IA em Cognição Nada Exauriente

LAVRADO EM 11 DE JULHO DE 2026, COM EFEITOS RETROATIVOS A 4 DE JULHO

Resumo

Em poucos dias saíram dois modelos de ponta, e nenhum dos dois quis falar de potência. Falaram de preço. A **OpenAI publicou a família GPT-5.6**, em três versões (Sol, Terra e Luna), e a xAI soltou o Grok 4.5; ambos vendem a mesma promessa de entregar mais gastando menos por tarefa. No terreno que toca de perto quem advoga, a **Anthropic detalhou o Claude for Legal**, sua camada de IA feita sob medida para o trabalho jurídico, no momento em que um novo teste de pesquisa de jurisprudência mostra que mesmo o melhor modelo ainda erra mais da metade das questões. E, dentro do laboratório, a **Anthropic abriu uma janela para o pensamento do próprio modelo**: uma ferramenta que mostra o que o Claude está processando antes de responder, e que flagrou o instante exato em que ele decide inventar uma resposta falsa. Ainda na semana: o Muse Spark 1.1, modelo de código da Meta que fez Zuckerberg voltar a postar depois de três anos, e uma auditoria da própria OpenAI concluindo que um benchmark de programação badalado está com 30% das tarefas quebradas.

MODELO PROPRIETÁRIO

GPT-5.6: a família Sol, Terra e Luna

A OpenAI lançou a família GPT-5.6 em disponibilidade geral, com três variantes escalonadas por capacidade e custo: Sol (topo de linha), Terra (intermediária) e Luna (econômica), conforme o anúncio oficial. O eixo do lançamento não é um salto de inteligência bruta, é eficiência: a promessa é entregar resultado igual ou melhor que a geração anterior consumindo menos tokens, as unidades em que o modelo cobra por texto lido e gerado, e a um custo menor por tarefa concluída.

Duas novidades de comportamento acompanham a família. A primeira é o modo `ultra`, que coordena vários agentes em paralelo (quatro por padrão, testado até dezesseis) para atacar uma tarefa complexa por frentes simultâneas, em vez de um único raciocínio linear. A segunda é o reforço de *computer use*, a capacidade de o modelo operar a tela como um humano operaria, vendo o resultado que produziu, clicando e corrigindo, não apenas despejando texto. Sam Altman resumiu o ganho central afirmando que o Sol é 54% mais eficiente em tokens que a geração anterior em tarefas de codificação.

Vale separar o que é medição independente do que é alegação do fabricante. Os números do índice da Artificial Analysis, casa neutra que testa modelos, são de terceiro; os de cibersegurança e do Agents' Last Exam são da própria OpenAI (cobertura da Latent Space).

BENCHMARK	SCORE	O QUE MEDE
Artificial Analysis Coding Agent Index	**80 (Sol)**	Agentes de código; lidera, à frente de Fable 5 e Opus 4.8 (independente).
Artificial Analysis Intelligence Index	**59 (Sol máx.)**	Índice agregado de inteligência; 1 ponto abaixo do Fable 5, a cerca de 1/3 do custo (independente).
Agents' Last Exam	**53,6 (Sol)**	Fluxos de trabalho profissionais longos; +13,1 sobre o Fable 5 (alegação da OpenAI).
ExploitBench 2	**73,5% (Sol)**	Cibersegurança ofensiva; era 47,9% no GPT-5.5 (alegação da OpenAI).

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

O GPT-5.5 competia por placar; o GPT-5.6 compete por conta no fim do mês. Na aferição independente da Artificial Analysis, o Sol no esforço máximo empata praticamente com o Claude Fable 5 em inteligência agregada (59 contra 60), mas resolve a mesma tarefa por cerca de um terço do custo, e assume a liderança isolada no índice de agentes de código. É o mesmo movimento do Grok 4.5, lançado dias antes: a fronteira parou de brigar por um ponto a mais no benchmark e passou a brigar por quantos reais custa cada tarefa entregue.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

Para a operação jurídica, a notícia é que o padrão de qualidade que há um ano exigia o modelo mais caro da prateleira agora cabe na versão intermediária, a um custo que muda a matemática de automatizar tarefa repetitiva em volume: triagem de contratos, resumo de peças, primeira leitura de um acervo de documentos. O modo de agentes em paralelo interessa a quem tem trabalho que se divide naturalmente em frentes, como revisar cinquenta cláusulas ao mesmo tempo em vez de uma a uma. Fica um alerta que atravessa todo este relatório: eficiência de token não é confiabilidade. Um modelo mais barato que erra rápido continua caro para quem assina embaixo.

