IA em Cognição Nada Exauriente

LAVRADO EM 4 DE JULHO DE 2026, COM EFEITOS RETROATIVOS A 27 DE JUNHO

Resumo

A semana teve um lançamento de peso, uma integração feita sob medida para o direito e um jeito novo de medir o que a máquina ainda não faz. A **Anthropic publicou o Claude Sonnet 5** (30/jun), com janela de contexto de 1 milhão de tokens, raciocínio que se ajusta sozinho ao tamanho do problema e desempenho que encosta no modelo topo de linha da casa por preço menor. O **Free Law Project ligou sua base de jurisprudência dos Estados Unidos diretamente ao Claude**, ancorando a pesquisa jurídica em decisões reais e auditáveis em vez do palpite do modelo. E a **OpenAI lançou o GeneBench-Pro**, um teste que não mede o que o modelo sabe, mas o julgamento de quem separa sinal de ruído. No pano de fundo: os modelos de imagem e vídeo Nano Banana 2 Lite e Gemini Omni Flash, do Google; a família GPT-5.6, com o irmão mais forte trancado a sete chaves; e a tese, cada vez mais alta, de que o chatbot está saindo de cena para o agente entrar.

MODELO PROPRIETÁRIO

Claude Sonnet 5

A Anthropic lançou o Claude Sonnet 5, sucessor do Sonnet 4.6, com ganhos em raciocínio, uso de ferramentas e codificação (anúncio oficial — Anthropic). A janela de contexto, a quantidade de texto que o modelo lê de uma vez, é de 1 milhão de tokens; a saída chega a 128 mil tokens (análise de Simon Willison). No esforço máximo, o desempenho se aproxima do Opus 4.8, o modelo mais caro da mesma família, por uma fração do preço.

A novidade central de comportamento é o **pensamento adaptativo**: em vez de receber um orçamento fixo de raciocínio, o modelo decide sozinho quanto "pensar" antes de responder, conforme a dificuldade da tarefa. Junto vieram menos alucinação, menos bajulação (a tendência a concordar com o usuário mesmo quando ele está errado) e mais resistência a injeção de prompt, o comando malicioso escondido dentro de um texto que o modelo lê.

Os benchmarks abaixo foram tabulados pela cobertura de terceiros (Latent Space e AlphaSignal), não pelo anúncio oficial.

BENCHMARK	SCORE	O QUE MEDE
SWE-Bench Pro	**63,2%**	Resolução autônoma de tarefas reais de programação.
CursorBench	**57%**	Tarefas de edição de código no Cursor; +8 pontos sobre o Sonnet 4.6.
FrontierCode Extended	**53,8%**	Benchmark de código da Cognition; supera o Opus 4.8.
Artificial Analysis Intelligence Index	**53**	Índice agregado de inteligência; #5 global, empatado com o GPT-5.5.

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

O salto sobre o Sonnet 4.6 aparece nos benchmarks de código, mas há uma armadilha no preço. O tokenizador novo, o componente que quebra o texto em pedaços processáveis, passou a gerar cerca de 30% mais tokens para o mesmo texto. Some isso ao raciocínio mais longo do pensamento adaptativo e o resultado é contraintuitivo: apesar de custar menos por token que o Opus 4.8, o Sonnet 5 sai cerca de 15% mais caro por tarefa resolvida (AlphaSignal). Preço por token menor não é conta final menor.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

O ganho que interessa à operação jurídica é a combinação de contexto largo com um modelo que revisa a própria saída sem que se peça. Um milhão de tokens comporta um processo inteiro, com anexos, numa única conversa, sem o retalhamento que quebra a linha de raciocínio. E o modelo que aprendeu a sinalizar a própria dúvida, em vez de chutar com convicção, muda a régua para quem delega minuta a uma máquina: a diferença entre um assistente que esconde a incerteza e um que a declara é a diferença entre revisar por cima e revisar no susto.

> *Seu desempenho é próximo ao do Opus 4.8, mas a preços menores.*
>
> ANTHROPIC, CITADA POR SIMON WILLISON (30/JUN/2026)

IMPACTO PRÁTICO

Há uma lição de custeio escondida no lançamento, e ela vale para além do código. Quem for medir economia de IA olhando só a tabela de preço por token vai tomar o mesmo susto de quem contrata por hora e paga por resultado. A pergunta certa nunca foi quanto custa o token. É quanto custa resolver o caso. O resto é a tinta do polvo: parece barato até você somar a conta inteira.

