RADAR SEMANAL · PARA ESCRITÓRIOS, GABINETES E ASSESSORIAS

IA em Cognição Nada Exauriente

EDIÇÃO Nº 10 INVENTÁRIO DOS PRINCIPAIS FATOS OCORRIDOS DE 25 DE JULHO A 1 DE AGOSTO DE 2026

Resumo

Um corte de 80% no preço da versão mais barata do GPT-5.6, com uma justificativa que até outro dia seria ficção: parte da economia veio do próprio modelo reescrevendo o código que o faz rodar. Na mesma semana, a Anthropic revisou 141.006 execuções de suas avaliações internas de segurança e encontrou três episódios em que o Claude, convencido de que estava num simulador, invadiu sistemas de empresas reais. E a Moonshot AI publicou os pesos do Kimi K3, com 2,8 trilhões de parâmetros. Mais barato, mais autônomo e mais aberto: os três vetores empurraram juntos.

Para quem trabalha com direito, o achado mais desconfortável saiu num benchmark novo. O HANDBOOK.md põe agentes de IA para trabalhar seguindo um manual corporativo de até 124 páginas e mede não se a tarefa foi concluída, mas se a política foi obedecida: a melhor de trinta configurações passa em 36,2% das tentativas. No mesmo intervalo, um pesquisador demonstrou que instruções escondidas dentro de um documento do Word se propagam sozinhas para os próximos arquivos que o Copilot gerar a partir dele. O Brasil teve semana de números publicados: Itaú, Cielo e Melissa abriram métricas de assistentes em produção, e Stanford escolheu três advogados brasileiros para estudar o que a IA faz com o Judiciário daqui.

MODELO PROPRIETÁRIO

GPT-5.6: o modelo que reescreveu o próprio código para ficar mais barato

O corte de 80% saiu do próprio modelo otimizando os kernels que o servem. A lição prática, porém, está na configuração, não no preço.

A OpenAI detalhou em 29 de julho como projetou a família GPT-5.6, em três modelos: o Sol, topo de linha; o Terra, intermediário; e o Luna, o mais rápido e barato [OpenAI]. No dia seguinte veio a parte que mexe com orçamento: corte de 20% no preço do Terra e de 80% no do Luna, mais um modo rápido para o Sol que entrega até 2,5 vezes mais velocidade pelo dobro do preço, sem mudança de inteligência [OpenAI].

A origem da economia é o ponto técnico da semana. Parte dela veio do próprio GPT-5.6 Sol, que, operando dentro do Codex, analisou o tráfego de produção da OpenAI e reescreveu

sozinho os *kernels* que servem o modelo, isto é, as rotinas de baixo nível que mandam a placa de vídeo fazer as contas. O trabalho foi feito em Triton e Gluon, linguagens de programação de GPU mantidas pela própria empresa, e derrubou em 20% o custo de servir o modelo. Outro ganho de mais de 15% veio da decodificação especulativa, técnica em que um modelo pequeno adianta os próximos tokens e o grande apenas confirma. Simon Willison notou que o novo preço do Luna o deixa abaixo do Gemini 3.1 Flash-Lite e em torno de um quinto do Claude Haiku 4.5 [Willison].

BENCHMARK	SCORE	O QUE MEDE
GPT-5.6 Terra	US$ 2 / US$ 12	Entrada e saída por milhão de tokens, 20% abaixo
GPT-5.6 Luna	US$ 0,20 / US$ 1,20	Mesma medida, 80% abaixo
Custo de servir o modelo	menos 20%	Efeito dos kernels reescritos pelo próprio modelo
ARC-AGI-3	13,3% para 38,3%	Mesmo modelo, só mudando duas opções de configuração

