IA em Cognição Nada Exauriente

LAVRADO EM 20 DE JUNHO DE 2026, COM EFEITOS RETROATIVOS A 13 DE JUNHO

Resumo

A semana foi de um modelo aberto que encostou nos proprietários e de duas demonstrações de que a IA já trabalha dentro de profissões reguladas. A **Z.ai publicou o GLM-5.2**, modelo de pesos abertos com licença permissiva que disputa de igual para igual com Claude Opus 4.8 e GPT-5.5 em programação, e ainda lê um milhão de tokens de uma vez. O escritório nativo de IA **Crosby lançou um benchmark de negociação de contrato**, o primeiro a medir como os modelos de fronteira conduzem uma negociação inteira, turno a turno, e não uma tarefa isolada. E a **OpenAI publicou na NEJM AI** um estudo em que seu modelo de raciocínio reabriu 376 casos de doenças genéticas sem diagnóstico e fechou 18 deles. No pano de fundo: a Xiaomi mostrou um modelo de 1 trilhão de parâmetros a mil tokens por segundo, uma safra de pesquisa sobre agentes que se reescrevem sozinhos, e uma leva de ferramentas de infraestrutura com foco em segurança. No Brasil, semana de intenção: o governo federal e a USP anunciaram iniciativas, ainda sem números de resultado.

MODELO OPEN-SOURCE

GLM-5.2: o modelo aberto que encostou nos proprietários

A Z.ai, laboratório chinês, liberou o GLM-5.2 com 753 bilhões de parâmetros e licença MIT, a mais permissiva que existe, que autoriza uso comercial sem pedir nada em troca (blog técnico — Z.ai no Hugging Face). O modelo é open-weights: os pesos, o miolo numérico treinado, ficam disponíveis para baixar e rodar na própria infraestrutura, sem depender da nuvem do fabricante. A janela de contexto, a quantidade de texto que ele lê de uma vez, saltou de 200 mil para 1 milhão de tokens, algo como milhares de páginas num único pedido.

Por dentro, é um MoE, sigla de Mixture of Experts: em vez de acionar todos os 753 bilhões de parâmetros a cada palavra, o modelo ativa só uma fração, cerca de 40 bilhões, escolhendo quais "especialistas" internos consultar. Sai mais barato sem perder capacidade. A novidade de arquitetura que faz o contexto de um milhão de tokens ser de fato utilizável se chama IndexShare: ela reaproveita o mesmo índice de atenção entre quatro camadas seguidas e corta em 2,9 vezes o custo de processamento por token nessa janela gigante.

BENCHMARK	SCORE	O QUE MEDE
SWE-bench Pro	**62,1%**	Correção autônoma de bugs reais de software; o Opus 4.8 faz 69,2% e o GPT-5.5, 58,6%.
Terminal Bench 2.1	**81,0%**	Tarefas conduzidas na linha de comando; era 63,5% no GLM-5.1, contra 85,0% do Opus 4.8.
FrontierSWE	**74,4%**	Engenharia de software de alta dificuldade; quase empata com os 75,1% do Opus 4.8.
AIME 2026	**99,2%**	Olimpíada de matemática dos EUA; era 95,3% na geração anterior.

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

O GLM-5.1 já era competente; o 5.2 fecha a distância para os modelos fechados de ponta. O salto mais visível está em programação de longa duração, as tarefas que exigem dezenas de passos encadeados: no Terminal Bench, a pontuação quase dobrou. Simon Willison, um dos comentaristas técnicos que o relatório acompanha, testou o modelo e o classificou como *provavelmente o modelo de pesos abertos só-texto mais poderoso disponível hoje* (análise de Willison). Ele anotou um defeito honesto: o GLM-5.2 gasta mais tokens de saída que os outros modelos abertos para chegar à resposta, um custo de verbosidade que o preço baixo por token, como se vê adiante, mais que compensa.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

Modelo aberto e permissivo significa que uma operação jurídica pode rodá-lo dentro da própria casa, sobre documentos sigilosos, sem mandar uma linha sequer para servidor de terceiro. O custo torna o argumento concreto: na medição do Artificial Analysis, o GLM-5.2 resolve uma tarefa por US$ 2,40, contra US$ 10,40 do Opus 4.8 e US$ 31 do Fable 5 (cobertura — Latent Space). Jeremy Howard, da fast.ai, disse usá-lo como equivalente prático ao Opus 4.8. Para quem precisa revisar dez mil contratos atrás de uma cláusula e não pode terceirizar a leitura, a conta entre alugar inteligência fechada e instalar inteligência própria acabou de mudar.