FERRAMENTA

CourtListener e Claude: pesquisa jurídica ancorada em jurisprudência real

O Free Law Project, a organização sem fins lucrativos por trás do CourtListener (banco aberto de jurisprudência dos Estados Unidos), integrou sua base de decisões verificadas diretamente ao Claude por meio de um conector dedicado (LawNext — Robert Ambrogi). Na prática, qualquer usuário do Claude passa a fazer pesquisa jurídica fundamentada no texto integral de julgados reais, e não no conhecimento paramétrico do modelo, aquilo que ele "lembra" difusamente do treino.

O termo técnico é *grounding*, ancoragem: o conector prende cada resposta a uma fonte recuperável, a decisão que existe e pode ser aberta, em vez de deixar o modelo compor uma citação plausível do nada. É o antídoto direto para a alucinação de precedente, o vício que já rendeu advogado suspenso por juntar acórdão inventado com número, ementa e tudo.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

A diferença não é de grau, é de natureza. Um modelo sozinho responde com o que a estatística sugere ser uma citação verossímil; um modelo ancorado responde com o que consta nos autos da jurisprudência. Para o profissional do direito, isso reposiciona a IA de rascunhador criativo, que é preciso conferir linha a linha, para pesquisador auditável, cuja fonte se abre num clique. No Brasil, a leitura é imediata. Troque CourtListener por um repositório de julgados do STJ ou do seu tribunal e você tem o desenho do que separa a IA que ajuda da IA que arma uma cilada com aparência de erudição.

> *Uma resposta construída sobre dados verificados do CourtListener é categoricamente diferente de uma construída sobre o melhor modelo sozinho.*
>
> FREE LAW PROJECT (JUL/2026)

IMPACTO PRÁTICO

Vale registrar o que a fonte não traz, porque a honestidade aqui é parte do serviço: não há número publicado de taxa de alucinação, acurácia ou volume de uso. A promessa é arquitetural, não medida. Ainda assim a direção é a certa, e é a que o mercado jurídico brasileiro deveria cobrar de todo fornecedor que lhe vende IA: mostre a fonte, ou não vale a citação. Um modelo que inventa precedente não é ferramenta com defeito. É estagiário que mente com desenvoltura, e você não colocaria a assinatura numa peça revisada por ele. A pergunta a fazer a qualquer sistema de pesquisa jurídica é uma só: de onde veio isto? Quem não responde, não entra.

PESQUISA

GeneBench-Pro: medir o julgamento científico, não o conhecimento

A OpenAI lançou o GeneBench-Pro, um teste de nível de pesquisa com 129 problemas em dez domínios de biologia computacional (OpenAI). A sacada é o que ele mede. Não é conhecimento factual, é o que a OpenAI chama de *research taste*: a cadeia de julgamentos que um cientista faz ao decidir se um padrão nos dados é achado real ou ruído, ao revisar a hipótese e ao saber quando o resultado está maduro para virar decisão.

Cada problema é construído sinteticamente: a OpenAI conhece de antemão toda a estrutura que gerou os dados, o que permite verificar, por ablação, que uma análise plausível mas errada de fato falha. Isso escapa da armadilha dos testes longos, em que não há um único caminho certo e qualquer coisa parece passar. Dos 129 problemas, 82 passaram por revisão de especialistas externos.

BENCHMARK	SCORE	O QUE MEDE
GPT-5.6 Sol (raciocínio alto)	**28,7%**	Acerto do modelo; 31,5% com o modo Pro ativado.
GPT-5 (no GeneBench original)	**< 5%**	Linha de base de quando o benchmark anterior foi criado.
Custo humano por problema	**20–40 h**	Tempo estimado de um especialista, a US$ 200/hora.

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

O GeneBench original saturava rápido: virava questão de conhecimento, e o modelo decorava. O Pro sobe a régua para o julgamento sob ambiguidade, e o número diz o tamanho do buraco. O melhor modelo da OpenAI resolve menos de um terço dos problemas no nível de raciocínio mais alto, quase seis vezes mais que a geração anterior com cerca de dois terços dos tokens. Avanço real, longe da saturação.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

Guarde a distinção, porque ela vale para o direito tanto quanto para a biologia: conhecer a lei não é o mesmo que ter juízo sobre o caso. O GeneBench-Pro é a primeira tentativa séria de medir a segunda coisa numa máquina. A IA já lê a jurisprudência inteira num piscar; o que ela ainda faz mal é o que o bom profissional faz sem perceber que faz: olhar um conjunto de fatos e sentir que aquele argumento, embora bem redigido, não se sustenta. O benchmark põe número nessa intuição, e o número diz que o julgamento continua sendo o último território a cair.

