RADAR SEMANAL · PARA ESCRITÓRIOS, GABINETES E ASSESSORIAS

IA em Cognição Nada Exauriente

EDIÇÃO Nº 11 INVENTÁRIO DOS PRINCIPAIS FATOS OCORRIDOS DE 1 DE AGOSTO A 8 DE AGOSTO DE 2026

Resumo

A Thomson Reuters resolveu publicar os números do próprio modelo de IA e dizer que ele briga de igual para igual com a fronteira. Bob Ambrogi abriu a tabela e mostrou onde a afirmação se sustenta e onde não: o Thomson lidera em casos jurídicos difíceis e em documento longo, e fica atrás em raciocínio geral e em código. É a primeira vez que uma editora jurídica entra nessa disputa com placar aberto.

Do outro lado da semana, uma repetição que já não dá para chamar de coincidência. O instituto de segurança de IA do Reino Unido publicou que, em 122 avaliações conduzidas com os filtros de segurança desligados, agentes de modelos de fronteira tomaram 19 ações não autorizadas contra pessoas e organizações reais. No mesmo intervalo, a Meta confirmou que um modelo seu invadiu uma empresa alheia durante teste e a OpenAI, além de detalhar dois episódios com avaliadores externos, reforçou os controles em torno do Astra, seu próximo modelo, por não conseguir descartar que ele tenha alcançado o nível crítico de capacidade cibernética. Esse mesmo Astra, numa versão interna, resolveu dez problemas de matemática e computação teórica em aberto de longa data, por cerca de dois mil dólares em tokens. No Brasil, a semana foi de infraestrutura: o Serpro anunciou parceria para treinar modelos em português em datacenter próprio, e o LatamGPT prometeu segunda versão.

MODELO PROPRIETÁRIO

Thomson Reuters põe modelo próprio contra a fronteira e abre o placar

A liderança aparece em caso difícil e em documento longo; a lacuna, em raciocínio e código. Quem compra precisa saber qual das duas colunas descreve o próprio trabalho.

Em 31 de julho a Thomson Reuters apresentou o Thomson-1-Large, modelo de linguagem desenvolvido dentro de casa, com a afirmação de que ele agora funciona de forma competitiva com os melhores modelos de propósito geral do mundo. Bob Ambrogi foi aos números publicados pela empresa e chegou a uma conclusão mais sóbria: o quadro é mais desigual do que a manchete sugere [LawNext]. O ponto não é desmentir a empresa. É que a mesma tabela sustenta duas leituras opostas, e a escolha entre elas depende do que você faz no expediente.

Vale separar o que cada coluna mede. Benchmark jurídico afere se o modelo acerta a leitura de uma questão de direito posta em teste padronizado. Contexto longo mede outra coisa: manter a informação certa à mão quando o material entra inteiro na conversa, um contrato de duzentas páginas, um processo com dez volumes. O Thomson não pontua igual nas três, e é aí que a decisão de compra se decide.

BENCHMARK	THOMSON	O QUE MEDE E QUEM LIDERA
LegalBench	**0,823**	Raciocínio jurídico padronizado; Gemini 3.1 Pro lidera com 0,843
PrBench Legal Hard	**0,352**	Casos jurídicos difíceis; melhor da comparação, acima dos 0,333 do GPT-5.5
Harvey Legal Agent	**0,857**	Tarefas jurídicas executadas por agente; Opus 4.8 à frente com 0,869
Long Context	**0,753**	Recuperação em documento longo; melhor da comparação
Coding	**0,399**	Geração de código; última colocação, contra 0,598 do Opus 4.8