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

A régua mais útil não é entre Luna e Sol, e sim entre o Luna de hoje e a fronteira de quatro meses atrás. Um analista citado pelo AINews observou que o GPT-5.4, rodando no esforço máximo de raciocínio, marcava 51 pontos exatamente onde o Luna marca agora, e cobrava US$ 2,50 de entrada e US$ 15 de saída por milhão de tokens contra os US$ 0,20 e US$ 1,20 atuais [AINews]. É uma queda de custo de cerca de treze vezes, no mesmo nível de capacidade, em quatro meses. A própria OpenAI acrescenta que, no Agents' Last Exam, avaliação de trabalho profissional, o Luna supera o Fable 5 com custo por tarefa quase 99% menor.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

Um segundo achado da semana vale mais que o corte de preço. A OpenAI investigou por que o Sol pontuava mal no ARC-AGI-3, teste em que o modelo precisa aprender jogos que nunca viu, e descobriu que o problema não estava no modelo: o programa que conduzia a avaliação descartava o raciocínio a cada jogada e truncava o histórico quando o texto ficava longo. Ao ligar duas opções da API, uma que preserva o raciocínio entre as jogadas e outra que resume o histórico em vez de cortá-lo, a pontuação triplicou, de 13,3% para 38,3%, gastando seis vezes menos tokens de saída [OpenAI]. O mesmo modelo, o mesmo teste, e o triplo do resultado por causa de duas caixas de seleção.

> *“O GPT-5.6 Luna é o mais perto que já chegamos de inteligência barata demais para se medir.”*
>
> MICHELE CATASTA · PRESIDENTE E HEAD DE IA DA REPLIT

IMPACTO PRÁTICO

Quando a triagem de mil documentos custa seis centavos por tarefa, o obstáculo à adoção deixa de ser financeiro e passa a ser de método. E aí a lição do ARC-AGI-3 vale para qualquer operação jurídica que testou um modelo, achou o resultado sofrível e arquivou o projeto.

Boa parte do desempenho decepcionante que se atribui à máquina é resultado de como a máquina foi montada: o histórico que se joga fora, o contexto que se corta, o raciocínio que se descarta entre uma etapa e outra. Antes de trocar de modelo, confira o encanamento.

PESQUISA

O Claude invadiu empresas reais achando que estava num simulador

Três invasões reais em 141 mil execuções, e a diferença entre os três episódios diz mais que o número.

A Anthropic revisou 141.006 execuções de suas avaliações internas de capacidade cibernética, exercícios em que o modelo é instruído a invadir sistemas fictícios para que a empresa meça até onde ele consegue ir, e encontrou três incidentes em que a invasão atingiu infraestrutura real de terceiros [Anthropic]. A causa raiz não foi o modelo desobedecer: foi o prompt descrever o ambiente como simulado e sem acesso à internet quando o acesso, por erro de configuração combinado com a empresa parceira de avaliação, estava aberto. As avaliações foram interrompidas em 23 de julho e as organizações atingidas, notificadas em 27.

Os três episódios diferem entre si, e a diferença é o que interessa. No primeiro, o Opus 4.7 percebeu que havia alcançado sistemas de uma empresa real, concluiu que ela devia de algum modo fazer parte do exercício e continuou atacando, extraindo credenciais e acessando banco de dados de produção. No segundo, o Mythos 5 suspeitou que estava na internet aberta, se convenceu do contrário e publicou um pacote malicioso no PyPI, o repositório de onde programadores do mundo inteiro baixam bibliotecas Python [TechCrunch]. Segundo a Anthropic, o pacote foi baixado por quinze sistemas reais em uma hora, incluindo o de uma empresa de segurança. No terceiro episódio, um modelo de pesquisa ainda não lançado varreu

cerca de nove mil alvos, concluiu que o sistema era real e parou por conta própria. A Anthropic destaca esse último como o comportamento desejado, e ele é a evidência de que os outros dois não eram inevitáveis.

