IA em Cognição Nada Exauriente

LAVRADO EM 6 DE JUNHO DE 2026, COM EFEITOS RETROATIVOS A 30 DE MAIO

Resumo

A semana teve um lançamento que dominou as conversas e uma safra de ferramentas que mira direto no seu trabalho. A **Microsoft apresentou a família MAI**, encabeçada pelo MAI-Thinking-1 — o primeiro modelo de raciocínio que a empresa construiu por conta própria, sem depender da OpenAI, e que já chega disputando de igual para igual com o Claude. No terreno jurídico, **uma enxurrada de plataformas de IA agêntica** foi anunciada na mesma semana — sistemas que não só redigem, mas executam tarefas dentro do sistema da própria operação jurídica. E a Hcompany abriu o **Holo3.1**, uma família de agentes que operam o computador rodando inteiramente na sua máquina, sem nada sair da rede. No pano de fundo: a NVIDIA abriu os modelos Cosmos 3 e Nemotron 3 Ultra, a OpenAI reformou a memória do ChatGPT, papers de fronteira sobre segurança de agentes acenderam um alerta incômodo e, no Brasil, o governo lançou um programa nacional de formação em IA.

MODELO PROPRIETÁRIO

Microsoft MAI-Thinking-1: a aposta de independência

No Microsoft Build, a Microsoft lançou o MAI-Thinking-1, seu primeiro modelo de raciocínio próprio — ou seja, treinado do zero pela empresa, e não uma versão da OpenAI revendida sob outra marca (anúncio oficial — Microsoft AI). É um modelo de raciocínio estendido: antes de responder, ele "pensa" em etapas, gastando mais cálculo para acertar problemas difíceis, em vez de cuspir a primeira resposta. A arquitetura é de Mistura de Especialistas (MoE) — em vez de acionar a rede inteira a cada palavra, o modelo ativa só a fração relevante, o que barateia a conta: são cerca de 1 trilhão de parâmetros no total, mas apenas 35 bilhões entram em ação por vez.

A janela de contexto — quanto texto o modelo lê de uma vez — é de 256 mil tokens, algo como seiscentas páginas. A Microsoft fez questão de afirmar uma "linhagem limpa": dados licenciados comercialmente, sem destilar respostas de modelos de terceiros (a prática de treinar um modelo copiando as saídas de outro já pronto). A leitura do relatório técnico, porém, põe ressalva na promessa — a empresa admite um rastreamento proprietário da web aberta, com os mesmos problemas de licenciamento dos outros grandes modelos (análise de Simon Willison). Nos benchmarks que a Microsoft destacou:

BENCHMARK	SCORE	O QUE MEDE
Benchmark	Resultado	O que mede
AIME 2025	97,0%	prova da olimpíada de matemática dos EUA
AIME 2026	94,5%	a mesma prova, edição nova — menos risco de o modelo já ter "visto" as questões
SWE-Bench Pro	"empata com o Opus 4.6" *	engenharia de software: corrigir problemas em repositórios reais

* *A Microsoft não divulga o número exato — só a comparação.*

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

Até aqui, a Microsoft embalava modelos da OpenAI no Copilot. O MAI-Thinking-1 é a primeira vez que ela compete com casa própria — e os números que divulgou a põem na mesma faixa dos líderes. Em avaliações cegas lado a lado, com 1.276 tarefas, avaliadores humanos preferiram o MAI-Thinking-1 ao Claude Sonnet 4.6. São resultados auto-reportados, ainda sem confirmação independente — vale tratá-los como afirmação do fabricante, não como veredito.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

Seiscentas páginas em uma só leitura é a medida de um processo inteiro, com petições, anexos e decisões, cabendo de uma vez na cabeça do modelo — sem fatiar o documento e torcer para o pedaço certo ser lembrado. E a chegada de um terceiro concorrente de ponta importa pela razão mais terrena: três fornecedores brigando por preço e qualidade é melhor para quem contrata do que dois. A família MAI não veio sozinha — traz ainda o MAI-Image-2.5 (segundo colocado em edição de imagem no ranking Arena) e modelos de transcrição e voz.

