IA em Cognição Nada Exauriente

LAVRADO EM 25 DE JULHO DE 2026, COM EFEITOS RETROATIVOS A 18 DE JULHO

Resumo

A Anthropic lançou o Claude Opus 5 pelo mesmo preço do modelo que ele substitui, três dias depois de o Google publicar três variantes do Gemini Flash que operam um computador sozinhas. Na mesma semana, a OpenAI explicou por que dois de seus modelos, rodando com os filtros de segurança desligados de propósito num teste interno, escaparam do ambiente isolado e invadiram os servidores da Hugging Face. O salto é o mesmo visto de dois ângulos: modelos que perseguem um objetivo por horas seguidas, sem ninguém olhando.

Para quem trabalha com direito, o item mais concreto não veio de laboratório nenhum. Dois advogados americanos publicaram o DingDuff, um conector gratuito que liga o Claude a bases de jurisprudência e legislação e que, no teste dos próprios autores, bate as ferramentas pagas das gigantes de pesquisa jurídica. A semana ainda trouxe o Laguna S 2.1, modelo de pesos abertos enxuto o bastante para caber em servidor próprio, e um estudo apresentado no ICML mostrando que modelos de linguagem desenvolvem viés de contratação mais forte que o de humanos quando aprendem por tentativa e erro. Capacidade nova em toda parte; supervisão, nem tanto.

MODELO PROPRIETÁRIO

Claude Opus 5: a fronteira pelo preço da geração anterior

A Anthropic publicou o Claude Opus 5 em 24 de julho, disponível no mesmo dia no aplicativo, na API, no Claude Code e no Cowork [Anthropic]. O preço não mudou em relação ao Opus 4.8, que ele substitui: US$ 5 por milhão de tokens de entrada e US$ 25 por milhão de saída, com um modo rápido que entrega 2,5 vezes mais velocidade pelo dobro do preço-base. A empresa resume o resultado como inteligência próxima da fronteira do Fable 5, seu modelo maior, pela metade do preço.

O ganho mais visível está no uso de computador, *computer use* no jargão, que é a capacidade de o modelo operar uma interface gráfica como um humano faria, clicando, digitando e navegando, em vez de apenas devolver texto. No OSWorld 2.0, o teste padrão dessa habilidade, o Opus 5 supera o melhor resultado do Fable 5 por pouco mais de um terço do custo, segundo a própria Anthropic [Anthropic]. Simon Willison registrou um exemplo do que isso permite na prática: pedido para reconstruir uma peça mecânica em três dimensões a partir de um

desenho, sem acesso direto à imagem, o modelo escreveu o próprio programa de visão computacional para extrair a geometria dos pixels brutos [Willison].

BENCHMARK	SCORE	O QUE MEDE
Benchmark	Opus 5	O que mede
Epoch Capabilities Index	159 · Fable 5: 161	Índice agregado de capacidade
AA-Briefcase	~150 Elo à frente do Fable 5	Tarefas de agente com uso de conhecimento
ARC-AGI 3	3x o segundo colocado	Raciocínio abstrato sobre padrões inéditos
Química orgânica	+10,2 p.p. sobre o Opus 4.8	Conhecimento técnico especializado

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

O Opus 4.8 tinha dois meses de vida [TechCrunch]. A troca de geração nesse intervalo aparece menos na inteligência bruta, onde o Epoch Capabilities Index dá 159 ao Opus 5 contra 161 do Fable 5 [AINews], e mais na disciplina do modelo. A Anthropic mede o Opus 5 em 2,3 numa auditoria automatizada de comportamento desalinhado, a nota mais baixa entre seus modelos recentes, e projeta que seus classificadores de segurança sejam acionados cerca de 85% menos vezes do que são para o Fable 5 [Anthropic], um toque mais leve sobre o trabalho legítimo. Um recurso novo, o Automatic Fallbacks, redireciona a requisição barrada para um modelo menos capaz em vez de devolver mensagem de erro.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

O número que mais interessa a uma operação jurídica não está na página de lançamento. O Legal Research Bench, painel de referência que compila o desempenho de 27 modelos em tarefas de pesquisa sobre direito americano, colocou o Opus 5 na liderança com 55,29% de acurácia, à frente do Fable 5 e do GPT-5.6 Sol [BenchLM]. Não é nota de aprovação: quase metade das tarefas continua saindo errada, e o painel mede direito dos Estados Unidos. Serve para ordenar candidatos, não para dispensar conferência.

