o operador do direito deveria olhar. O Chat GOV.BR atendendo 60% das demandas sozinho não é promessa de palestra; é a antecâmara do balcão que sua estrutura vai encontrar do outro lado, no protocolo, na intimação, na resposta automática. A pergunta deixou de ser se a IA chega ao cotidiano jurídico. Ela já está atendendo no guichê.

O recorte dos últimos sete dias termina aqui. Até a próxima! ■

Dennis Martins · Adriana Aneli · Fábio Rodrigues Fazuoli

CURADORIA E EDIÇÃO · 30 DE MAIO DE 2026