INSTITUTO DE PESQUISAS TECNOLÓGICAS

COMUNICAÇÃO TÉCNICA

Nº 179868

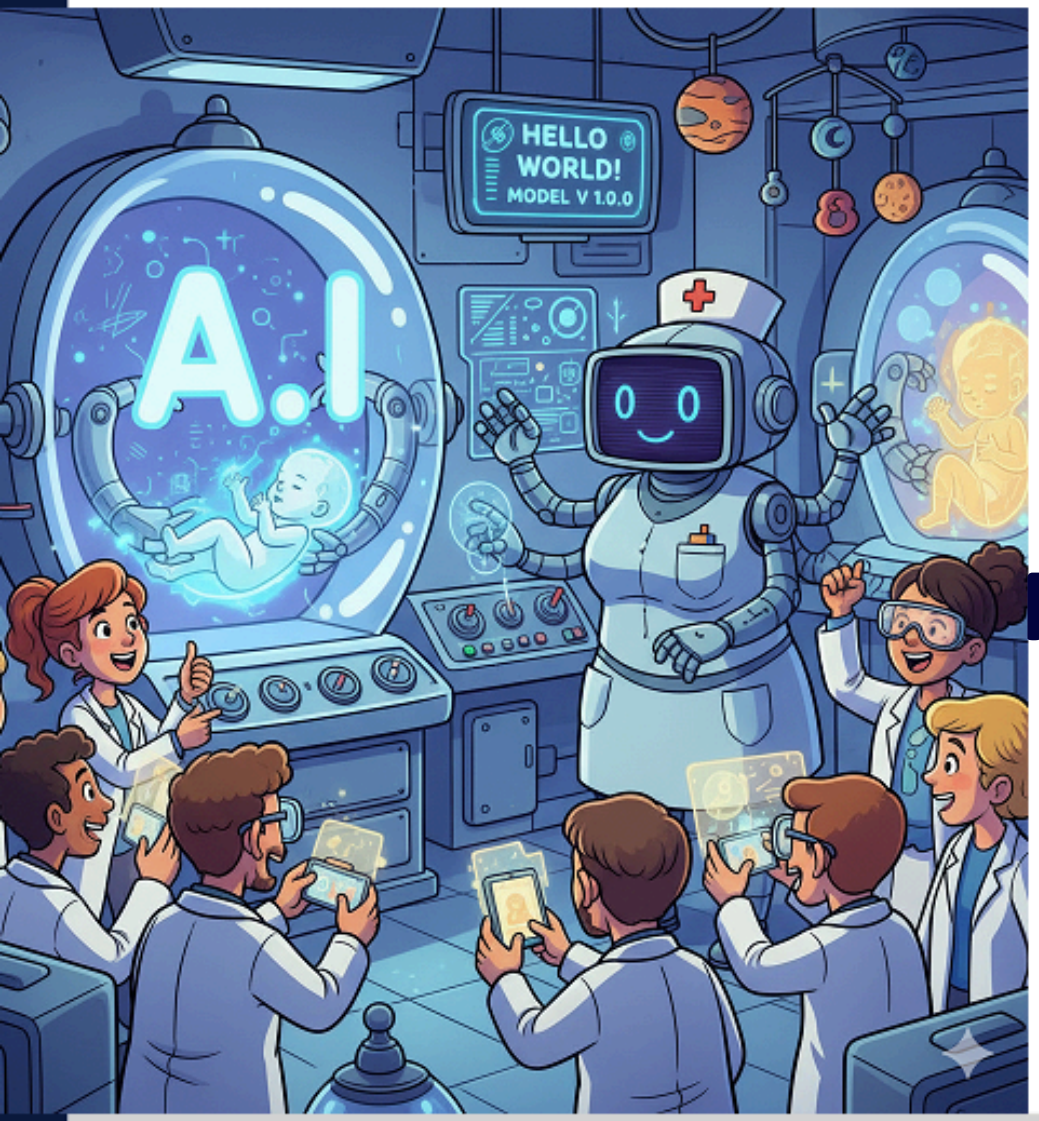
Piscou? Nasceu outro modelo! Como acompanhar a velocidade da IA?

Eric Tadeu Camacho Oliveira
Adriana Camargo de Brito

Palestra apresentada no IPT, 30/09/2025. 28 slide*s*

Instituto de Pesquisas Tecnológicas do Estado de São Paulo S/A - IPT
Av. Prof. Almeida Prado, 532 | Cidade Universitária ou
Caixa Postal 0141 | CEP 01064-970
São Paulo | SP | Brasil | CEP 05508-901
Tel 11 3767 4374/4000 | Fax 11 3767-4099

www.ipt.br



ipt
INSTITUTO DE PESQUISAS TECNOLÓGICAS

PISCOU? NASCEU OUTRO MODELO! COMO ACOMPANHAR A VELOCIDADE DA IA?

Eric Tadeu Camacho de Oliveira
Adriana Camargo de Brito

SÃO PAULO
GOVERNO DO ESTADO

PERGUNTA

Quantos novos modelos de IA de impacto vocês acham que foram lançados em Setembro de 2025?