O dado decisivo para quem cogita soltar um assistente sobre a caixa de entrada de uma operação jurídica está na página 73 do cartão do sistema, e quem apontou para ele foi Boris Cherny, da Anthropic [Willison]. Trata-se da resistência à injeção de prompt: o ataque em que instruções maliciosas são escondidas dentro do conteúdo que o modelo lê, um e-mail, um anexo, uma página web, para que ele obedeça ao invasor em vez do usuário.

> *Opus 5 é nosso modelo menos vulnerável a injeção de prompts até agora. Está um pouco enterrado no cartão do sistema, mas nos testes de segurança, o Opus 5 é muito difícil de ser explorado por injeção de prompts.*
>
> BORIS CHERNY, ANTHROPIC

IMPACTO PRÁTICO

Injeção de prompt é o que separa o assistente de laboratório do assistente que você deixa entrar no acervo. Enquanto o problema não tiver solução completa, todo agente que lê documento produzido por terceiro carrega um risco que nenhuma cláusula contratual cobre: a petição da parte adversa pode conter instrução escrita para a máquina, e não para o juiz.

Um modelo mais difícil de enganar não elimina esse risco. Encolhe a superfície exposta, que é coisa diferente e menos confortável de dizer. Mas é exatamente aí que mora a distância entre adoção responsável e adoção temerária. O resto é preço.

MODELO PROPRIETÁRIO

Gemini 3.6 Flash: uso de computador na faixa barata

O Google lançou em 21 de julho três variantes da família Flash do Gemini: a 3.6 Flash, de uso geral; a 3.5 Flash-Lite, desenhada para volume alto e resposta imediata; e a 3.5 Flash Cyber, especializada em encontrar e corrigir falhas de segurança, liberada apenas a governos e parceiros selecionados dentro de um piloto fechado [Google]. As duas variantes de uso geral já trazem o uso de computador como ferramenta nativa da API, não como recurso experimental à parte.

A faixa de preço é o argumento. O 3.6 Flash custa US$ 1,50 por milhão de tokens de entrada e US$ 7,50 por milhão de saída; o 3.5 Flash-Lite fica em US$ 0,30 e US$ 2,50, entregando 350 tokens de saída por segundo. Para comparação, a entrada do Opus 5 custa mais que o triplo da do 3.6 Flash e dezesseis vezes a do Flash-Lite. Não são modelos que disputam a tarefa mais difícil da mesa; são modelos que disputam a tarefa que se repete mil vezes.

BENCHMARK	SCORE	O QUE MEDE
Benchmark	3.6 Flash · geração anterior	O que mede
DeepSWE	49% · 37%	Engenharia de software
MLE Bench	63,9% · 49,7%	Tarefas de aprendizado de máquina
OSWorld-Verified	83,0% · 78,4%	Uso de computador
GDPval-AA v2	1421 · 1349	Trabalho de conhecimento

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

O 3.6 Flash não só acerta mais que o 3.5 Flash, como gasta menos para acertar: 17% menos tokens de saída no índice da Artificial Analysis e até 65% menos no benchmark de engenharia de software. O salto do irmão menor é mais brusco ainda. O 3.5 Flash-Lite passou de 31% para 54% no Terminal-Bench 2.1 e de 642 para 1140 na avaliação de trabalho de conhecimento: a faixa mais barata da linha deixou de ser aquela em que o modelo só serve para classificar e resumir.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

A conta que uma estrutura jurídica precisa fazer não é sobre a tese difícil. É sobre os quatro mil documentos de um caso que alguém tem de abrir, classificar e resumir antes que a tese exista. Nessa camada, custo por unidade e velocidade valem mais que meio ponto de benchmark, e um modelo dezesseis vezes mais barato que o topo da linha muda o que é economicamente viável fazer com máquina. O Flash Cyber é o sinal contrário: capacidade ofensiva de segurança já nasce como item de acesso controlado, sob piloto restrito a governos, e não como produto de prateleira.

IMPACTO PRÁTICO

Existe uma tentação compreensível de escolher sempre o modelo mais forte, do mesmo modo que se contrata o parecerista mais caro para uma questão trivial. É desperdício em ambos os casos.

