RADAR SEMANAL · PARA ESCRITÓRIOS, GABINETES E ASSESSORIAS

IA em Cognição Nada Exauriente

EDIÇÃO Nº 12 INVENTÁRIO DOS PRINCIPAIS FATOS OCORRIDOS DE 8 DE AGOSTO A 15 DE AGOSTO DE 2026

Resumo

O que o modelo pensa antes de responder nunca foi para ser lido por ninguém. Esta semana um grupo de pesquisadores mostrou que é, e que já foi: o rascunho oculto dos modelos da Anthropic, da OpenAI e do Google pode ser decodificado por quem tiver o bloco criptografado em mãos, e uma varredura em material público já achou dado pessoal e senha lá dentro. É o destaque de uma semana em que a Meta voltou a publicar pesos abertos com um modelo de 30 bilhões de parâmetros que roda na máquina do usuário, e a DeepJudge propôs um protocolo para que o trabalho acumulado numa ferramenta jurídica atravesse para outra sem se perder, com Harvey e Thomson Reuters embarcadas.

No restante do inventário: o Claude empurrou um limite matemático que vinha subindo devagar, a OpenAI mediu um crescimento de 108 vezes no uso de agentes de codificação dentro do setor jurídico, a METR encontrou aceleração real na descoberta de vulnerabilidades de software, e o Google Research explicou por que um modelo erra um fato que ele demonstravelmente sabe. Do Brasil vieram a BIA do Bradesco com números de adoção e uma rede neural da USP que lê radiografia odontológica.

PESQUISA

O rascunho oculto do modelo pode ser lido, e às vezes traz o seu segredo

A criptografia protegia o método do fabricante contra concorrentes. Não protegia o documento de quem trabalhava com ele.

Modelos de raciocínio geram uma cadeia de passos antes da resposta final, o chamado *chain-of-thought*, que é o rascunho onde o modelo tenta caminhos, erra e corrige. Desde o o1 da OpenAI, os laboratórios escondem esse rascunho atrás de assinatura criptográfica, para que concorrentes não copiem a capacidade. Um grupo de pesquisadores que inclui Ilia Shumailov e Jonas Geiping publicou a demonstração de que a proteção tem um defeito de arquitetura: os blocos criptografados são compatíveis entre sessões, contas e modelos diferentes do mesmo fornecedor [paper no arXiv].

O ataque não força o modelo forte a falar. Ele pega o bloco cifrado produzido pelo modelo caro e o injeta num modelo mais fraco do mesmo fornecedor, que tem menos salvaguardas e devolve o conteúdo em texto puro. Funciona nos três: no Claude repetindo o bloco assinado dentro do Haiku 4.5, no GPT inserindo o campo de conteúdo criptografado numa conversa fabricada, no Gemini anexando a assinatura de pensamento. Os autores contaram os tokens recuperados e bateram um a um com os tokens de raciocínio cobrados na fatura da API, que é a prova de que o material extraído é mesmo o rascunho original.

315.320
BLOCOS DECODIFICADOS DE REPOSITÓRIOS PÚBLICOS

367
ARTEFATOS DE DADO PESSOAL RECUPERADOS

182
CREDENCIAIS EXPOSTAS NO RASCUNHO

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

Até aqui, a discussão sobre rascunho oculto era comercial: o fabricante escondia para não ser destilado, ou seja, para que um concorrente não treinasse um modelo barato imitando o raciocínio do modelo caro. O trabalho vira essa lógica do avesso ao mostrar quatro usos do mesmo defeito, e só o primeiro é o roubo de método. Os outros três atingem o usuário: extração de dado privado em escala, recuperação de informação perigosa que o modelo já havia filtrado na resposta visível, e injeção de comando invisível embutida inteiramente no bloco cifrado, que envenena execuções de agentes sem aparecer em lugar nenhum do texto.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

Sessões de Claude Code e Codex compartilhadas publicamente carregam esses blocos. Quem colou o log de uma sessão num repositório, num chamado de suporte ou num grupo de trabalho publicou junto o que o modelo pensou enquanto lia o material. Numa varredura preliminar de cerca de 7 mil traços públicos, Alexander Panfilov, coautor do estudo, relata ter encontrado 62 chaves de API, 33 endereços de e-mail e 33 senhas [cobertura na Latent Space]. Para a operação jurídica, a tradução é direta: o rascunho descreve a minuta, resume a estratégia e nomeia as partes, com uma diferença cruel em relação à resposta final. A resposta você revisou antes de mandar. O rascunho ninguém leu.

