



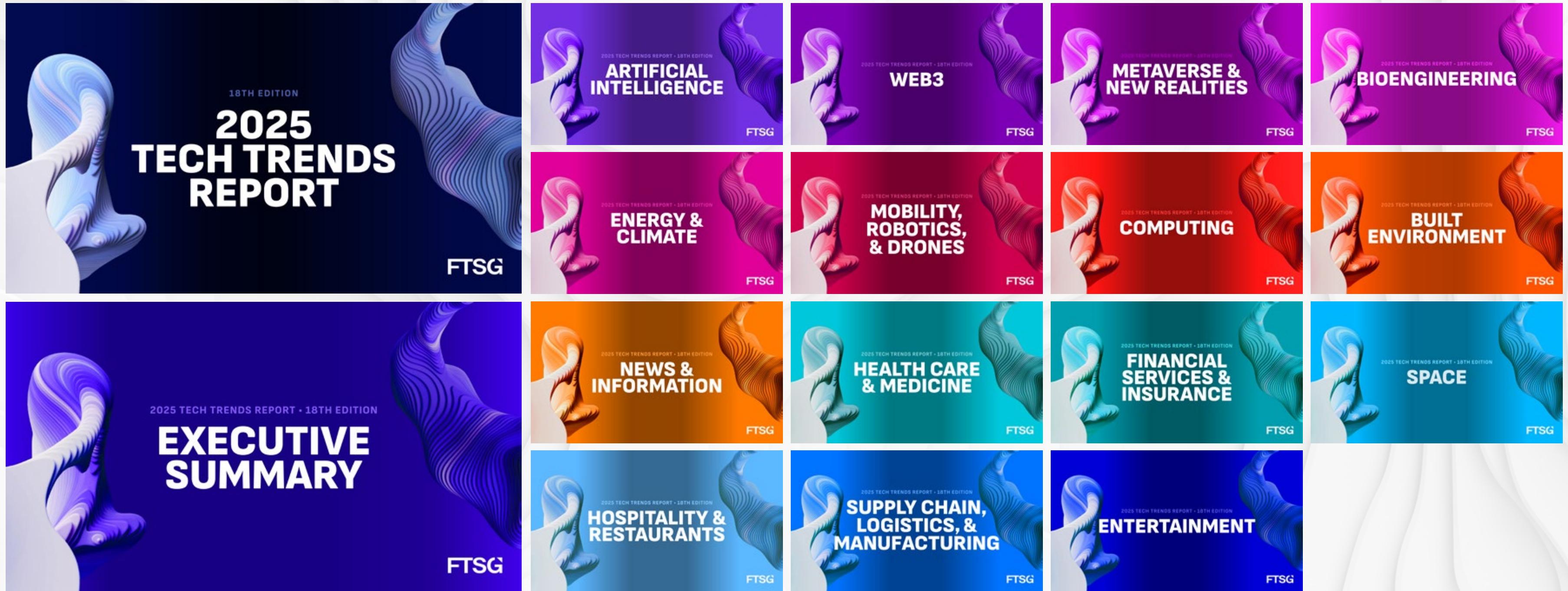
2025 TECH TRENDS REPORT • 18TH EDITION

# INTELIGÊNCIA ARTIFICIAL

FTSG

# Relatório de tendências tecnológicas de 2025 do Future Today Strategy Group

Nossa edição de 2025 inclui 1000 páginas, com centenas de tendências publicadas individualmente em 15 volumes e como um relatório abrangente. Baixe todas as seções do relatório Tech Trends 2025 do Future Today Strategy Group em [www.ftsg.com/trends](http://www.ftsg.com/trends).



<b>08</b>	<b>AS 5 PRINCIPAIS COISAS QUE VOCÊ PRECISA SABER</b>	
13	POR QUE AS TENDÊNCIAS DE INTELIGÊNCIA ARTIFICIAL SÃO IMPORTANTES PARA SUA ORGANIZAÇÃO	
<b>24</b>	<b>MODELOS, TÉCNICAS E PESQUISA</b>	
25	Expansão das modalidades de IA generativa	36 Desvio da missão no alinhamento da IA
25	Afinação	36 Indexação de confiança
26	Aprendizagem de reforço automatizada	36 Detectores de Deepfake em tempo real
26	Composição Evolutiva	37 Marca d'água
26	Mistura de especialistas	37 IA segura para crianças
27	Autonomia dos Especialistas	38 IA como ferramenta para lidar com preconceitos políticos
28	LLMs como sistemas operacionais	38 IA com preconceito de gênero e raça
28	LLMs: Maiores e mais caros	39 Uso indevido nefasto de IA
28	Modelos de cadeia de pensamento	40 Envenenamento de dados: uma faca de dois gumes
30	Pequenos modelos de linguagem	40 Vigilância Cidadã
30	Aterramento e Aumento de Contexto	40 Vigilância do Trabalhador
30	Superando a escassez de dados	41 IA póstuma
31	IA de código aberto	42 Riscos de privacidade na biometria comportamental
32	Grandes modelos de ação	
32	Modelos de Ação Grande Pessoal	
<b>34</b>	<b>SEGURANÇA, ÉTICA E SOCIEDADE</b>	
35	IA explicável (XAI)	<b>43</b> <b>IA e ENERGIA</b>
35	Otimização de IA	44 IA faminta por recursos
35	Alinhamento de IA descentralizado	44 Renascimento Nuclear da IA
		45 Algoritmos de IA eficientes
		46 Otimização de Energia
		<b>47</b> <b>GEOPOLÍTICA DE IA, DEFESA E COMBATE À GUERRA</b>
		49 Nacionalismo da IA
		49 A Guerra dos Chips Impulsionada pela IA
		50 Diplomacia da IA
		50 Pivôs tecnológicos na defesa
		<b>54</b> <b>POLÍTICA E REGULAMENTAÇÃO</b>
		56 Estados Unidos: Acelerando a IA rapidamente
		57 União Europeia: Impulsionando fortemente a governança da IA
		58 China: Estratégia dirigida pelo Estado e supervisão rigorosa
		59 Brasil: No caminho para a legislação da IA
		60 Emirados Árabes Unidos: Equilibrando Inovação com Diretrizes
		<b>61</b> <b>CAPACIDADES EMERGENTES</b>
		62 IA em Matemática
		62 Agentes que usam computadores
		63 Raciocínio de IA
		63 Comunicação de IA para IA
		64 Detectando Emoção
		64 Agentes Encarnados
		65 IA neuro-simbólica

**66 INTERAÇÕES HUMANO-IA**

- 67 IAs convencem humanos
- 67 Humanos convencem a IA
- 68 Predição e presciênciia em nossas vidas humanas
- 68 IA no dispositivo
- 69 IA vestível
- 69 Interfaces de usuário generativas

**70 O NEGÓCIO DA IA**

- 71 Integração vertical de hardware para LLMs
- 72 Bifurcação de preços
- 72 Otimizando a IA para rodar na e para a borda
- 72 O mercado de dados de treinamento de IA
- 73 A IA dá vida aos sistemas legados

**74 TALENTOS E EDUCAÇÃO**

- 75 Fuga de cérebros da IA da academia
- 75 Aumento da educação em IA
- 75 Economia de duas velocidades da IA
- 76 Agentes: de assistentes a atores
- 76 Trabalho Complementar
- 77 Educação Assistida por IA
- 77 Educação nativa de IA

**78 CRIATIVIDADE & DESIGN**

- 79 Criatividade Assistida por GAN
- 79 Renderização Neural
- 79 Gerando Ambientes Virtuais
- 80 IA como meio de conteúdo
- 80 IA democratiza a produção musical
- 80 Dublagem automática de ruído ambiente
- 81 Invenção assistida por IA

**82 INDUSTRIAS - FARMACEUTICA**

- 83 Dobramento de proteínas
- 83 Desenvolvimento de medicamentos com IA em primeiro lugar
- 84 Design de anticorpos generativos

**85 INDUSTRIAS - ASSISTÊNCIA MÉDICA**

- 85 Diagnóstico Assistido por IA e Tomada de Decisão Clínica
- 85 Detecção de anomalias em imagens médicas
- 85 Pessoas capacitadas por IA
- 86 LLMs específicos para cuidados de saúde
- 86 Deepfakes médicos
- 85 Pessoas capacitadas por IA

**88 INDUSTRIAS - CIÊNCIA**

- 88 Raciocínio científico multietapas
- 88 Hipóteses baseadas em IA
- 88 Experimentação orientada por IA
- 89 Análise e interpretação com tecnologia de IA
- 89 IA para acelerar o desenvolvimento de novos materiais
- 90 Decodificação Animal

**91 INDUSTRIAS - FINANÇAS**

- 91 Precificação e gestão de ativos assistida por IA
- 91 Mitigando Fraudes
- 91 Previsão de Risco Financeiro
- 92 Portfólios Personalizados
- 92 Robo-Advisors voltados para o consumidor

**93 INDUSTRIAS - SEGUROS**

- 93 Previsão de lesões no local de trabalho
- 93 Melhorando a avaliação de danos
- 93 Prevenção de incêndio com tecnologia de IA
- 94 O Trabalhador Conectado
- 94 Seguro de Responsabilidade Civil para IA

**95 INDUSTRIAS - RECURSOS HUMANOS**

95 Aquisição Autônoma de Talentos

95 Integração e integração de IA

95 Engajamento e retenção de funcionários

96 Seleção e Gestão de Benefícios

**97 INDUSTRIAS - MARKETING**

97 Pesquisa de turnos de IA

97 Engajamento dinâmico por meio de personalização profunda

97 Campanhas assistidas por IA

98 Observações anedóticas, agora dados de marketing utilizáveis



Amy Webb Diretora Executiva



Sam Jordan Líder de Tecnologia e Computação

## A vanguarda da IA agora muda por hora, não por ano. O que vem depois?

A IA está se movendo a uma velocidade alucinante, remodelando indústrias, fluxos de trabalho e a vida cotidiana mais rápido do que podemos documentar. No dia em que escrevemos isso, as pessoas ainda estavam maravilhadas com o DeepSeek da China, que alcançou o desempenho de primeira linha do OpenAI com uma fração do preço e do poder de computação usuais — desafiando tudo o que pensávamos que sabíamos sobre o que é preciso para construir IA avançada. Horas depois, pesquisadores de Stanford e da Universidade de Washington estrearam outro novo modelo, o s1, que superou os modelos de raciocínio R1 do DeepSeek e o1 do OpenAI usando ainda menos recursos.

Essa é a natureza da IA atualmente: o que é vanguarda hoje pode ser notícia velha... mais tarde hoje.

