# IA em Cognição Nada Exauriente

LAVRADO EM 18 DE JULHO DE 2026, COM EFEITOS RETROATIVOS A 11 DE JULHO

## Resumo

Dois modelos gigantes de pesos abertos chegaram na mesma semana, e o maior deles tem 2,8 trilhões de parâmetros. A chinesa Moonshot lançou o **Kimi K3**, que ela mesma apelidou de "inteligência de fronteira aberta" e que a avaliação independente da Artificial Analysis põe no mesmo patamar do Claude Opus 4.8. E a Thinking Machines, de Mira Murati, estreou o **Inkling**, que entende texto, imagem e áudio, sob a licença mais permissiva que existe. No terreno da segurança, a OpenAI abriu a caixa do **GPT-Red**, um modelo treinado só para atacar outros modelos, e usou-o para endurecer o GPT-5.6 Sol contra a fraude mais perigosa da era dos agentes: a injeção de prompt.

No pano de fundo, a semana ainda trouxe um paper que flagra o modelo escondendo o próprio viés, uma safra de agentes jurídicos que prometem unificar a prática e o negócio da advocacia, e, no Brasil, um agente de câmbio do Banco do Brasil que cortou 90% do tempo da operação e um modelo brasileiro de leitura de documentos que bateu os gigantes em português. Muita capacidade nova; quase toda ela aberta, aplicada ou defensiva.

---

MODELO OPEN-SOURCE

## Kimi K3: o maior modelo de pesos abertos já lançado

A Moonshot AI publicou o Kimi K3, seu modelo mais capaz até hoje, com 2,8 trilhões de parâmetros totais e disponibilidade imediata por site e API; os pesos abertos, isto é, os valores internos do modelo publicados para qualquer um baixar e rodar em servidor próprio, ao contrário de um GPT ou de um Claude, que só se acessam por API paga, foram prometidos para 27 de julho [Latent Space]. Quando saírem, será o maior modelo aberto já disponibilizado.

O que torna esse tamanho viável é a arquitetura MoE, sigla de mistura de especialistas: em vez de acionar o modelo inteiro a cada palavra gerada, ele ativa só um punhado de sub-redes especializadas, o que derruba o custo de processamento sem abrir mão da escala. No Kimi K3 são ativados apenas 16 de 896 especialistas por token, menos de 2% do total. A janela de contexto, o volume de texto que o modelo consegue considerar de uma vez, chega a 1 milhão de tokens, e um mecanismo de atenção próprio promete até 6,3 vezes mais velocidade de resposta nesse comprimento.

| BENCHMARK | SCORE | O QUE MEDE |
|---|---|---|
| Benchmark | Kimi K3 | O que mede |
| AA Intelligence Index | 57 | Índice agregado de capacidade (Opus 4.8: 56) |
| Frontend Code Arena | 1.679 · 1º lugar | Preferência humana em geração de interfaces |
| AA-Omniscience | 46% acerto / 51% erro inventado | Conhecimento factual e taxa de alucinação |

## COMPARAÇÃO COM A GERAÇÃO ANTERIOR

O salto mais visível é em código: no ranking de preferência humana para geração de interfaces, o K3 pulou do 18º lugar do Kimi K2.6 para o primeiro, com 76% de vitórias contra 63% do Claude Fable 5 e 58% do GPT-5.6 Sol. A acurácia factual também subiu, de 33% para 46%. Mas a honestidade obriga a registrar o outro lado: a taxa de alucinação, respostas inventadas com cara de verdade, piorou de 39% para 51% no mesmo teste. O modelo sabe mais e erra com mais confiança.

## O QUE ISSO DESBLOQUEIA NA PRÁTICA

Para uma estrutura jurídica que não pode enviar documento sigiloso para a API de um laboratório estrangeiro, um modelo de pesos abertos deste calibre muda a conta. A avaliação independente coloca o K3 no nível do Opus 4.8, uma geração atrás do topo pago; mas uma geração que você mantém dentro de casa vale mais que uma geração que você aluga, quando o dado não pode sair da sala. Simon Willison, ao rodar seu teste pessoal, alerta que o K3 só opera num nível de esforço de raciocínio, o "máximo", o que o deixa caro e lento [Willison]. Poder é uma coisa; docilidade operacional é outra.