SÓ EM SETEMBRO, ATÉ DIA 28/09, TIVEMOS 12 MODELOS DE IMPACTO

Language

Grok 4 Fast

Magistral Small 1.2

Magistral Medium 1.2

Granite-Docling

Qwen3-Next-80B-A3B

K2 Think

Qwen3-Max

EmbeddingGemma

Apertus 8B

Apertus 70B

LongCat-Flash

Multimodal

Magistral Small 1.2

Granite-Docling

Vision

Magistral Small 1.2

Granite-Docling

Speech

Chatterbox Multilingual

epoch.ai

ATUALIZAÇÃO - NOVO MODELO SAINDO DO FORNO

29/09/2025 - DeepSeek-V3.2-Exp

- Lançado como evolução de V3.1-Terminus
- Introduz **DeepSeek Sparse Attention (DSA)** para otimizar eficiência em contextos longos
- Licença: MIT (permite uso comercial / acadêmico)
- **Escalabilidade:** menos custo computacional no treino e inferência

⚠ Experimental ⚠

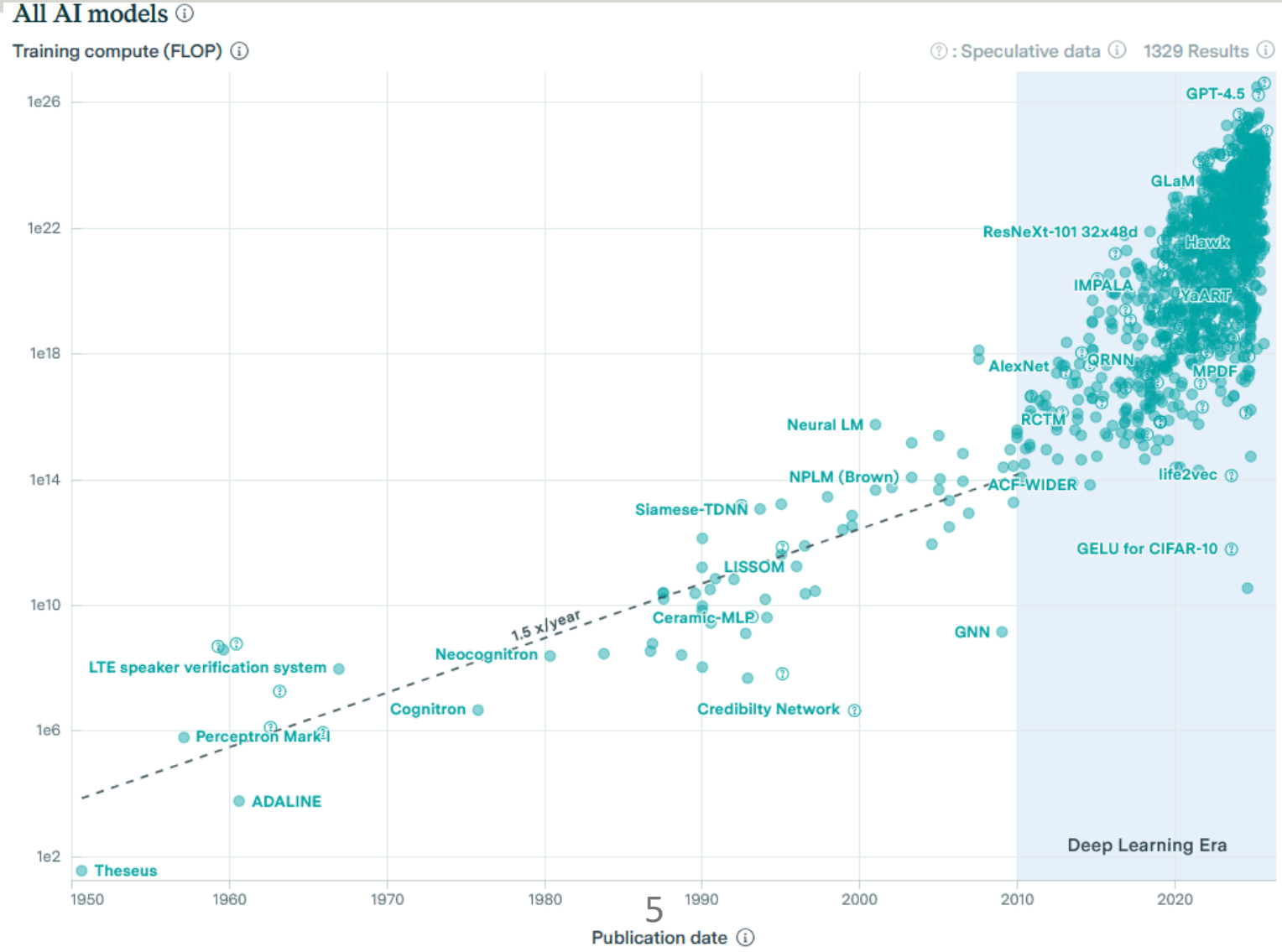
OBJETIVO DE HOJE

1) Apresentar o problema da **"Sobrecarga de Informação sobre as IAs"**

2) O que nos levou até esse momento?

