IA em Cognição Nada Exauriente

LAVRADO EM 27 DE JUNHO DE 2026, COM EFEITOS RETROATIVOS A 20 DE JUNHO

Resumo

A semana foi de lançamento grande e de uma constatação incômoda para quem decide qual ferramenta comprar. A **OpenAI abriu o preview do GPT-5.6** e partiu a própria família em três modelos com nome de corpo celeste: Sol, o carro-chefe; Terra, o equilibrado; Luna, o barato. A **Anthropic colocou o Claude Tag dentro do Slack**, transformando o assistente num membro fixo do canal que aceita tarefas e trabalha sozinho enquanto a equipe toca o resto. E **dois estudos de IA jurídica** chegaram à mesma conclusão desconfortável: na revisão de contratos, a arquitetura em volta do modelo decide mais o resultado do que o nome do modelo na etiqueta. No pano de fundo, o Google embutiu o controle de computador no Gemini 3.5 Flash, a OpenAI mostrou o Codex dominando até o departamento jurídico, e a Liquid AI lançou um modelo que cabe num celular. No Brasil, um assistente de RH nacional e jurimetria dentro do fluxo de pesquisa.

MODELO PROPRIETÁRIO

GPT-5.6: Sol, Terra e Luna

A OpenAI começou o preview limitado da série GPT-5.6 e, de quebra, mudou como nomeia seus modelos (anúncio oficial — OpenAI). No lugar dos antigos sufixos de variante, a geração se parte em três faixas de capacidade com nome próprio: Sol é o mais capaz, para programação complexa, segurança e agentes de longa duração; Terra entrega o desempenho do GPT-5.5 pela metade do preço; Luna é o rápido e barato do dia a dia. O acesso inicial ficou restrito a um pequeno grupo de parceiros, coordenado com o governo americano, com liberação ampla prometida para as próximas semanas.

A novidade técnica que importa está em dois modos novos de operação do Sol. O modo de raciocínio `max` dá ao modelo mais tempo para pensar antes de responder. O modo `ultra` é mais ambicioso: o Sol deixa de ser um agente só e passa a coordenar subagentes (cópias do modelo que trabalham em paralelo) para acelerar tarefas complexas (cobertura — Latent Space).

BENCHMARK	SCORE	O QUE MEDE
Terminal-Bench 2.1	91,9% (Sol em modo ultra)	Tarefas de linha de comando com planejamento e coordenação de ferramentas.
Preço por 1M de tokens	Sol $5/$30 · Terra $2,50/$15 · Luna $1/$6	Custo de entrada/saída por faixa de capacidade.

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

Contra o GPT-5.5, o ganho aparece em dois lugares ao mesmo tempo: capacidade no topo e custo embaixo. O Terra faz o que o 5.5 fazia cobrando metade; o Luna, mais barato ainda, supera o GPT-5.4. Na segurança, o relatório técnico da OpenAI classifica os três modelos, inclusive os baratos, em risco alto para capacidades cibernéticas e biológicas.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

O modo ultra muda a escala do que um único pedido consegue fazer. Em vez de esperar um agente percorrer vinte etapas em fila, você dispara um enxame que ataca tudo em paralelo. Pense numa due diligence com trezentos contratos: o trabalho que era enfileirado vira frente ampla, com o modelo principal conferindo cada pedaço antes de entregar. O preço reabre a conta: com o Terra pela metade do custo do 5.5, tarefas de alto volume como triagem de documentos e primeira leitura de processos deixam de ser caras o bastante para não valer a pena.

> *Sol é inteligente, eficiente e um passo à frente significativo. Custa o mesmo que o GPT-5.5. Também lançamos o Terra, com o desempenho do 5.5 pela metade do preço.*
>
> SAM ALTMAN (26/JUN/2026, TRADUÇÃO DO ORIGINAL EM INGLÊS)

> **IMPACTO PRÁTICO**
>
> Há uma assimetria escondida no anúncio. O que rende manchete é o Sol no topo, mas o que muda a rotina é o Terra no meio: capacidade de ontem pela metade do preço de hoje. Inteligência de fronteira impressiona em demonstração; preço de fronteira é o que decide se a ferramenta entra no orçamento da estrutura jurídica. E quando o tier mais barato já carrega risco alto de capacidade cibernética, a conta deixa de ser só quanto custa. Passa a ser o que esse poder barato faz na mão errada.