IMPACTO PRÁTICO

Por anos, o modelo aberto foi o primo pobre: bom para estudar, fraco para produzir. Essa hierarquia está se desfazendo. Quando um modelo que você baixa de graça e roda no seu próprio servidor chega a poucos pontos do melhor modelo pago do mundo, a pergunta deixa de ser "qual é o mais inteligente" e passa a ser "de quem é a chave do cofre onde meus dados vão entrar". Para escritórios, gabinetes e assessorias que lidam com sigilo por ofício, soberania sobre o dado vale mais que dois pontos num benchmark. O cofre que você controla é sempre mais seguro que o cofre alugado.

PESQUISA

Crosby mede a IA negociando contrato, turno a turno

O Crosby, escritório de advocacia construído em torno de IA, lançou o Multi-turn Negotiation Bench, apelidado de Redline (Artificial Lawyer). É o primeiro benchmark público que avalia não uma tarefa solta de redigir cláusula, mas o fluxo inteiro de uma negociação contratual de um advogado comercial sênior: entender o contexto do negócio, ler a alavancagem entre as partes e se adaptar ao longo de vários turnos de troca de minuta.

A diferença é metodológica e importa. Medir se o modelo redige uma cláusula correta testa memória e gramática jurídica; medir se ele conduz uma negociação testa julgamento sequencial, a capacidade de mudar de estratégia quando a outra parte muda a dela. O Redline pontua essa sequência de decisões num ambiente de negociação real.

BENCHMARK	SCORE	O QUE MEDE
ChatGPT 5.5	50,5%	Negociação contratual multi-turno no Redline; líder entre os modelos testados.
Claude Fable 5	47,3%	Mesmo teste; segundo colocado.
Gemini 3.5 Flash	45,1%	Mesmo teste; terceiro.
Claude Opus 4.8	44,4%	Mesmo teste; último entre os quatro modelos de fronteira.

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

Não havia geração anterior: o vácuo era o estado da arte. A avaliação de IA jurídica vinha se contentando com tarefas pontuais, do tipo "resuma este parecer" ou "ache a cláusula de foro". O Redline é o primeiro a cobrar do modelo o que um advogado faz de verdade numa mesa de negociação, que é encadear dezenas de microdecisões sob informação incompleta. E o achado

central é desconfortável para o hype e tranquilizador para o profissional: humanos ainda superam a IA em *encontrar rotas novas de acordo*, a parte criativa de destravar um impasse.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

A semana foi inteira nesse compasso, e não só no benchmark. A BlackBoiler lançou o Veris, suplemento do Word que combina marcação determinística treinada no histórico da própria firma com IA generativa acionada só onde precisa (LawNext). A Relativity comprou a Gavel para levar automação de minutas para dentro do Word (LawNext). O denominador comum é claro: a IA jurídica está saindo da demonstração e entrando no lugar onde o advogado de fato trabalha, que é o editor de texto sobre o contrato.

> *O próximo estágio da IA contratual será moldado pela consistência, pela governança e pela execução eficiente em custo, não apenas pela geração de linguagem.*
>
> DANIEL BRODERICK, CEO DA BLACKBOILER (16/JUN/2026)

IMPACTO PRÁTICO

Leia a tabela de novo, com calma. O melhor modelo do mundo acerta metade de uma negociação contratual. Metade. Quem vende IA jurídica como substituto do advogado está vendendo um carro que vira à esquerda em metade das curvas. O número não diz "a máquina chegou para te aposentar"; diz "a máquina chegou para fazer o trabalho braçal, e o julgamento continua sendo seu". O relatório do uso real do Harvey, ferramenta de IA jurídica adotada em produção, mostra usuário avançado economizando onze horas por semana (Legal IT Insider). Onze horas que voltam para onde a IA ainda erra metade das vezes: a hora de decidir. A ferramenta poda o galho seco para você cuidar da árvore.

PESQUISA

OpenAI reabre 376 casos de doença rara e fecha 18

Pesquisadores do Boston Children's Hospital, ligado a Harvard, e da OpenAI usaram o modelo de raciocínio o3 Deep Research para reanalisar 376 casos genéticos pediátricos que tinham resistido a anos de investigação especializada e seguiam sem diagnóstico (estudo na NEJM AI). O modelo levantou hipóteses com as evidências linkadas para um especialista revisar. Depois da confirmação clínica em laboratório, 18 casos ganharam diagnóstico, um rendimento adicional de 4,8% sobre o que anos de análise humana não tinham fechado.