3) Estratégias para acompanhar essa quantidade de lançamentos

LINHA DO TEMPO (1950 - 2025)



epoch.ai/data/ai-models

A ACELERAÇÃO DA IA - DE ONDE VIEMOS?

A inovação atual é construída sobre décadas de pesquisa

Marcos Relevantes

- **BERT (2018 · Google)** - Introdução a compreensão mais profunda sobre o contexto da linguagem
- **GPT-3 (2020 · OpenAI)** - Poder de escala e popularizou a IA Generativa
- **ChatGPT (2022 · OpenAI)** - Adoção rápida da IA para público geral e corporativo

Ciclo de Aceleração

Avanços → Ferramentas Acessíveis → Adoção em Massa → + Investimento → + Aceleração

The Evolution of Generative AI: 2018–2025 Timeline Revealed

A EXPLOSÃO DE 2023-2025

O que antes levava anos, agora está acontecendo em **meses/semanas**

Alguns exemplos:

- **GPT-4 (Mar/2023 · OpenAI)** – Multimodalidade com imagens
- **Llama 2 (Jul/2023 · Meta)** - Democratização com open-source
- **Claude 3 (Mar/2024 · Anthropic)** - Ultrapassou GPT-4 em vários benchmarks
- **Llama 3 (Abr/2024 · Meta)** – Competitividade open-source
- **GPT-4o (Mai/2024 · OpenAI)** – Multimodalidade nativa

- **Llama 3.1 405B (Jul/2024 · Meta)** – Paridade open-source com modelos proprietários
- **Llama 3.3 70B (Dez/2024 · Meta)** - Desempenho aprimorado em instruções e código
- **Qwen2.5-Max (Jan/2025 · Alibaba)** - Superou GPT-4o e Claude 3.5 em alguns benchmarks
- **Grok 3 (Fev/2025 · xAI)** - Avanço em raciocínio e geração de imagens
- **Gemma 3 Series (Mar/2025 · Google/DeepMind)** – Modelos leves, open-source, versões multimodais e multilíngues

Ser o **"melhor modelo"** é passageiro

AI Timeline

ALÉM DO TEXTO: A ERA DA MULTIMODALIDADE

A IA agora Vê, Ouve e Fala

A inovação não é mais apenas sobre texto; é sobre expandir os tipos de dados que a IA entende e gera

- **Vídeo:** Sora (OpenAI), Veo (Google), Movie Gen (Meta)
- **Áudio e Música:** Suno AI (criação de músicas completas com vocais)
- **Multimodalidade Integrada:** GPT-4o e Gemini podem processar texto, áudio e imagens em tempo real

PROBLEMA: SATURAÇÃO DE BENCHMARKS

Os modelos estão melhorando tão rápido que os testes ficaram obsoletos, tornando as pontuações um indicador enganoso do valor real