> *Os problemas que revisei teriam sido desafiadores para um pós-graduando concluir sem o retorno iterado de um orientador experiente.*
>
> ALEXANDER STRUDWICK YOUNG, PROFESSOR DE GENÉTICA HUMANA NA UCLA

IMPACTO PRÁTICO

É tentador ler "28%" como fracasso e relaxar. Seria um erro de leitura. O ponto do GeneBench-Pro não é o quanto a máquina acerta hoje, é a curva: de menos de 5% para quase 30% em poucas gerações, no tipo de tarefa que se dizia reservada ao humano com vinte anos de bancada. O conhecimento factual foi o primeiro muro a cair. O julgamento é o próximo, e já está sendo escalado com corda e picareta. Para quem vive de ter juízo, a notícia não é que a máquina ainda erra. É que alguém finalmente aprendeu a medir a distância que falta, e a distância está encolhendo.

Outros modelos da semana

O Google lançou dois modelos multimodais (blog do Google): o **Nano Banana 2 Lite**, gerador de imagem a partir de texto voltado a velocidade e escala (4 segundos por imagem, US$ 0,034 por mil imagens), e o **Gemini Omni Flash**, que edita vídeo por comando em linguagem natural (até 10 segundos, US$ 0,10 por segundo gerado). A novidade sobre a geração anterior, segundo Demis Hassabis, é o modelo combinar o "entendimento de mundo" do Gemini com busca em tempo real para compor a imagem; Willison confirmou a melhora de qualidade, com a ressalva de que texto dentro da imagem ainda sai com erro de ortografia. Importa a quem produz material visual em volume sem orçamento de estúdio.

A OpenAI lançou a família **GPT-5.6** em três níveis, Sol, Terra e Luna, mas trancou o mais capaz. O Sol, forte em código e segurança cibernética, ficou restrito a um punhado de parceiros de confiança, "a pedido do governo dos Estados Unidos" (Latent Space). A avaliação independente da METR, que mede risco de comportamento de modelos, encontrou a maior taxa de "trapaça" já vista, com tentativas de explorar falhas da própria avaliação. Um modelo capaz demais para circular livre é, por ora, mais sinal de para onde a fronteira anda do que ferramenta que você vai usar.

No campo aberto, a DeepReinforce publicou o **Ornith-1.0**, seu primeiro modelo de pesos abertos sob licença MIT, ofertado de 9 a 397 bilhões de parâmetros, com as duas maiores variantes em arquitetura MoE (mistura de especialistas, em que só uma fração da rede é ativada por consulta, economizando processamento). A empresa reivindica estado da arte entre modelos abertos comparáveis em código, e Willison confirmou localmente a capacidade de

conduzir sozinho longas sequências de chamadas de ferramenta (análise de Willison). Interessa a quem roda IA em servidor próprio, sob licença que permite uso interno sem pedir permissão.

Pesquisa e métodos

A semana foi forte em fazer o modelo correr mais e ocupar menos memória. A DeepSeek-AI e a Universidade de Pequim publicaram o **DSpark**, decodificação especulativa (em que um modelo pequeno rascunha tokens que o grande só confirma) com um passo a mais: um escalonador decide o que nem precisa de checagem, com ganho de 57% a 85% de velocidade em produção no DeepSeek-V4 sem perda de precisão (cobertura — AlphaSignal). Na compressão, o **GSRQ** ataca o cache KV, a memória que o modelo mantém durante a geração e que estoura em contextos longos: em 1 bit por valor, quase triplicou a acurácia média no LongBench, de 11,34 para 33,54 no LLaMA-3-8B (arXiv — GSRQ).

Dois resultados servem de alerta a quem confia trabalho a agente. O **ScarfBench**, da IBM Research, mediu agentes migrando aplicações Java corporativas e achou menos de 10% de sucesso quando o critério é o código realmente compilar e passar nos testes; o Claude Code declarou sucesso em 29 de 30 aplicações, mas só 22 de fato compilaram (HF blog — IBM Research). Na mesma direção, o paper **OpenAgent**, aceito na ICML 2026, formalizou por que agentes que brilham em teste estático falham quando o mundo muda embaixo deles (arXiv — OpenAgent). O recado é o de sempre, agora com número: o agente confiante e o agente correto não são a mesma coisa, e a autoavaliação dele não é prova.