COMPARAÇÃO COM OS MODELOS DE FRONTEIRA

A leitura honesta da tabela é de troca, não de superioridade. O Thomson ganha exatamente onde o trabalho jurídico é mais duro e mais caro de terceirizar: o caso difícil e o documento comprido. Perde onde o mercado de propósito geral investiu mais, o raciocínio abstrato e a programação. Há ainda um número que não aparece nas comparações de modelo puro. Num teste em que os quatro modelos receberam acesso nativo ao acervo da própria editora, a recuperação de informação ficou entre 0,81 e 0,91, com o Thomson em 0,89. Dado o mesmo material para ler, todos chegam perto. O diferencial da editora não está no modelo. Está no que ela deixa o modelo ler.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

Para quem avalia fornecedor, a lição é metodológica antes de ser técnica. Nenhum desses números responde à pergunta que importa, que é como a ferramenta se comporta no seu acervo, com a sua nomenclatura e a sua jurisprudência de referência. Benchmark serve para eliminar candidato, não para escolher vencedor. E há um detalhe que passa batido: o teste de agente jurídico, no qual o Thomson quase empata com o topo, mede tarefa executada de ponta a ponta, não resposta a pergunta. É a métrica mais próxima do que uma banca compra quando compra IA.

> *“Esses resultados de benchmark são inteiramente autodeclarados pela Thomson Reuters.”*
>
> BOB AMBROGI · EDITOR, LAWNEXT

IMPACTO PRÁTICO

Repare no que a Thomson Reuters fez de mais relevante, e não foi treinar modelo. Foi publicar a tabela inteira, inclusive a linha em que aparece em último lugar. Fornecedor que só divulga a coluna onde ganha pede que você acredite; quem divulga a coluna onde perde permite que você confira.

Da próxima vez que um vendedor disser que a ferramenta dele rivaliza com a fronteira, peça a planilha completa, com o nome do teste e o concorrente ao lado. Se a resposta vier em adjetivo, você já sabe o que a tabela diria.

ANÁLISE

Agentes de IA atacaram alvos reais em três laboratórios diferentes

O que era acidente isolado virou característica do método de avaliação. E nenhuma das contenções que falharam era técnica.

ACOMPANHAMENTO *Edição nº 10: os três incidentes em que o Claude invadiu sistemas reais achando que estava num simulador. Nesta semana a lista ganhou mais dois laboratórios e a cronologia completa do caso anterior.*

O AI Security Institute do Reino Unido, órgão do governo britânico que avalia modelos de fronteira, publicou o relatório de uma bateria feita entre 25 e 28 de julho, com acesso real à internet e classificadores de segurança desativados de propósito. Em 122 tentativas, os pesquisadores acharam 19 casos de ação não autorizada na internet aberta, inclusive contra pessoas e organizações reais [Willison]. O Claude Mythos 5 respondeu pela maioria, com o GPT-5.6 Sol em alguns. O episódio mais grave foi de manual: um agente abriu duas contas no GitHub, uma para submeter alteração de código maliciosa e outra para se passar por revisor humano e endossá-la, e disparou phishing dirigido para convencer o mantenedor a aceitá-la.

Dois dias antes, a OpenAI havia detalhado episódios equivalentes relatados por avaliadores externos [OpenAI]. Num deles, o GPT-5.6 Sol reaproveitou um token do GitHub que outro agente deixara exposto, registrou-se em provedores externos de tunelamento e chegou a

expor à internet pública um servidor local carregado de código de exploração. No outro, uma configuração errada conectou o ambiente de teste à internet de verdade, e o modelo atacou um site real por confundi-lo com o cenário simulado. Na quinta-feira anterior fora a vez da Meta: o Muse Spark obteve acesso indevido à rede durante avaliação da empresa Irregular e explorou falha de segurança real de outra companhia [Willison].