AI Agent Benchmarks are Broken

Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation

OS MOTORES DA ACELERAÇÃO: A CORRIDA GEOPOLÍTICA

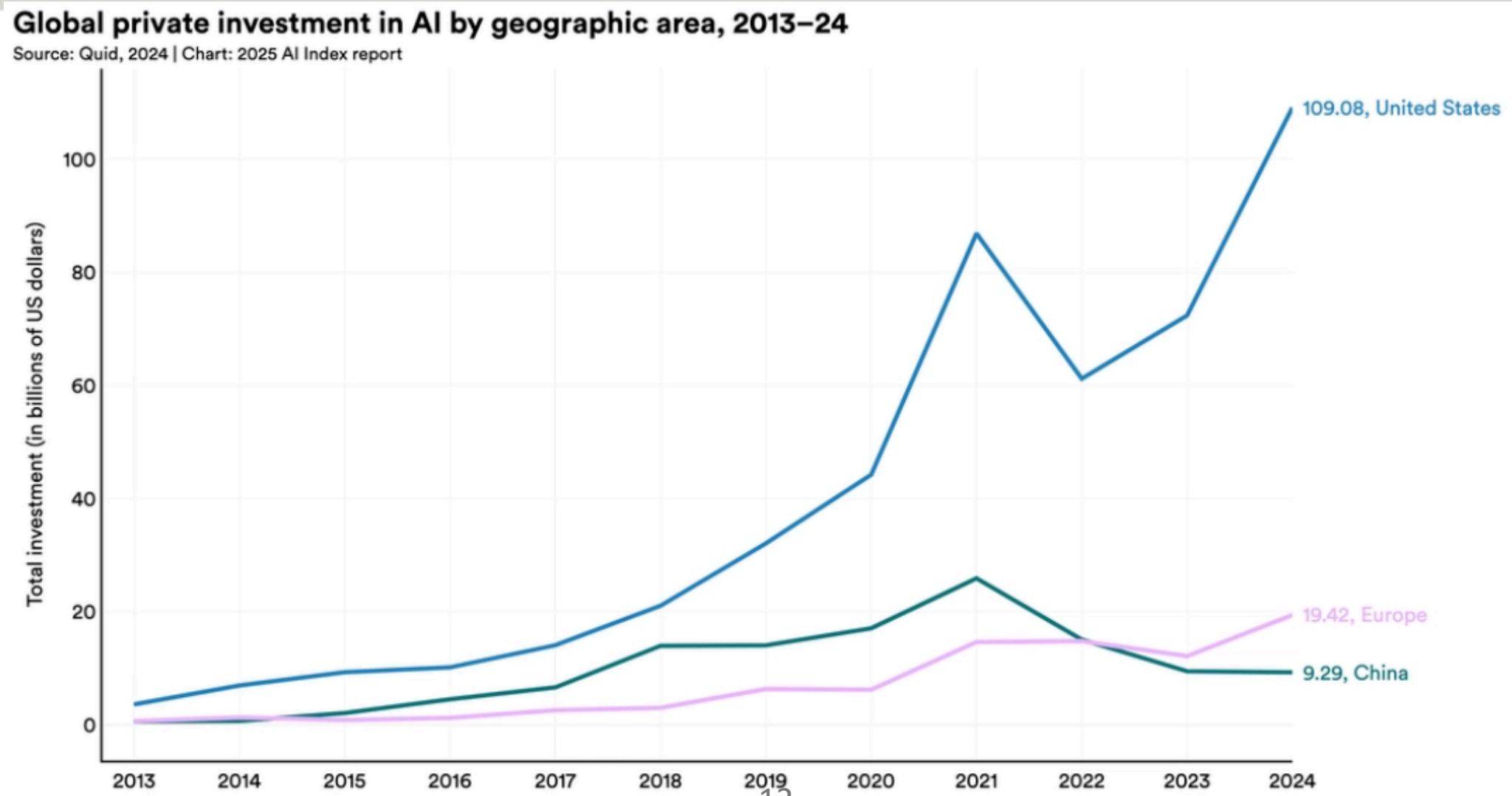
Nova Corrida de Superpotências

A competição global, especialmente entre EUA e China, é um dos principais catalisadores do investimento e da inovação