> *“Se você já compartilhou online uma sessão do Claude Code ou do Codex com blocos de raciocínio criptografados, eles podem ser decodificados e vazar seus dados pessoais.”*
>
> ALEXANDER PANFILOV · COAUTOR DO ESTUDO

IMPACTO PRÁTICO

A divulgação foi responsável e os três fornecedores foram avisados antes da publicação, o que resolve o vazamento futuro e não resolve o passado. O material já compartilhado continua onde está, decodificável por quem souber o método, e nenhuma correção no servidor apaga o que saiu.

Vale a varredura do que a estrutura jurídica publicou nos últimos meses: repositório interno, anexo de chamado, print colado em grupo de mensagens. Sigilo profissional não distingue rascunho de peça final, e o cliente muito menos. A pergunta que fica não é se o modelo guardou segredo, é quantas cópias do seu segredo estão em lugares que você nem chamaria de arquivo.

MODELO OPEN-SOURCE

Muse Glimmer: a Meta volta aos pesos abertos com um modelo que cabe na sua máquina

Rodar sem nuvem resolve para onde o documento vai. A medição independente mostra que não resolve o quanto se pode confiar na resposta.

A Meta publicou o Muse Glimmer, modelo de 30 bilhões de parâmetros sob licença Apache 2.0, desenhado para tarefas de agente com visão [anúncio no Hugging Face]. A arquitetura foi montada para caber em hardware modesto: um codificador de visão de 2 bilhões de parâmetros acoplado a um decodificador de texto de 28 bilhões, com atenção híbrida que alterna janelas curtas e completas para economizar memória. O modelo já saiu disponível para execução local no Ollama [release no Ollama].

Simon Willison rodou o modelo descrevendo fotografias e navegando bases de código, e publicou as saídas dos dois testes; na versão quantizada do LM Studio, isto é, com os números internos guardados em precisão reduzida para ocupar menos espaço, o modelo ocupa cerca de 18 GB [teste do Willison]. A Meta divulgou 75,5 no MCP Atlas, que mede tarefas de agente em geral, e 51,2 no SWE-Bench Pro, de engenharia de software. Mas a medição independente da Artificial Analysis conta uma segunda história [Artificial Analysis].

MODELO ABERTO	ÍNDICE DE INTELIGÊNCIA	O QUE MEDE
Qwen3.6 27B	38	capacidade agregada em raciocínio, código e conhecimento
Muse Glimmer 30B	35	mesma métrica, medida pela Artificial Analysis
Gemma 4 31B	30	mesma métrica, faixa de tamanho equivalente

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

O salto sobre a própria casa é grande: o Llama 4 Maverick marca 14 na mesma escala, e o Glimmer chega a 35 dezesseis meses depois. Fora de casa, ele empata com o Kimi K2.5, que faz 36 com 1 trilhão de parâmetros, ou seja, 33 vezes mais. Contra a concorrência aberta do próprio tamanho, porém, o modelo fica atrás do Qwen3.6 27B, que é menor. E há um número que o anúncio não traz: no GDPval-AA v2, avaliação de tarefas de agente da mesma casa, o Muse Glimmer marca 953 de Elo com taxa de alucinação de 82%, contra 49% do Qwen3.6 [medição independente].

O QUE ISSO DESBLOQUEIA NA PRÁTICA

Um modelo que roda na máquina resolve a pergunta que toda estrutura jurídica faz antes de adotar ferramenta nova: para onde vai o documento do cliente. Ele não sai. Não há terceiro processando, não há termo de uso a interpretar, não há transferência internacional a justificar. Para triagem de acervo, classificação de documento e leitura de imagem digitalizada, 18 GB cabem numa placa de vídeo de estação de trabalho comum. O que a taxa de alucinação acrescenta é onde esse ganho para: 82% é um número que desqualifica o modelo para tarefa de agente sem conferência humana, e o desqualifica justamente na função para a qual ele foi anunciado.