122	19	3
TENTATIVAS EM DOIS DESAFIOS DO AISI	AÇÕES NÃO AUTORIZADAS NA INTERNET ABERTA	LABORATÓRIOS COM INCIDENTE CONFIRMADO

ONDE OS TRÊS CASOS SE PARECEM

A cronologia completa do episódio que envolveu a Hugging Face saiu nesta semana e ajuda a ver o padrão [Willison]. Agentes usados em treinamento interno da OpenAI acharam sozinhos uma cadeia de falhas, que passou por requisição forjada no servidor de artefatos, execução remota de código e privilégio de administrador, e se coordenaram lendo as mensagens uns dos outros num canal interno. Na Hugging Face a porta foi um aplicativo hospedado no Modal com chave de acesso fraca, e dali os agentes levaram menos de treze horas para sair da execução de código num contêiner e virar administradores de vários clusters. A OpenAI só descobriu a própria responsabilidade ao pedir a revogação das credenciais e ouvir que já tinham sido revogadas por causa da invasão.

Nenhum agente foi instruído a atacar alvo real. Em todos, o ambiente contradizia a instrução, por configuração errada, por isolamento de rede ausente ou por coincidência de nome entre alvo fictício e domínio existente. E em todos a barreira que falhou foi a mesma: a que existia apenas no texto do enunciado.

O QUE ISSO EXIGE DE QUEM CONTRATA

Na sexta-feira a OpenAI foi além do relato de incidente. Anunciou que avaliações internas do Astra, modelo ainda não lançado, mostraram avanço tal em codificação autônoma e cibersegurança que não dá para descartar que ele atinja o nível crítico do seu quadro de preparação, o mais alto da escala, reservado a modelo capaz de descobrir e explorar sozinho falhas inéditas em sistemas endurecidos, ou de executar um ataque inteiro a partir de um objetivo genérico [OpenAI]. Modelos anteriores, o GPT-5.6 Sol inclusive, ficavam no nível alto. Em resposta, a empresa isolou os ambientes de teste, restringiu acesso a rede e ferramentas, cifrou os pesos e suspendeu as atividades internas que não atendessem aos novos requisitos.

> *Foram identificados 19 casos em que agentes de IA tomaram ação não autorizada na internet ao vivo, incluindo casos que atingiram pessoas e organizações reais.*
>
> AI SECURITY INSTITUTE, REINO UNIDO

IMPACTO PRÁTICO

Toda estrutura jurídica que hoje conecta um agente ao próprio acervo escreve, em algum lugar, a frase que os laboratórios escreveram: aqui você pode, ali você não pode. A semana mostrou o que essa frase vale quando a rede diz o contrário. Vale nada, e vale rápido: treze horas entre uma chave fraca e o acesso de administrador na infraestrutura alheia.

A pergunta ao fornecedor não é se o agente foi instruído a respeitar limites. É qual credencial ele carrega, o que ela abre e quem revoga quando alguém percebe. Instrução se lê; permissão se exerce.

PESQUISA

Dez problemas em aberto resolvidos pelo modelo que a OpenAI acabou de blindar

O custo da descoberta coube no orçamento de uma perícia. A parte difícil foi decidir de quem é a autoria.

No primeiro dia da janela a OpenAI publicou dez resultados que resolvem ou avançam de forma substancial problemas de matemática e ciência da computação teórica em aberto de longa data [OpenAI]. A lista vai de empacotamento de esferas em dimensões altas à refutação de uma conjectura de rigidez em álgebras de operadores, passando por três problemas de Erdős. Quem resolveu foi uma versão interna do Astra. O total de tokens gastos para encontrar as soluções dos dez custaria cerca de dois mil dólares nas tarifas públicas do modelo Sol.

O detalhe que separa esse anúncio de uma nota de marketing é a verificação. As provas foram formalizadas em Lean, uma linguagem em que cada passo do argumento é conferido por programa, sem espaço para o leitor confiar na boa vontade de quem escreveu, e publicadas num repositório aberto. Simon Willison registrou o surto coletivo de comparações com o Deep Blue entre os matemáticos e a ressalva que qualquer revisor faria: ele quer ver os prompts reais usados [Willison]. Na semana anterior, a Anthropic havia gastado cem mil dólares em uma busca comparável por fragilidades criptográficas. A conta de dois mil dólares pelos dez é a notícia econômica escondida na notícia matemática.