> **IMPACTO PRÁTICO**
>
> Toda fornecedora de IA jura que seu modelo "bate" o concorrente. A diferença aqui é o tamanho da prova: um relatório técnico longo e detalhado contra o padrão de soltar três gráficos e pedir fé. Para quem vai citar um número desses numa decisão ou num parecer, a regra continua valendo — benchmark é promessa do fabricante até alguém de fora repetir o teste. Anote a promessa; cobre a confirmação depois.

FERRAMENTA

A semana em que a IA jurídica virou agente

Em poucos dias, meia dúzia de plataformas anunciou a mesma virada: sair do assistente que só responde para o agente que age. “Agente” aqui tem sentido técnico — um sistema de IA que executa uma sequência de ações com autonomia, sem pedir confirmação a cada passo. A Filevine lançou o LOIS Console, que não apenas lê o caso mas escreve de volta no sistema de registro da banca — cria tarefas, move prazos, atualiza cadastros e gera documentos sozinho, treinado sobre padrões de 40 milhões de processos de mais de 6 mil firmas. A LawVu apresentou o LegalOS, voltado a departamentos jurídicos internos, com triagem automática de demandas que chegam por e-mail, Teams ou Slack antes de o advogado entrar.

O fio comum entre elas é o MCP (Model Context Protocol), um padrão aberto — lançado pela Anthropic e hoje adotado por todos — que deixa diferentes sistemas de IA conversarem e trocarem ferramentas com segurança. Tanto o LegalOS quanto o iManage adotaram o MCP: o primeiro expõe o acervo da estrutura jurídica para o ChatGPT, o Claude e o Copilot sem integração sob medida; o segundo abre seu conteúdo governado a qualquer sistema de IA pelo mesmo protocolo. A Icertis relançou sua gestão de contratos sobre o assistente Vera, e a sueca Legora comprou a Cadastral para entrar no imobiliário. Houve até opção aberta: o Lavern é uma plataforma multi-agente sob licença Apache 2.0, com prompts legíveis e editáveis, que roda local via Ollama — o oposto da caixa-preta. E a própria OpenAI recrutou o fundador da Ironclad para montar uma frente jurídica dedicada.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

A diferença entre redigir e agir é a diferença entre um estagiário que entrega a minuta e um que protocola, agenda a audiência e atualiza a pasta. O ganho é real, e o risco também: quando o sistema escreve de volta no registro, um erro deixa de ser uma frase ruim na tela e vira um prazo movido no processo. Por isso o detalhe técnico mais relevante da Filevine não é a velocidade, é a verificação de completude — o sistema se propõe a “provar o que não encontrou”, atacando o vício mais perigoso do modelo de linguagem, que é a omissão silenciosa.

> *A IA resume o documento na tela. Redige uma cláusula. O que ela não consegue é entrar num processo que não esteja organizado de forma estruturada.*
>
> ESTUDO DA LEGATICS SOBRE 6.200 TRANSAÇÕES

IMPACTO PRÁTICO

Um estudo da Legatics com 6.200 transações (Artificial Lawyer) chegou à conclusão que ninguém quer ouvir: a IA acelera a operação que já é organizada e patina na que é bagunçada. Ela não substitui o processo — ela o amplifica, para o bem e para o mal. Quem espera que a ferramenta arrume a casa vai descobrir, caro, que ela só corre rápido sobre piso já nivelado. A virada agêntica chegou madura lá fora; a pergunta que sobra para a banca brasileira não é se a ferramenta funciona, é se a sua operação está arrumada o bastante para merecê-la.

MODELO OPEN-SOURCE

Holo3.1: o agente que não sai da sua sala

A Hcompany abriu a família Holo3.1, agentes de **computer use** — modelos que enxergam a tela e operam o computador como um humano, clicando, digitando e navegando por aplicativos. A novidade que importa não é só a competência, é onde ela roda: tudo na própria máquina (blog técnico — Hugging Face). O modelo vem em quatro tamanhos e, pela primeira vez na linha, com versões quantizadas — quantização é o processo de comprimir o modelo para ocupar menos memória e rodar em hardware comum, com perda mínima de qualidade.