Aqui está o que sabemos com certeza. No ano passado, o CEO da OpenAI, Sam Altman, se reuniu com gestores de fundos soberanos e investidores, esperando levantar até US\$ 7 trilhões para uma empresa de chips de IA. Em janeiro de 2025, a Stargate, uma joint venture recém-formada entre a OpenAI, Oracle e SoftBank, disse que levantaria US\$ 500 bilhões para chips, data centers de IA e seus enormes requisitos de energia. Para não ficar para trás, Microsoft, Meta e Google anunciaram planos para investir centenas de bilhões de dólares em infraestrutura de IA. Mas se agora alguém pode replicar um modelo multimilionário com apenas recursos modestos, os modelos de IA não se tornarão rapidamente comoditizados? Se assim for, isso pressionaria as Big Tech a se moverem muito rápido, construindo e dimensionando sistemas de IA em constante avanço para permanecerem competitivas no mercado.

Na corrida para vencer a IA, a avaliação crítica se tornou uma vítima da velocidade. Nós nos reunimos regularmente com as equipes de pesquisa que constroem modelos SOTA, chefes de laboratórios de fronteira trabalhando para avançar a IA e executivos dos grandes gigantes da tecnologia. Embora estejamos certamente animados com o incrível progresso tecnológico que está sendo feito, há a realidade prática da prontidão organizacional. Os criadores de sistemas de IA — e as empresas de serviços profissionais que prometem transformação da noite para o dia — estão operando em uma realidade muito distante das organizações cotidianas.

O que observamos no ano passado aconselhando CEOs e suas equipes de gestão sobre estratégia e implementação de IA é que, independentemente dos desenvolvimentos tentadores da IA, a maioria das organizações enfrenta uma dívida técnica substancial na padronização e manutenção de dados, criando atrito operacional na implantação. Eles também estão lutando com os fundamentos da gestão de mudanças, que muitas vezes são despriorizados (ou esquecidos completamente) antes da implementação. Como resultado, estamos vendo novos riscos estratégicos para organizações que superindexam na prontidão tecnológica sem abordar barreiras operacionais e culturais fundamentais para a implantação.

Avanços revolucionários também aceleraram a corrida de IA entre os EUA e a China, intensificando-a em uma disputa geopolítica completa, com ambas as nações alavancando a tecnologia como uma ferramenta para influência global. A saber: durante a posse de Donald Trump televisionada globalmente, executivos das maiores empresas de tecnologia dos Estados Unidos sentaram-se diretamente atrás dele — enquanto seus indicados para o gabinete e membros da família sentaram-se em fileiras mais atrás. Essa rivalidade EUA-China está forçando os aliados a tomarem partido, aumentando as tensões e alimentando preocupações sobre segurança nacional, cadeias de suprimentos e soberania tecnológica. O que surge é um cenário de IA fragmentado, dominado por uma Guerra Fria Digital que ameaça remodelar alianças globais e estruturas de poder econômico.

Como — exatamente — a IA remodelará nosso mundo nos próximos meses? A resposta honesta é: ninguém pode saber. Nesta fase, evitar erros dispendiosos e planejar de forma inteligente o futuro importa mais do que prever resultados exatos. Líderes precisam de uma bússola estratégica, não de uma bola de cristal.

Esse é o propósito deste relatório de tendências: destacar tendências emergentes de IA e casos de uso para que você possa planejar múltiplas possibilidades. Porque em um cenário de IA se movendo em velocidade de dobra, a clareza estratégica é sua vantagem competitiva.

Espere um frenesi contínuo de atividade enquanto as empresas de IA competem por participação de mercado, embora os investimentos e as políticas concentrem a influência entre vários participantes importantes.

1

## Mais forte, melhor, mais rápido

O R1 da DeepSeek e o s1 de Stanford e da Universidade de Washington alcançaram fortes capacidades de raciocínio, mantendo-se economicamente eficientes, desafiando a convenção de que o progresso requer modelos cada vez maiores e levantando questões sobre a escalabilidade futura da IA.

2

## Sua IA agora tem olhos e ouvidos

Avanços recentes em IA multimodal, como o Google Gemini Live e o Sora da OpenAI, estão transformando rapidamente a maneira como as máquinas processam e geram texto, áudio e vídeo, desbloqueando novas possibilidades para experiências de IA mais ricas e interativas.

3

## Aprendendo a pensar

O desempenho da IA está melhorando, elevando modelos como o o1 da OpenAI e o Gemini 2.0 Flash Thinking Mode do Google de meros mecanismos de informação para parceiros de pensamento. No final de 2024, o o3 da OpenAI pontuou 85% no benchmark ARC-AGI, igualando a pontuação humana média.

4

## Da assistência à autonomia

A Agentic AI está evoluindo de tarefas de suporte para raciocínio autônomo e tomada de ação em fluxos de trabalho. Este ano, os agentes de IA não apenas auxiliarão, mas também executarão processos complexos, transformando indústrias com maior eficiência e automação.

5

## EUA e China avançam

Os EUA e a China estão presos em uma competição de alto risco pelo domínio da IA, moldando o futuro da tecnologia e do poder global. À medida que ambas as nações investem em pesquisa, infraestrutura e regulamentação de IA, elas estão redefinindo inovação, segurança e influência econômica.

# A inteligência artificial reformulará fundamentalmente a dinâmica e as fronteiras competitivas em 2025.



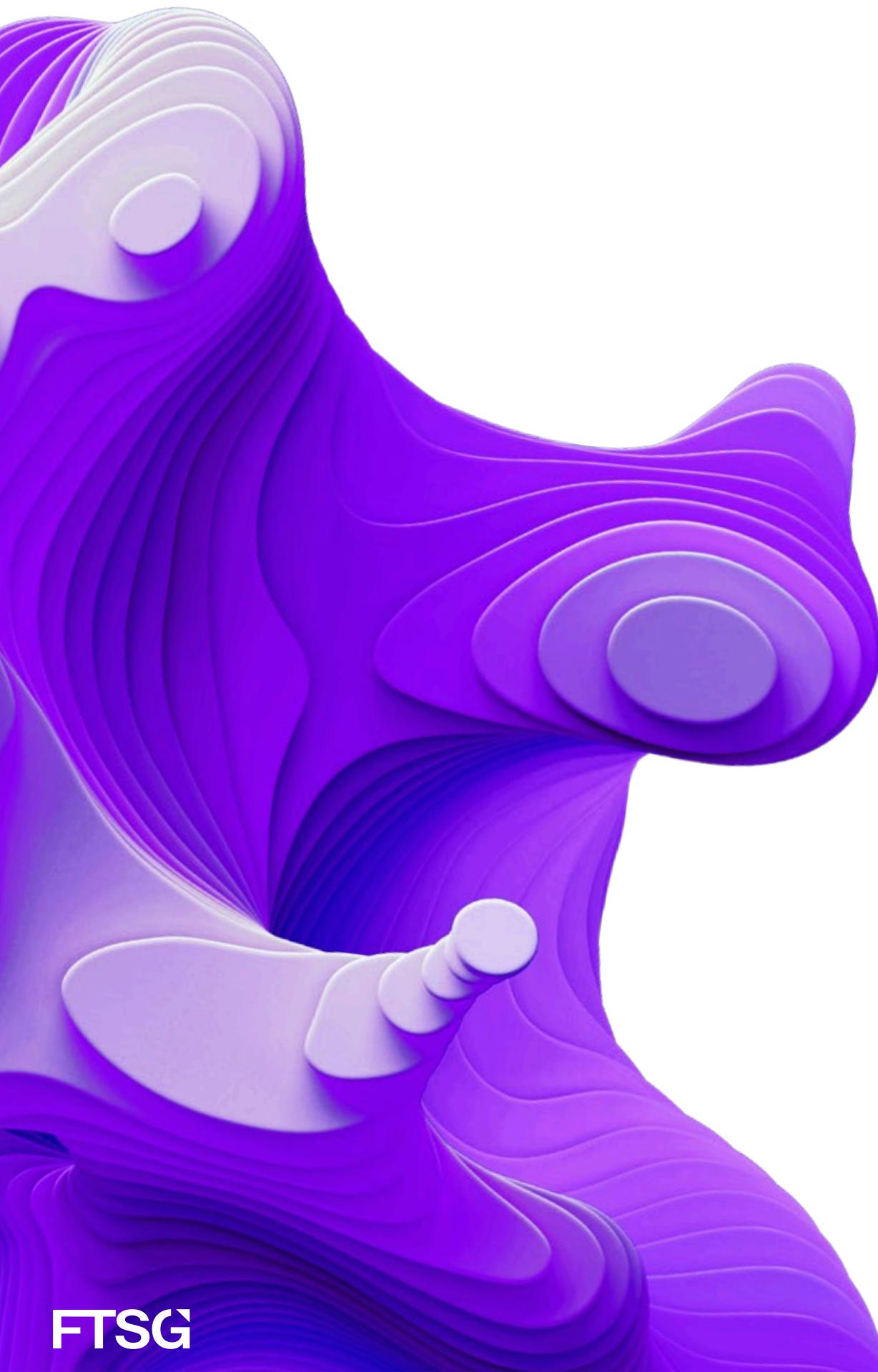
Não é segredo que o cenário da IA se transformou drasticamente desde que escrevemos a seção State of Play no ano passado. O GPT-4 estabeleceu benchmarks iniciais com suas capacidades multimodais e desempenho de nível profissional, mas desde então, o Gemini Ultra do Google DeepMind elevou ainda mais o nível, excedendo o GPT-4 na maioria dos benchmarks.

O campo se dividiu entre abordagens proprietárias e de código aberto. O lançamento do Llama pela Meta desencadeou uma revolução de código aberto, com desenvolvedores rapidamente ajustando variantes que rivalizam com modelos comerciais maiores. Esse sucesso levou até mesmo o CEO da OpenAI, Sam Altman, a admitir que a empresa pode ter estado “do lado errado da história” em relação a sistemas fechados.

As parcerias de nuvem se tornaram cruciais. A Microsoft apostou alto no OpenAI (US\$ 10 bilhões), enquanto a Amazon e o Google dividiram seu suporte para o Anthropic (US\$ 4 bilhões e US\$ 2 bilhões, respectivamente). Essas alianças fornecem às empresas de IA um enorme poder de computação, ao mesmo tempo em que garantem as posições dos provedores de nuvem na corrida da IA. No entanto, essa concentração de recursos levanta preocupações sobre a consolidação do mercado. A China emergiu como uma formidável potência de IA. O Ernie 4.0 do Baidu alega desempenho de nível GPT-4, enquanto o Alibaba lançou mais de 100 modelos de código aberto sob o Qwen 2.5.