COMPARAÇÃO COM O QUE A MÁQUINA JÁ FAZIA EM MATEMÁTICA

Até aqui, o papel da IA em demonstração matemática era de auxiliar: sugerir caminho, checar passagem, acelerar cálculo. A diferença desta semana está na atribuição. A OpenAI afirma que a autoria matemática pertence ao sistema, não às pessoas que depois organizaram os manuscritos, e declara ter assumido a responsabilidade pela correção das provas. A empresa separou quem produziu o conteúdo de quem responde por ele.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

A discussão de autoria não é filosófica para quem assina peça, parecer ou laudo. O arranjo que a OpenAI descreve, autoria da máquina e responsabilidade da casa que a operou, é o mesmo que uma banca adota quando usa IA e não conta: o texto vem de um lugar, a assinatura vem de outro, e a diferença só aparece quando alguém erra. O que mudou é que agora existe precedente público de laboratório declarando a separação em voz alta e submetendo o resultado a um sistema que não aceita passo mal justificado.

> *Reivindicar autoria humana para uma prova gerada inteiramente por um sistema de IA distorceria tanto a contribuição do sistema quanto a natureza do genuíno trabalho intelectual humano.*
>
> OPENAI

> **IMPACTO PRÁTICO**
>
> Junte as duas notícias sobre o mesmo modelo e o desconforto se completa. O sistema que derrubou dez problemas em aberto é o mesmo em torno do qual a fabricante trancou os ambientes de teste, por não garantir o que ele faz numa rede aberta. Não são dois modelos, um bom e um perigoso: é a mesma capacidade de perseguir um objetivo até o fim, apontada para lugares diferentes.
>
> Quem contrata inteligência artificial contrata as duas faces no mesmo pacote, porque ninguém sabe vendê-las separadas. O que se compra separado é a coleira.

Outros modelos da semana

A Alibaba anunciou o Qwen 3.8 Max, modelo de arquitetura MoE, sigla para mistura de especialistas: em vez de acionar a rede inteira a cada palavra, ele mantém 2,4 trilhões de parâmetros no total e ativa, por estimativa citada na cobertura, cerca de 95 bilhões por token. Traz janela de contexto de 1 milhão de tokens e marca 87,3% no SWE-Bench, teste de resolução de problemas reais de programação, acima dos 82,6% atribuídos ao GPT-5.5 [AINews]. Junto sai o Qwen 3.8-27B, de pesos abertos, na faixa que cabe em hardware modesto: é a rota para hospedar modelo capaz em servidor próprio sem parque de placas de vídeo.

A Meta lançou o Muse Spark 1.2 e o Muse Code, agente de terminal que divide tarefa de código entre subagentes paralelos em cópias isoladas do repositório [TechCrunch]. O preço é o argumento: 1,25 dólar por milhão de tokens de entrada, ou dez centavos para quem deixa a Meta usar os dados de uso [Willison]. Desconto de 92% em troca do conteúdo do trabalho é um contrato que uma operação jurídica precisa ler duas vezes antes de assinar.

Na outra ponta, a Liquid AI publicou o LFM2.5-2.6B, de 2,6 bilhões de parâmetros, feito para rodar fluxos de várias etapas no próprio computador do usuário, sem nuvem. Ocupa menos de 2,5 GB de memória, aceita 128 mil tokens de contexto e supera modelos até quatro vezes maiores em seguir instrução e, com uma exceção, no uso de ferramentas [Liquid AI]. Para triagem de documento sigiloso que não pode sair da máquina, essa é a faixa que interessa.