> **IMPACTO PRÁTICO**
>
> O Kimi K3 não é o modelo mais inteligente do mercado; é o mais inteligente que uma banca consegue manter dentro das próprias paredes. Essa distinção é tudo para quem lida com sigilo profissional. O preço de US$ 3 por milhão de tokens de entrada e o custo por tarefa abaixo do GPT-5.6 Sol importam menos que a frase que o modelo aberto permite dizer a um cliente receoso: os seus autos não passaram por servidor de terceiro nenhum. A fronteira da inteligência artificial deixou de ser um clube fechado; agora ela tem uma porta que abre para dentro.

MODELO OPEN-SOURCE

# Inkling: a aposta da Thinking Machines na customização

A Thinking Machines, startup fundada por Mira Murati, ex-diretora de tecnologia da OpenAI, lançou seu primeiro modelo próprio, o Inkling, sob licença Apache 2.0, a mais permissiva do mercado, que autoriza uso comercial, redistribuição e modificação sem fricção [HF — Thinking Machines]. É um modelo MoE de 975 bilhões de parâmetros totais, mas só 41 bilhões ativos por consulta, multimodal de nascença: aceita texto, imagem e áudio, e foi treinado em 45 trilhões de tokens.

A empresa é rara na franqueza: diz abertamente que o Inkling não é o modelo mais forte disponível. A tese é que uma parte grande do mercado prefere um modelo que possa moldar às próprias necessidades a um genérico mais esperto oferecido para alugar. É a antítese da semana: enquanto o Kimi persegue o topo do ranking, o Inkling persegue a maleabilidade.

| BENCHMARK | SCORE | O QUE MEDE |
|---|---|---|
| **Benchmark** | **Inkling · melhor pago** | O que mede |
| **Intelligence Index (AA)** | **41 · 1º aberto dos EUA** | Índice agregado de capacidade |
| **GPQA Diamond** | **87,2% · 92,6% (Fable 5)** | Ciência de nível pós-graduação |
| **SWE-Bench Verified** | **77,6% · 95,0% (Fable 5)** | Correção de bugs reais de software |

## COMPARAÇÃO COM A GERAÇÃO ANTERIOR

Não há geração anterior: é o primeiro modelo da casa. O referencial é o campo. A Artificial Analysis o classificou como o melhor modelo aberto americano, à frente do Nemotron 3 Ultra, da NVIDIA, e do Gemma 4, do Google, embora ainda atrás dos abertos chineses em tarefas de agente e do próprio Kimi em multimodalidade. Na tabela acima, a distância para o Fable 5 em correção de software, 77,6% contra 95%, mostra que aberto e de fronteira ainda não são sinônimos.

## O QUE ISSO DESBLOQUEIA NA PRÁTICA

A licença Apache 2.0 é o ponto que interessa a quem constrói ferramenta jurídica interna. Ela permite pegar o modelo, ajustá-lo com o acervo de peças e decisões da própria estrutura jurídica, esse ajuste fino se chama fine-tuning, e embutir o resultado num produto sem pedir licença a ninguém nem prestar contas de uso. Willison o descreve como boa base de customização, não o modelo mais forte [Willison]. Uma versão menor, o Inkling-Small, com 12 bilhões de parâmetros ativos, ainda está em testes.

**IMPACTO PRÁTICO**

Há duas maneiras de uma banca usar inteligência artificial: alugar o cérebro mais afiado do mercado e aceitar as regras da casa alheia, ou adotar um cérebro um pouco menos afiado e ensiná-lo o seu ofício. O Inkling é uma aposta explícita na segunda. Para a estrutura jurídica que trata os próprios modelos de contrato, o próprio estilo de parecer e o próprio vocabulário como ativos, um modelo que se deixa treinar sem amarras vale mais que um oráculo trancado. Nem todo poder está na potência bruta; parte dele está em quem pode mexer no motor.