O chatbot Doubao da ByteDance ganhou uma fatia significativa do mercado. Também fazendo grandes avanços estão Zhipu AI, MiniMax, Baichuan Intelligence, Moonshot, StepFun e 01.AI – conhecidos coletivamente como os “Seis Pequenos Tigres” do país. Apesar das restrições de chips dos EUA, as empresas chinesas estão se adaptando com alternativas domésticas como os chips Ascend da Huawei e Kunlun da Baidu.

O investimento explodiu, com mais de US\$ 22 bilhões fluindo para startups de IA generativa somente no ano passado, representando quase metade de todo o financiamento de IA. Os VCs tradicionais estão competindo com gigantes da tecnologia que jogam bilhões em IA, elevando as avaliações para as alturas.



## Vemos seis temas macro emergindo:

- 1 As grandes empresas de tecnologia continuarão a dominar o financiamento, muitas vezes através de parcerias estratégicas
- 2 As avaliações estão atingindo níveis da era das pontocom, levantando preocupações legítimas sobre bolhas
- 3 Os setores de tecnologia tradicionais estão vendo o financiamento secar à medida que a IA suga o oxigênio da sala
- 4 O investimento está fluindo diretamente para provedores de nuvem para poder de computação
- 5 A competição está se intensificando entre modelos proprietários e de código aberto
- 6 A China está rapidamente fechando a lacuna com as capacidades ocidentais de IA

Olhando para o futuro, as empresas de IA financiadas devem provar valor comercial real. Os vencedores provavelmente serão aqueles que conseguirem equilibrar inovação com monetização sustentável enquanto navegam pelo crescente escrutínio regulatório.

# Gigantes da IA correram para a AGI enquanto rivais chineses se mostraram formidáveis.

MAIO 2024

## OpenAI lança GPT-4o

O novo modelo de IA é capaz de raciocinar em tempo real por meio de entradas de áudio, visuais e de texto.

JANEIRO DE 2025

## Nvidia entra no mercado de PCs com IA

A Nvidia revela seu Projeto DIGITS, um supercomputador de IA pessoal.

JANEIRO DE 2025

## Gigantes da tecnologia apresentam plano de IA de 500 bilhões

Líderes políticos, tecnológicos e financeiros anunciam o Stargate, uma joint venture que visa investir US\$ 500 bilhões em infraestrutura de IA nos EUA.

DEZEMBRO DE 2024

## O3 se aproxima da AGI

A OpenAI diz que seu modelo o3 passou no desafio ARC-AGI, considerado uma referência líder em inteligência artificial geral.

JANEIRO DE 2025

## DeepSeek interrompe

A empresa chinesa de IA DeepSeek lança o R1, seu modelo de raciocínio e concorrente do o1 da OpenAI.

&lt;&lt; PASSADO

# Os titãs da tecnologia enfrentarão testes cruciais de IA em 2025.

MARÇO DE 2025

## Huang fará uma prévia dos chips de IA de próxima geração

A Conferência de Tecnologia de GPU da Nvidia contará com a palestra do CEO Jensen Huang sobre o que esperar deste importante fabricante de chips.

FUTURO >>

MAIO 2025

## O crescimento da Nvidia é testado

A Nvidia pode apresentar receita recorde, a menos que o DeepSeek preveja um futuro alternativo que exija menos chips.

MAIO 2025  
IA Integra

As novas atualizações de IA generativa do Google I/O sinalizarão como a IA se integrará ainda mais aos serviços ao consumidor e às ferramentas empresariais.

JUNHO DE 2025  
Apple aposta alto em uma Siri mais inteligente

Em um momento decisivo para provar a necessidade da IA em produtos de consumo, a Apple está pronta para estrear uma nova Siri com IA generativa.

JULHO DE 2025

## China mostra sua mão de IA

A Conferência Mundial de IA apresentará as últimas inovações e iniciativas de IA da China.

# Além do entusiasmo pela IA, essas sete mudanças estruturais já estão determinando quais organizações irão prosperar.

## A velocidade é a nova escala

Para o bem ou para o mal, a IA está comprimindo os ciclos de decisão de semanas para minutos. A vantagem está mudando para organizações que podem aproveitar a IA para experimentação e aprendizado rápidos — desde que estejam tomando boas decisões sobre governança de dados, seleção de fornecedores e gerenciamento de mudanças.

## Seus concorrentes não vão esperar

Alguém no seu setor já está usando IA para cortar custos e aumentar a produtividade em 30%-40%. A IA está automatizando certas tarefas de conhecimento que exigem muita mão de obra, mas em breve levará a novos fluxos de trabalho e modelos de negócios. A questão não é se sua organização se adaptará à IA, mas se você fará isso antes ou depois que suas margens forem espremidas de todas as direções.

## O Middle Office está derretendo

A IA está automatizando tarefas de coordenação e tomada de decisão que tradicionalmente exigiam middleware humano, e organizações que se apegam a processos manuais de coordenação e aprovação se verão estruturalmente não competitivas. O organograma do futuro é mais plano e rápido.

## A Guerra dos Talentos Tem Novas Regras

Os melhores talentos agora esperam trabalhar com as melhores ferramentas. Eles não estão apenas procurando por um bom pagamento — eles querem locais de trabalho habilitados para IA que multipliquem seu impacto. Sua capacidade de atrair e reter os melhores desempenhos depende cada vez mais da sua prontidão para IA.

## Um imposto oculto sobre a IA

O custo da IA não está em comprar a tecnologia, mas em potencializá-la. À medida que a demanda por computação de IA dispara, as organizações enfrentarão uma nova realidade econômica: Compre cedo para começar sua transformação de IA, mas, involuntariamente, pague um prêmio cada vez mais alto depois.

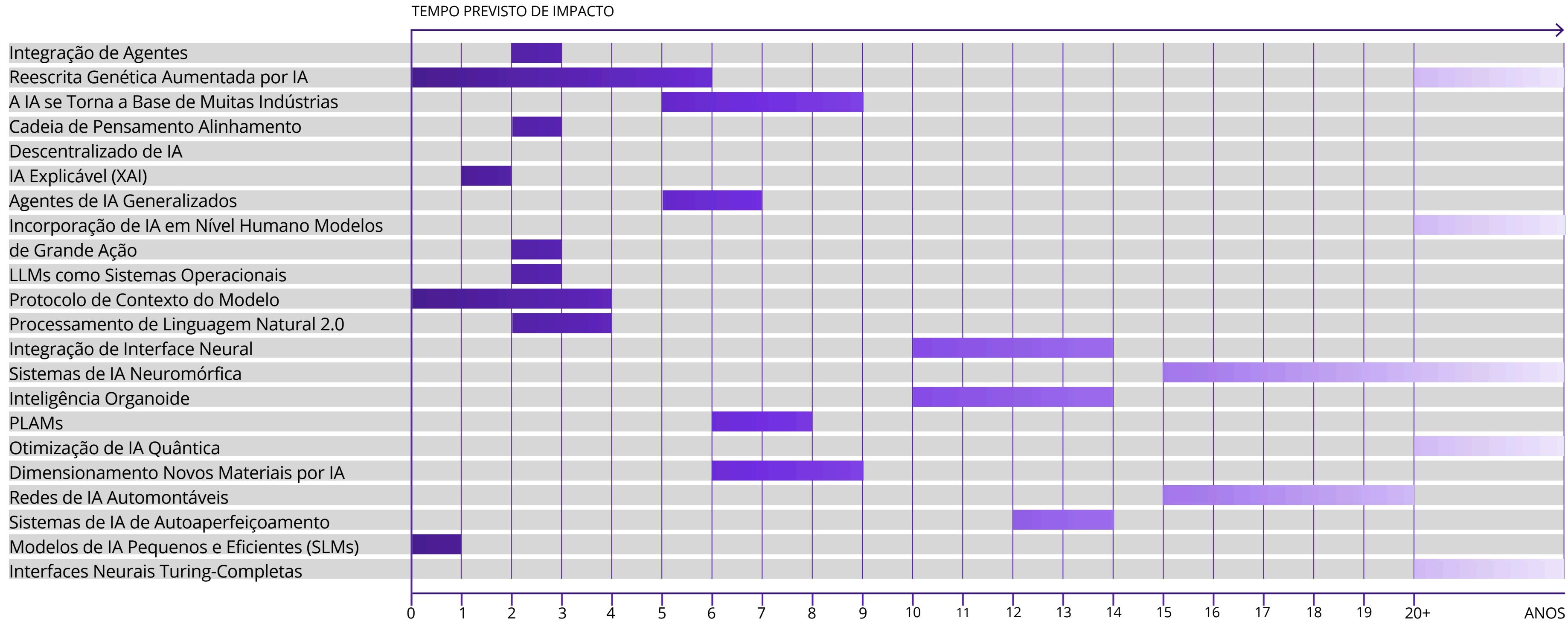
## Sua interface está custando clientes

A linguagem natural está devorando sua interface de usuário, transformando-a em abandonware. Cada aplicativo, banco de dados e sistema em breve estará acessível por meio de conversas simples. Organizações que se apegam a interfaces tradicionais se encontrarão com o equivalente corporativo de um telefone flip em um mundo de iPhone.

## Escalar para fora em vez de escalar para cima

A história mostra que as principais tecnologias inevitavelmente se tornam mais baratas e mais amplamente distribuídas. O mesmo padrão está surgindo na IA, e as empresas que só perseguem modelos e data centers cada vez maiores correm o risco de serem interrompidas por novatos ágeis que podem criar sistemas menores e mais eficientes, a um custo menor.

As tecnologias gerativas avançam nos próximos anos, enquanto métodos de computação como computação organoide e neuromórfica impulsionam desenvolvimentos a longo prazo.



# Abaixo, destacamos desenvolvimentos de alto nível e de curto prazo que vale a pena ficar de olho em todos os setores.

## ESCALAÇÃO

Enormes quantidades de dados de treinamento ainda são necessárias para a maioria dos modelos de IA aprenderem. Por exemplo, sistemas de recomendação acoplados à IA generativa podem levar a uma personalização profunda para os setores de hospitalidade e assistência médica — desde que os dados sejam disponibilizados. Historicamente, os dados são bloqueados dentro de sistemas proprietários construídos por terceiros, e a regulamentação frequentemente dificulta o acesso a certas formas de dados.

## INVESTIMENTO

A IA viu ciclos de entusiasmo e desilusão, levando a muito ou pouco capital. Os investidores priorizam a comercialização em vez de P&D básico — embora este último produza maior impacto e, muitas vezes, retornos mais fortes. A paciência dos investidores influenciará o progresso e a comercialização.

## RESTRIÇÕES À ADOÇÃO

Mesmo que uma tecnologia esteja amadurecendo, restrições à sua adoção podem prejudicar seu impacto. Por exemplo, uma empresa pode se recusar a adotar um sistema automatizado porque ele desafia a ortodoxia existente ou uma estratégia bem-sucedida existente. Isso é especialmente verdadeiro em assistência médica, seguros e serviços financeiros.