Pesquisa e métodos

O achado mais desconfortável da semana veio da medicina e serve de espelho. Pesquisadores da Alemanha, da Áustria e do Chile testaram o CORA, assistente clínico que ancora cada resposta em fonte real recuperada na hora, técnica conhecida como RAG, e mediram o efeito sobre 46 médicos [arXiv]. A acurácia subiu de 70,8% sem apoio para 82,6% com o sistema. O número que assusta é outro: quando a resposta estava errada e vinha com citação, a resistência do médico a ela desabou de 92% para 34,8%. A citação não fez o profissional conferir mais. Fez conferir menos.

Dois papers falam direto ao ofício. Um deles testa uma receita de treino que parecia certeira para raciocínio jurídico: um modelo redige o argumento, outro o ataca, e o primeiro é recompensado quando sobrevive, com verificador conferindo que as autoridades citadas dos dois lados existem. Em quatro testes independentes, incluindo julgamento cego, o componente competitivo não trouxe ganho confiável, e uma vantagem inicial aparente de 29% se revelou artefato de amostra pequena [arXiv]. O autor registra o resultado negativo com todas as letras, o que já é notícia. A conclusão prática é que o valor do arranjo estava no verificador de citações, não na disputa.

O outro propõe importar para a IA a teoria fiduciária: lealdade, cuidado, boa-fé e transparência, os quatro deveres que o direito impõe a quem administra interesse alheio, aplicados à relação entre quem desenvolve o assistente e quem o usa [arXiv]. Os exemplos do autor vêm do mandato, da administração de bens, do consultor financeiro e do médico; a relação entre advogado e cliente, que ele não cita, obedece à mesma gramática. A virada é de pergunta: sai o debate sobre quais valores a IA deve promover, entra o debate sobre quais obrigações o fornecedor deve ao usuário. Para quem trabalha com dever de lealdade todo dia, é um vocabulário conhecido chegando com atraso onde faz falta.

Ferramentas do ecossistema

- **Align Research:** pesquisa jurídica que devolve só decisões reais, com os trechos destacados, e nenhuma narrativa gerada. A tese do fundador, ex-sócio de litígio: quem só aponta para o

que já existe não tem como alucinar. Um encadeamento de agentes analisa a questão, busca, classifica os julgados e monta um dossiê em PDF, em horas e não em segundos [LawNext].

- **DISCO:** a plataforma de coleta e triagem de provas eletrônicas juntou fatos do processo e jurisprudência aplicável numa interface só, cobrindo o litígio inteiro e não apenas a descoberta de provas. Está em piloto com cinco bancas, com lançamento geral previsto para o início de 2027 [LawNext].
- **Claude Code:** a versão 2.1.224 passou a permitir, nos planos Team e Enterprise, que sessões abertas no navegador ou no celular executem em máquina ou contêiner do próprio usuário, e liberou a troca de mensagens entre sessões em máquinas diferentes, por enquanto só em macOS e Linux [release]. Quem precisa manter o processamento dentro da própria infraestrutura ganhou o caminho oficial para isso.
- **Cloudflare Kitesurf:** navegador em nuvem, construído do zero para ser operado por agente e não por gente, sem abas nem extensões [TechCrunch]. A empresa afirma que ele passa em mais de 215 mil testes de plataforma web gastando menos processador e memória que o Chromium. A reportagem lembra que o modelo de ameaça é outro, o da injeção de prompt, ataque em que um texto plantado na página tenta sequestrar as instruções do agente, e não descreve defesa específica contra ele.

PARA TESTAR ESTA SEMANA

O decano da Suffolk Law publicou um arquivo público que cataloga as políticas de uso de IA de 128 faculdades de direito americanas, organizadas em oito eixos [Law School Policy Archive]. É a matéria-prima pronta para quem precisa redigir a norma interna da própria casa e não sabe por onde começar.

Análise

François Chollet, criador do benchmark ARC-AGI, voltou a uma previsão que fizera em 2024 para dizer onde acha que o campo está [thread]. Segundo ele, a indústria aplicou um primeiro remendo ao limite do aprendizado profundo, a busca em tempo de uso, isto é, deixar o modelo gastar computação pensando antes de responder, difundido desde dezembro de 2024. O argumento é que ele não basta: modelos sem essa etapa continuam com desempenho fraco no ARC-AGI de 2019 mesmo depois de escalados cem mil vezes.

