RADAR SEMANAL · PARA ESCRITÓRIOS, GABINETES E ASSESSORIAS

IA em Cognição Nada Exauriente

EDIÇÃO Nº 13 INVENTÁRIO DOS PRINCIPAIS FATOS OCORRIDOS DE 15 DE AGOSTO A 22 DE AGOSTO DE 2026

Resumo

Um sistema da Nvidia resolveu todos os 183 níveis do conjunto público do ARC-AGI-3, teste em que a máquina entra num jogo sem manual e precisa descobrir as regras jogando. O modelo por dentro não é da Nvidia e não foi treinado para a tarefa: é o Claude Opus 5, comprado pronto. O que a casa construiu foi o que vai em volta. É o achado mais consequente de uma semana em que o primeiro benchmark formal da revisão final de contrato reprovou os modelos de fronteira na tarefa mais rotineira do trabalho jurídico, e um modelo aberto de 27 bilhões de parâmetros, que cabe num arquivo de 17 gigabytes, empatou com a fronteira proprietária numa medição independente.

No restante do inventário: a OpenAI parou por duas semanas o treinamento dos próprios modelos depois que um sistema em desenvolvimento encostou no patamar crítico de capacidade cibernética, o Pew Research mediu quanto da web publicada depois do ChatGPT carrega marca de autoria por máquina, e quatro plataformas jurídicas internacionais ligaram os próprios acervos a assistentes de IA. Do Brasil vieram números de atendimento com agentes na TIM e na Cacau Show, e uma pesquisa da USP que explica por que o chatbot concorda tanto com você.

PESQUISA

O andaime, não o modelo: a Nvidia resolve os 183 níveis do ARC-AGI-3

O ganho não veio de treinar melhor, veio de organizar melhor a volta. Quem contrata ferramenta de IA está comparando a peça errada.

A Nvidia publicou uma arquitetura chamada AVO que embrulha um modelo de linguagem já existente numa camada de ferramentas, memória e regras operacionais, o que a engenharia de agentes chama de *harness*, o andaime que transforma um modelo bruto em agente capaz de executar tarefa de ponta a ponta. Com o Claude Opus 5 por dentro, o AVO completou os 25 ambientes do conjunto público do ARC-AGI-3, resolvendo os 183 níveis em 6.624 ações, com pontuação máxima na métrica RHAE, que combina conclusão da tarefa com eficiência de movimentos em relação a humanos de primeira vez [blog da Nvidia]. O ARC Prize, que mantém

o benchmark, reporta cerca de 30% para o Claude Opus 5 avaliado sozinho. A Nvidia adverte, no próprio post, que os dois números saíram de configurações diferentes e não medem o ganho isolado do andaime: servem para ilustrar que avaliar só o modelo não descreve o desempenho do agente completo.

A peça central do AVO é um agente supervisor que observa o agente principal e o tira do caminho errado quando ele entra num beco sem saída. O mesmo princípio apareceu num paper independente na mesma semana. O LEGO-RL treina o modelo por reforço, isto é, por tentativa e recompensa, dentro do andaime em que ele vai realmente rodar, em vez de num ambiente de laboratório separado, e mede o resultado em três harnesses de codificação diferentes no SWE-bench Verified, o teste que mede resolução de problemas reais de software [paper no arXiv].

ANDAIME	SWE-BENCH VERIFIED	O QUE MEDE
OpenHands SDK	**64,0% para 70,4%**	resolução de problemas reais de software antes e depois do LEGO-RL
Claude Code	**62,4% para 68,2%**	mesma métrica, mesmo modelo, outro andaime
OpenCode	**57,2% para 66,6%**	mesma métrica, andaime com o menor ponto de partida

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

Até o ano passado, a conversa sobre capacidade de IA era uma conversa sobre o modelo: quantos parâmetros, quanto contexto, qual score no teste de raciocínio. O modelo era o produto, e tudo o mais era acessório de integração. Os números do LEGO-RL viram essa hierarquia do avesso na medida certa, sem exagero: antes de qualquer treino, o mesmo modelo já rende diferente conforme a moldura, quase sete pontos separando o melhor andaime do pior. Treinado dentro de cada uma delas, sobe de seis a nove pontos em todas. O progresso está migrando do laboratório que treina o modelo para quem monta a volta.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

Para quem contrata, a consequência é desagradável: a pergunta que a estrutura jurídica costuma fazer ao fornecedor, que é qual modelo está por trás da ferramenta, mede a variável menos determinante. O fornecedor que responde apenas o nome do modelo respondeu quase nada. O que separa um produto do outro é o que o sistema faz quando o caminho não dá certo, quem revisa o próprio trabalho antes de devolver, e o que acontece com a tarefa quando ela não cabe numa sessão. É a diferença entre contratar quem faz o trabalho e contratar quem o confere antes de ele sair.