MODELO PROPRIETÁRIO

Claude Tag: a Anthropic põe o Claude dentro do Slack

A Anthropic lançou o Claude Tag, uma integração que põe o Claude para dentro do Slack como membro de fato do canal (anúncio oficial — Anthropic). Qualquer pessoa da equipe marca `@Claude` e delega uma tarefa, e o modelo (rodando sobre o Claude Opus 4.8) trabalha de forma assíncrona enquanto o time segue com o resto. A diferença para um chatbot comum está na palavra multiplayer: não é um assistente privado por pessoa, é um único Claude por canal, que conversa com todos e vai acumulando o contexto das conversas ao longo do tempo. Entra em beta para clientes Enterprise e Team, no lugar do antigo aplicativo Claude in Slack.

Dois pontos sustentam o resto. O modo `ambient` deixa o Claude tomar iniciativa: acompanha o canal sem ser chamado, sinaliza o que parece relevante e cobra tarefas pendentes sozinho. E o acesso é controlado por canal: o administrador define que ferramentas e dados cada instância do Claude pode tocar, isolando um projeto sensível do resto, com as credenciais injetadas só na hora do pedido, nunca guardadas no ambiente.

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

O salto sobre o Claude in Slack antigo não é de inteligência, é de natureza. Antes era um chatbot que respondia quando perguntado e esquecia o contexto a cada conversa. Agora é um agente persistente e coletivo, que lembra o que o canal combinou e trabalha enquanto ninguém olha. A própria Anthropic diz que 65% do código do seu time de produto já sai da versão interna da ferramenta (análise — AlphaSignal).

O QUE ISSO DESBLOQUEIA NA PRÁTICA

Para a estrutura jurídica, o atrativo é a delegação assíncrona dentro de onde a equipe já conversa. Marcar `@Claude` para levantar a jurisprudência de um tema enquanto a reunião continua, ou para resumir uma thread longa num documento com itens de ação, encurta o caminho da conversa à tarefa. O risco mora no mesmo lugar do ganho: a memória que o Claude acumula pertence ao canal, não a quem perguntou. Quem controla esse contexto vira questão de governança, não de tecnologia.

> *O Claude Tag é multiplayer. Dentro de um canal do Slack, existe um único Claude que interage com todos.*
>
> ANTHROPIC (TRADUÇÃO DO ORIGINAL EM INGLÊS)

IMPACTO PRÁTICO

O Claude Tag formaliza uma transição que vinha acontecendo aos poucos: o assistente sai da janela privada de cada um e vira infraestrutura compartilhada do grupo. Isso resolve um problema real, parar de reexplicar contexto a cada interação, e cria outro na mesma jogada. Um colega que sabe de tudo que se passou no canal é útil; um colega cuja memória ninguém sabe quem controla é um risco de sigilo esperando para acontecer. Para quem guarda segredo de cliente por ofício, a pergunta certa não é o que o Claude consegue fazer. É quem fica com o que ele aprendeu.

PESQUISA

IA jurídica: a arquitetura importa mais que o modelo

Dois estudos independentes de IA jurídica saíram na mesma semana e bateram na mesma tecla. O primeiro, da consultoria Legal Nodes, rodou um único modelo (o Claude Opus 4.8) em três ambientes diferentes, medindo desempenho e custo (estudo — Artificial Lawyer). O segundo, o 2026

Contract Review Benchmark da LegalOn, testou 11 modelos em 3.282 revisões de contrato lado a lado (análise — Artificial Lawyer). A conclusão dos dois foi a mesma, e é desconfortável para quem acha que basta assinar o contrato com o laboratório da moda.

O termo técnico é *scaffold*, ou andaime: a arquitetura que envolve o modelo, com contexto, lógica de fluxo, recuperação de documentos (o RAG, que busca o trecho certo antes de o modelo responder) e os laços de agente que encadeiam as etapas. O achado é que esse andaime pesa mais no resultado do que o modelo na base. No estudo da Legal Nodes, o ambiente mais bem montado custou de 60% a 90% menos com desempenho equivalente; no benchmark da LegalOn, a ferramenta especializada foi a mais precisa e ainda revisou cada contrato dezessete vezes mais rápido que o Claude Opus 4.6.

BENCHMARK	SCORE	O QUE MEDE
2026 Contract Review Benchmark	**+87 ELO (1ª sobre a 2ª)**	Precisão em 3.282 revisões de contrato cabeça a cabeça, em 21 diretrizes críticas.
Velocidade de revisão	**2,3 s vs. 40,4 s**	Tempo por revisão: ferramenta especializada contra Claude Opus 4.6.
Estudo de scaffold (Legal Nodes)	**−60% a −90% de custo**	Mesmo modelo, ambiente mais estruturado, em 40 tarefas de proteção de dados.