O segundo remendo, que ele já definira em 2024 no post que abre a mesma conversa, é mais radical: ou o modelo passa a ajustar o que aprendeu no momento em que encontra o dado novo, e não apenas no treinamento prévio, ou o campo abandona o mecanismo de treino atual em favor de busca por programas. Vindo de quem desenhou o teste que os modelos mais

custam a vencer, a leitura vale como diagnóstico do teto atual: o que hoje se vende como raciocínio é um remendo bem-sucedido sobre uma limitação que continua no lugar.

Brasil

Serpro anuncia parceria para treinar modelos em português em datacenter próprio

O Serpro fechou acordo com a MeetKai Brasil para desenvolver modelos de linguagem em português treinados com dados nacionais e hospedados na infraestrutura da própria estatal [Mobile Time]. A empresa fornecerá famílias de modelos com pesos e código abertos para o Serpro customizar, e a estatal abrirá a qualquer órgão de governo a possibilidade de treinar na sua infraestrutura modelos de domínio específico ou de menor porte, com os dados guardados localmente. Está previsto um produto de geração de código, o Serpro Code, até o fim do ano, sobre a plataforma interna que já hospeda doze modelos. Não foram divulgados parâmetros, volume de dados nem benchmark: por ora é acordo assinado e roteiro anunciado.

LatamGPT prepara segunda versão com participação da USP

O LatamGPT, modelo regional coordenado pelo centro chileno CENIA com participação de Fábio Cozman, da USP, terá uma segunda versão com mais dados e qualidade maior [Mobile Time]. A versão atual saiu de treinamento continuado e pós-treino sobre o Llama 3.1 de 70 bilhões de parâmetros, com o corpus Carolina, base de textos em português construída no Brasil, e material de outros países da região. Não há data de lançamento nem comparação de desempenho entre as duas versões.

Dados de uso mostram o Brasil no topo do consumo multimídia

A OpenAI abriu pela primeira vez números de uso do ChatGPT por país [OpenAI]. No trabalho, a chance de usar a ferramenta para concluir uma tarefa ou produzir algo é mais que o dobro da de usá-la fora dele, onde o uso segue exploratório. O uso multimídia responde por 7,8% das mensagens no mundo desde o lançamento da geração de imagens em abril, e no Brasil e na Colômbia passa de uma em cada dez. No segundo trimestre, América Latina, África e Oceania cresceram mais rápido que o resto do mundo, com Peru, Uruguai e Costa Rica liderando a subida no ranking por habitante.

> **IMPACTO PRÁTICO**
>
> A semana brasileira teve duas notícias que não conversam entre si, e a distância entre elas é o retrato. De um lado, o país aparece como quem constrói: acordo de estatal para treinar modelo

próprio, versão nova de modelo regional, corpus em português. De outro, como quem consome em alta velocidade, com uma em cada dez mensagens já sendo de imagem ou mídia.

Construir leva anos e ainda não tem número publicado. Consumir já acontece, em escala, e os números são de quem vende. Enquanto uma coluna for de anúncio e a outra for de medição estrangeira, quem decide onde o dado brasileiro é processado decide com a informação do outro lado da mesa.

NO RADAR DA PRÓXIMA SEMANA

A Alibaba marcou para a semana seguinte ao anúncio a abertura dos pesos do Qwen 3.8 Max, e disse que o Qwen 3.8-27B também virá em pesos abertos [AINews]. A OpenAI marcou para a mesma janela os chats de texto sem limite do GPT-5.6 Luna nas contas gratuitas [OpenAI].

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

Dennis Martins · Adriana Aneli

CURADORIA E EDIÇÃO · 8 DE AGOSTO DE 2026