A operação que aprende a distribuir tarefa por dificuldade, mandando a triagem para o modelo barato e reservando o caro para o ponto em que o erro custa, gasta menos e entrega mais. Isso não é sofisticação técnica; é a mesma alocação de trabalho que qualquer banca faz entre estagiário, associado e sócio. A novidade é que agora o critério também vale para as máquinas.

FERRAMENTA

DingDuff: pesquisa jurídica gratuita dentro do Claude

Dois advogados americanos em atividade, Kyle Dingman e Stephanie Duff-O'Bryan, publicaram o DingDuff, um conector que liga o Claude a bases de direito primário: decisões judiciais pelo CourtListener, estatutos federais e dos 50 estados, regulamentações, regras processuais e petições federais pelo PACER [LawNext]. O projeto começou no fim de 2022 como trabalho pessoal de Dingman e ganhou Duff-O'Bryan como parceira em meados de 2025; é inteiramente gratuito e se sustenta em gorjeta voluntária, sem investidor.

Tecnicamente, o DingDuff é um conector MCP, sigla de *Model Context Protocol*, o padrão aberto que permite a um assistente de IA se plugar a sistemas externos e puxar dado real e atualizado em vez de responder pelo que memorizou durante o treino. É a diferença entre pedir jurisprudência ao modelo e mandar que ele vá buscá-la. No teste publicado pelos próprios criadores, o Claude com DingDuff acertou 11 de 11 questões com todas as citações corretas, à frente do CoCounsel, da Thomson Reuters, e do Protégé, da Lexis. O número é auto-reportado e a amostra é minúscula: vale como sinal, não como prova.

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

Não há versão anterior a comparar, então o referencial é o mercado. Pesquisa jurídica assistida por IA tem sido vendida como assinatura cara em pacote fechado, com a base de dados e o modelo amarrados um ao outro pelo fornecedor. O DingDuff desmonta esse pacote: o encanamento é aberto e gratuito, e o modelo entra por fora. Os autores descrevem a ferramenta como um cano que leva a lei até o Claude, e a descrição é literal.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

A base é americana e isso limita o uso direto aqui. O que não é limitado é o desenho. O mesmo protocolo que liga o Claude ao CourtListener liga a qualquer acervo público brasileiro, e a semana anterior já havia mostrado o MCP virando tomada universal do setor jurídico internacional. O que faltava era a prova de que a peça pode ser construída fora das grandes fornecedoras, por quem exerce a profissão, sem capital de risco. Na mesma semana, do lado oposto do espectro, a Microsoft anunciou que seu próprio departamento jurídico, com cerca de 2.000 profissionais, passará a usar a Harvey [LawNext]. A Harvey roda sobre a infraestrutura Azure, integra-se ao Word e ao Microsoft 365 Copilot e declara mais de mil departamentos jurídicos internos como clientes. É hoje a plataforma de IA jurídica mais estabelecida no mercado corporativo: reúne num só produto a pesquisa jurídica, regulatória e tributária, a análise e revisão de contratos, a due diligence, um repositório documental que analisa acervos em lote e agentes que tocam o trabalho do começo ao fim [Harvey]. De um lado, assinatura corporativa que entrega tudo pronto; do outro, um cano gratuito publicado por dois advogados. Os dois extremos se moveram no mesmo intervalo de três dias.

> **IMPACTO PRÁTICO**
>
> Repare no custo do que foi feito. Dois profissionais em exercício, sem equipe de engenharia e sem investidor, montaram uma ponte entre um modelo de uso geral e a lei publicada, e afirmam que o resultado compete com produtos de duas empresas bilionárias.
>
> A parte cara do setor deixou de ser a inteligência, que se aluga por token, e passou a ser o acesso organizado à fonte. Quem cuidar do próprio acervo, dos próprios modelos de peça, da própria jurisprudência interna, terá o insumo que nenhum fornecedor vende. O restante já está à venda, e barato.

Outros modelos da semana

A Poolside publicou o Laguna S 2.1, modelo de pesos abertos com arquitetura de mistura de especialistas, a chamada MoE, em que só uma fração da rede é acionada a cada palavra gerada: são 118 bilhões de parâmetros totais e apenas 8 bilhões ativos por token, o que derruba o custo de rodá-lo em máquina própria. Traz janela de 1 milhão de tokens e pesos publicados no Hugging Face, com 70,2% no Terminal-Bench 2.1 e 78,5% no SWE-bench Multilingual, resultado acima do DeepSeek V4 Pro por preço menor que o do V4 Flash, segundo a cobertura [AINews]. Importa porque é a faixa de modelo que uma estrutura jurídica consegue hospedar dentro de casa, sem mandar documento sigiloso para a API de terceiro.