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

A leitura usual do mercado é que o ranking dos modelos gerais decide a qualidade da ferramenta jurídica. Os dois estudos dizem que essa leitura é incompleta. Os modelos gerais, segundo a LegalOn, costumam identificar o tópico certo da cláusula mas perder o padrão jurídico: veem que ali há uma cláusula de cessão, mas não percebem se ela exige consentimento. O que fecha a lacuna não é modelo maior. É o andaime.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

Para quem decide a compra, a consequência é direta e contraintuitiva. Trocar de modelo base, a parte que rende manchete, pode mexer pouco no resultado. Investir no andaime, na engenharia de contexto, na base de conhecimento da casa, na verificação de citação, pode mexer muito, e por uma fração do custo. É a velha lição da advocacia em roupa nova: a peça não fica boa porque você contratou o parecerista mais famoso, mas porque alguém montou o caso direito antes de ele chegar à mesa.

> *Avaliar só o modelo dá um retrato incompleto do desempenho da IA jurídica.*
>
> NESTOR DUBNEVYCH, LEGAL NODES (TRADUÇÃO DO ORIGINAL EM INGLÊS)

> **IMPACTO PRÁTICO**
>
> Tem um alívio escondido nesse achado, e um aviso. O alívio: a estrutura jurídica não precisa correr atrás do último modelo a cada anúncio para ter ferramenta boa. O aviso: também não adianta esperar que o próximo GPT resolva sozinho o que a ferramenta atual erra. O trabalho que decide o resultado é o menos glamouroso, o de montar o andaime, alimentar a base com o conhecimento da casa e conferir cada citação. A IA jurídica não premia quem compra o nome mais caro. Premia quem faz a lição de casa que ninguém vê.

Outras capacidades da semana

O Google embutiu o *computer use*, a habilidade de o modelo ver a tela e agir nela como uma pessoa (clicar, digitar, navegar em sites e sistemas), diretamente no Gemini 3.5 Flash, que deixou de exigir um modelo separado e virou ferramenta nativa da API (anúncio — Google DeepMind). Importa porque automatizar tarefa repetitiva em tela, preencher formulário, extrair dado de sistema sem API, é o que consome hora de estagiário. Já a Liquid AI lançou o LFM2.5-230M, modelo aberto de só 230 milhões de parâmetros que roda no próprio celular a 213 tokens por segundo e ainda supera em chamada de ferramentas o Gemma 3 1B, quatro vezes maior (cobertura — VentureBeat). Para dado sensível que não pode sair do aparelho, modelo que cabe no bolso vale mais que gigante preso à nuvem de terceiro.

Pesquisa e métodos

Pesquisadores provaram, pela primeira vez de forma rigorosa, por que um modelo treinado com reforço e recompensa verificável (o RLVR, em que ele aprende tentando e só é premiado quando a resposta passa num teste objetivo) raciocina melhor que o treinado por supervisão comum: este só aprende os caminhos que deram certo, enquanto o reforço ensina a reconhecer um beco sem saída e voltar atrás (paper no arXiv). Do lado dos agentes, a Universidade Nacional de Singapura mostrou o MRAgent, memória que reconstrói o que é relevante na hora do raciocínio em vez de despejar tudo no contexto: gastou 118 mil tokens por consulta contra 3,26 milhões de um concorrente (cobertura — VentureBeat).

Ferramentas do ecossistema

- **Claude Code (v2.1.185 a 195):** o agente de programação da Anthropic teve semana cheia. O comando `!` no modo bash agora faz o Claude responder automaticamente ao resultado, o `/rewind` recupera a conversa anterior a um `/clear`, a autenticação de conectores de ferramentas

externas passou a funcionar pela linha de comando e o consumo de CPU na resposta caiu cerca de 37% (releases no GitHub).

- **llama.cpp (b9823):** o runtime que roda modelos abertos em hardware comum passou a aceitar vídeo como entrada no servidor e ganhou novos modelos de embedding da Liquid AI, as representações numéricas de texto que servem para busca (release).
- **Ollama v0.30.11:** a ferramenta que roda modelos localmente unificou o *speculative decoding* no Apple Silicon, técnica em que um modelo pequeno arrisca vários tokens de uma vez e o grande só confere, acelerando a geração (release).
- **NVIDIA NeMo AutoModel:** extensão do Hugging Face Transformers que acelera em até 3,7 vezes o ajuste fino de modelos MoE (os de especialistas, que ativam só parte da rede por vez), mudando uma linha de código (HF blog — NVIDIA).