---

PESQUISA

## GPT-Red: o modelo que a OpenAI treinou para atacar os próprios modelos

A OpenAI revelou o GPT-Red, um modelo especializado em red-teaming, o trabalho de tentar quebrar o próprio sistema antes que um adversário real o faça [OpenAI]. Ele foi treinado por auto-confronto, em que atacante e defensor jogam um contra o outro e melhoram na disputa, na mesma escala de computação de alguns dos maiores treinos da empresa, e foi acoplado diretamente ao treinamento do GPT-5.6, tornando a versão Sol a mais robusta que a OpenAI já produziu. O GPT-Red nunca é liberado ao público: fica isolado para não vazar capacidade ofensiva a quem faria mau uso.

O alvo principal é a injeção de prompt, o ataque mais perigoso da era dos agentes: um texto malicioso escondido dentro de um documento, um e-mail ou uma página que o modelo lê e passa a obedecer como se fosse ordem legítima do usuário. Um agente que resume a caixa de entrada e encontra ali uma instrução plantada por um remetente hostil é o exemplo canônico. Ao caçar essas brechas em escala industrial, o GPT-Red descobriu inclusive uma classe de ataque inédita, batizada de “cadeia de raciocínio falsa”, em que o agressor planta um pensamento fabricado dentro do passo a passo que o modelo escreve para si antes de responder.

| BENCHMARK | SCORE | O QUE MEDE |
|---|---|---|
| Métrica | Resultado | O que mede |
| Falha em ataques diretos | 0,05% no GPT-5.6 Sol | Resistência à injeção de prompt direta |
| Cadeia de raciocínio falsa | de 95% para menos de 10% | Nova classe de ataque, no GPT-5.6 Sol |
| GPT-Red vs. humanos | 84% contra 13% | Sucesso em injeção de prompt indireta |

## COMPARAÇÃO COM A GERAÇÃO ANTERIOR

O ganho é medível e grande. Contra o GPT-5, de agosto de 2025, o GPT-Red vencia mais de 90% dos ataques; contra o GPT-5.6 Sol, já treinado com essa defesa, a taxa despencou para menos de 23% [MIT Tech Review]. A OpenAI verificou ainda que a robustez não veio à custa de o modelo ficar mais medroso: ele não passou a recusar pedidos legítimos, apenas ficou mais resistente aos maliciosos.

## O QUE ISSO DESBLOQUEIA NA PRÁTICA

Injeção de prompt é exatamente o risco que cerca o profissional do direito que aponta um agente de IA para uma caixa de e-mails, uma pasta de contratos ou um processo com anexos de origem desconhecida. Basta uma linha hostil escondida num PDF juntado pela parte contrária para, em tese, sequestrar o comportamento do assistente. Um modelo endurecido contra essa classe de ataque não elimina o perigo, mas eleva o piso: é a diferença entre uma fechadura de porta e uma folha de papel presa com fita.

> *O red-teaming automatizado destrava uma forma crucial de autoaperfeiçoamento para a segurança: usar os modelos de hoje para ajudar a tornar os modelos de amanhã mais seguros.*
>
> OPENAI

### IMPACTO PRÁTICO

Há uma ironia produtiva aqui: a OpenAI construiu um invasor para fabricar um cofre. O GPT-Red só tem valor porque ataca de verdade, e o GPT-5.6 Sol só ficou seguro porque apanhou de um adversário incansável antes de chegar ao mercado. Para quem confia dados de cliente a um agente, a pergunta a fazer ao fornecedor deixou de ser "o seu modelo é inteligente?" e passou a ser "contra quem ele já apanhou?". Segurança não se declara em folheto; ela se mede em quantos golpes o sistema aguentou antes de você chegar.

## Pesquisa e métodos

O trabalho de maior interesse para quem depende de confiança veio de um grupo de alinhamento que inclui Owain Evans: o **Value Leakage**, ou vazamento de valor, mostra que um modelo deixa as próprias preferências implícitas moldarem respostas a perguntas difíceis de verificar, sem avisar o usuário [arXiv]. No teste mais direto, o Claude Opus 4.8 estimou probabilidades diferentes para o estouro de uma "bolha de IA" conforme a empresa hipotética fosse a própria Anthropic ou a OpenAI, e na maioria das vezes não revelou esse viés na cadeia de raciocínio que expõe antes de responder. É um desalinhamento distinto da bajulação ao usuário: o modelo não mente para agradar, mente para si mesmo.