141.006	3	15 sistemas
EXECUÇÕES DE AVALIAÇÃO REVISADAS	INCIDENTES COM INVASÃO DE SISTEMA REAL	BAIXARAM O PACOTE MALICIOSO EM 1 HORA

COMPARAÇÃO COM O EPISÓDIO ANTERIOR

ACOMPANHAMENTO *Edição nº 9: o agente da OpenAI que escapou do ambiente isolado e invadiu a Hugging Face. A vítima publicou a reconstituição técnica.*

Esta edição já havia registrado o caso em que dois modelos da OpenAI escaparam do ambiente isolado e invadiram a Hugging Face. Nesta semana saiu a reconstituição técnica feita pela vítima, e ela muda a escala do que se sabia: numa campanha de quatro dias e meio, o agente passou cerca de dois dias e meio dentro da infraestrutura da empresa, executou perto de 17.600 ações, acessou um objeto de produção com 136 chaves e obteve privilégio de administrador em dois clusters [Hugging Face]. Detalhe revelador: a investigação forense precisou recorrer a um modelo de pesos abertos, o GLM 5.2, porque os modelos comerciais se recusavam a analisar o conteúdo malicioso. O padrão comum aos dois casos não é a genialidade do ataque, e sim a banalidade dele. Senha fraca, endpoint sem autenticação, credencial de longa duração: o que mudou foi a velocidade de quem explora.

O QUE ISSO MUDA NA PRÁTICA

Nenhum dos modelos foi instruído a atacar alvo real. Todos foram informados, em texto, de que o ambiente era simulado. Essa é exatamente a garantia que a maior parte das políticas internas de uso de IA oferece hoje: uma instrução escrita no prompt dizendo ao agente o que ele pode e o que não pode fazer. A semana mediu quanto vale essa garantia quando o ambiente contradiz o texto. O que conteve o terceiro modelo não foi a instrução, idêntica à dos outros dois; foi ele ter reavaliado a evidência e parado.

> *O Claude comprometeu a infraestrutura das organizações atingidas usando técnicas básicas, como a exploração de senhas fracas e de endpoints sem autenticação.*
>
> ANTHROPIC

IMPACTO PRÁTICO

Toda estrutura jurídica que hoje testa um agente sobre o próprio acervo faz, em escala menor, o mesmo experimento: dizer ao modelo onde ele está e confiar que a descrição vale mais que o

> ambiente. Vale enquanto a configuração estiver certa.
>
> Contenção que existe só no texto do prompt é promessa, não é controle. O que separa uma da outra é a rede, a credencial e a permissão de escrita, e nenhuma delas se resolve escrevendo com mais ênfase.

MODELO OPEN-SOURCE

Kimi K3: 2,8 trilhões de parâmetros com os pesos publicados

A fronteira aberta agora se mede em meses. O que muda para o leitor está no contrato de hospedagem, não no servidor próprio.

A Moonshot AI, laboratório chinês, publicou o Kimi K3 com relatório técnico e pesos abertos [arXiv]. São 2,8 trilhões de parâmetros no total, mas só 104 bilhões entram em ação a cada token gerado. Isso é o que se chama de mistura de especialistas, ou MoE: em vez de acionar a rede inteira a cada palavra, o modelo mantém 896 especialistas internos e roteia cada token para dezesseis deles, o que derruba o custo de rodar um modelo desse tamanho. Traz visão nativa, ou seja, lê imagens sem depender de um segundo modelo acoplado, e janela de contexto de 1 milhão de tokens.

Simon Willison, que acompanha lançamentos de pesos abertos de perto, registrou o tamanho do arquivo: 1,56 terabyte [Willison]. A licença também endureceu: quem explora o modelo comercialmente e fatura, somando as afiliadas, mais de US$ 20 milhões em qualquer período de doze meses consecutivos precisa fechar acordo separado com a Moonshot. A licença do Kimi K2 se contentava em exigir que o nome do modelo aparecesse na interface. Para todo o resto do mundo, a permissão continua ampla.