IMPACTO PRÁTICO

Nenhum comparativo de modelo publicado nos últimos meses mede o que a semana mostrou ser decisivo. Os rankings de inteligência comparam motores; o que chega à mesa do profissional do direito é o veículo inteiro, com direção, freio e alguém no volante.

Na próxima demonstração de ferramenta jurídica, peça para ver a tarefa que dá errado, não a que dá certo. O andaime só aparece quando o modelo tropeça.

PESQUISA

ContractScrub: os modelos de fronteira reprovam na revisão final de contrato

A tarefa parecia feita sob medida para a máquina. O resultado mostra que benchmark geral não prevê desempenho em ofício específico.

Um grupo de pesquisadores publicou o ContractScrub, primeiro benchmark formal para medir modelos de linguagem naquela conferência final de contrato que precede a assinatura: caçar termo definido usado fora da definição, remissão que aponta para a cláusula errada, redação inconsistente entre trechos que deveriam dizer a mesma coisa. Os contratos do teste foram construídos à mão por advogados experientes, com os erros plantados por categoria [paper no arXiv].

A expectativa era de bom desempenho. A tarefa combina exatamente o que os modelos de fronteira alegam fazer bem: leitura de documento longo, checagem de consistência interna e reconhecimento de entidades nomeadas. O resultado foi o oposto. Apenas um modelo alcançou recall médio macro de 0,75, ou seja, encontrou três quartos dos erros existentes, e os demais ficaram abaixo disso. Recall é a medida que interessa aqui porque o erro que escapa é o que vai para a assinatura.

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

Não é caso isolado. Na mesma semana, um estudo com casos reais da Corte Europeia de Direitos Humanos testou se um modelo de ponta produz raciocínio juridicamente significativo ao prever o desfecho de um caso, e concluiu que o texto sai bem-formado por fora e vazio por dentro [estudo no arXiv]. O achado metodológico do mesmo trabalho é o que deveria acender a luz amarela: os avaliadores automáticos, isto é, modelos usados para julgar a saída de outros modelos, concordaram fortemente entre si e fracamente com avaliadores humanos treinados. Eles são consistentes, e consistentemente distantes de quem entende do assunto.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

A leitura errada desses dois papers é a de que a máquina não serve para contrato. A leitura correta é sobre onde colocar a conferência. Um sistema que encontra três quartos dos erros é útil quando entra antes da revisão humana e inútil quando entra no lugar dela, e a diferença entre os dois arranjos não está no software, está no fluxo de trabalho que a estrutura jurídica desenhou em volta. Vale também para a compra: benchmark geral de modelo, aquele número redondo que o fornecedor exibe no slide, não prevê desempenho na tarefa específica do seu contrato. Só o teste no seu material prevê.

> *"O modelo produz análises estruturalmente completas, mas substantivamente rasas."*
>
> AMOGH RAINA, ILIAS CHALKIDIS, DANIEL HERSHCOVICH E HENRIK PALMER OLSEN · AUTORES DO ESTUDO

IMPACTO PRÁTICO

A parte incômoda não é a nota baixa, é o descolamento. Os mesmos modelos que passeiam por provas de raciocínio geral tropeçam na conferência de remissão de cláusula, que é trabalho que estagiário do segundo ano faz com atenção e café.

O que isso ensina sobre compra de tecnologia jurídica cabe em uma linha: exija a demonstração no seu contrato, não no contrato do fornecedor. Documento de vitrine foi escolhido para ganhar.

MODELO OPEN-SOURCE

Qwen 3.8 27B: um arquivo de 17 gigabytes empata com a fronteira

O modelo que cabe numa estação de trabalho comum alcançou a nota do modelo que roda em data center. O preço aparece no relógio, não na fatura.

A Alibaba publicou o Qwen 3.8 27B, modelo de pesos abertos com capacidade de visão, ou seja, com os arquivos do modelo disponíveis para download e execução em máquina própria, sem passar por servidor de terceiro. Simon Willison o rodou em casa na versão quantizada, aquela em que os números internos do modelo são guardados com menos precisão para ocupar menos espaço, e o arquivo resultante tem 17 gigabytes [teste do Willison]. O contexto máximo é de 262.144 tokens, o equivalente a algumas centenas de páginas lidas de uma vez.