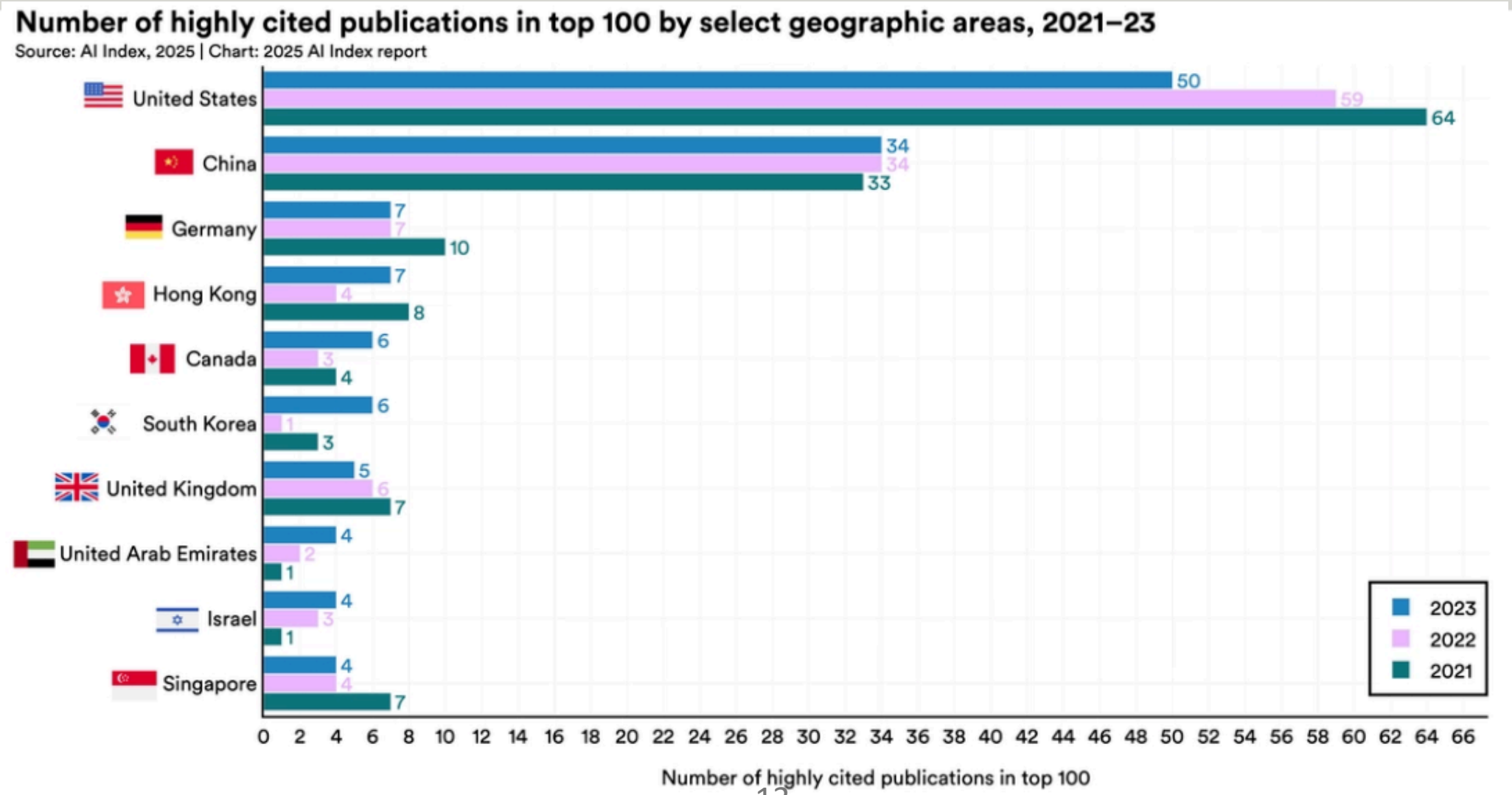
EUA vs China

- Investimento Privado
- Produção de Modelos
- Qualidade

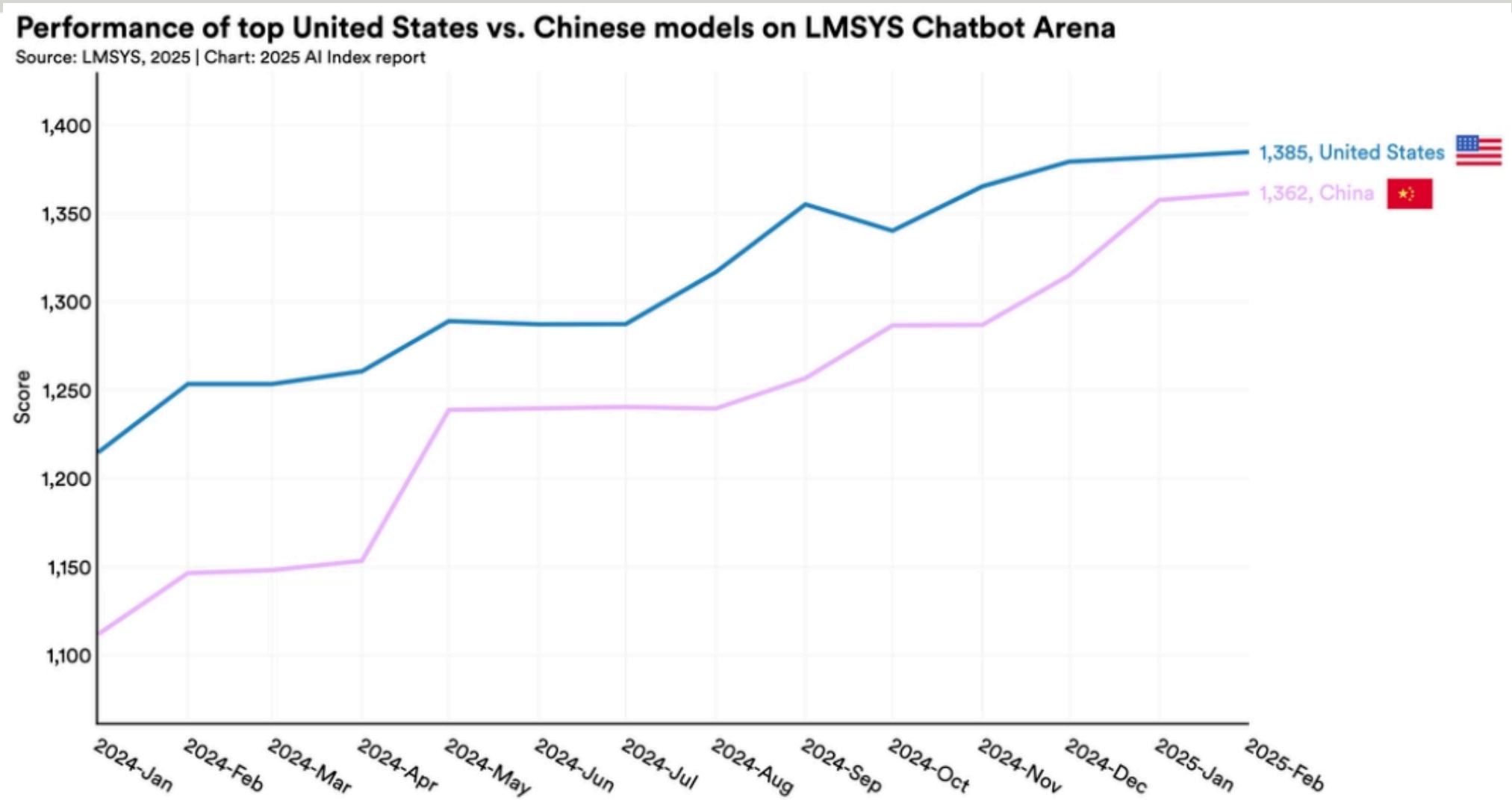
INVESTIMENTO PRIVADO EM IA



NÚMERO DE PUBLICAÇÕES



PERFORMANCE DOS MODELOS EUA - CHINA



Estratégias:

- **EUA/Ocidente:** Foco em um ambiente **"pró-inovação"** liderado pelo setor privado
- **China:** Abordagem **"top-down"** dirigida pelo Estado, com metas claras para liderança global até 2030