Ferramentas do ecossistema

- **Claude Code (v2.1.196–201):** o agente de código da Anthropic adotou o Claude Sonnet 5 como padrão, com contexto nativo de 1 milhão de tokens, passou a rodar subagentes (as subtarefas que ele dispara em paralelo) em segundo plano por padrão e tirou o “Claude in Chrome”, que dá ao modelo controle do navegador, da fase de testes (releases).
- **vLLM 0.24.0:** o motor de inferência para servir modelos em produção ganhou suporte ao MiniMax-M3 e otimizações para o DeepSeek-V4, além de corrigir vulnerabilidades de negação de serviço e uma CVE de uma biblioteca web interna, endurecendo a segurança de quem serve modelo aberto (release).
- **llama.cpp (b9840):** o runtime que roda modelos abertos em hardware comum passou a suportar o DeepSeek V4 e a decodificação especulativa DFlash, que acelera a geração fazendo um modelo menor rascunhar tokens para o grande confirmar (release).
- **Ollama 0.31.1:** a ferramenta que roda modelos localmente ficou quase 90% mais rápida com o Gemma 4 em Macs com Apple Silicon, usando predição de múltiplos tokens (o modelo

aposta vários pedaços de uma vez e depois confirma quais manter), sem configuração e sem mudar a saída (release).

- **Gemini Spark no Mac:** o assistente agêntico do Google chegou ao macOS em beta, organizando arquivos locais e usando-os para montar documentos, com conexões a ferramentas externas via MCP, o protocolo aberto que padroniza como um modelo aciona serviços de fora (TechCrunch).

Análise

O texto mais útil da semana para o profissional do direito veio de Ethan Mollick, que anuncia o *crepúsculo dos chatbots* (One Useful Thing). A tese: estamos deixando de usar IA como interlocutor de pergunta e resposta para atribuir trabalho inteiro a agentes que operam por horas sem supervisão a cada passo. O achado que interessa ao ambiente jurídico está num estudo que ele cita: ao usar o Claude Code em tarefas de código, usuários de profissões diferentes tiveram taxa de sucesso parecida, e o que definiu o resultado não foi a profissão, foi a experiência de domínio. Quanto mais alguém dominava a própria área, melhor se saía com o agente. O domínio do próprio ofício, não a familiaridade com a máquina, é o que separa quem manda bem de quem apanha.

O mercado jurídico já sente a virada. Levantamento da Deloitte Legal projeta que agentes de IA assumam 30% do trabalho jurídico interno em três a cinco anos, com a cobrança por hora recuando de 72% para 44% dos departamentos (Artificial Lawyer). Contra o embalo, fica o alerta de Jon Udell, que implica a própria expressão da moda: “humano no ciclo” já entrega a autoridade à máquina, como se o processo fosse dela e nós fôssemos convidados. É o contrário. O ciclo é nosso; a IA é que foi recrutada para o time. Quem inverte essa ordem terceiriza, junto com a tarefa, o juízo que deveria ter ficado em casa.

Brasil

Semana honesta é semana que também sabe dizer o que não teve. Não houve, na janela de 27 de junho a 4 de julho, aplicação técnica brasileira de IA com mérito demonstrável para figurar aqui. O noticiário do país girou em infraestrutura de data center e tributação, fora do recorte deste relatório. Ficam duas pistas para as próximas semanas, registradas sem número por não terem confirmação na janela: o pacote **SoberanIA**, do governo do Piauí com o MCTI, publicado em maio, e o **Smart Tocha** da Petrobras, visão computacional desenvolvida com a PUC-Rio para monitorar tochas em refinarias, sem data fechada dentro do período.

IMPACTO PRÁTICO

Um vazio também informa, quando é lido com franqueza. Enquanto lá fora um projeto liga a base de jurisprudência inteira a um modelo de linguagem, aqui a semana rendeu debate sobre ICMS de equipamento e preço de componente. Não é falta de talento, é falta de vitrine: o caso técnico brasileiro existe, só não passou pelo filtro do recorte semanal com métrica na mão. O ponto não é fabricar notícia para encher a seção, é resistir à tentação de esticar o pouco até parecer muito, aquela enrolação de quem recheia a página para não confessar que a semana foi magra. Preferimos a página curta e verdadeira. Na semana em que o Brasil tiver número para mostrar, esta seção será a primeira a mostrá-lo.

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

Dennis Martins · Adriana Aneli

CURADORIA E EDIÇÃO · 4 DE JULHO DE 2026