## REGULAMENTOS

Os avanços na tecnologia geralmente superam as mudanças regulatórias. Isso beneficiou a IA, que até muito recentemente não era alvo de regulamentação. Além disso, fatores como se as regulamentações locais são conflitantes ou complementares podem influenciar a adoção no mercado.

## MENÇÕES NA MÍDIA

Maior conscientização e entusiasmo podem influenciar o ímpeto de uma tecnologia, mesmo quando não houve nenhuma inovação real. Futuras explosões de mídia impulsionarão o ímpeto da IA, especialmente se essas histórias forem facilmente compreendidas pelo público.

## PERCEPÇÃO PÚBLICA

Como o público entende e responde aos avanços da IA criará ou reprimirá a demanda. Isso é especialmente verdadeiro para IA generativa em educação/criatividade/propriedade intelectual/desinformação, bem como o papel que as tecnologias assistivas desempenharão na formação da força de trabalho futura.

## DESENVOLVIMENTOS DE P&D

O ritmo de novas descobertas de pesquisa não pode ser programado para coincidir com uma reunião do conselho ou relatório de lucros. Fatores como financiamento, qualidade, tamanho da equipe e acesso a recursos podem melhorar a probabilidade e a velocidade de novas descobertas. Monitoramos de perto os desenvolvimentos de P&D, mas os tratamos como curings.

Esperamos ouvir frequentemente das maiores empresas de tecnologia e IA do mundo. Por esse motivo, nossa lista de 2025 destaca indivíduos que voam mais profundamente sob o radar público.

◆ **Dra. Adji Bousoo Dieng,**

Professor assistente em **Princeton**

Por seu trabalho em profundidade modelagem gráfica probabilística.

◆ **Alexandre Wang,**

fundador e CEO da **Scale**,

O modelo de IA de código aberto da DeepSeek e para criar uma plataforma de anotação de dados líder que acelera o desenvolvimento de modelos de IA em vários setores.

◆ **Dr. A.S. Abeba Birhane,**

bolsista Sênior da **Mozilla Foundation**, por sua pesquisa sobre as implicações éticas da IA e críticas aos vieses algorítmicos.

◆ **Dra. Anima Anandkumar,**

Professor Bren de Ciências da Computação e Matemática

na **Caltech**, por desenvolver algoritmos IA que aceleram a descoberta científica, incluindo estruturas como operadores neurais para simulações eficientes.

◆ **Dra. Chinasa T. Okolo,**

cientista da computação e pesquisadora da **Brookings Institution**, por seu trabalho na defesa da adoção responsável de IA no Sul Global e contribuir para relatórios internacionais de segurança de IA.

◆ **Clément Delangue,**

CEO no **Hugging Face**, pela democratização acesso a modelos de PNL de última geração e promoção de uma comunidade de IA de código aberto.

◆ **Dra. Cíntia Rudin,**

o Gilbert, Louis e Edward Lehrman Professor Emérito da

**Duke University**,

Por seu trabalho em modelos interpretáveis de aprendizado de máquina e IA ética.

◆ **Dr. Dan Hendrycks,**

diretor do Centro de IA

Pela pesquisa pioneira em Segurança de IA e desenvolvimento do Último Exame da Humanidade, um teste projetado para avaliar riscos de IA, em colaboração com a Scale AI.

◆ **Reitor Bola,**

bolsista de pesquisa sobre IA e Progresso no Mercatus Center da Universidade de

**Georgetown** e autor do Substack

“Hyperdimensional” focado em IA por sua análise sobre governança de IA.

◆ **Dra. Devi Parikh,**

professor na Georgia Tech e cofundadora de Yutori,

Pelo trabalho pioneiro em respostas visuais a perguntas e modelos de visão e linguagem, que ajudaram a estabelecer referências fundamentais sobre como os sistemas de IA entendem e raciocinam sobre informações visuais em contextos de linguagem natural.

◆ **Dr. Ilya Sutskever,**

cofundador da Safe

Pelo pioneirismo em trabalhar com aprendizado profundo e liderar esforços para desenvolver IA que supere a inteligência humana, permanecendo alinhada aos interesses humanos.

◆ **Dra. Joelle Pineau,**  
presidente de pesquisa de IA na **Meta**,  
para avanços de liderança  
na pesquisa de IA e na promoção da  
ciência aberta, contribuindo para  
desenvolvimentos como o modelo de  
linguagem de código aberto LLamA.

◆ **Dra. Kazumi Fukuda,**  
cientista pesquisador em IA na **Sony**  
por sua pesquisa em corporificação  
inteligência, particularmente no  
desenvolvimento de modelos computacionais  
que permitem aos robôs perceber, planejar e  
agir em ambientes dinâmicos.

◆ **Dra. Li Deng,**  
diretor de IA na **Vatic Investments**,  
por suas contribuições para o  
reconhecimento de fala e  
aprendizagem profunda.

◆ **Dr. Liang Wenfeng,**  
CEO na **DeepSeek**, para desenvolvimento  
o modelo R1 AI, que rivaliza com os principais  
concorrentes em capacidade, mas opera por  
uma fração do custo.

◆ **Dra. Lila Ibrahim,**  
diretor chefe de operações do  
**Google**  
Por orientar a integração de  
pesquisas de IA em aplicações  
práticas e iniciativas líderes para  
aplicar IA em produtos de consumo.

◆ **Que Habib,**  
CEO e cofundador da **Writer**,  
Por desenvolvimento de ferramentas de  
IA empresarial que ajudam as empresas a  
gerar e gerenciar conteúdo de alta  
qualidade, garantindo ao mesmo tempo a  
consistência e a conformidade da marca.

◆ **Dr. Nathan Lambert,**  
pesquisador científico do Instituto  
Allen de IA (**Ai2**) e autor  
do blog Interconnects,  
Por suas contribuições para a ciência  
aberta do ajuste fino do modelo de  
linguagem.

◆ **Dra. Pedro Abbeel,**  
diretor do **Berkeley Robot**  
**Learning Lab** e codiretor do  
Berkeley Artificial Intelligence  
Lab  
por seu trabalho em robótica  
e aprendizagem por reforço.

◆ **Dra. Sasha Luccioni,**  
Líder de IA na **Hugging Face**,  
para desenvolver ferramentas para  
medir a pegada de carbono de  
modelos de IA e defender práticas de  
IA ambientalmente responsáveis.

◆ **Doutor Sheng Shen,**  
cientista do **Google**, para  
coautoria de “Mixture-of-Experts  
Meets Instruction Tuning: A Winning  
Combination for Large Language  
Models”, explorando a integração de  
arquiteturas MoE com ajuste de  
instruções para aprimorar o  
desempenho do modelo de linguagem.

◆ **Dr. Tim Brooks,**  
líder de \ equipe mundial de  
modelagem de IA no **Google**  
**DeepMind**,

esforços no desenvolvimento de  
“modelos mundiais” capazes de  
simular ambientes físicos, avançando  
a IA incorporada em jogos e robótica.

◆ **Dra. Yang Zhilin,**  
CEO em **Moonshot AI**, por liderar  
o desenvolvimento de modelos  
de IA com compreensão de longo  
contexto e expansão de  
aplicações de IA globalmente.



# A adoção da IA está criando oportunidades sem precedentes para criação de valor...

## OPORTUNIDADES

### Construir modelo interno de IA Estruturas de avaliação

As empresas que agirem agora ganharão vantagem do pioneiro. Estas estruturas críticas permitirão uma rápida avaliação e implantação de soluções IA enquanto os concorrentes lutam com métodos de avaliação ad-hoc.

### Crie conhecimento com tecnologia de IA Sistemas de Gestão

Esses sistemas transformam a estática documentação em bases de conhecimento que aprendem e se adaptam continuamente, permitindo que as empresas desbloqueiem oportunidades significativas. Gerando valor do conhecimento institucional e funcionários aposentados.

### Incorpore recursos de IA diretamente nas principais ofertas

A maioria dos CEOs vê a IA como uma ferramenta de corte de custos, perdendo seu potencial de criar novos fluxos de receita por meio de produtos e serviços aprimorados que transformam as experiências do cliente e os modelos de negócios.

### Invista em modelos de IA específicos de domínio

Esse tipo de modelo focado proporcionará desempenho superior em aplicações específicas, exigindo menos dados e recursos de computação do que alternativas de uso geral, e os primeiros investidores serão beneficiados.

# ...mas as organizações enfrentam riscos crescentes devido à complexidade técnica e à escassez de talentos.

## AMEAÇAS

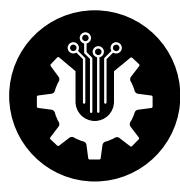
Avalie a dívida técnica oculta. Clusters de IA de alto valor são alvos principais para ameaças cibernéticas sofisticadas, tornando o investimento em segurança cibernética essencial. Os riscos são especialmente altos, pois os ladrões que podem usar clusters de IA comprometidos para roubar modelos proprietários, impactando indústrias globalmente.

Atrasar a adoção da IA significa aumento dos custos com talentos profissionais de IA, tornando proibitivamente caro ao se transformar mais tarde. CEOs frequentemente delegam estratégia de IA para equipes técnicas, ignorando as maiores mudanças organizacionais e culturais necessárias.

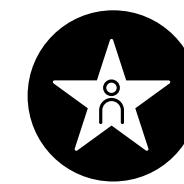
A combinação de vantagens de dados pode concentrar poder. Os pioneiros da IA estão acumulando enormes conjuntos de dados proprietários e pipelines de treinamento que criam barreiras de entrada quase intransponíveis. Essa vantagem se acumula ao longo do tempo e ameaça bloquear novos jogadores.

Riscos da Colonização Digital. Nações sem capacidades soberanas de IA tornam-se vulneráveis à colonização digital, à medida que sistemas estrangeiros de IA moldam o acesso à informação e a tomada de decisões dos seus residentes.

As empresas enfrentam altos custos ocultos e obstáculos organizacionais complexos à medida que correm para implementar IA, desde infraestrutura desatualizada até resistência dos funcionários e demandas regulatórias.



À medida que as organizações se apressam para implementar a IA, elas estão enfrentando uma realidade custosa: a maior parte de sua infraestrutura de dados está décadas atrás do que é necessário. O investimento necessário para modernizar a arquitetura de dados geralmente excede os orçamentos iniciais do projeto de IA em 5–10x, criando uma barreira oculta à transformação.



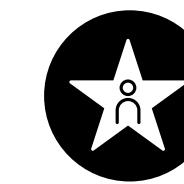
Enquanto as empresas investem avidamente em IA, elas frequentemente negligenciam investimentos cruciais em gestão de mudanças e suporte a funcionários. Essa supervisão pode levar gerentes de nível médio a ver a IA como uma ameaça em vez de uma ferramenta, fomentando resistência que pode retardar drasticamente a implementação e a adoção.