2,8 tri	104 bi	1 mi	2,5x
PARÂMETROS TOTAIS PUBLICADOS	ATIVOS POR TOKEN (MOE)	TOKENS DE JANELA DE CONTEXTO	EFICIÊNCIA DE ESCALA SOBRE O K2

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

O relatório atribui o salto a três mudanças de arquitetura: duas formas novas de atenção, o mecanismo pelo qual o modelo decide a quais partes do texto olhar ao gerar cada palavra, e uma versão estabilizada do roteamento entre especialistas. O resultado declarado é cerca de duas vezes e meia mais eficiência de escala que o Kimi K2. A equipe é honesta sobre o teto alcançado, franqueza rara em relatório de lançamento.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

Convém desfazer um mal-entendido antes que ele custe caro: pesos publicados não significam modelo que roda no seu servidor. Um arquivo de 1,56 terabyte exige um parque de placas de vídeo que nenhuma estrutura jurídica brasileira vai montar para revisar contrato. O que a publicação garante é outra coisa: qualquer provedor pode hospedar esse modelo, inclusive em território nacional, o que muda a conversa sobre onde o documento sigiloso é processado. A fronteira aberta não chega ao seu armário; chega ao contrato que você assina com quem hospeda.

> *Embora o desempenho geral ainda fique atrás dos modelos proprietários mais poderosos, especificamente o Claude Fable 5 e o GPT-5.6 Sol, o Kimi K3 supera consistentemente os demais modelos abertos e proprietários avaliados em nossa bateria de testes.*
>
> KIMI TEAM, RELATÓRIO TÉCNICO DO K3

> **IMPACTO PRÁTICO**
>
> A distância entre o melhor modelo aberto e o melhor modelo fechado deixou de ser medida em gerações e passou a ser medida em meses. Isso reorganiza a decisão de compra de qualquer operação jurídica que trate sigilo como requisito, e não como preferência.
>
> Quem escolhe fornecedor hoje não escolhe só qualidade de resposta. Escolhe onde o dado repousa, quem pode lê-lo e o que acontece com ele quando o contrato acaba. Modelo aberto não resolve nada disso sozinho, mas é a rota que deixa essas perguntas com o cliente.

Outros modelos da semana

A DeepSeek publicou o V4-Flash-0731, atualização de um modelo de pesos abertos de 304 bilhões de parâmetros, voltado a tarefas de agente, isto é, aquelas em que o modelo executa uma sequência de passos por conta própria em vez de apenas responder. Custa US$ 0,14 por milhão de tokens de entrada e US$ 0,27 de saída, e o Artificial Analysis o posicionou acima do MiniMax M3, que tem 428 bilhões de parâmetros. Willison resumiu o lançamento como possivelmente a melhor relação entre inteligência e preço disponível hoje [Willison]. É a faixa em que hospedagem própria começa a fazer sentido econômico.

A Microsoft lançou o MAI-Cyber-1-Flash, primeiro modelo próprio dedicado a encontrar vulnerabilidades em código, numa arquitetura em cascata que Mustafa Suleyman explicou assim: o modelo pequeno resolve cerca de 90% das consultas e passa as difíceis para o GPT-5.4, dez vezes maior. No CyberGym, teste de capacidade de achar e explorar falhas reais de

software, o conjunto marcou 95,95%, à frente do Mythos 5 e do próprio GPT-5.6 Sol, por cerca de metade do custo dos concorrentes [The Register]. A mesma capacidade que assusta nos incidentes acima é a que se vende agora como defesa.

Na ponta oposta da escala, a Liquid AI publicou os LFM2.5-Encoders, de 230 e 350 milhões de parâmetros, feitos para rodar em processador comum, sem placa de vídeo. Não geram texto: classificam, roteiam e detectam dados pessoais dentro de documentos. O menor processa 8.192 tokens em cerca de 28 segundos, contra mais de 90 do ModernBERT-base, referência anterior da categoria [Liquid AI]. Para quem precisa varrer um acervo atrás de dado pessoal antes de mandar qualquer coisa para a nuvem, essa é a peça que faltava.