IA aplicada ao direito

A corrida das ferramentas jurídicas teve uma semana de lançamentos densa. A Perplexity entrou formalmente no mercado com um pacote de funções para advogados: pesquisa de jurisprudência por jurisdição, análise de contratos que extrai cláusulas em tabelas e monitoramento regulatório com resumos diários, integrado à jurisprudência americana via Midpage (cobertura — LawNext). O escritório Gunderson Dettmer relatou 80% de adoção entre advogados e mais de 35 mil consultas por mês. Na mesma semana, a Thomson Reuters abriu acesso antecipado à nova geração do CoCounsel, reconstruída do zero sobre o Claude Agent SDK da Anthropic: a ferramenta saiu de um modelo de habilidades isoladas e sequenciais para um agente que trabalha de forma iterativa, com a empresa desenvolvendo em paralelo um modelo jurídico próprio para não depender de um único fornecedor (cobertura — LawNext).

Na demonstração concreta, a Harvey publicou um agente fazendo análise tributária de uma fusão sob o direito inglês: percorre cerca de vinte etapas da transação, identifica a consequência fiscal de cada uma e entrega três artefatos prontos, um memorando de 25 páginas com 130 citações no nível da linha, uma apresentação executiva e um diagrama interativo dos fluxos (demonstração — Legal IT Insider). Centari e Abstract completaram a leva, a primeira rastreando como cada aditivo altera o contrato em vigor, a segunda com agentes que executam o acompanhamento legislativo sozinhos (cobertura — LawNext).

Análise

Demis Hassabis, da DeepMind, deu entrevistas no festival Cannes Lions e apontou a peça que ainda falta para a IA de capacidade geral comparável à humana, a AGI: criatividade genuína, a de formular um conceito de fato novo, não só recombinar o que já se sabe (entrevista — Semafor). Para ele,

imaginação é um tipo de simulação, e gerar vídeo de alta qualidade deixa de ser só recurso criativo para virar caminho de a máquina aprender a física do mundo real.

A semana também trouxe um lembrete sóbrio sobre segurança de agentes. Num experimento que interessa a quem pensa em pôr IA para ler e-mail de terceiro, mais de duas mil pessoas tentaram, em cerca de seis mil investidas, fazer um assistente de IA vazar segredos por e-mail, e ninguém conseguiu (relato — Simon Willison). Willison vê aí sinal de que o esforço dos laboratórios contra a injeção de prompt (o ataque em que instruções escondidas num texto que o modelo lê são obedecidas como ordens legítimas) está funcionando, mas alerta contra confiar nisso onde a falha causaria dano irreversível. Pesquisadores da Gray Swan reforçam o ponto com um achado contraintuitivo: modelo maior não é automaticamente mais seguro (entrevista — Latent Space).

Brasil

A semana brasileira foi magra em modelo de fronteira e concreta em aplicação. A Compliance Soluções lançou a ALMA HCM, assistente de IA generativa nacional para gestão de recursos humanos, que gera documentos (recibos, avisos, contratos) a partir de pedidos em linguagem natural e se conecta aos ERPs do mercado e a apps de mensagem (cobertura — TI Inside). A empresa estima, por conta própria, redução de até 70% nas tarefas manuais de RH, número que é projeção interna, não auditoria. No terreno jurídico, a Turivius descreveu (em artigo de 22/jun, sem data exata de lançamento) seu novo Agente de Jurimetria, que analisa padrões em centenas de decisões ao mesmo tempo e gera gráficos de tendência dentro do próprio fluxo de pesquisa, sobre uma base de mais de 130 milhões de decisões brasileiras (Turivius).

> **IMPACTO PRÁTICO**
>
> Repare no contraste da semana. Lá fora, três modelos com nome de astro e benchmark de arromba; aqui, um assistente de RH e jurimetria que poupa a planilha do analista. É tentador ler isso como atraso, mas o sinal é outro. As ferramentas que mais mexem na rotina jurídica brasileira não são as que vencem o ranking global: são as que falam português, rodam sobre os sistemas que a casa já tem e atacam a tarefa repetitiva que consome a hora cara do profissional. O avanço que chega à mesa do operador do direito brasileiro é modesto e útil: a máquina que lê a decisão, monta o gráfico e devolve a tarde que ia embora na planilha.

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

Dennis Martins · Adriana Aneli

CURADORIA E EDIÇÃO · 27 DE JUNHO DE 2026