O que chamou atenção foi a medição independente. O índice de inteligência da Artificial Analysis combina nove avaliações de tarefas agênticas, código, raciocínio científico e conhecimento numa única escala [metodologia da Artificial Analysis]. Nele, o Qwen marcou 52, empatado com o GPT-5.6 Luna e a um ponto de modelos dezenas de vezes maiores [registro de Willison].

MODELO	ÍNDICE ARTIFICIAL ANALYSIS	O QUE MEDE
Qwen 3.8 27B	**52**	posição na escala agregada do índice
GPT-5.6 Luna	**52**	mesma escala, modelo proprietário de acesso por API
GLM-5.2, 753 bilhões de parâmetros	**53**	mesma escala, modelo aberto 28 vezes maior
DeepSeek V4 Pro, 1,7 trilhão de parâmetros	**53**	mesma escala, modelo aberto 60 vezes maior

COMPARAÇÃO COM A GERAÇÃO ANTERIOR

Há um ano, a distância entre o que cabia numa máquina de mesa e o que rodava em data center era a distância entre rascunho e trabalho. Ela encolheu para um ponto numa escala de cem. O que não encolheu foi o tempo. O modelo vem configurado no nível máximo de raciocínio estendido, o modo em que ele escreve para si mesmo uma longa cadeia de passos antes de responder, e nesse ajuste padrão Willison registrou um único teste em que a máquina gastou 21 minutos e 22.276 tokens de raciocínio para produzir 3.223 tokens de resposta. Ele recomenda baixar o nível para uso prático.

O QUE ISSO DESBLOQUEIA NA PRÁTICA

Rodar o modelo na própria máquina responde de uma vez a pergunta que trava a adoção em estrutura jurídica: para onde vai o documento do cliente. Ele não vai a lugar nenhum. Não há terceiro processando, não há termo de uso a interpretar, não há transferência internacional a justificar. Dezessete gigabytes cabem numa placa de vídeo de estação de trabalho comum, e a nota independente diz que a qualidade não é mais de brinquedo. O que muda é o ritmo: quem espera resposta em segundos precisa reconfigurar o modelo antes de julgar se ele serve.

> *Um arquivo de 17 GB conseguir fazer tudo isso nas minhas máquinas domésticas é um milagre.*
>
> SIMON WILLISON

> **IMPACTO PRÁTICO**
>
> O modelo aberto deixou de ser a opção de quem não podia pagar a fronteira e virou a opção de quem não pode entregar o documento. São motivos diferentes, e o segundo é o que interessa a quem tem dever de sigilo.
>
> A ressalva é do tamanho do ganho. Um modelo que pensa vinte minutos antes da primeira linha não perde para a nuvem em qualidade, perde em paciência, e a paciência de quem tem prazo processual correndo é curta. Configure antes de concluir que não presta.

Outros modelos da semana

> **ACOMPANHAMENTO** *Edição nº 12: o GLM-5.3 entrou aqui como o modelo que melhorou sem ganhar um parâmetro, só com mais pós-treinamento. O efeito colateral desse treino apareceu depois.*

No anúncio de 14 de agosto, a Z.ai registra que a mesma rodada de pós-treinamento que elevou o desempenho em codificação elevou junto, e mais rápido do que a equipe previa, a capacidade de encontrar falhas de segurança em código alheio. Numa varredura conduzida com equipes de segurança, o modelo localizou 2.436 vulnerabilidades em 269 projetos de código aberto [post da Z.ai]. A capacidade defensiva e a ofensiva saem do mesmo treino, e ninguém pediu a segunda.

A Liquid AI publicou os checkpoints DSpark para três modelos da família LFM2.5: modelos de rascunho pequenos, na faixa de 300 milhões de parâmetros, que propõem os próximos tokens depressa para o modelo grande apenas conferir num único passe, em vez de gerar um a um. O ganho chega a 3,18 vezes numa GPU H100 e 2,87 vezes num MacBook Pro, com saída idêntica à do modo normal por construção do método [anúncio no Hugging Face]. É aceleração sem troca: não há qualidade a negociar, e nada do que já foi validado precisa ser reavaliado.