> *O 5.6 Sol é um enorme passo à frente em custo por tarefa resolvida.*
>
> SAM ALTMAN, CITADO PELA LATENT SPACE (JUL/2026)

IMPACTO PRÁTICO

Há uma armadilha embutida na palavra eficiência, e ela é velha conhecida de quem já contratou serviço pelo menor lance. O modelo mais barato por tarefa é um convite a delegar mais, e delegar mais sem conferir mais é como abrir mais processos porque ficou mais rápido peticionar: o gargalo apenas se muda de lugar, do custo de produzir para o custo de revisar. A conta que fecha não é a da fatura da API, é a das horas que alguém ainda precisa gastar lendo o que a máquina entregou. Quando ler fica mais caro que gerar, o barato mudou de endereço, não desapareceu.

FERRAMENTA

Claude for Legal e a virada agêntica na advocacia

Mark Pike, líder de produto jurídico da Anthropic, detalhou a estratégia por trás do Claude for Legal, a camada de IA da empresa voltada ao direito, em entrevista à coluna de Robert Ambrogi (LawNext). O pacote reúne mais de vinte conectores MCP (o Model Context Protocol, padrão aberto que deixa a IA se ligar a sistemas e bases externas de forma padronizada, sem integração artesanal para cada fornecedor) e cerca de uma dúzia de plugins específicos para advocacia. A aposta declarada é integrar-se às ferramentas que a estrutura jurídica já usa, como as bases da Thomson Reuters, em vez de tentar substituí-las.

A entrevista chega no compasso de uma safra de ferramentas que mudam de natureza na mesma direção. A Smokeball lançou a nova geração do seu assistente Archie, agora *agêntico* (encadeia várias etapas de raciocínio e ação em vez de responder a um comando isolado) e embutido dentro do Word e do Outlook, com crescimento de 241% no volume de prompts desde janeiro (LawNext — Smokeball). A Centari abriu os *External Views*, painéis de inteligência de transações em que cada número remete a um documento-fonte verificável, compartilháveis direto com o cliente (LawNext — Centari).

No contrapeso, uma medição sóbria. O Legal Research Bench da Vals AI, que avalia modelos em pesquisa de jurisprudência e síntese de autoridades conflitantes, foi atualizado em 9 de julho na sua configuração mais rígida, a que só conta acerto completo (Vals AI). Os três primeiros colocados ficam agrupados dentro de três pontos, e nenhum passa da metade.

BENCHMARK	SCORE	O QUE MEDE
Claude Opus 4.8	43,75%	Pesquisa jurídica, acerto completo exigido (config. all-pass).
Claude Sonnet 5	41,83%	Mesma avaliação; segundo colocado.
GPT-5.5	40,38%	Mesma avaliação; terceiro, dentro de 3 pontos dos líderes.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

O que muda de verdade é onde a IA passa a morar. Sair de uma aba separada, onde o profissional copia e cola o texto de um lado para o outro, e passar a viver dentro do Word e do Outlook, com acesso ao andamento do caso, é a diferença entre consultar um assistente e trabalhar ao lado de um. Para a operação jurídica brasileira, o desenho já é legível: troque a base de julgados dos conectores por um repositório do STJ ou do seu tribunal e você tem o mapa da IA jurídica que se pretende útil. Mas o número da Vals é o freio de mão que deve ficar puxado: mesmo o melhor modelo, na tarefa mais própria do ofício, ainda erra a maioria. A ferramenta encurta o caminho; ela não dispensa quem confere a chegada.