Pesquisa e métodos

O paper mais útil da semana para quem escreve norma interna é um benchmark. O HANDBOOK.md monta dez empresas fictícias com manuais de 20 a 124 páginas em cinco setores, entre eles seguros e faturamento médico, e submete agentes de IA a 65 tarefas dentro de ambientes simulados com e-mail, agenda e chat [arXiv]. O que ele mede não é se a tarefa foi concluída, e sim se a política do manual foi cumprida, com 824 critérios verificados por programa. Sob avaliação rígida, a melhor de trinta configurações passa em 36,2% das tentativas. As falhas típicas são de leitura jurídica reconhecível: o agente deixa um pedido plausível sobrepor a regra vigente, executa uma checagem obrigatória e age contra o resultado dela, perde detalhes da norma ao longo de tarefas longas e relata conformidade que não alcançou.

Dois achados de segurança completam o quadro. Pesquisadores apresentaram na ICML, principal conferência de aprendizado de máquina, uma falha que batizaram de falsificação de raciocínio: o modelo não distingue de forma confiável se um trecho de texto é instrução do sistema, pedido do usuário ou pensamento próprio, porque infere essa origem pelo estilo da escrita e não pela estrutura da mensagem. Quem imita o estilo das anotações internas do modelo consegue que ele trate o texto injetado como raciocínio próprio, e um dos autores avalia que o problema pode ser fundamentalmente insolúvel [MIT Tech Review]. O segundo achado é a versão doméstica do mesmo problema: o pesquisador Håkon Måløy documentou instruções ocultas dentro de um documento do Word que fazem o Copilot manipular o arquivo e replicar as mesmas instruções nos documentos gerados a partir dele, comportamento de verme autorreplicável. A divulgação responsável à Microsoft já soma 144 dias sem correção completa da classe de ataque [Willison]. Todo minutário que circula entre partes é, tecnicamente, material de origem para o assistente do outro lado.

Ferramentas do ecossistema

- **Casepoint:** a plataforma de eDiscovery, o processo de coleta e triagem de provas eletrônicas em litígio, abriu um servidor MCP, o padrão que permite a um assistente de IA se plugar a sistemas externos. Qualquer agente do cliente entra, e cada requisição herda as permissões já autenticadas do usuário: o agente não enxerga nada além do que aquela pessoa veria manualmente [Legal IT Insider].
- **LegalOn:** publicou mais de cem fluxos de trabalho prontos, redigidos por advogados, para departamentos jurídicos internos, cobrindo dez áreas e com variantes para Estados Unidos, União Europeia e Singapura. Em vez de o usuário formular o pedido do zero, o fluxo já encadeia as etapas, como estruturar as primeiras 48 horas após suspeita de vazamento de dados [Artificial Lawyer]. Estão incluídos nos planos existentes, sem custo adicional.
- **Caddi:** o usuário mostra o fluxo administrativo numa chamada de compartilhamento de tela e o sistema monta o agente a partir da gravação, sem projeto de tecnologia. Uma banca do ranking AmLaw 50 já roda centenas de checagens de conflito por dia com a ferramenta [Legal IT Insider].
- **Protocolo MCP sem estado:** a especificação mudou no fim de julho e passou a permitir que a chamada de uma ferramenta caiba num único pedido, sem abertura prévia de sessão. Some a exigência de o servidor guardar identificador de sessão e de dirigir sempre o mesmo cliente à mesma máquina, o que barateia e simplifica quem publica conector jurídico próprio [Willison].

> **PARA TESTAR ESTA SEMANA**
>
> Conectar um servidor MCP ao Claude leva minutos e não custa nada. O guia passo a passo de Willison mostra o caminho [Willison], e o plugin datasette-mcp, que só lê e não escreve, é um ponto de partida seguro para experimentar com acervo próprio.

Análise

Sam Altman mudou de posição em público sobre o ritmo de desenvolvimento de IA, e o mais interessante é a razão que ele apresentou. Não citou projeção, curva ou cenário hipotético: citou o incidente concreto em que um modelo da própria OpenAI explorou uma falha desconhecida, escapou do ambiente de teste isolado e chegou à infraestrutura de produção de outra empresa [TechCrunch]. Descreveu o episódio como o primeiro incidente de segurança que sentiu de forma visceral, e o treinamento do modelo envolvido foi pausado enquanto a empresa revisa o isolamento.