BENCHMARK	SCORE	O QUE MEDE
Benchmark	**Holo3 → Holo3.1**	O que mede
AndroidWorld (35B)	**67% → 79,3%**	tarefas reais em aplicativos de celular
AndroidWorld (4B/9B)	**58% → 72%**	as mesmas tarefas, em modelos que cabem em hardware modesto

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

Sobre o Holo3, de abril, o salto é de mais de 25% em todos os tamanhos, e as versões comprimidas em OSWorld ficam a apenas dois pontos da versão de precisão cheia — ou seja, encolher o modelo quase não custou desempenho. O Holo3.1 também passou a falar a língua dos frameworks de agente de terceiros, integrando-se a sistemas já montados, em vez de exigir o seu próprio.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

Aqui está o ponto que fala direto ao operador do direito: um agente que opera o computador sem nada sair da rede do escritório, gabinete ou assessoria. Sigilo profissional e proteção de dados pessoais (a LGPD) deixam de ser obstáculo intransponível à automação quando o

documento do cliente não trafega para um servidor de terceiros. A tarefa repetitiva — preencher um sistema, extrair dados de um portal de tribunal, conferir uma planilha — pode ser delegada a uma máquina que nunca tira o arquivo de dentro de casa.

> **IMPACTO PRÁTICO**
>
> Por anos, a escolha pareceu binária: ou você manda os dados do cliente para a nuvem de alguém e ganha automação, ou guarda o sigilo e faz tudo na mão. O agente local desfaz o falso dilema. Não é mágica nem é para amanhã — exige máquina razoável e mão técnica para montar. Mas pela primeira vez a privacidade deixou de ser o preço da eficiência.

Outros modelos da semana

A NVIDIA abriu o Cosmos 3 e o Nemotron 3 Ultra — o primeiro, um modelo multimodal aberto (lida com texto, imagem, vídeo e áudio juntos) que lidera vários rankings de geração; o segundo, um modelo de linguagem de pesos abertos com 550 bilhões de parâmetros, posicionado como o melhor open-weight dos EUA. “Pesos abertos” quer dizer que qualquer um pode baixar e rodar o modelo no próprio servidor, sem depender de uma API paga — o que, para dados sensíveis, muda o cálculo. A OpenAI reformou a memória do ChatGPT com o Dreaming, que passa a sintetizar sozinho o histórico de conversas em segundo plano e ficou barato o bastante (cerca de cinco vezes menos cálculo) para chegar aos usuários gratuitos. E a JetBrains lançou o Mellum2, modelo aberto de código em arquitetura MoE (12 bilhões de parâmetros, só 2,5 bilhões ativos) com inferência mais de duas vezes mais rápida que pares do mesmo porte — feito para tarefas de apoio como roteamento e sumarização, não para ser a estrela.

Pesquisa e métodos

O paper mais inquietante da semana é o Safety Paradox: testando 30 modelos abertos mais o GPT-5 e o Claude 4.6, os autores mostram, com prova formal, que quanto melhor o modelo reconhece um pedido perigoso, mais preciso fica um ataque que explora justamente esse reconhecimento. Alinhamento de segurança melhor pode, paradoxalmente, abrir uma porta nova — um lembrete de que “modelo seguro” não é um selo definitivo. Em agentes, o Retrospective Harness Optimization mostrou um agente que se auto-otimiza relendo as próprias tentativas passadas: a taxa de acerto no SWE-Bench Pro subiu de 59% para 78% numa única rodada, sem nenhum avaliador externo.

Na fronteira do raciocínio puro, o Benchmarks in Leipzig reuniu 49 matemáticos para criar 100 questões de nível de pesquisa; o número de perguntas que nenhum modelo resolvia caiu de 41

para 2 ao longo de três rodadas, com os modelos de "raciocínio pesado" dominando a fase final. E o Less is MoE achou um jeito de comprimir modelos MoE pela metade preservando a capacidade — com a descoberta curiosa de que remover só 12 de 1,35 milhão de dimensões certas já derruba todo o desempenho em matemática, sinal de quão concentrado é o conhecimento desses modelos.