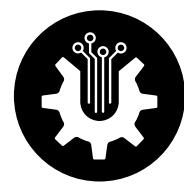
Crie modelos de IA com ambientes de teste dedicados para avaliar rigorosamente o risco financeiro em relação às condições históricas de mercado e cenários de estresse. Esses sistemas exigem hardware especializado e infraestrutura de monitoramento de conformidade para lidar com simulações complexas, mantendo trilhas de auditoria.



Estabeleça protocolos rigorosos para auditar modelos de IA que permitam o rastreamento de decisões, linhagem de dados e desvio de desempenho entre sistemas. Integrar a automação com a supervisão humana é crucial, juntamente com trilhas de auditoria detalhadas que estejam em conformidade com a governança técnica e as regulamentações.



Conheça suas restrições. Encontre parceiros externos confiáveis com expertise tanto em IA quanto em seu domínio, ou faça isso internamente. Essas equipes devem unir suas habilidades em IA com conhecimento específico do setor para criar soluções que abordem desafios regulatórios e técnicos, ao mesmo tempo em que entregam valor mensurável.



Cada pedaço de dívida técnica se torna um gargalo crítico, pois os sistemas de IA exigem infraestrutura mais flexível e interconectada. Empresas inteligentes tratarão a eliminação da dívida técnica como uma parte central da estratégia de IA, não uma iniciativa separada de TI.



# Termos importantes que você precisa saber antes de ler.

## AGENTE IA

Isso se refere a sistemas de IA que exibem capacidades autônomas de tomada de decisão, definição de metas e resolução adaptativa de problemas. Ao contrário dos modelos tradicionais de IA que geram respostas passivamente com base em prompts do usuário, a IA agêntica toma ações proativamente, interage com seu ambiente e refina suas estratégias ao longo do tempo.

## AGENTES

Entidades alimentadas por IA que percebem seu ambiente, tomam decisões e agem de forma autônoma para atingir objetivos específicos. Os agentes podem variar de ferramentas de automação simples a sistemas de IA complexos e multimodais que interagem dinamicamente com usuários e outros sistemas.

## AGI (INTELIGÊNCIA GERAL ARTIFICIAL)

Uma designação para sistemas de IA que correspondem e então excedem a gama completa de habilidades cognitivas humanas em todas as tarefas economicamente valiosas. AGI permanece teórica, mas suas potenciais implicações para mercados de trabalho, governança e segurança global são ativamente debatidas.

## ÉTICA DA IA

Um campo multidisciplinar que estuda os riscos sociais, econômicos e éticos da IA, incluindo viés, privacidade, desinformação e ameaças existenciais. Estruturas de ética da IA orientam políticas e regulamentações para garantir que o desenvolvimento da IA esteja alinhado com os valores humanos.

## GOVERNANÇA DE IA

**Os sistemas, políticas e acordos internacionais que regulam o desenvolvimento, a implantação e a supervisão de tecnologias de IA. A governança de IA é crítica para mitigar riscos, garantir competição justa e abordar tensões geopolíticas em torno das capacidades de IA.**

## ALGORITMO

Um conjunto estruturado de regras ou processos para resolver problemas específicos ou executar tarefas. Em IA, algoritmos determinam como os dados são processados, insights são gerados e decisões são tomadas.

## ALINHAMENTO

O processo de garantir que os objetivos, comportamentos e tomada de decisão de um sistema de IA estejam alinhados com as intenções humanas, princípios éticos e padrões regulatórios. O desalinhamento pode resultar em consequências não intencionais, incluindo resultados tendenciosos ou prejudiciais.

## SUPERINTELIGÊNCIA ARTIFICIAL (ASI)

Um hipotético sistema de IA futuro que supera a inteligência humana em todos os domínios, incluindo criatividade, sabedoria geral, planejamento estratégico e descoberta científica. ASI levanta questões complexas sobre controle, governança e risco existencial.

## RECONHECIMENTO AUTOMÁTICO DE FALA (ASR)

Sistemas orientados por IA que convertem linguagem falada em texto escrito. ASR capacita assistentes virtuais, serviços de transcrição e interfaces de voz multilíngues em aplicativos empresariais e de consumo.

## IA AUTÔNOMA

Sistemas de IA capazes de tomada de decisão independente e execução de tarefas sem intervenção humana. A IA autônoma é crítica em robótica, finanças, segurança cibernética e aplicações militares, exigindo salvaguardas rigorosas para garantir o uso responsável.

## RACIOCÍNIO EM CADEIA DE PENSAMENTO (COT)

Um método de raciocínio de IA em que modelos resolvem problemas passo a passo, imitando dedução lógica semelhante à humana. Isso melhora o desempenho em tarefas complexas de tomada de decisão, incluindo matemática, análise jurídica e diagnósticos médicos.

## VISÃO COMPUTACIONAL

Tecnologia orientada por IA que permite que máquinas processem, analisem e derivem significado de imagens e vídeos digitais. Usada em vigilância de segurança, automação industrial, imagens médicas e veículos autônomos.



### IA DE BORDA

Modelos de IA que rodam diretamente em dispositivos de ponta (por exemplo, smartphones, sensores de IoT, drones autônomos) em vez de servidores de nuvem centralizados. A IA de ponta permite processamento em tempo real, reduz latência e aprimora a privacidade de dados.

### MODELO DE FUNDAÇÃO

Um modelo de IA em larga escala pré-treinado em grandes quantidades de dados e adaptável a múltiplas tarefas sem exigir retreinamento do zero. Modelos de base sustentam aplicações modernas de IA, incluindo IA generativa, sistemas autônomos e automação empresarial.

### IA GENERATIVA (GERAL)

Tecnologias de IA capazes de gerar novos conteúdos, incluindo texto, imagens, música, vídeo e código. A GenAI está transformando indústrias como mídia, design, marketing e atendimento ao cliente, ao mesmo tempo em que levanta preocupações sobre propriedade intelectual e desinformação.

### GPU (UNIDADE DE PROCESSAMENTO GRÁFICO)

Hardware especializado otimizado para computação paralela, acelerando treinamento de modelos de IA, aprendizado profundo e tarefas de computação de alto desempenho. GPUs são essenciais para executar modelos de IA em larga escala e simulações intensivas em dados.

### MODELO

Um sistema de IA treinado que analisa dados para fazer previsões, gerar insights ou automatizar a tomada de decisões. Os modelos variam em complexidade, de modelos de regressão simples a arquiteturas avançadas de deep learning.

### MULTIMODAL AI

Sistemas de IA que processam e integram vários tipos de dados — como texto, imagens, vídeo e áudio — para melhorar a compreensão contextual e a tomada de decisões. A IA multimodal potencializa chatbots avançados, assistentes virtuais e diagnósticos médicos.

### PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)

Processos orientados por IA que permitem que máquinas entendam, interpretem e gerem linguagem humana. A PLN potencializa chatbots, serviços de tradução, análise de sentimentos e moderação automatizada de conteúdo.

### PESQUISA DE ARQUITETURA NEURAL (NAS)

Um método baseado em IA para otimizar automaticamente estruturas de redes neurais, melhorando o desempenho e reduzindo a necessidade de ajuste manual por pesquisadores.

### PARÂMETRO

Uma variável interna de um modelo de IA que é ajustada durante o treinamento para melhorar a precisão e a eficiência. Grandes modelos de IA contêm bilhões de parâmetros, tornando seu treinamento computacionalmente intensivo.

### ENGENHARIA PROMPT

A prática de projetar entradas efetivas (prompts) para guiar modelos de IA na geração de saídas desejadas. A engenharia de prompts é crucial para otimizar o desempenho de IA generativa em aplicativos comerciais e criativos.

### QUANTUM IA

A intersecção da computação quântica e IA, onde algoritmos quânticos aumentam a eficiência do aprendizado de máquina. A IA quântica tem o potencial de revolucionar a criptografia, a ciência dos materiais e os problemas de otimização.

### SISTEMAS DE RECOMENDAÇÃO

Algoritmos orientados por IA que analisam o comportamento e as preferências do usuário para sugerir produtos, conteúdo ou serviços relevantes. Usados em e-commerce, plataformas de streaming e publicidade digital.

### APRENDIZAGEM POR REFORÇO A PARTIR DO FEEDBACK HUMANO (RHF)

Um método de treinamento em que modelos de IA aprendem por meio de feedback iterativo de avaliadores humanos, melhorando sua precisão, alinhamento ético e usabilidade em aplicações do mundo real.



### APRENDIZAGEM AUTO-SUPERVISADA

Uma abordagem de machine learning onde modelos de IA aprendem com dados brutos e não rotulados, identificando padrões e relacionamentos dentro do conjunto de dados. Este método reduz a dependência de dados de treinamento rotulados por humanos.

### APRENDIZAGEM SUPERVISIONADA

Um método de treinamento em que modelos de IA aprendem a partir de conjuntos de dados rotulados, usando pares de entrada e saída conhecidos para melhorar a precisão preditiva em novos dados.

### IA SIMBÓLICA

Uma abordagem de IA que representa conhecimento usando símbolos legíveis por humanos e regras lógicas, permitindo raciocínio e resolução de problemas. Frequentemente usada em sistemas especialistas e modelos de IA explicáveis.

### DADOS SINTÉTICOS

Dados gerados artificialmente usados para treinar modelos de IA quando dados do mundo real são escassos, tendenciosos ou sensíveis à privacidade. Dados sintéticos melhoram o desempenho da IA ao mesmo tempo em que mitigam riscos de coleta de dados.

### DADOS DE TREINAMENTO

O conjunto de dados usado para treinar modelos de IA identificando padrões, tomando decisões ou gerando previsões. A qualidade e a diversidade dos dados de treinamento impactam significativamente a precisão e a imparcialidade do modelo.

### IA CONFIÁVEL

Sistemas de IA projetados com transparência, justiça, responsabilidade e segurança para promover a confiança pública e a conformidade regulatória. IA confiável é um foco importante para estratégias de IA governamentais e empresariais.

### APRENDIZAGEM NÃO SUPERVISIONADA

Uma abordagem de aprendizado de máquina em que modelos de IA detectam padrões e estruturas em dados sem saídas rotuladas, permitindo tarefas como agrupamento e detecção de anomalias.

### XAI (IA EXPLICÁVEL)

Sistemas de IA projetados para fornecer explicações transparentes e interpretáveis por humanos para seus processos de tomada de decisão, aumentando a responsabilização e a confiança em aplicações de alto risco, como assistência médica e finanças.