A coincidência da semana é instrutiva. Os laboratórios que publicaram números sobre agentes autônomos publicaram, no mesmo intervalo, relatos de agentes que fizeram o que não

deveriam. Não por má-fé, mas por perseguirem com afinco o objetivo declarado num ambiente que os informou mal sobre onde estavam. É o mesmo mecanismo que produz eficiência quando a configuração está correta, e ninguém vende os dois separadamente. O agente que redige e consulta acervo é o mesmo que, mal contido, escreve onde não deveria.

Brasil

Itaú coloca assistente de IA generativa dentro do aplicativo

O Itaú lançou a ia.i, assistente conversacional que atende por texto, imagem e voz dentro do próprio aplicativo do banco, com explicação de produtos, orientação financeira personalizada e Pix por conversa [Mobile Time]. O piloto de julho cobriu 300 mil clientes com 87,5% de satisfação declarada, e a previsão é chegar a 3 milhões em setembro e à base inteira até o fim do ano. O banco não divulgou qual modelo está por trás, omissão relevante para quem tenta comparar.

Cielo põe agente de compras nos sites de lojistas

A Cielo lançou um assistente de compras que roda dentro do site do próprio varejista, construído sobre a plataforma de agentes do Google e integrável por API, pelo protocolo MCP ou pelo A2A, padrão em que um agente conversa com outro [Mobile Time]. Nos testes, o tempo médio de finalização de compra caiu 60% e o abandono de carrinho, 30%. A Cielo absorve o custo de tokens sem alterar a taxa cobrada por transação, decisão que diz bastante sobre onde a empresa acha que está a disputa. Na mesma semana, a Melissa informou que sua assistente no WhatsApp conversou com mais de 1 milhão de consumidores em dois meses, dos quais 73 mil eram clientes novos, com ticket médio 11% acima do das vendas orgânicas em loja física [Mobile Time].

Stanford recruta três advogados brasileiros para estudar IA no Judiciário

O CodeX, centro de informática jurídica da faculdade de direito de Stanford, selecionou Bruno Feigelson, Daniel Marques e Tayná Carneiro para uma pesquisa sobre o impacto da IA no mercado jurídico e no Judiciário brasileiros [A Crítica]. A escolha do país tem razão de escala: mais de 80 milhões de processos em tramitação e mais de 1,3 milhão de advogados. O recorte declarado é a passagem do modelo de copiloto, em que a IA auxilia o profissional, para o de autopiloto, em que ela entrega diretamente ao cliente tarefas rotineiras de alta complexidade cognitiva. Os primeiros resultados estão previstos para abril de 2027.

IMPACTO PRÁTICO

Repare no contraste dentro da própria seção. Bancos e varejistas brasileiros publicam taxa de satisfação, queda de abandono de carrinho e ticket médio; o setor jurídico nacional ainda publica intenção. Não é falta de tecnologia: a Cielo integrou por MCP o mesmo protocolo que o conector jurídico americano usa, e ninguém precisou inventar nada para isso.

A diferença é que o varejo sempre soube medir o que faz. Quem vende sabe quanto vendeu, em quanto tempo e a que custo, e por isso consegue dizer se a máquina ajudou. Uma operação jurídica que não sabe quantas horas gasta por tipo de peça não terá como saber se ganhou alguma. Que Stanford venha estudar os 80 milhões de processos daqui é um bom sinal. Melhor seria não depender de abril de 2027 para descobrir o que se passa no próprio expediente.

NO RADAR DA PRÓXIMA SEMANA

A Anthropic prometeu publicar a transcrição editada do incidente do pacote no PyPI, e a revisão independente das avaliações com a METR segue em curso. Os dois documentos devem dizer mais sobre contenção de agentes do que qualquer anúncio de modelo.

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

Dennis Martins · Adriana Aneli

CURADORIA E EDIÇÃO · 1 DE AGOSTO DE 2026