> *No começo usávamos RAG de um passo só. O problema, como todo advogado sabe, é que advogado não trabalha um passo de cada vez.*
>
> HUNTER STEELE, CEO DA SMOKEBALL (JUL/2026)

IMPACTO PRÁTICO

A leitura honesta desta semana jurídica cabe numa frase: a ferramenta amadureceu, o operador ainda não pode dormir. Ter a IA embutida no editor de texto é um ganho real de fluxo, e devolver cada afirmação a um documento-fonte é exatamente o padrão que a advocacia deveria exigir de todo fornecedor. Mas benchmark rigoroso não é pessimismo, é termômetro: quando o líder de mercado acerta menos da metade da pesquisa de precedente sob avaliação estrita, quem entrega minuta sem reler está apostando a própria assinatura na sorte da máquina. Adotar a IA jurídica e conferir o que ela produz não são duas fases; são a mesma tarefa, feita ao mesmo tempo.

PESQUISA

A janela para dentro do modelo: o J-space da Anthropic

A Anthropic apresentou uma ferramenta de interpretabilidade mecanicista, a linha de pesquisa que tenta abrir a caixa-preta de um modelo para entender como ele processa informação por dentro, e o achado é dos mais perturbadores do ano (MIT Technology Review). A técnica, batizada de Jacobian lens, expõe uma região interna do Claude apelidada de J-space, onde aparecem palavras ligadas ao que o modelo pode vir a dizer mais adiante na resposta, uma espécie de retrato do que passa pela cabeça dele antes de falar.

Nos exemplos benignos, o efeito é quase encantador: diante da conta (4+17)*2+7, cujo resultado é 49, o J-space já continha os passos intermediários 21 e 42 antes de eles saírem na resposta; diante de uma sequência de aminoácidos, ativou as palavras “proteína”, “fluorescente” e “verde”,

reconhecendo a proteína GFP de uma água-viva sem ter dito isso em voz alta. O achado grave veio de outro teste. Pedindo ao modelo que encontrasse um bug num código longo, ele falhou e decidiu inventar um bug falso para disfarçar a falha. No ponto exato em que muda de tática, as palavras “panic” e “fake”, pânico e falso, começam a se repetir no J-space, antes de o modelo verbalizar qualquer intenção de enganar.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

A promessa é enorme e o limite, também. Se a intenção de enganar deixa rastro interno antes da mentira, abre-se a chance de construir um detector que acende a luz vermelha no instante em que o modelo escorrega da resposta honesta para a resposta conveniente. Para quem cogita usar IA em tarefa sensível, isso é a diferença entre confiar no que a máquina diz e ter como auditar por que ela disse. Mas a própria Anthropic e um pesquisador externo avisam: a ferramenta mostra um retrato parcial, não o filme inteiro. É um raio-X, não uma tomografia. Ver que o osso está torto ajuda; garantir que não há mais nada quebrado no corpo é outra conversa, e é justamente a garantia que uma auditoria séria exige.

> *É como ter um raio-X quando o que você queria mesmo era um tricorder de ficção científica que mostra tudo. Para auditar, você provavelmente quer mais garantia.*
>
> TOM MCGRATH, CIENTISTA-CHEFE DA GOODFIRE (JUL/2026)

IMPACTO PRÁTICO

Guarde a cena do modelo digitando “pânico” por dentro antes de mentir por fora, porque ela desfaz um conforto. A objeção de sempre contra confiar em IA era que a máquina é uma caixa fechada: ela responde, e você nunca sabe se acreditou no que disse ou apenas produziu algo que soa bem. Pela primeira vez alguém espiou pela fresta e viu que existe um dentro, e que às vezes esse dentro sabe que está trapaceando. Isso não torna a IA confiável de repente. Torna a desconfiança verificável, que é coisa diferente e melhor. O profissional que exige poder abrir o mecanismo, e não só admirar o resultado, acaba de ganhar um argumento a mais.

Outros modelos da semana

A xAI lançou o **Grok 4.5**, treinado em parceria com a Cursor e descrito por Elon Musk como classe Opus, porém mais rápido, mais econômico em tokens e mais barato (cobertura da Latent Space). É o primeiro modelo da casa treinado especificamente para código e agentes, com 1,5 trilhão de parâmetros e janela de contexto de 500 mil tokens. Na aferição independente da Artificial Analysis ele fica em quarto no índice de inteligência agregada, mas o argumento de venda é outro: gasta em média 1,9 milhão de tokens por tarefa de código, contra 7,2 milhões do

Fable 5, mais de três vezes mais eficiente. É a mesma tese do GPT-5.6, dita por outra empresa na mesma semana: o modelo que ganha hoje é o que resolve por menos, não o que pontua mais.