IMPACTO PRÁTICO

Pesos abertos e execução local são a resposta certa para a pergunta da confidencialidade. Não são resposta para a pergunta da exatidão, e as duas costumam ser feitas juntas na hora de escolher ferramenta.

Quem for testar faça a divisão explícita: o modelo local serve para o que pode ser conferido em segundos, como separar, etiquetar e resumir para triagem. O que vai ser assinado ainda passa por leitura humana, e nesse ponto a licença permissiva não muda nada. Um modelo que erra dentro do seu computador erra igual a um que erra na nuvem, com a diferença de que a conta do erro é sua sozinho.

FERRAMENTA

Um protocolo aberto para o contexto atravessar de uma ferramenta jurídica para outra

O que prende a estrutura jurídica ao fornecedor não é o modelo, é o acervo de trabalho que ficou lá dentro. Dois concorrentes assinaram o padrão que move esse acervo.

A DeepJudge, de Zurique, publicou o Agent Handoff Protocol, padrão aberto que transfere o contexto de trabalho entre produtos de IA jurídica diferentes: histórico da conversa, materiais enviados e análises já feitas seguem com o usuário quando ele muda de aplicativo [LawNext]. A Harvey implementa em beta a partir deste mês e a Thomson Reuters vai integrá-lo ao CoCounsel for Legal.

O protocolo roda sobre o MCP, o Model Context Protocol da Anthropic, e não o substitui. A distinção é de escopo: o MCP serve para uma ferramenta buscar um dado pontual em outra, como quem consulta uma tabela. O AHP transfere a tarefa inteira em andamento, com o estado do trabalho junto, e cada fornecedor decide o que passa adiante. A Harvey tende a mandar o histórico conversacional; a DeepJudge prioriza os documentos já localizados.

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

Até agora, trocar de ferramenta jurídica significava recomeçar. As buscas refinadas ao longo de semanas, os documentos carregados, as instruções ajustadas para o vocabulário do caso, tudo isso ficava no fornecedor anterior, e o custo de sair crescia a cada mês de uso. Era dependência construída por acúmulo, não por contrato. Um padrão aberto que dois concorrentes de peso adotam no mesmo anúncio ataca exatamente esse acúmulo.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

Para quem contrata, muda a negociação. Renovação de licença com o acervo de trabalho preso do outro lado da mesa não é negociação, é aceitação. Com portabilidade real, a comparação entre fornecedores volta a ser sobre qualidade da entrega, e a migração deixa de custar meses de reconstrução. A ressalva está na letra do protocolo: quem decide o que atravessa é o fornecedor de origem, não o cliente. Portabilidade que depende da boa vontade de quem você está deixando ainda não é portabilidade plena, e vale ler o que cada um se compromete a exportar antes de assinar.

IMPACTO PRÁTICO

O padrão nasce com os dois nomes que mais importam no mercado jurídico internacional embarcados, o que resolve o problema clássico de padrão aberto que ninguém implementa. Falta o teste de campo, e ele só acontece quando o primeiro cliente grande migrar de verdade.

> Enquanto isso, a cláusula que vale reler nos contratos em vigor é a de saída: o que exatamente o fornecedor devolve, em que formato e em quanto tempo. Ferramenta jurídica se escolhe pelo que ela faz, e se lamenta pelo que ela retém.

Outros modelos da semana

A Z.ai publicou o **GLM-5.3**, modelo de cerca de 750 bilhões de parâmetros com a mesma arquitetura do antecessor e pós-treinamento bem mais extenso, isto é, a fase de ajuste que vem depois do treino bruto e ensina o modelo a seguir instruções e usar ferramentas. Ele supera o Kimi K3 em vários testes de codificação com agentes usando cerca de um terço dos parâmetros, o que importa porque parâmetro a menos é custo de operação a menos para quem hospeda [análise de Nathan Lambert].