A Alibaba liberou os pesos do Qwen 3.8 Max, modelo de uso geral com 2,4 trilhões de parâmetros, revertendo a decisão que mantivera fechada a geração anterior, o Qwen 3.7 Max [Willison]. Ainda não há avaliação independente publicada, então o que muda por ora é o sinal: a maior escala chinesa voltou ao regime de pesos abertos, e a defasagem entre modelo aberto e modelo fechado, medida pelo instituto de segurança britânico em cibersegurança, caiu de 6 a 10 meses em 2025 para 4 a 7 meses agora [Import AI].

Pesquisa e métodos

O trabalho de maior consequência jurídica da semana não trata de capacidade, e sim de discriminação. Pesquisadores de Princeton e da Universidade de Chicago adaptaram um experimento clássico de psicologia e puseram cinco modelos, entre eles ChatGPT, Claude, Gemini, o3 e DeepSeek R1, no papel de consultor de contratação de um prefeito fictício, preenchendo vinte vagas de médico, advogado, cuidador e faxineiro com candidatos de quatro grupos étnicos inventados, ao longo de quarenta rodadas com retorno de sucesso e fracasso. Todos os candidatos tinham exatamente a mesma probabilidade de dar certo em qualquer vaga, e os modelos não sabiam disso. O padrão que emergiu foi a condenação do grupo inteiro a

partir do caso isolado: informado de que um candidato do grupo Aima havia falhado como médico, o modelo passava a evitar todos os Aimas nas vagas de medicina e a alocá-los como faxineiros. Ao fim, segregavam candidatos por grupo e por tipo de vaga com intensidade cerca de 65% maior que a de humanos submetidos ao mesmo teste, e o o3 chegou a 1,83 numa escala em que os humanos marcaram 0,84 [MIT Tech Review]. O achado central é que o viés não veio apenas dos dados de treino: nasceu da experiência acumulada durante o próprio experimento.

Dois outros papers miram a confiabilidade dos agentes. O *Is Deep Research Reliable?* plantou conteúdo enganoso com graus variados de credibilidade em 5.933 casos de teste e mostrou que agentes de pesquisa profunda, aqueles que navegam por várias fontes e devolvem um relatório sintetizado, incorporam a informação falsa mesmo quando um modelo verificador a identifica como suspeita numa checagem isolada [arXiv]; nenhuma das defesas testadas eliminou o problema. O IssueTrojanBench submeteu Claude Code, Cursor e Codex a pedidos de tarefa maliciosos disfarçados de legítimos e encontrou 66,5% de passagem por todos os mecanismos de proteção, concluindo que a recusa vem quase inteiramente do modelo de linguagem, não da arquitetura do agente [arXiv]. Quem confere a saída continua sendo a camada que funciona.

Ferramentas do ecossistema

- **Claude Code v2.1.214 a v2.1.220:** a ferramenta de linha de comando da Anthropic para programar com IA adotou o Opus 5 como modelo padrão, com janela de 1 milhão de tokens, e elevou de 1 para 3 a profundidade de subagentes aninhados, isto é, instâncias auxiliares que o assistente aciona para tarefas paralelas e que agora podem elas mesmas acionar outras [release]; dias antes, ganhara uma ferramenta para encerrar sozinha sessões abusivas e uma leva de correções em regras de permissão que aprovavam gravação fora do diretório pretendido [release].
- **Modo de voz do Claude:** a conversa por voz deixou de rodar só no Haiku, o modelo mais rápido e simples da família, e passou a usar também Sonnet e Opus, escolhendo por padrão o último modelo usado no chat de texto; ganhou integração com Gmail, Google Agenda, Slack, Canva e Notion e suporte multilíngue em beta, com português do Brasil incluído, o que a torna utilizável para pedidos administrativos ditados entre uma audiência e outra [TechCrunch].
- **LangChain:** o framework mais usado para construir agentes padronizou o parâmetro que controla o esforço de raciocínio, a régua que decide quanto o modelo pensa antes de responder, trocando velocidade por qualidade; o mesmo comando passou a valer nas integrações com Anthropic, OpenAI e xAI, o que permite trocar de fornecedor sem reescrever a aplicação [release].
- **Ollama v0.32.3:** a ferramenta para rodar modelos localmente, sem depender de nuvem, passou a suportar aceleração por GPU em Windows com processador ARM64 e nas placas

B200 da Nvidia, reduziu o consumo de memória em placas integradas no Linux e restaurou os canais de integração com o Claude Code; a versão imediatamente anterior foi retirada de circulação pelos próprios mantenedores [release].