Pesquisa e métodos

Dois trabalhos mediram o limite dos agentes. O AI4AI-Bench testou se um agente consegue reescrever o próprio algoritmo de treinamento, e não apenas ajustar parâmetros: em 29 configurações de seis sistemas, a nota média foi 0,166 numa escala em que 0,1 é o algoritmo original do repositório e 1,0 é o ótimo da tarefa [paper no arXiv]. E um estudo publicado em julho, repercutido nesta semana, deu a agentes de fronteira seis dias e cerca de três mil dólares em créditos para avançar duas perguntas de pesquisa inéditas: os dois trabalhos produzidos foram rejeitados pelos autores originais das perguntas [paper no arXiv]. Os agentes executam bem e julgam mal, que é a mesma divisão de trabalho que aparece no ContractScrub.

Um terceiro estudo explica por que gastar mais computação na hora de responder rende menos do que se supõe. Comparando cinco famílias de técnicas em cinco domínios, incluindo direito, os autores separaram gerar candidatos de escolher entre eles: gerar funciona, escolher falha, porque o modelo que pontua as respostas correlaciona-se com a qualidade real em apenas 0,12 [paper no arXiv]. Na ponta oposta, outro grupo ensinou o modelo a decidir sozinho, no primeiro token da resposta, se o problema pede raciocínio longo, curto ou nenhum, cortando 41% do comprimento médio e mantendo a acurácia praticamente no mesmo patamar [paper no arXiv]. É o remédio direto para o excesso de deliberação que o Qwen exibiu.

Fora dos agentes, o Pew Research mediu autoria por máquina em cerca de 500 mil páginas em inglês coletadas ao longo de cinco anos: entre as amostradas em julho de 2026, cerca de 10% trazem sinais significativos de texto gerado por IA, e entre as publicadas depois do lançamento do ChatGPT o índice passa de um terço [Pew Research]. Nas amostras de 2026, a dispersão por domínio é o dado que interessa a quem produz prova: cerca de 10% em .com, 4,6% em .org e cerca de 1% em .edu e .gov.

Ferramentas do ecossistema

- O **NetDocuments**, plataforma de gestão documental do mercado jurídico, ganhou a Tabular Review, que transforma um lote de documentos em tabela na qual cada documento vira linha e cada pergunta do usuário vira coluna, com a citação de cada célula ligada ao trecho de origem, e a extração automática de autoridades legais, que identifica todas as citações jurídicas do material e monta um mapa navegável com link para o inteiro teor no CourtListener [LawNext]. Resolve a parte mais lenta da due diligence, que é montar a planilha comparativa à mão.
- O **Clio Docket**, base de peças processuais herdada da compra da vLex, passou a operar dentro do Clio Work, o ambiente de trabalho com IA da mesma casa, com acesso a mais de um bilhão de documentos judiciais e alerta de nova movimentação sem trocar de ferramenta [LawNext]. A pesquisa deixa de ser etapa separada do acompanhamento do caso.
- A **Thomson Reuters** conectou o CoCounsel e outras ferramentas suas ao acervo do iManage por meio do MCP, o protocolo que padroniza como um assistente de IA acessa uma base externa, preservando controles de acesso e barreiras éticas entre clientes durante a busca [Legal IT Insider]. O documento flui direto de um sistema ao outro, e o produto do trabalho volta sozinho ao repositório de origem.
- A **Trellis**, que reúne mais de 2,5 bilhões de registros de mais de 3.500 tribunais estaduais americanos de primeira instância, lançou um plugin para o ChatGPT equivalente ao conector que já mantinha para o Claude [LawNext]. O movimento das quatro plataformas é o mesmo: abrir o acervo ao assistente que o profissional já usa, em vez de disputar a tela com ele.

PARA TESTAR ESTA SEMANA

Abra o CORS Chat, página única que Simon Willison publicou em 15 de agosto e que roda inteiramente no navegador, aponte para um provedor com acesso ao Qwen 3.8 27B e peça a ele a leitura de uma cláusula [CORS Chat]. O teste leva quinze minutos, não exige instalar nada e serve para sentir na prática o custo do raciocínio estendido: no ajuste padrão do modelo, Willison registrou minutos inteiros de deliberação antes da primeira linha da resposta.

Análise

A OpenAI publicou em 18 de agosto que um modelo em desenvolvimento, chamado internamente de Astra, apresentou evidência preliminar de atingir o patamar crítico de capacidade cibernética previsto na política interna que classifica risco antes do lançamento. A resposta foi parar por duas semanas o treinamento por reforço dos modelos destinados a lançamento e suspender a maior rodada de treino planejada [OpenAI]. A empresa descreve também o monitoramento que passou a rodar sobre a cadeia de raciocínio dos modelos, o rascunho de passos que antecede a resposta, a um custo declarado de cerca de 20% do processamento monitorado. Um laboratório que gasta um quinto da própria capacidade vigiando o que o modelo pensa está dizendo, sem dizer, que não confia inteiramente no que treinou.