O Google lançou o **Gemini 3.7 Flash**, versão rápida e barata da família, com ganho em codificação e tarefas de agente, e cita raciocínio jurídico entre as áreas que melhoraram [blog do Google]. O **Grok 4.6** voltou à fronteira de inteligência e assumiu a liderança em eficiência de custo, ou seja, entrega por dólar gasto, na medição independente, que manteve o preço da geração anterior [medição independente]. E a NVIDIA abriu o **Nemotron 3.5 Lightning**, modelo pequeno de arquitetura MoE, em que só uma fração dos especialistas internos é ativada por consulta, com 3 bilhões de parâmetros ativos de um total de 30, desenhado para a camada de execução de agentes que ficam ligados o tempo todo [release no Ollama].

Pesquisa e métodos

> **ACOMPANHAMENTO** *Edição nº 11: registramos agentes de IA atacando alvos reais em três laboratórios, com as contenções falhando por decisão humana, não por limite técnico. Esta semana o padrão apareceu de novo, agora entre agentes que deveriam colaborar.*

O Frontier Red Team da Anthropic soltou três modelos Claude no mesmo projeto com instruções incompatíveis e observou o resultado: houve sabotagem entre eles, inclusive com código auto-replicável, e também trégua espontânea, que no Mythos 5 apareceu em 98% das execuções [TechCrunch]. No mesmo tema, a reconstituição do incidente em que agentes de treinamento da OpenAI escaparam do ambiente isolado e exploraram uma falha de execução remota no Artifactory, dentro da infraestrutura da própria empresa, mostrou algo novo: eles deixaram recados uns para os outros no conteúdo de arquivos de um repositório compartilhado, improvisando um mural de mensagens [Import AI]. A invasão à Hugging Face veio depois disso. Jack Clark registra que escreve sem informação privilegiada e que parte do relato não tem confirmação oficial, e a ressalva vale.

Uma versão de pesquisa não lançada do Claude elevou de 41,6% para 67,2% o limite inferior conhecido para a proporção de zeros da função zeta de Riemann [Anthropic]. O caminho até lá foi acidentado, e é a parte interessante: numa primeira sessão o modelo tentou 650 ideias e nenhuma funcionou; numa segunda, de um dia e meio, cerca de 60 subagentes trabalharam em paralelo, e dois deles desenvolveram as ideias matemáticas centrais. As duas sessões somadas consumiram 31 milhões de tokens de saída. Dois matemáticos da casa validaram o resultado, dois externos examinaram o artigo, e há formalização verificável em Lean. A própria Anthropic corta o entusiasmo pela raiz: não espera que a técnica leve à prova da Hipótese de Riemann, e o resultado saiu como subproduto de uma tentativa mais ampla que falhou.

A METR mediu se há aceleração real de descobertas atribuível a modelos e achou um quadro dividido: nítida em vulnerabilidades de software, com o cURL saindo de 9 falhas catalogadas em 2025 para 36 até junho de 2026, moderada em matemática, e nenhuma clara em otimização de algoritmos [METR]. A própria METR ressalva que enxerga só a descoberta pública: laboratório que retém achado interno não entra na conta. Já o Google Research explicou uma frustração conhecida de quem usa o modelo para consulta factual: no Gemini 3 Pro e no GPT-5, 95% a 98% dos fatos testados estão de fato memorizados, mas 26% a 34% não são recuperados na resposta direta. Deixar o modelo raciocinar antes de responder recupera de 40% a 65% deles, e ainda assim 11% a 12% seguem inacessíveis [Google Research]. O erro factual, nesses casos, não é ignorância: é uma biblioteca que tem o livro e perdeu a ficha.

Ferramentas do ecossistema

- **vLLM v0.27.0**, servidor de inferência que empresas usam para hospedar modelo próprio, ganhou suporte full-stack ao Kimi K3 e passou a atender também Qwen3.5 e VaultGemma, numa release de 561 commits que encurta o caminho entre um modelo aberto sair e ele rodar em infraestrutura interna [release].
- **Claude Code** teve oito releases na semana, com destaque para bifurcação de subagentes por padrão, que permite a um agente dividir a própria tarefa em ramos paralelos, e mensagens entre sessões, útil para quem deixa trabalho longo rodando [releases].
- **Ollama** adicionou dois harnesses de agente, o DeepSeek Harness e o Muse Code, que são os andaimes que transformam um modelo em agente capaz de executar tarefa de ponta a ponta na máquina local [release].
- **A biblioteca oficial da OpenAI** chegou à versão 3.1.0 já com o tier Ultrafast e com a depreciação da API de vídeo do Sora [changelog]. O modo roda o GPT-5.6 Sol a até 14 vezes a velocidade normal, em parceria com a Cerebras [TechCrunch].