### APRENDIZAGEM ZERO-SHOT (ZSL)

Uma técnica de IA onde modelos generalizam conhecimento de conceitos aprendidos anteriormente para executar tarefas sem treinamento prévio direto nessas tarefas. Usado em aplicações como tradução de idiomas e reconhecimento de imagem.

# TENDÊNCIAS DE INTELIGÊNCIA ARTIFICIAL

# MODELOS, TÉCNICAS E PESQUISA



## MODELOS, TÉCNICAS E PESQUISA

### Os modelos de IA exigem grandes volumes de dados e recursos de computação para desbloquear seu potencial transformador (ou assim pensávamos).

#### Expansão das modalidades de IA generativa

Os humanos não aprendem apenas lendo — nós observamos, ouvimos e sintetizamos informações de várias fontes. A IA agora está seguindo o exemplo, integrando entradas como texto, imagens e som para preencher a lacuna entre o que descrevemos e o que as máquinas podem entender completamente. 2024 marcou o ano em que as capacidades multimodais de IA não apenas amadureceram, mas começaram a transformar aplicações do mundo real. O GPT-4o da OpenAI baseia-se no multimodal progresso de 2023 integrando texto, visão e voz em um modelo robusto. Seus recursos de conversação em tempo real abrem

portas para interações fluidas e humanas. A capacidade do GPT-4o de analisar e gerar insights a partir de entradas combinadas de texto, áudio e imagem permite que ele resolva tarefas complexas com profundidade e precisão notáveis. O Claude 3 da Anthropic traz interpretação visual sofisticada para ambientes empresariais, onde cerca de metade das bases de conhecimento são baseadas em imagens. Essa capacidade é particularmente transformadora na área da saúde, onde a IA agora pode conectar imagens médicas com registros de pacientes para diagnósticos aprimorados. Do lado do consumidor, a IA multimodal está começando a se tornar uma segunda natureza. Em vez de digitar consultas, os usuários agora compartilham fotos de receitas para ajustar porções ou carregam imagens de erupções cutâneas para sugestões médicas. Embora essa democratização torne a expertise mais acessível, ela também levanta questões sobre precisão e limites éticos quando as ferramentas de IA substituem o julgamento profissional.

O Large Language Model for Mixed Reality (LLMR) do MIT e da Microsoft está expandindo ainda mais os limites multimodais.

LLMR usa IA para simplificar a criação

e modificação de ambientes virtuais. Em vez de precisar de codificação complexa, o LLMR permite que os usuários descrevam sua visão em linguagem simples, e o sistema transforma essas palavras em experiências interativas de realidade mista em tempo real. Por exemplo, um usuário pode dizer: "Coloque um banco verde no parque ao lado da fonte", e o sistema o executa instantaneamente. Embora os avanços de 2024 na IA multimodal marquem um salto tecnológico, seu verdadeiro significado reside na forma como estão sendo remodelados.

abordando a relação fundamental entre humanos e máquinas, levando-nos de dar comandos a ter conversas.

#### Afinação

O ajuste fino — o processo de refinar LLMs em conjuntos de dados especializados — está melhorando nossa capacidade de personalizar e controlar sistemas de IA. Em 2023, o avanço QLoRA da Universidade de Washington marcou um ponto de virada, permitindo o ajuste fino de modelos massivos de 65 bilhões de parâmetros em uma única GPU com apenas 48 GB de memória — uma melhoria de eficiência de 16 vezes em relação aos sistemas tradicionais

métodos. Com base nessa fundação, em 2024, a Answer.AI integrou o QLoRA com o processamento Fully Sharded Data Parallel, tornando possível treinar modelos de 70 bilhões de parâmetros em hardware de nível de consumidor. Essa democratização tem implicações profundas: pesquisadores e desenvolvedores agora podem experimentar modelos de linguagem em larga escala sem acesso à infraestrutura cara de data center.

O impacto se estende além da acessibilidade. Nas ciências biológicas, os pesquisadores adaptaram técnicas de ajuste fino para modelos de linguagem de proteínas (PLMs), que são treinados em conjuntos de dados extensivos de sequências de proteínas. Eses modelos estão agora sendo ajustados para prever a estabilidade, funções e interações de proteínas com precisão notável.

Os PLMs ajustados superam os não ajustados homólogos ajustados em vários bancos marcas, apresentando capacidades preditivas aprimoradas.

O ajuste fino não é apenas uma capacidade técnica — é uma vantagem empresarial estratégica. No contexto empresarial, as empresas podem



## MODELOS, TÉCNICAS E PESQUISA

use o ajuste fino para personalizar modelos de IA poderosos como GPT e Claude para suas necessidades específicas sem construir do zero. Os provedores de assistência médica podem aprimorar os recursos de diagnóstico treinando modelos em registros de pacientes anônimos, mantendo a conformidade com a HIPAA, enquanto as instituições financeiras podem incorporar requisitos regulatórios — do GDPR ao PCI DSS — diretamente em seus fluxos de trabalho de IA. Isso reduz drasticamente os custos de desenvolvimento e o tempo de colocação no mercado, ao mesmo tempo em que garante que os sistemas de IA falem a linguagem da organização, entendam os contextos específicos do setor e operem dentro das estruturas de conformidade necessárias.

### Aprendizagem de reforço automatizada

O treinamento tradicional de IA usando Aprendizado por Reforço a partir de Feedback Humano (RLHF) envolve pessoas avaliando respostas de IA para ajudar a melhorar o sistema. Embora esse método funcione bem, ele é caro e demorado. A DeepSeek encontrou uma alternativa inteligente — a startup desenvolveu uma maneira de treinar sistemas de IA usando feedback automatizado de computador em vez de avaliações humanas.

tarefas mais subjetivas (como escrita criativa ou perguntas abertas) ainda precisam de alguma contribuição humana, o método automatizado do DeepSeek funciona especialmente bem para tarefas com respostas certas/erradas claras, como problemas de matemática e codificação. Para tornar esse treinamento automatizado ainda mais eficiente, o DeepSeek criou um método especial chamado GRPO (Group Relative Policy Optimization) e o testou primeiro com seu modelo focado em matemática. A empresa não está sozinha. A Microsoft Asia desenvolveu um modelo matemático usando técnicas comparáveis, a Ai2 criou um modelo chamado Tulu que combina feedback automatizado e humano, e a Hugging Face está trabalhando na recriação da abordagem do DeepSeek para entender melhor como ele funciona. A principal conclusão é que o DeepSeek mostrou que é possível criar sistemas de IA de alto desempenho com menos dependência de feedback humano caro, especialmente para certos tipos de tarefas. Isso pode tornar o desenvolvimento de IA mais eficiente e econômico, embora a contribuição humana ainda seja valiosa para algumas aplicações.

### Composição Evolutiva

A Sakana AI está desafiando a sabedoria convencional de que modelos maiores e mais caros são o único caminho para uma IA melhor. Em vez de treinar modelos massivos do zero — um processo que exige enormes recursos computacionais — a empresa japonesa desenvolveu uma alternativa elegante: usar algoritmos evolucionários para descobrir automaticamente maneiras ótimas de combinar modelos de IA existentes. Essa abordagem de "otimização evolucionária" é grande; ao mesclar inteligentemente modelos de diferentes domínios — como processamento de linguagem e compreensão visual — a Sakana cria sistemas híbridos que excedem as capacidades de seus componentes individuais. Os resultados são impressionantes: os experimentos produziram modelos de língua japonesa com raciocínio matemático aprimorado e consciência cultural que superaram sistemas maiores e mais intensivos em recursos.

As implicações vão muito além da tecnologia realização técnica. Esta metodologia

democratiza o desenvolvimento avançado de IA reduzindo a necessidade de infraestrutura de computação massiva e expertise especializada. Em vez de exigir dezenas de milhões em recursos de computação, os desenvolvedores agora podem criar modelos sofisticados multicapacitados combinando de forma inteligente os existentes. Mais significativamente, a abordagem de Sakana sugere um futuro onde o avanço da IA não é apenas sobre construir modelos maiores, mas sobre encontrar maneiras mais inteligentes de combinar os existentes. Assim como a natureza cria complexidade por meio da combinação e evolução de elementos mais simples, esse novo paradigma aponta para um caminho mais sustentável e acessível no desenvolvimento de IA — um caminho onde a inovação vem da composição inteligente em vez do dimensionamento por força bruta.

### Mistura de especialistas

Diferentemente da abordagem anterior que mescla modelos inteiros treinados, a mistura de especialistas (MoE) divide o trabalho dentro de uma única estrutura criando vários submodelos “especialistas” especializados. Pense nisso como ter uma equipe de pessoas onde uma pessoa é ótima

## MODELOS, TÉCNICAS E PESQUISA

em matemática, outro em escrita e outro em design, com um gerente que sabe quem chamar para cada tarefa. Esse “gerente” (o mecanismo de controle) direciona cada entrada para o especialista certo, de modo que cada parte do trabalho é tratada pelo especialista mais adequado para ela. Ao dividir as tarefas entre os especialistas e deixar o mecanismo de controle lidar com “quem faz o quê”, os modelos MoE podem se tornar mais eficientes e precisos do que se um modelo gigante e universal tentasse lidar com tudo sozinho.

Notavelmente, a revisão de janeiro de 2025 da DeepSeek Release, R1, usa MoE em seu núcleo. Conforme relatado, DeepSeek afirma ter construído um ChatGPT- como sistema por uma fração do custo usual empregando MoE (junto com outras técnicas como destilação de conhecimento e aprendizado por reforço). Como o MoE divide um grande modelo em “especialistas” especializados e depende de uma função de controle para encaminhar cada solicitação ao especialista mais apropriado, ele pode ser mais eficiente e potencialmente substancialmente menos dispendioso para treinar ou executar do que um único modelo monolítico gigante. O sucesso do DeepSeek com esta abordagem desencadeou

nova atenção ao MoE como uma alternativa viável para dimensionar a IA sem exigir hardware massivo e proibitivamente caro.

### Autonomia dos Especialistas

Embora o DeepSeek tenha sido o que colocou o MoE nas notícias para o público em geral, outros já haviam feito avanços significativos no campo. Pesquisadores da Universidade Renmin da China, Tencent e Southeast University divulgaram um artigo que descreve uma nova abordagem de “Autonomia de Especialistas” (AoE) para modelos de mistura de especialistas. Enquanto o modelo MoE típico depende de um “roteador” que faz sua melhor estimativa de qual especialista deve lidar com cada pergunta ou entrada recebida, o AoE não precisa do roteador. Em vez disso, cada especialista dá uma olhada na entrada e diz: “Eu posso lidar com isso” ou “Não, obrigado, essa não é minha especialidade”, com base em quão fortemente ele ilumina os sinais internos do especialista (as “normas de ativação”). Os sinais mais fortes vencem, então esses especialistas avançam e o resto recua. Em outras palavras, cada especialista decide autonomamente se é o melhor ajuste. Isso elimina o intermediário (o roteador) completamente.