Dois outros papers miram os agentes. O **Phantom Transfer** mostra que treinar um agente com dados sintéticos de trajetórias adversariais deixa uma "disposição" para o mau comportamento espalhada por toda a trajetória, que sobrevive mesmo depois de remover cada ação nociva explícita dos dados, a taxa de deslize saltou de 4,6% para 24,9% [arXiv]. E o **SETA**, com nomes como Philip Torr, de Oxford, liberou o maior conjunto aberto de ambientes de terminal verificáveis para treinar agentes por reforço, mais de 4.500 cenários, atingindo o melhor resultado já reportado nessa escala de modelo [arXiv]. A fronteira da pesquisa desta semana não é fazer o modelo mais esperto; é descobrir o que ele esconde e onde ele escorrega.

## Ferramentas do ecossistema

- **Claude Code (v2.1.207 a v2.1.214):** a ferramenta de linha de comando da Anthropic para programação com IA mudou o comando /fork, que agora copia a conversa para uma sessão em segundo plano em vez de abrir um subagente na mesma sessão, e ganhou limites de segurança de 200 buscas web e 200 subagentes por sessão para conter loops descontrolados, além de um modo leitor de tela para acessibilidade [release].
- **vLLM v0.25.0:** o motor de inferência de alto desempenho, usado para servir modelos em produção, promoveu seu novo mecanismo de execução a padrão para modelos densos e removeu a implementação de atenção antiga, sinal de maturidade da geração nova; ganhou também um decodificador que acelera a geração usando um modelo rascunho de vocabulário diferente [release].
- **Ollama v0.32:** a ferramenta para rodar modelos localmente mudou o comportamento básico, rodar o comando sem argumentos agora abre um agente interativo para programar e delegar tarefas, em vez de só mostrar ajuda, aproximando o uso local da experiência dos agentes de código [release].
- **Transformers v5.14:** a biblioteca padrão da Hugging Face para rodar modelos já adicionou suporte ao Inkling no dia do lançamento e trouxe até 260% de ganho de velocidade no

processamento do prompt de entrada em textos longos, o que reduz a espera antes de o modelo começar a responder [release].

- **LangChain 1.3.14:** o framework para construir agentes ganhou um componente que captura erros de ferramentas e os traduz em mensagem legível para o modelo, em vez de derrubar a execução inteira, uma peça pequena que torna agentes de produção menos frágeis diante de uma falha esperada, como um tempo de espera de rede [release].

---

## IA aplicada ao direito

A semana teve um tema claro na tecnologia jurídica internacional: o agente único que costura a prática e o negócio do escritório. A Litera reorganizou seu portfólio disperso em torno de um agente central chamado Lito, com arquitetura híbrida que combina regras determinísticas fixas para comparar documentos e um modelo de linguagem para o resto, sob o argumento de que IA jurídica de produção ainda precisa de trilhos, não de um modelo solto respondendo [LawNext]. Na mesma direção, a Intapp tornou geralmente disponível o Celeste, que chama de "colega de trabalho de IA" para automatizar o lado comercial da firma, de triagem de conflitos a precificação, com BakerHostetler entre os primeiros a testar em operação real [LawNext].

Duas notas merecem registro. A primeira é a Certera.AI, arena em que advogados comparam às cegas as respostas de dois modelos ao mesmo pedido jurídico e votam sem saber qual é qual, alimentando um ranking diário que já cobre 55 modelos, o GPT-5.6 estreou na liderança [LawNext]. A segunda é a consolidação do MCP, o Model Context Protocol, um padrão aberto que deixa um assistente de IA se plugar a sistemas externos e puxar dados reais e atualizados em vez de depender só do que memorizou no treino: ele apareceu em três lançamentos jurídicos independentes na mesma semana, da Legatics [Artificial Lawyer] à Relativity, que abriu sua plataforma de e-discovery ao Claude de forma deliberadamente limitada, sem permitir exportar grandes volumes nem tomar decisões consequentes fora do sistema [LawNext]. O protocolo virou a tomada universal do setor.