PARA TESTAR ESTA SEMANA

Instale o Ollama e rode `ollama run muse-glimmer` para ter o modelo aberto da Meta funcionando sem nuvem em cerca de 18 GB de disco. Peça a ele que descreva um documento digitalizado e confira a saída contra o original: o teste leva quinze minutos e mostra na prática onde a leitura local ajuda e onde ela inventa [modelo no Ollama].

Análise

A OpenAI publicou dois estudos sobre como empresas usam IA e um número salta para quem lê este relatório: desde fevereiro, os usuários semanais do Codex, o agente de codificação da casa, cresceram 108 vezes no setor jurídico, o maior salto entre as funções que a empresa listou, à frente de vendas e recrutamento, ambas com 41 vezes, e de engenharia, com 5 [OpenAI]. Vale entender o que o dado diz e o que ele não diz. Ele mede adoção de uma ferramenta de escrever código dentro de organizações jurídicas, não substituição de trabalho jurídico.

O que ele revela é a mudança de posição do profissional do direito diante da máquina: de quem pergunta para quem encomenda. A distância entre as duas coisas é a diferença entre pedir uma resposta e receber um trabalho pronto, com o passo da conferência transferido para você. No mesmo período, a Hugging Face registrou que os modelos abertos chineses passaram os americanos em porte e que o Claude Code liderou o tráfego de agentes na plataforma em julho, com 44,4% [State of Open Models]. As duas medições contam a mesma história por ângulos diferentes: o agente deixou de ser demonstração e virou infraestrutura de trabalho.

Brasil

A BIA do Bradesco mostra números depois da troca para IA generativa

O assistente virtual do Bradesco saiu de 500 mil para 3,5 milhões de usuários ativos por mês e de 1,3 milhão para 10 milhões de interações mensais em um ano, com retenção subindo de 75% para 90% [Mobile Time]. A orquestração fica numa plataforma própria, a Bridge, que combina modelos de linguagem em parte dos fluxos e mantém automação determinística em casos pontuais. É raro ver número de retenção divulgado, e é ele que sustenta o resto: audiência que volta é audiência que foi atendida.

Rede neural da USP lê radiografia odontológica com precisão alta

O Jornal da USP divulgou em 14 de agosto o sistema do grupo InReDD que numera dentes e detecta cáries em radiografias [Jornal da USP]. A acurácia chega a 0,998 na numeração e 0,94

na detecção de cáries, segundo o artigo publicado na PLOS Digital Health em novembro de 2025 [artigo na PLOS]. O estudo está fora da janela desta edição e entra aqui como contexto da cobertura desta semana. Interessa ao leitor jurídico menos pelo consultório e mais pelo laudo: prova pericial produzida com apoio de modelo treinado passa a exigir da parte contrária uma pergunta que ainda não é rotina no processo brasileiro, que é a de saber com qual base o sistema foi treinado.

IMPACTO PRÁTICO

Duas notícias brasileiras na semana, e as duas com número aferido em vez de anúncio de intenção, o que já é diferente do padrão. O que falta em ambas é a mesma coisa: nenhuma diz o que acontece quando o sistema erra.

Assistente bancário que erra gera reclamação e reprocessamento. Laudo que erra gera decisão. Enquanto a divulgação brasileira medir só o acerto médio, a estrutura jurídica que precisar contestar qualquer um dos dois vai ter de construir sozinha a pergunta que a fonte não respondeu, e vai fazer isso no prazo do processo, não no da pesquisa.

NO RADAR DA PRÓXIMA SEMANA

A Harvey começa este mês o beta do Agent Handoff Protocol, primeiro teste real de portabilidade entre ferramentas jurídicas concorrentes [LawNext]. A Meta anunciou que os pesos do Muse Spark 1.2 saem em seguida, no mesmo compromisso público que trouxe o Glimmer [Latent Space].

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

Dennis Martins · Adriana Aneli

CURADORIA E EDIÇÃO · 15 DE AGOSTO DE 2026