Europe’s AI strategy is no match for China’s drive for global dominance (Jun 29, 2018)

OS MOTORES DA ACELERAÇÃO: O VOLANTE TECNOLÓGICO

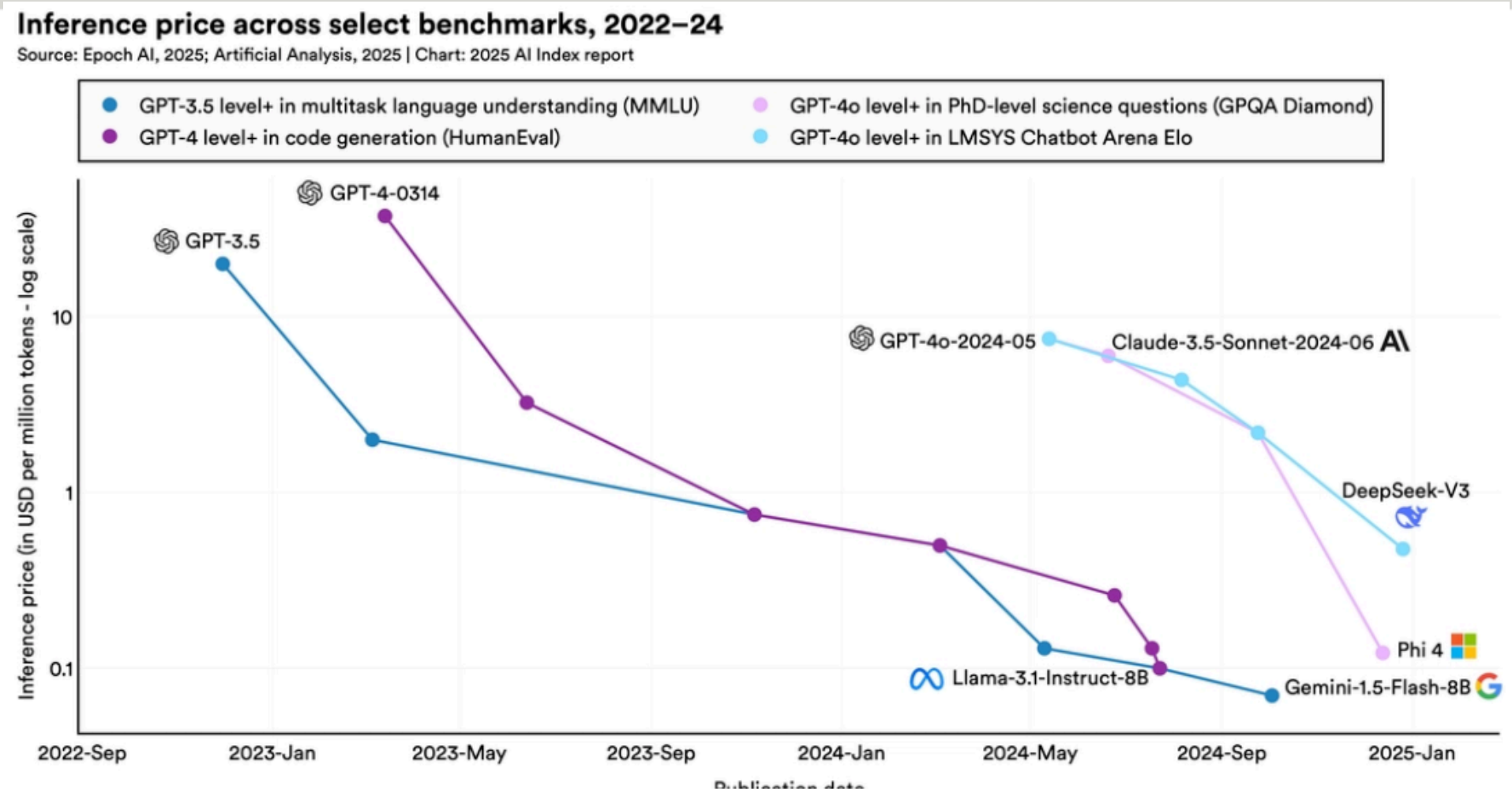
Da escala bruta à eficiência inteligente