O detalhe que importa para qualquer profissão de risco está no desenho. O o3 Deep Research é um modelo de raciocínio, treinado para encadear passos de inferência em vez de cuspir uma resposta imediata, e foi usado como camada de explicação sobre os dados genômicos: ele

conectou sintoma, padrão de herança e literatura científica em justificativas auditáveis. O modelo não decidiu nada. Produziu hipótese fundamentada; o médico confirmou ou descartou.

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

Não foi um caso isolado de sorte. Na mesma semana a OpenAI mostrou o mesmo padrão na química: ligado a um laboratório automatizado, o modelo melhorou de forma autônoma o rendimento de uma reação difícil usada em medicamentos, de 16,6% para 25,2% de aproveitamento médio (OpenAI). E o GPT-5.5 Instant, a versão rápida e gratuita do modelo, passou a empatar com os modelos de raciocínio caros em perguntas de saúde, derrubando em 71% a taxa de respostas com erro de fato (OpenAI). O salto da geração anterior não é "o modelo ficou mais esperto"; é "o modelo virou colaborador confiável em campo regulado".

O QUE ISSO DESBLOQUEIA NA PRÁTICA

O direito e a medicina compartilham a mesma exigência: a resposta precisa ser rastreável até a fonte, e a decisão final tem dono humano com responsabilidade pessoal. O que o estudo da NEJM AI demonstra é justamente esse arranjo funcionando, a IA como pesquisador incansável que vasculha milhões de variantes e devolve hipóteses com a evidência anexada, deixando o juízo para quem responde por ele. Troque "variante genética" por "precedente" e "prontuário" por "autos", e você tem o modelo de uso de IA que uma operação jurídica séria deveria perseguir.

> **IMPACTO PRÁTICO**
>
> Dezoito crianças passaram anos sem nome para a própria doença, e uma máquina ajudou a encontrá-lo em semanas. É fácil ler isso como milagre tecnológico e errar a lição. O que curou não foi a IA sozinha: foi a IA gerando pista auditável e o especialista confirmando no laboratório. Tire qualquer uma das duas pontas e o sistema desaba, ou em diagnóstico fantasma, ou em lentidão de sempre. A IA que entra em profissão regulada pela porta certa não é a que decide no seu lugar. É a que mostra o trabalho e deixa a assinatura com você.

Outros modelos da semana

A Xiaomi apresentou o MiMo-V2.5-Pro-UltraSpeed, modelo de 1 trilhão de parâmetros que gera mil tokens por segundo, velocidade rara nessa escala (Import AI). O truque é combinar quantização FP4, que reduz a precisão numérica dos pesos para economizar memória e acelerar a conta, com decodificação especulativa, em que um modelo pequeno chuta os próximos tokens e o grande só confirma. Roda em 8 GPUs comuns, o que tira modelos de trilhão de parâmetro do clube exclusivo dos clusters gigantes. Já a Poolside liberou o Laguna M.1, modelo MoE de 225 bilhões de parâmetros com 23 bilhões ativos e contexto de 256 mil

tokens, afinado para programação agêntica, isto é, para guiar agentes que executam tarefas de código em vários passos (Latent Space), ampliando a oferta de modelos abertos para quem automatiza desenvolvimento.

Pesquisa e métodos

O tema quente foi o agente que melhora a si mesmo sem retreinar o modelo. Uma série de trabalhos descreve o harness auto-melhorável: em vez de mexer nos pesos, caros e arriscados de alterar, o sistema reescreve a camada ao redor do modelo, as instruções, as ferramentas e a lógica de tentativa e erro, mantendo só as mudanças que não quebram tarefas já resolvidas (AlphaSignal). Os ganhos chegam a 21 pontos percentuais em benchmarks de agente, com o modelo congelado. É capacidade barateada por engenharia de contexto, não por mais GPU.

Do lado do que interessa diretamente ao sigilo profissional, o benchmark MosaicLeaks mostrou que agentes de pesquisa vazam informação confidencial por efeito mosaico: cada consulta isolada parece inócua, mas a combinação delas revela o segredo, e treinar a privacidade no modelo reduz o vazamento em mais de três vezes, coisa que um simples aviso no prompt não resolve (Hugging Face). E surgiram dois trabalhos acadêmicos voltados ao direito: o LegalWorld, ambiente de simulação para agentes jurídicos baseado em 75 mil sentenças reais, que conclui que nenhum modelo atual lidera ao mesmo tempo em consulta, redação e sustentação (arXiv), e o LegalHalluLens, que audita a alucinação jurídica por tipo de erro e mede o viés entre omitir e inventar (arXiv). Os dois confirmam pela ciência o que o benchmark do Crosby mostrou pela prática: a máquina ajuda muito e ainda erra onde dói.