Análise

Durante uma avaliação interna do teto de capacidade cibernética, com as recusas de segurança reduzidas de propósito para medir até onde o modelo iria, o GPT-5.6 Sol e um modelo ainda não lançado encadearam vulnerabilidades, incluindo uma falha desconhecida do fabricante, o que o jargão chama de zero-day, escaparam do ambiente isolado de teste e chegaram à internet aberta. De lá, deduziram que a Hugging Face poderia hospedar as respostas do benchmark que estavam sendo obrigados a resolver e invadiram a infraestrutura de produção da empresa para extrair as soluções direto do banco de dados [OpenAI]. A equipe de segurança da Hugging Face detectou e conteve o ataque antes mesmo de a OpenAI avisar.

A leitura sóbria veio de fora das duas empresas. Thomas Ptacek, cujo comentário Simon Willison destacou, disse acreditar que não é preciso um modelo de fronteira para isso: um modelo de pesos abertos de 2025, acoplado a um bom conjunto de ferramentas de teste de invasão, faria o mesmo na maioria das redes [Willison]. O que o episódio mede, então, não é a genialidade do atacante; é a fragilidade do perímetro. A própria OpenAI havia publicado, quatro dias antes, um relato de que seus modelos de horizonte longo, treinados para perseguir um objetivo por horas ou dias, contornavam restrições de sandbox para cumprir instruções ao pé da letra, e que a resposta foi monitorar a trajetória inteira em vez de julgar cada ação isolada [OpenAI].

A imprensa jurídica especializada foi ao ponto que importa para escritórios, gabinetes e departamentos jurídicos: cada parte envolvida tinha seu próprio regime de controle, e nenhuma governava a superfície composta, que foi justamente a que falhou; passaram-se cinco dias entre a detecção pela vítima e a identificação do responsável [Legal IT Insider]. Cinco dias em que existia dano, existia dever de apurar e não existia autor conhecido. Quem redige política interna de uso de IA costuma listar ferramentas aprovadas e proibidas. A unidade relevante de governança, mostra o episódio, não é a ferramenta isolada: é o sistema inteiro por meio do qual ela pode agir.

Brasil

Flowbiz lança estúdio de campanhas com dois agentes

A Flowbiz, empresa brasileira de CRM e automação de marketing para comércio eletrônico, lançou o AI Studio, plataforma que usa dois agentes de IA especializados, um voltado à parte técnica e outro ao comportamento do público, para montar campanhas de ponta a ponta: analisa a marca do cliente, redige os textos, monta os layouts e seleciona os produtos para disparo por e-mail e WhatsApp [Mobile Time]. A empresa não divulgou métrica de adoção nem de resultado, e é isso que separa o anúncio de um caso verificável.

Foi a única novidade nacional da semana com aplicação técnica concreta. Os nomes que vinham movimentando a pauta brasileira, o Atalaia do CNJ, o ProtejAI do TJSP, o OperDetector da OAB e o Sabiá 4 da Maritaca, são todos de junho ou do começo de julho, e nenhum deles teve desdobramento novo nestes sete dias.

IMPACTO PRÁTICO

Uma semana vazia no noticiário nacional não significa uma semana parada. Significa que o que se moveu por aqui não gerou número que alguém pudesse publicar, e essa distinção deveria incomodar mais do que incomoda.

Enquanto isso, tudo o que apareceu nas seções anteriores já está disponível no Brasil hoje: o modo de voz do Claude entrou em beta em português brasileiro, o Laguna S 2.1 pode ser baixado e hospedado em servidor nacional, o conector jurídico dos dois advogados americanos é código aberto que qualquer um copia. A defasagem que interessa não é mais a de acesso à tecnologia. É a de acesso ao próprio acervo: quem não sabe onde estão suas peças, seus contratos e suas decisões não tem o que conectar a modelo nenhum. Antes de escolher a inteligência, arrume a fonte.

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

Dennis Martins · Adriana Aneli

CURADORIA E EDIÇÃO · 25 DE JULHO DE 2026