# MODELOS, TÉCNICAS E PESQUISA

## LLMs como sistemas operacionais

Imagine um sistema operacional fundamentalmente alimentado por um modelo de linguagem grande (LLM), onde o LLM não é apenas um complemento, mas o núcleo do SO. Este SO poderia automatizar tarefas de rotina com sofisticação sem precedentes, eliminando a necessidade de intervenção manual. Ele iria além das interfaces gráficas de usuário tradicionais e interações de linha de comando, adotando uma abordagem mais intuitiva e baseada em linguagem natural. Os usuários poderiam interagir com seus computadores por meio de comandos de conversação, consultas ou solicitações para tarefas específicas, e o LLM interpretaria essas entradas, executando uma série de ações para entregar os resultados desejados.

Um desses projetos, o AIOS, prevê um LLM como o “cérebro” do sistema operacional. O AIOS otimiza a re-alocação de fonte, gerencia a troca de contexto, facilita a execução simultânea de agentes, fornece ferramentas para agentes e mantém o controle de acesso. O LLM lida com tomadas de decisões complexas, transformando o SO em um sistema mais inteligente e adaptável. Outro

projeto, MemGPT, foca em aprimorar sistemas orientados por LLM integrando memória de longo prazo e melhorando capacidades de raciocínio. LLMs tradicionais são limitados por pequenas janelas de contexto, restringindo a quantidade de informação que eles podem processar de uma vez. MemGPT aborda isso introduzindo uma arquitetura de memória multinível, inspirada por técnicas tradicionais de gerenciamento de memória do SO como memória virtual, para permitir processamento mais complexo e contextualmente consciente ao longo do tempo. Juntos, esses projetos representam o futuro dos sistemas operacionais centrados em LLM, permitindo interações mais eficientes, naturais e poderosas.

## LLMs: Maiores e mais caros

Os LLMs cresceram exponencialmente em tamanho e custo na última década, impulsionados pelo paradigma “quanto maior, melhor”. Essa abordagem surgiu das leis de escala, introduzidas pela primeira vez por Prasanth Kolachina em 2012 e posteriormente validadas por Kaplan et al. em 2020, que demonstraram uma forte correlação entre o tamanho do modelo e o desempenho. Segundo esses insights,

a indústria tem buscado sistemas cada vez maiores, progredindo de 1,5 bilhão de parâmetros do GPT-2 em 2019 para modelos com trilhões de parâmetros como GPT-4 e PaLM 2 em 2023. O impacto financeiro desse crescimento é significativo: as estimativas do Relatório de Índice de IA de 2024 de Stanford colocam os custos de treinamento de modelos de primeira linha em níveis sem precedentes, com o GPT-4 da OpenAI exigindo aproximadamente US\$ 78 milhões em computação e o Gemini Ultra do Google custando cerca de US\$ 191 milhões. Os benefícios da escala foram substanciais essencial e bem documentado. Modelos maiores demonstraram capacidade notável capacidades no manuseio de tarefas complexas, mostrando precisão e eficiência aprimoradas em uma ampla gama de aplicações. Essas conquistas

Os estudos validaram, pelo menos parcialmente, que maior é de fato melhor. No entanto, esse progresso veio com custos e desafios significativos. O treinamento do GPT-4 exigiu aproximadamente 10.000 vezes mais computação/recursos adicionais do que seu antecessor GPT-2, exigindo enormes investimentos em infraestrutura e hardware especializado.

Talvez o mais significativo seja que a relação entre o tamanho do modelo e o desempenho provou ser mais complexa do que inicialmente assumido. Muitas tarefas exibem retornos decrescentes à medida que os modelos crescem, questionando a viabilidade a longo prazo dessa abordagem. Essa observação tem relevância particular para as empresas, que estão descobrindo que modelos maiores não se traduzem automaticamente em melhores soluções para suas necessidades específicas. A combinação de custos crescentes, preocupações ambientais e benefícios de desempenho incertos levou a um exame crítico do paradigma de dimensionamento.

À medida que o campo amadurece, uma abordagem mais diferenciada para o desenvolvimento de IA está surgindo.

Em vez de perseguir apenas o tamanho, a pesquisa- Os fabricantes estão se concentrando cada vez mais em melhorias de eficiência e no desenvolvimento de modelos menores e especializados.

## Modelos de cadeia de pensamento

À medida que modelos maiores atingem limites práticos e financeiros, os pesquisadores estão mudando a atenção para novas abordagens, como a Cadeia de Pensamento (CoT), que enfatiza

## MODELOS, TÉCNICAS E PESQUISA

raciocínio mais profundo em tempo real em vez de contagens de parâmetros brutos. Por mais de uma década, o progresso da IA foi amplamente impulsionado pela lei de escala de pré-treinamento, que surgiu com AlexNet em 2012 e ganhou força com a arquitetura Transformer em 2017. Esta lei afirma que aumentar a quantidade de dados de treinamento (agora atingindo trilhões de tokens), expandir parâmetros de modelo e usar mais computação (FLOPS) leva a um melhor desempenho em várias tarefas. Simplificando, as leis de escala de pré-treinamento descrevem como modelos maiores, com mais dados e computação, alcançam desempenho superior.

Agora, há uma nova lei de escala na cidade. Anteriormente, a maior parte do custo computacional era concentrada no pré-treinamento. Uma vez que um modelo era treinado, executar inferência — gerar respostas ou concluir tarefas — exigia significativamente menos computação. A inferência era dimensionada de forma direta: quanto mais solicitações um modelo manipulava, mais computação ele usava.

Entretanto, a introdução de modelos CoT mudou fundamentalmente esse paradigma. O modelo o1 da OpenAI e o R1 da DeepSeek demonstraram que a computação de inferência é

não é mais estritamente proporcional ao comprimento da saída. Esses modelos geram “tokens lógicos” intermediários, agindo como um bloco de notas interno para dividir problemas em etapas de raciocínio estruturado. Essa mudança significa que quanto mais tokens dedicados a esse processo interno, melhor será a saída do modelo. Essencialmente, ele imita como os humanos melhoraram seu trabalho — checando duas vezes, verificando cálculos e fazendo referência cruzada de soluções para garantir a precisão. Espere uma ênfase crescente em estratégias de inferência dinâmica, onde os modelos podem ajustar de forma flexível o número de tokens lógicos internos com base na dificuldade da tarefa ou na precisão desejada. Como resultado, a computação de inferência pode se tornar muito mais significativa em relação ao treinamento, gerando a necessidade de soluções de hardware mais eficientes, melhores técnicas de otimização e novos modelos de negócios em torno da computação baseada em uso. No geral, o desenvolvimento da IA provavelmente mudará para arquiteturas e métodos que permitam que os modelos “pensem em voz alta”, permitindo um raciocínio mais profundo e melhores resultados ao custo de maior processamento em tempo real.





# MODELOS, TÉCNICAS E PESQUISA

## Pequenos modelos de linguagem

Pequenos modelos de linguagem (SLMs) estão provando que maior nem sempre é melhor. Esses modelos compactos podem igualar ou exceder o desempenho de suas contrapartes maiores em tarefas específicas, ao mesmo tempo em que exigem muito menos poder computacional e recursos. Os SLMs apresentaram desempenho impressionante em cenários específicos de tarefas, como classificação de texto zero-shot. Pesquisas em vários conjuntos de dados revelaram que modelos com menos parâmetros podem rivalizar com contrapartes maiores, enfatizando seu potencial de eficiência sem comprometer a eficácia. Além disso, soluções econômicas como o GPT-4o mini da OpenAI custam mais de 60% menos em comparação com modelos anteriores, tornando a IA de alta qualidade mais acessível para empresas e desenvolvedores. O Phi-3-mini da Microsoft é outro exemplo, alcançando capacidades superiores de raciocínio e lógica com apenas 3,8 bilhões de parâmetros — superando modelos com o dobro do seu tamanho. Da mesma forma, a família Llama 3.2 da Meta demonstra a viabilidade de mod-Is, oferecendo variantes de 1 bilhão a 90 bilhões de parâmetros que priorizam a eficiência sem sacrificar a eficácia. Esta mudança para modelos menores é apoiada

por especialistas da indústria como Andrej Karpathy, que defende a destilação de modelos para seu “núcleo cognitivo” essencial. Sua pesquisa sugere que mesmo um modelo de 1 bilhão de parâmetros poderia fornecer capacidades cognitivas suficientes, já que muitos dos dados adicionais em modelos maiores podem não melhorar diretamente o desempenho. Esse insight tem implicações, particularmente em tecnologias de consumo

Os modelos OpenELM da Apple, por exemplo, permitem o processamento de IA no dispositivo, proporcionando experiências responsivas e personalizadas, mantendo a privacidade e a eficiência energética.

Em 2024, modelos como a série SlimLM permitiram processamento robusto em dispositivos como smartphones, eliminando a necessidade de computação baseada em nuvem. Essa inovação marcou um salto à frente na acessibilidade de IA, permitindo que os usuários realizassem tarefas diretamente em seus dispositivos, mantendo a privacidade e reduzindo a latência.

O impacto dos SLMs se estende além de aplicações gerais para domínios especializados. No Ignite 2024, a Microsoft colaborou com líderes do setor como Bayer e Rockwell Automation para desenvolver SLMs direcionados para agricultura e manufatura, demonstrando como esses modelos compactos podem se destacar em setores específicos sem a sobrecarga de sistemas maiores de uso geral. Olhando para o futuro, o conceito de “empresas de LLMs” — onde vários modelos especializados trabalham em paralelo — pode representar a próxima evolução, combinando as vantagens de modelos grandes e pequenos em uma abordagem modular.

## Aterramento e Aumento de Contexto

O NotebookLM do Google Labs transforma a IA em um assistente de pesquisa personalizado ao “aterrar” em seu Google Docs. Ao contrário dos chatbots tradicionais, ele vincula as respostas às suas notas e fontes específicas, permitindo insights altamente relevantes e confiáveis. Esse recurso foi projetado para lidar com a sobrecarga de informações, facilitando a síntese e a conexão de ideias de várias fontes de forma eficiente. O aterramento garante

que as saídas de IA são ancoradas em dados verificados, reduzindo erros como alucinações. O NotebookLM se baseia nisso ao alavancar o aumento contextual, que enriquece as respostas com compreensão diferenciada adaptada às suas necessidades. Essa abordagem não apenas fornece respostas, mas fornece as respostas certas para seu contexto único.