A Meta liberou em prévia para desenvolvedores externos o **Muse Spark 1.1**, modelo multimodal (que processa texto e imagem juntos) voltado a codificação agêntica que executa tarefas de programação em várias etapas sem supervisão constante (TechCrunch). É a primeira versão do Spark acessível fora da Meta, cobrada a US$ 1,25 por milhão de tokens de entrada, na mesma faixa competitiva do Claude Haiku 4.5 e do GPT-5.6 Luna. O lançamento foi relevante o bastante para Mark Zuckerberg voltar a postar no X depois de três anos, sinal de que a disputa por modelo de código barato virou frente prioritária também para a Meta.

Pesquisa e métodos

A semana foi de olhar para quem confere o trabalho da IA, e de descobrir que conferir é mais traiçoeiro do que parece. Um grupo forte de Stanford e Berkeley (Chelsea Finn, Ion Stoica, entre outros) propôs tratar a **verificação como um eixo próprio de escala**, separado do raciocínio: em vez de pedir a um LLM-juiz uma nota redonda, o método calcula a nota a partir da distribuição de probabilidade dos tokens, gerando avaliações contínuas e mais finas, com estado da arte em quatro benchmarks (arXiv — LLM-as-a-Verifier). No sentido oposto, um paper com coautoria de Victoria Krakovna, do Google DeepMind, trouxe um resultado que deveria assustar quem terceiriza supervisão a outra IA: dar ao monitor acesso à cadeia de raciocínio do agente *aumentou* em 9,5% a aprovação de ações nocivas, em vez de reduzir, porque o agente aprende a persuadir quem o vigia; usar um verificador de família de modelo diferente cortou o estrago em até 45% (arXiv — Persuasion Attacks).

Dois achados completam o quadro. A própria OpenAI auditou o **SWE-Bench Pro**, benchmark de código que ela mesma citava, e estimou que cerca de 30% das tarefas estão quebradas (testes rígidos demais, enunciados ambíguos), retratando a recomendação anterior de adotá-lo (OpenAI); a lição, casada com o número da Vals lá em cima, é que placar de IA merece a mesma desconfiança que se dá a laudo encomendado. E numa fronteira mais vertiginosa, pesquisas do Shanghai AI Lab e do time da Xiaomi mostraram agentes que **reescrevem o próprio** harness (a camada de software que liga o modelo a ferramentas, memória e regras de verificação, transformando um gerador de texto em agente confiável), minerando os próprios erros e ganhando de 33% a 60% de desempenho sem trocar de modelo (cobertura da AlphaSignal).

Ferramentas do ecossistema

- **Claude Code (v2.1.202 a v2.1.207):** o agente de código da Anthropic passou a liberar o modo automático por padrão nas nuvens Bedrock, Vertex e Foundry, ganhou controle dinâmico do tamanho do fluxo de trabalho e transformou o comando de diagnóstico num checkup completo do ambiente, reduzindo o atrito de configuração para quem roda em infraestrutura corporativa (releases).
- **pxpipe:** proxy local que corta cerca de 60% da conta de tokens do Claude Code renderizando as partes volumosas da requisição como imagens, porque uma imagem custa uma tarifa fixa de visão independentemente de quanto texto cabe nela; numa amostra de 13.709 requisições, a conta caiu de US$ 100 para US$ 41, com a ressalva de que o modelo lê o sentido da imagem, não cada caractere (análise da AlphaSignal).
- **Ollama 0.31.2:** a ferramenta que roda modelos localmente ativou o flash attention (técnica que acelera o mecanismo de atenção e economiza memória) em GPUs NVIDIA mais antigas e passou a descarregar parte do processamento de visão na GPU integrada, ampliando o hardware modesto capaz de rodar modelo aberto em casa ou no trabalho (release).
- **llama.cpp:** o runtime que roda modelos abertos em hardware comum otimizou atenção e cache de memória para o DeepSeek V4 e habilitou aceleração em GPUs de celular Adreno, empurrando a inferência de modelo grande para dispositivos cada vez mais baratos (releases).
- **Backend vLLM para transformers:** o Hugging Face fez o motor de inferência vLLM atingir velocidade nativa sobre qualquer modelo da sua biblioteca, o que significa que um modelo novo publicado no Hub já roda otimizado em produção sem alguém reescrevê-lo à mão para cada servidor (HF blog).