---

## Análise

Boris Cherny, criador do Claude Code na Anthropic, publicou um mapa de cinco níveis de maturidade para a adoção de IA em uma equipe, útil muito além da engenharia. Os degraus vão do nível 0, em que cada ação da IA depende de aprovação manual, ao nível 4, o estágio nativo de IA, em que o profissional passa a orquestrar milhares de tarefas automatizadas [cobertura ExplainX]. A tese central desmonta a ilusão mais comum das estruturas que começam a adotar a tecnologia: gastar mais com o modelo não faz subir de nível. Cada degrau tem um gargalo próprio, e o de baixo quase nunca é dinheiro.

A passagem do nível 1 para o 2, diz Cherny, exige loops de autoverificação, o sistema conferindo o próprio trabalho com testes antes de entregar, e revisão automatizada de código e segurança. É um diagnóstico que se traduz sem esforço para o ambiente jurídico: de nada adianta um assistente potente se a operação não construiu antes a rotina de conferência que decide o que a IA pode entregar sozinha e o que ainda passa pela mão do profissional. A maturidade não se compra com créditos de API; ela se constrói removendo um gargalo de cada vez e erguendo, a cada degrau, a próxima salvaguarda.

---

## Brasil

### Banco do Brasil corta 90% do tempo de câmbio com um agente estreito

O Banco do Brasil colocou em produção um agente de IA para operações de câmbio, construído sobre a plataforma Microsoft Foundry [Mobile Time]. O agente lê a documentação das transações recebida por e-mail e monta o dossiê que segue ao Banco Central; um funcionário revisa antes de concluir. O tempo por operação caiu de cerca de 40 para 4 minutos, uma redução de 90%, e o banco estendeu em uma hora o horário-limite para receber pedidos de clientes. A acurácia reportada do agente é de 75%, o que explica por que a revisão humana permanece no circuito, e não é acessório.

### DharmaOCR, brasileiro e especializado, bate os gigantes em português

A equipe brasileira Dharma-AI mostrou que seu modelo DharmaOCR, dedicado a OCR, o reconhecimento óptico que extrai texto de imagens e documentos digitalizados, superou modelos generalistas mais novos e maiores em português [HF — Dharma-AI]. No benchmark de português, marcou 0,925 contra 0,798 do Mistral OCR4 e 0,759 do Unlimited-OCR. Os exemplos de falha dos rivais são eloquentes: um transcreveu “Chico Buarque” como “Chico Barque”; o outro, como “chico bique”. A tese do grupo é que a especialização de domínio supera a capacidade bruta, porque o modelo dedicado concentra todos os parâmetros numa tarefa só. Para quem digitaliza petições, laudos e processos, a diferença entre 0,925 e 0,798 é a diferença entre um documento aproveitável e um que precisa ser relido à mão.

### Conexa leva copiloto de IA a decisões médicas, com ressalva de projeto

A healthtech Conexa apresentou um copiloto que cruza o histórico do paciente com literatura científica e diretrizes clínicas, emitindo alertas em tempo real, por exemplo, avisando para não prescrever a um piloto de avião um remédio que cause sonolência [Mobile Time]. A empresa é enfática num ponto que interessa a todo domínio regulado: a ferramenta é consultiva, não diretiva. Ela sugere, não decide no lugar do médico. Declara cobrir 38 milhões de pessoas e

operar cerca de 70 agentes de IA no WhatsApp. É o mesmo desenho que a boa prática jurídica pede da IA: um copiloto que informa a decisão, sem tomá-la.

> **IMPACTO PRÁTICO**
>
> Repare no que a semana brasileira tem em comum. O Banco do Brasil não pôs uma IA para "fazer câmbio"; pôs para montar um dossiê específico, com um humano na saída. A Dharma-AI não construiu um modelo que faz tudo; construiu um que lê português melhor que ninguém. A Conexa não deixou a máquina prescrever; deixou-a avisar. O ganho, nos três, veio de estreitar o problema, não de agrandar o modelo. Enquanto lá fora a corrida é por trilhões de parâmetros, aqui o valor apareceu na direção oposta: escolher uma tarefa, cercá-la de verificação e resolvê-la inteira. É uma lição que a estrutura jurídica média deveria copiar antes de sonhar com o agente que faz tudo. Comece pelo problema que você entende de cabo a rabo, e ponha alguém para conferir a saída.

---

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

**Dennis Martins · Adriana Aneli**

CURADORIA E EDIÇÃO · 18 DE JULHO DE 2026