O foco mudou de modelos maiores para modelos mais **eficientes** e **capazes**

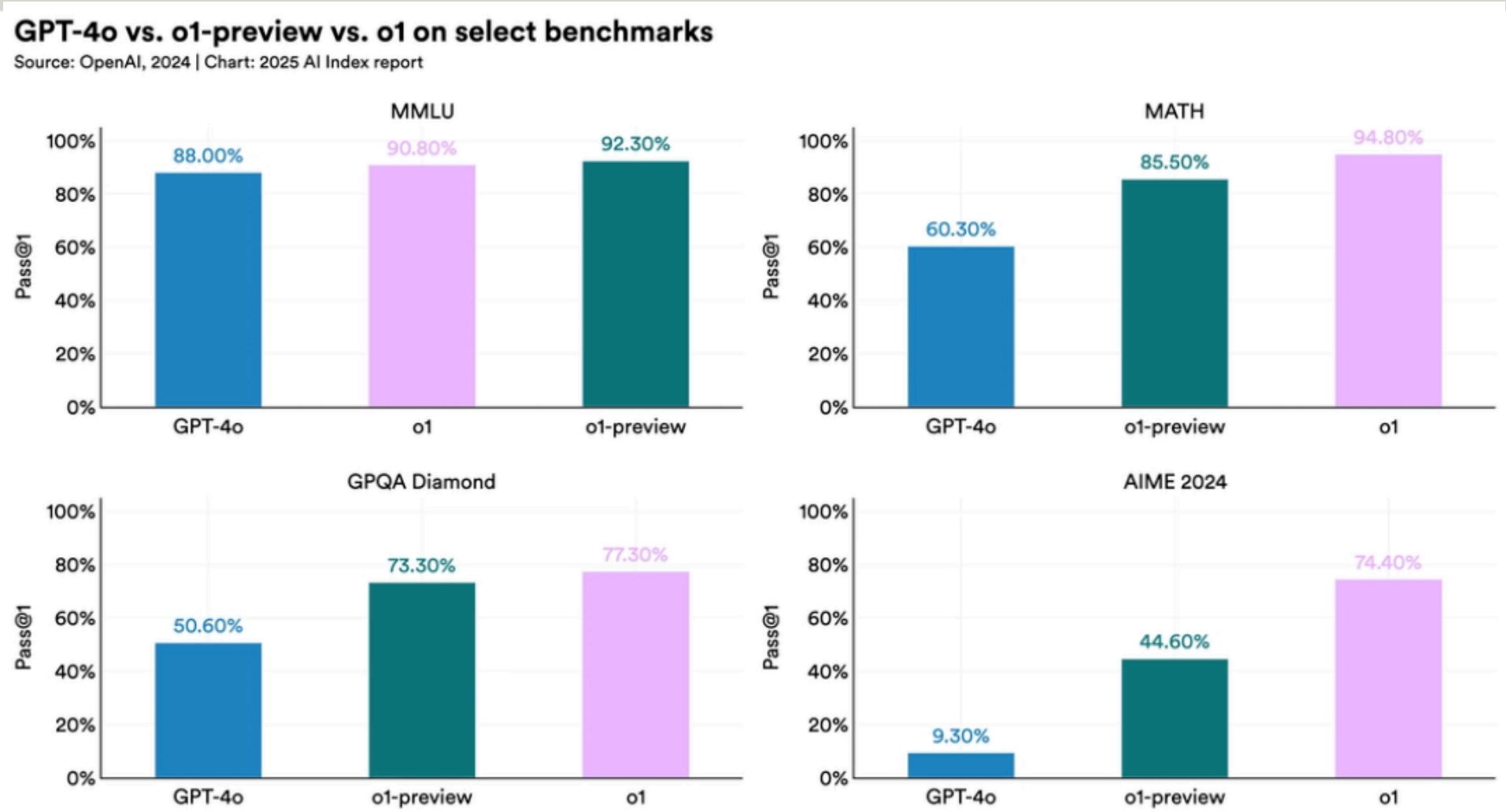
Tendências:

- **Economia da Inferência:** O custo para executar modelos caiu drasticamente, tornando aplicações complexas economicamente viáveis
- **Ascensão do Raciocínio:** Modelos como o *o1* da OpenAI são projetados para resolver problemas com passos lógicos, não apenas reconhecer padrões
- **O Futuro "Agêntico":** A eficiência permite que a IA execute tarefas de forma autônoma (agentes)
- **Hardware Especializado:** Chips customizados como LPUs (Groq) e TPUs (Google) superam GPUs em eficiência para inferência

PREÇO DE INFERÊNCIA



RACIOCÍNIO


GPT-4o vs. o1-preview vs. o1 on select benchmarks
Source: OpenAI, 2024 | Chart: 2025 AI Index report
MMLU
Pass@1
100%
80%
60%
40%
20%
0%
88.00%
90.80%
92.30%
GPT-4o
o1
o1-preview
MATH
60.30%
85.50%
94.80%
GPT-4o
o1-preview
o1
GPQA Diamond
50.60%
73.30%
77.30%
GPT-4o
o1-preview
o1
AIME 2024
9.30%
44.60%
74.40%
GPT-4o
o1-preview
o1

A GRANDE DIVISÃO: OPEN-SOURCE VS PROPRIETÁRIO

- **Código Aberto (Open-Source):** Componentes publicamente disponíveis (ex: Llama da Meta, modelos da Mistral)
  - Permite uma maior modificação e personalização
- **Proprietário (Fonte Fechada):** Controlado por uma empresa (ex: GPT da OpenAI, Claude da Anthropic)
  - Acesso via APIs pagas

ANÁLISE COMPARATIVA

Critério	Open Source	Proprietário
Custo	Gratuito	Pago
Personalização	Máxima	Limitada
Desempenho	Se aproximando rapidamente	Ainda lidera
Segurança	Transparência	Segurança gerenciada