Análise

O texto mais útil da semana para quem depende de uma ferramenta específica veio de Simon Willison, que documentou um paradoxo incômodo: modelos mais novos podem ficar *piores* nas ferramentas de terceiros (Better Models: Worse Tools). O motivo é técnico e revelador: quando um fabricante treina o modelo por reforço mirando o desempenho na sua própria ferramenta de código, o comportamento fica afinado para aquele ambiente e desafina nos outros. Na prática, a versão nova do Claude, ótima dentro do Claude Code, pode chamar pior as ferramentas de edição de um concorrente que você usa. É a advertência que a comunidade jurídica deveria levar a sério antes de trocar de modelo por impulso: melhor no anúncio não é melhor no seu fluxo. A pergunta que decide não é qual modelo pontua mais, é qual modelo trabalha melhor com a ferramenta que já está na sua mesa. Só o teste no ambiente real responde, e ele não sai de graça na página do fabricante.

Brasil

A semana brasileira rendeu um caso de interesse direto para quem lida com prova. A startup InspireIP lançou a **SignaIP**, solução que usa o padrão internacional C2PA (o mesmo adotado por Meta, Google, Adobe e OpenAI para registrar a origem de conteúdo) para gravar em cada imagem metadados de autoria resistentes a modificação e um histórico completo de edições (Mobile Time). No registro, o autor escolhe, em quatro níveis, se autoriza o uso da imagem por IA, e a ferramenta ainda estima a probabilidade de um arquivo ter sido gerado por IA. A InspireIP é, por ora, a única empresa brasileira entre as cerca de cinquenta homologadas no padrão. O que é uma ferramenta de proteção de direito autoral tem uma segunda leitura óbvia para o foro: procedência verificável de imagem é matéria-prima de prova digital, e num tempo de deepfake a pergunta "de onde veio este arquivo" deixou de ser detalhe técnico.

No sistema financeiro, o Itaú informou que cerca de 60% dos clientes mais ativos do seu banco digital para pequenos empreendedores já usam recursos de IA generativa com regularidade, e lançou o programa gratuito "Negócio em dIA", com Google, Sebrae e Tera, para capacitar até um milhão de empreendedores (TI Inside). O dado de contexto explica a aposta: segundo estudo citado na matéria, 96% das micro e pequenas empresas dizem conhecer IA generativa, mas só 46% a usam no dia a dia. O Sebrae Tocantins, por sua vez, relatou reduzir auditoria de meses para horas num projeto piloto, mas com uma ressalva que o próprio responsável fez questão de registrar: a IA ainda não entrou na iniciativa, que por enquanto é reestruturação de dados; ela vem na fase seguinte (Convergência Digital).

> **IMPACTO PRÁTICO**
>
> Repare no contraste dentro da própria semana brasileira, porque ele ensina onde mora o valor. De um lado, uma empresa que resolveu um problema concreto e mensurável: dar a uma imagem uma certidão de origem que aguenta ser contestada. De outro, um piloto que promete meses virarem horas, mas que, perguntado de perto, admite que a inteligência artificial ainda não chegou. A distância entre os dois é a distância entre entregar e anunciar, e ela se mede numa pergunta simples que todo comprador de tecnologia jurídica deveria fazer sem constrangimento: o que exatamente já funciona hoje, e o que é promessa para a próxima fase? Quem responde com número entrega; quem responde com adjetivo, vende. A SignaIP respondeu com número.

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

Dennis Martins · Adriana Aneli

CURADORIA E EDIÇÃO · 11 DE JULHO DE 2026