No dia anterior, o presidente da casa, Greg Brockman, publicou o outro lado da moeda: pediu ao ChatGPT Work, rodando o GPT-5.6 Sol já disponível ao público, que auditasse o próprio site pessoal. Em cerca de 15 minutos ele encontrou 13 falhas, de registro de domínio mal configurado a tráfego sem criptografia entre dois provedores, e em uma hora corrigiu todas navegando sozinho pelos painéis [OpenAI]. A mesma capacidade que obrigou a pausa é a que faz a auditoria. Não são dois assuntos.

Fechando a semana das declarações, o presidente-executivo da Anthropic, Dario Amodei, fez uma autocrítica incomum numa troca pública sobre o ceticismo do público com IA, reproduzida por Simon Willison [citação registrada por Willison].

> *A crítica mais precisa às empresas de IA, incluindo a Anthropic, é que ainda não entregamos as nossas grandes promessas de benefício ao mundo. A esta altura, dizer que a IA vai curar o câncer é mais clichê do que inspirador, e a maioria das pessoas acha que é enganoso. O que vai funcionar é de fato curar o câncer.*
>
> DARIO AMODEI · PRESIDENTE-EXECUTIVO, ANTHROPIC

Brasil

TIM leva cobrança por agentes ao WhatsApp e publica o ganho

A operadora informou que o sistema de cobrança por IA no WhatsApp, em operação desde janeiro, já alcançou 2 milhões de clientes, com ganho de 7,7 pontos percentuais na recuperação de dívidas e um quarto das negociações fechadas fora do horário comercial [Mobile Time]. A arquitetura é multiagente: um agente conduz o contato e a negociação, e agentes secundários cuidam do esclarecimento da conta e da argumentação. Para quem atua em direito do consumidor, o número que importa não é o da recuperação, é o do horário: cobrança negociada às onze da noite é cobrança feita sem testemunha do outro lado do balcão.

Cacau Show mede o atendimento com agente no pico da Páscoa

A empresa divulgou os números da operação de atendimento em redes sociais durante a Páscoa, período que concentra um quinto do faturamento anual em poucos dias, com governança híbrida entre operação humana e IA [Mobile Time]. O tempo médio de resposta caiu de 8 horas, em 2024, para 26 minutos.

163 mil
RESPOSTAS NA PÁSCOA, ANTE 35 MIL

72%
DE RESOLUÇÃO, ANTE 59%

26 min
DE ESPERA MÉDIA, ANTE 8 HORAS

Pesquisa da USP explica por que o chatbot concorda tanto com você

Seth Jacobowitz, da Faculdade de Filosofia, Letras e Ciências Humanas da USP, publicou na revista AI & Society um estudo sobre a bajulação dos assistentes de conversa, sustentando que ela não é defeito de acabamento [Jornal da USP]. A bajulação, sustenta o estudo, emerge da própria arquitetura dos sistemas e é funcional para eles. O trabalho identifica três funções nessa concordância: manter a conversa no tema do usuário, dar ao assistente uma personalidade previsível e criar dependência cognitiva. Jacobowitz chama o efeito de mediocridade mutuamente reforçada, em que nem o humano desafia a máquina, nem a máquina desafia o humano o bastante.

IMPACTO PRÁTICO

Junte as três notícias e aparece um desenho só. As duas primeiras medem volume, velocidade e resolução, que é o que a operação comercial precisa saber. A terceira mede a única coisa que a operação jurídica precisa saber antes de adotar a mesma tecnologia: se a máquina discorda de você quando você está errado.

É a diferença entre o balcão e o parecer. No balcão, concordar resolve o atendimento e encerra a interação com o cliente satisfeito. No parecer, concordar é o defeito: quem consulta quer

saber o que não viu, e a resposta que devolve a própria tese em linguagem mais bonita não é assessoria, é eco com timbre profissional.

NO RADAR DA PRÓXIMA SEMANA

A Z.ai anunciou que publica os pesos abertos do GLM-5.3 duas semanas após o lançamento, prazo que vence nos próximos dias [post da Z.ai]. A OpenAI prometeu para setembro o documento técnico do processamento privado de segurança, o mecanismo que promete detectar abuso sem dar a funcionários da empresa acesso ao conteúdo do cliente [OpenAI].

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

Dennis Martins · Adriana Aneli

CURADORIA E EDIÇÃO · 22 DE AGOSTO DE 2026