Ferramentas do ecossistema

- Claude Code ganhou modelo de reserva (até três alternativas acionadas em ordem quando o principal está sobrecarregado), garantindo que o trabalho não pare quando um servidor congestiona.
- vLLM 0.22 amadureceu o suporte ao modelo aberto chinês DeepSeek-V4, com ganho de quase 29% de latência — o motor que faz esses modelos rodarem em escala ficou mais rápido e barato.
- Transformers 5.10 incorporou o Gemma 4 unificado, do Google, que dispensa módulos separados de visão e áudio e projeta pixels e som direto no modelo — arquitetura mais simples, mesma competência.
- Ollama passou a rodar o Gemma 4 12B localmente, melhorando a quantização em chips Apple — mais um tijolo na pilha de "IA capaz sem mandar dado para fora".
- A nova CLI do Hugging Face foi redesenhada para agentes de código e corta o consumo de tokens em até seis vezes nas tarefas complexas — economia direta na conta de quem automatiza.

Análise: o assistente virou agente, e a conta da responsabilidade chegou

Ethan Mollick, que escreveu o livro sobre "co-inteligência" (a IA como parceira do humano), publicou um ensaio admitindo que o paradigma mudou: de assistente que ajuda, a IA virou agente que age sozinho e, em várias tarefas de valor, supera o humano. A semana deu a ele razão de sobra — e também o contraponto sombrio. Um estudo da Andon Labs, que põe agentes para tocar negócios reais por semanas, registrou modelos mentindo, recusando reembolsos e até formando cartéis de preço sem ninguém mandar — e indícios de que percebem quando estão sendo testados. E a Meta viu seu agente de suporte entregar contas do Instagram a um atacante que só pediu educadamente, sem golpe sofisticado.

Não por acaso, contenção virou pauta. A Anthropic detalhou como isola o Claude para que credenciais nunca entrem no ambiente onde o modelo age, e a OpenAI lançou um "modo de bloqueio" que corta a saída de dados para barrar vazamento por manipulação. A lição para o

jurídico é direta: quando um agente age, surge a pergunta de quem responde pelo ato. Auditabilidade — o registro de cada decisão que o agente tomou — deixa de ser refinamento técnico e vira requisito de quem vai prestar contas.

Brasil

Programa nacional quer formar até 180 pesquisadores em IA

Foi lançado o Industr.IA, programa de pesquisa aplicada em IA para a indústria, executado pelo Instituto Atlântico e pelo iRede com coordenação da Softex. São cerca de R$ 46 milhões em quatro anos para formar de 130 a 180 pesquisadores e criar ao menos 15 núcleos de pesquisa pelo país, cobrindo IA generativa, modelos de linguagem, visão computacional e edge AI (IA que roda no próprio dispositivo, sem nuvem). É aposta em capital humano, não um produto — e por isso mesmo a mais estrutural da semana brasileira.

TIM e Embraer aplicam IA em casos concretos

A TIM passou a usar IA para vigiar o consumo de energia de suas 136 usinas: o sistema monta o padrão esperado de cada unidade e compara com o faturado para flagrar inconsistências de medição (sem percentual de economia divulgado). E a Atech, do grupo Embraer, fechou parceria com o ICMC/USP para prever trajetórias de voo em quatro dimensões com aprendizado de máquina, mirando o controle de tráfego aéreo — projeto ainda em início, sem resultados medidos.

> **IMPACTO PRÁTICO**
>
> O contraste da semana é didático. Lá fora, meia dúzia de plataformas jurídicas agênticas e três modelos de ponta no mesmo intervalo de sete dias. Aqui, aplicações pontuais e um programa para formar gente — nenhum modelo brasileiro novo, nenhuma ferramenta jurídica nacional na lista. Não é motivo para lamento, é diagnóstico: o Brasil está investindo onde precisa investir primeiro, que é a base. Mas vale registrar a distância, porque a corrida não espera quem está montando a equipe enquanto os outros já largaram.

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

Dennis Martins · Adriana Aneli

CURADORIA E EDIÇÃO · 6 DE JUNHO DE 2026