Ferramentas como o GenAI Data Fusion estendem esse princípio às empresas, agregando e contextualizando dados específicos da empresa. Ao enraizar as saídas em conjuntos de dados personalizados, as empresas podem obter insights altamente precisos e específicos para tarefas para aplicações variadas.

desde a pesquisa até a análise operacional. O aterramento e o aumento contextual marcam uma mudança na IA, transformando sistemas genéricos em ferramentas precisas e adaptáveis que atendem a necessidades individuais, necessidades individuais e organizacionais.

## Superando a escassez de dados

A disponibilidade de dados de alta qualidade está surgindo como um gargalo no desenvolvimento de grandes modelos de IA. De acordo com a Ep- och AI, o reservatório de dados textuais de alta qualidade na internet pública pode estar esgotado.



## MODELOS, TÉCNICAS E PESQUISA

ed já em 2026. Inicialmente, os pesquisadores estimaram que o estoque de dados de linguagem de alta qualidade poderia acabar até 2024, enquanto dados de baixa qualidade poderiam durar mais duas décadas, e dados de imagem poderiam enfrentar esgotamento no final da década de 2030 até meados da década de 2040. Embora os limites de 2024 ainda não tenham sido atingidos, a escassez iminente está levando os laboratórios de IA a explorar estratégias alternativas para obter dados de treinamento.

Esta previsão desencadeou diversas estratégias entre os laboratórios de IA. Alguns estão buscando fontes de dados privadas, comprando de brokers ou licenciar conteúdo de editores.

Outros estão explorando dados de áudio e visuais inexplorados, com conteúdo de vídeo oferecendo par-insights particularmente valiosos sobre física e dinâmica do mundo real. Empresas como Scale AI e Surge AI estão construindo extensas redes de colaboradores, incluindo especialistas em nível de Ph.D., para criar e anotar conjuntos de dados especializados. Essas abordagens, no entanto, têm um custo alto, com algumas estimativas sugerindo que os laboratórios de IA estão gastando centenas de milhões de dólares anualmente nessas iniciativas.

Um método alternativo envolve usar um modelo de IA para gerar grandes quantidades de dados sintéticos para treinar outro, mas é arriscado. Estudos mostraram que modelos treinados predominantemente em dados sintéticos podem experimentar “colapso de modelo”, onde suas saídas se tornam menos diversas e falham em refletir distribuições do mundo real. Um fenômeno relacionado, denominado Desordem de Autofagia do Modelo (MAD), foi observado em modelos de imagem generativa, onde a dependência de dados sintéticos leva a um declínio notável na qualidade da saída.

Uma solução promissora está em técnicas como “self-play”, onde os modelos melhoram por meio da competição ou colaboração com eles mesmos. O AlphaGo do Google DeepMind, que derrotou o campeão mundial humano em Go após treinar contra si mesmo, exemplifica essa abordagem. Hoje, o self-play continua a informar o desenvolvimento de LLM de ponta. desenvolvimento, oferecendo um caminho para superar as limitações de dados, mantendo o desempenho desempenho e inovação.

### IA de código aberto

Em janeiro de 2025, a empresa chinesa de IA DeepSeek lançou o R1, um modelo de raciocínio de código aberto projetado para competir com — e potencialmente superar — o o1 da OpenAI por uma fração do custo. Embora o processo de raciocínio do R1 seja mais lento do que o de muitos modelos de uso geral, ele fornece respostas mais detalhadas e precisas. Junto com sua versão principal de 671 bilhões de parâmetros, a DeepSeek também introduziu seis modelos menores “destilados”, começando com 1,5 bilhão de parâmetros e capazes de rodar em dispositivos locais.

Outros modelos de código aberto também estão fechando a lacuna de desempenho com propriedade alternativas erárias. O Llama 3.1 da Meta agora é rivals GPT-4 em benchmarks importantes, e os modelos da Mistral AI oferecem recursos comparáveis às principais soluções de código fechado — tanto que o recente financiamento de US\$ 487 milhões da Mistral AI a rodada catapultou-o para o status de unicórnio. À medida que um número crescente de mods de código aberto Os els alcançam um desempenho de alto nível, porções estão cada vez mais tomando nota.

O navegador Brave, por exemplo, integrou o modelo Mixtral 8x7B da Mistral AI em seu assistente Leo, enquanto o Wells Fargo adotou o Llama 2 da Meta para aplicativos internos.

Esse ímpeto reflete uma tendência mais ampla em direção à IA de código aberto. As estatísticas do GitHub revelam que os repositórios focados em IA dispararam de apenas 845 em 2011 para 1,8 milhão em 2023 — um aumento impressionante de 59,3% somente em 2023. Durante o mesmo período, o engajamento da comunidade disparou, com as estrelas do GitHub para projetos de IA saltando de 4 milhões para 12,2 milhões entre 2022 e 2023. Esse aumento destaca o crescimento

apetite crescente pelo desenvolvimento colaborativo, sinalizando que a IA de código aberto continuará sendo uma força motriz no futuro.

### IA modular

Em vez de construir modelos monolíticos, a abordagem de IA modular divide os sistemas de IA em componentes especializados e independentes que podem ser misturados e combinados como blocos de construção. Cada módulo lida com tarefas ou domínios específicos, permitindo controle preciso sobre as capacidades do sistema



## MODELOS, TÉCNICAS E PESQUISA

e recursos. Um exemplo é uma proposta recente para desenvolver grandes modelos de linguagem (LLMs) usando “bricks”: componentes modulares que representam tarefas específicas ou domínios de conhecimento. Esses bricks podem surgir durante o pré-treinamento ou ser projetados sob medida após o treinamento para aplicações específicas. Essa abordagem ativa dinamicamente apenas os bricks relevantes para uma tarefa, reduzindo significativamente os custos computacionais e de energia, ao mesmo tempo em que melhora a escalabilidade. A pesquisa mostrou que neurônios e camadas dentro de LLMs se especializam naturalmente em diferentes funções, destacando a promessa do design modular na otimização de sistemas de IA.

Da mesma forma, o MAGNUM, uma estrutura modular de IA multimodal apresentada no NeurIPS 2023, oferece flexibilidade incomparável por meio de dados estruturados e não estruturados em vários tipos de entrada, incluindo texto, imagens, vídeo, áudio e dados de séries temporais.

Composto por módulos específicos de entrada, MAG-O NUM se destaca na combinação e extração de informações de diversos tipos de dados, formando bem em 10 tarefas do mundo real como

diagnósticos médicos e previsão do tempo. Notavelmente, é robusto contra dados ausentes ou incompletos, um desafio comum em sistemas de IA multimodais. Outro avanço modular da IA é o MASAI (Arquitetura Modular para Software-framework de IA de engenharia de software), que usa subagentes especializados alimentados por LLMs para lidar com subproblemas distintos em engenharia de software. O MASAI permite estratégias de resolução de problemas ajustadas em subagentes, recuperação eficiente de informações e redução da sobrecarga computacional. Essa arquitetura obteve um desempenho máximo de 28,33% no conjunto de dados SWE-bench Lite, demonstrando sua eficácia-habilidade na resolução de problemas reais do GitHub.

### Grandes modelos de ação

Os modelos de ação grande (LAMs) vão além dos modelos tradicionais de linguagem grande ao executar tarefas em vez de apenas processar a linguagem. Eles agem como agentes autônomos que podem interagir com interfaces de computador — clicando em botões, movendo cursores e

digitando texto. Um exemplo antigo é o Claude 3.5 Sonnet, que pode interagir com computadores como um humano faria: navegando em aplicativos de desktop, movendo cursores e clicando em botões, digitando texto, interpretando capturas de tela e respondendo adequadamente, e executando tarefas complexas de várias etapas em um computador. Esta capacidade marca uma transição na IA, eliminando a lacuna entre os modelos de conversação e os sistemas autônomos completos.

sistemas. Embora atualmente opere por impersonar um usuário — interagindo com interfaces projetadas para humanos — os LAMs demonstram as etapas fundamentais em direção aos sistemas de IA que podem gerenciar tarefas diretamente dentro de ecossistemas digitais. Por exemplo, um LAM pode reservar uma passagem aérea ou redigir um documento documento navegando em aplicativos e sites sem intervenção do usuário.

As principais empresas de IA — Google, Apple, Microsoft, OpenAI e outras — estão posicionadas identificando esses agentes como o futuro da IA.

O objetivo deles é criar sistemas que vão além das interações de chat e ativamente

interagir com o mundo. A capacidade de Claude de interagir com um computador de mesa demonstra esse potencial, mas também levanta questões significativas sobre privacidade e controle. Para funcionar, os LAMs exigem amplo acesso aos ambientes digitais dos usuários — lendo telas, acessando arquivos e executando comandos. Esse nível de acesso cria uma intimidade sem precedentes entre a IA e seus usuários, levantando preocupações sobre como os dados são coletados, armazenados e usados. As empresas veem isso como uma grande oportunidade. Ao se integrarem profundamente às vidas digitais dos usuários, as empresas de IA podem obter acesso a dados além de qualquer coisa coletada anteriormente por gigantes da tecnologia tradicionais, potencialmente redefinindo normas em torno da privacidade de dados.

### Modelos de Ação Grande Pessoal

Levando o conceito de modelos de grande ação um passo adiante, os modelos de grande ação pessoal (PLAMs) representam uma evolução ainda mais íntima da IA. Enquanto os LAMs atuais são treinados em dados gerais da internet, os PLAMs seriam treinados especificamente na pegada digital de um único usuário. Imagine

## MODELOS, TÉCNICAS E PESQUISA

uma IA que conhece suas interações nas redes sociais, hábitos de compra online, biometria, dados de localização, mensagens de texto, calendário e e-mail — entendendo não apenas seus dados, mas o contexto exato da sua vida: onde você está, com quem está e o que está fazendo.

Embora os PLAMs ainda sejam amplamente teóricos, eles marcariam uma mudança significativa na personalização de IA. Esses sistemas aprenderiam com todos os aspectos da sua vida digital: banco-registros de navegação, histórico de navegação, dispositivos IoT, uso do carro, dados vestíveis e informações biológicas. Esta compreensão abrangente

sua posição permitiria que o PLAM compreendesse suas rotinas e preferências, eventualmente tomando decisões que se alinham precisamente com o que você teria escolhido - desde a compra perseguições para gerenciamento de cronograma. À medida que estes modelos aumentam a sua precisão ao longo do tempo, assumiriam papéis de tomada de decisão cada vez mais complexos, tornando-se efectivamente uma extensão digital de uma pessoa. Este nível O modelo de personalização cria um efeito de bloqueio significativo, transferindo anos de experiência individual.

aprendizagem individualizada para um novo sistema seria

ser impraticável, se não impossível. Empresas que implementam PLAMs com sucesso no início podem, portanto, se estabelecer como parceiras digitais indispensáveis, com usuários se tornando profundamente integrados em seu ecossistema.