Vale registrar uma promessa com asterisco: a startup Subquadratic afirma ter quebrado o gargalo quadrático da atenção, o custo que quadruplica a cada vez que você dobra o tamanho do texto, com ganho de 56 vezes em velocidade e contexto de até 12 milhões de tokens (MIT Technology Review). Testes de terceiros confirmam os números, mas o modelo reaproveita pesos de outro já treinado, então a prova definitiva não veio. Se confirmado, é a diferença entre ler um processo e ler um acervo inteiro num pedido só.

Ferramentas do ecossistema

- **vLLM v0.23.0:** o motor de inferência mais usado para servir modelos ganhou descarregamento do cache de contexto em várias camadas de memória, o que permite atender contextos longos sem estourar a GPU, além de suporte ao DeepSeek-V4 (release).
- **Claude Code v2.1.183:** a ferramenta de programação por IA no terminal passou a bloquear sozinha comandos destrutivos como apagar histórico do git ou derrubar infraestrutura

quando você não pediu, proteção que vale ouro no modo automático, em que a IA age sem confirmar cada passo (release).

- **Claude Connectors:** a Anthropic liberou integrações nativas do Claude com Google Drive, Word e Slack que respeitam as permissões originais de cada sistema, ou seja, documento restrito continua restrito para a IA, requisito de conformidade para ambiente jurídico (Lightjur).
- **Ollama v0.30.9 e v0.30.10:** a ferramenta que roda modelos localmente, sem internet, corrigiu uma falha grave que truncava a saída de agentes de código e passou a acelerar a família Command A da Cohere em chips Apple Silicon (release).
- **llama.cpp b9731:** a base C++ que roda modelos quantizados em hardware comum acelerou em 12 vezes o cálculo das probabilidades de token, aliviando a latência de servidores locais sob carga (release).

Análise

Yann LeCun, pioneiro do aprendizado profundo e hoje à frente da AMI Labs, voltou a cutucar a aposta dominante (entrevista à CNBC). A tese técnica: os grandes modelos de linguagem manipulam palavras com maestria mas não constroem um modelo interno de como o mundo funciona, e o caminho promissor são os world models, sistemas que aprendem a prever consequências de uma ação antes de agir. LeCun fala com conflito de interesse declarado, já que sua empresa levantou mais de um bilhão de dólares apostando nessa rota. Mas o recado merece ser ouvido mesmo assim: a arquitetura que domina hoje não é, por decreto, a última.

Brasil

Foi uma semana de anúncio de intenção, não de entrega com números, e é assim que vai ficar registrada.

Banco de reuso de IA no governo federal

O Laboratório de Inteligência Artificial do Governo Federal vai ativar no segundo semestre um banco de reuso de soluções de IA construídas por órgãos públicos, com mais de 80 soluções já inscritas (Convergência Digital). O que muda é o combate à duplicação: um mapeamento mostrou órgãos diferentes reconstruindo a mesma solução em paralelo, e o banco existe para parar esse desperdício. Importa porque reuso de software no setor público é dinheiro de imposto que deixa de ser gasto duas vezes.

USP e UFERSA capacitam 300 professores

A USP de São Carlos e a UFERSA abriram 300 vagas em curso gratuito de 60 horas sobre IA aplicada à educação básica, voltado a professores da rede pública, de agosto a dezembro (Convergência Digital). A novidade é o foco em uso prático em sala, não em explicar o que a tecnologia é, com metade das vagas reservada a mulheres. Importa porque capacitar quem forma é a alavanca mais barata de difusão de uma tecnologia num país inteiro.

> **IMPACTO PRÁTICO**
>
> Toda semana o relatório olha para o Brasil esperando uma métrica de resultado, e toda semana, nas semanas mornas como esta, encontra promessa bem-intencionada. Não é demérito: banco de reuso e formação de professor são fundação, e fundação não aparece em benchmark. Mas vale a franqueza. Enquanto a Z.ai mede seu modelo em quatro casas decimais e o Crosby cronometra negociações turno a turno, a notícia daqui ainda é a inscrição aberta e o lançamento previsto. A distância entre anunciar e medir é exatamente a distância que o país precisa fechar.

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

Dennis Martins · Adriana Aneli

CURADORIA E EDIÇÃO · 20 DE JUNHO DE 2026