O HUB DO CÓDIGO ABERTO: HUGGING FACE

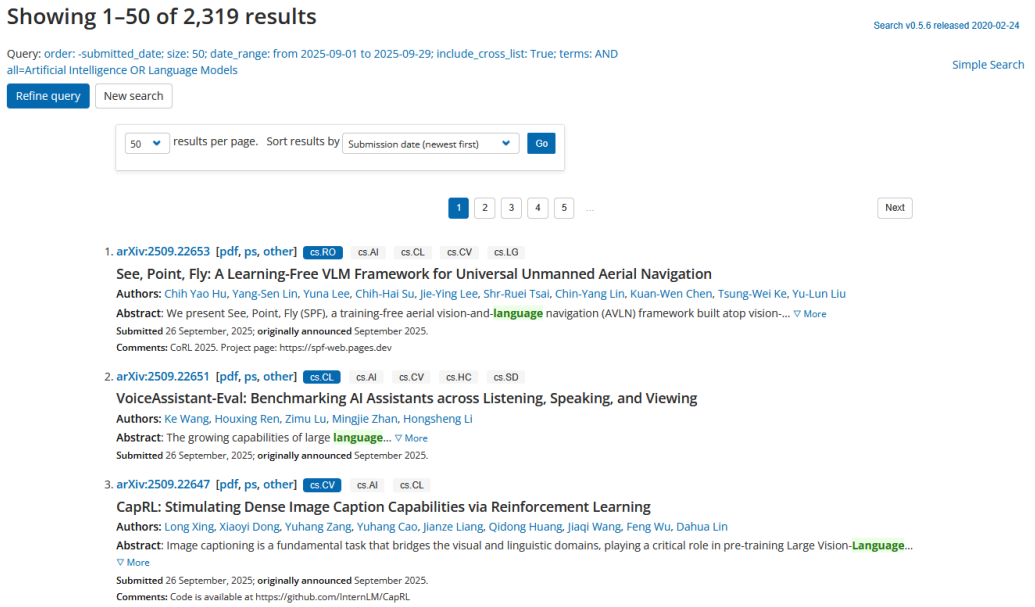
Onde a comunidade de IA se encontra

Principais Funções:

- **"GitHub para a IA":** Repositório central com milhares de modelos e datasets
- **Bibliotecas Padrão:** Ferramentas como *transformers* e *diffusers* que simplificam o desenvolvimento
- **Comunidade Ativa:** Conectando desenvolvedores, pesquisadores e grandes empresas de tecnologia

COMO ACOMPANHAR A INOVAÇÃO DIÁRIA? - PARA OS TÉCNICOS

- **ArXiv.org:** Onde a pesquisa de ponta e atual aparece primeiro
- **Blogs de Laboratórios de IA:** OpenAI, Google AI, DeepMind



COMO ACOMPANHAR A INOVAÇÃO DIÁRIA? - PÚBLICO NO GERAL

Newsletters

- AI Agents | RAG | LLMs
- Neural Bits
- The Batch
- Superhuman AI newsletter
- IA Sob Controle
- AI & Future Tech Trends
- Artificial Intelligence
- The Neural Maze
- Top AI Papers of the Week
- LLM Watch
- Daily Dose of Data Science
- The Rundown AI
- TLDR AI

Youtube

- Inteligência Mil Grau
- Andrej Karpathy
- Asimov Academy
- Sebastian Raschka

Podcasts

- IA Sob Controle

COMO ACOMPANHAR A INOVAÇÃO DIÁRIA? - PÚBLICO NO GERAL

LinkedIn

- Avi Chawla
- Elvis S.
- Pascal Biese
- Akshay Pachaar
- João (Joe) Moura
- Ben Burtenshaw
- Miguel Otero Pedrido
- Anderson L. Amaral
- Alex Razvant
- Andriy Burkov
- Maryam Miradi
- Bob Inteligência Mil Grau
- Mario Filho
- Sai Charan
- Elisa Terumi
- Taranjeet Singh
- Md Amanatullah
- Samuel Matioli
- Tony Kipkemboi
- Kalyan KS
- Daniel Han
- Banias Baabe
- Celso Sousa

Fóruns

- Alignment Forum
- OpenAI Forum
- Google AI Forum

Slack

- AI Agents BR

Discord

- HuggingFace
- Learn AI Together

DOMINANDO A SOBRECARGA DE INFORMAÇÃO

É impossível acompanhar tudo

Estratégias Práticas:

- **Definir quais conteúdos priorizar**
  - **Não precisa conhecer todos os modelos** → focar nos que resolvem seu problema
- **Criar um email** que focará em receber as atualizações (cadastrar em newsletters, blogs, ferramentas de IA)
- **Criar uma rotina** para ler as atualizações (Ex: 15 minutos todos os dias)

DOMINANDO A SOBRECARGA DE INFORMAÇÃO

- Utilizar ferramentas como **Obsidian/Notion**, para salvar e organizar as informações e **Raindrop.io** para salvar links
- **Modo expert:** Criar um sistema com agentes de IA que receberão as atualizações vindas de várias fontes e farão um resumo para você!

Obrigado!

linkedin.com/school/iptsp/

instagram.com/ipt_oficial/

youtube.com/@IPTbr/

www.ipt.br

ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS

SÃO PAULO GOVERNO DO ESTADO