Avanço e custos da fronteira em modelos de IA

Dossiê de aprofundamento — análise vertical do tema da semana

Vitor Wilher[1]

6 de junho de 2026

[1]Bacharel e Mestre em Economia pela UFF, Candidato ao PhD em Economia pela EPGE/FGV. É Especialista em Ciências de Dados e Inteligência Artificial Generativa pela PUC-Rio. Atualmente, exerce a função de Data Tech Lead na Análise Macro (http://analisemacro.com.br). Saiba mais em https://github.com/vitorwilher.

Índice

Avanço e custos da fronteira em modelos de IA

Corrida pelo topo: novos modelos de fronteira e ranking competitivo

A semana consolidou um novo equilíbrio precário no topo do ranking de modelos de fronteira. A Alibaba lançou o Qwen3.7-Max posicionando-o explicitamente como modelo preferencial para trabalho agêntico de longo prazo em texto, com janela de contexto de 1 milhão de tokens de entrada e 64 mil de saída, a 208,3 tokens por segundo. Na Artificial Analysis Intelligence Index, alcançou pontuação 56,6, situando-se em quinto/sétimo lugar dependendo da configuração de raciocínio dos concorrentes — atrás de Gemini 3.1 Pro Preview (57,2) e à frente de Gemini 3.5 Flash em raciocínio alto (55,3). É hoje o LLM chinês mais inteligente segundo esse índice, mas ainda trás dos líderes americanos da OpenAI, Anthropic e Google (The Batch).

Do lado americano, a Anthropic acelerou o ciclo de lançamentos com o Claude Opus 4.8, que segundo benchmarks da própria empresa supera GPT-5.5 e Gemini 3.1 Pro em coding agêntico, uso de computador, análise financeira e no Humanity's Last Exam. O salto mais celebrado é em confiabilidade: o modelo é quatro vezes menos propenso a deixar passar erro no próprio código, segundo testes citados, e o preço foi mantido (US$ 5/milhão tokens de entrada, US$ 25/milhão de saída). No GraphWalks, marca 85,9% no subset 256K e 68,1% em 1M — contra 76,9% e 40,3% do 4.7. Paralelamente, o MiniMax M3 surgiu como novo modelo de ponta chinês, open-source com janela de 1M e entrada multimodal (AIWhisperBR; TechDrops; AiDrop).

A economia oculta da fronteira: gastos massivos e crescimento de 2.000%

Um paper de economistas afiliados à University of Virginia, Anthropic e Bank of Canada — destacado por Jack Clark no Import AI — sustenta que o "PIB da IA" nos EUA atingiu aproximadamente US$ 250 bilhões em 2025, crescendo cerca de 2.600% ao ano em termos reais ajustados por qualidade. O gasto nominal em compute saltou de US$ 37 bi (2023) para US$ 90 bi (2024) e US$ 219 bi (2025); a capacidade bruta de compute cresce acima de 200% ao ano graças a ganhos de eficiência de chip. A tese central — alarmante para policymakers — é que essa explosão é praticamente invisível nas estatísticas convencionais de GDP, porque preços por unidade de capacidade caem quase tão rápido quanto a saída ajustada por qualidade sobe. A analogia de Clark é direta: economistas estão como o público de *Jaws* sem ouvir a trilha sonora — o tubarão está na água, mas os dados de superfície dizem que tudo está normal (Import AI 459).

Capex: a escalada do investimento em infraestrutura

A semana trouxe sinais inequívocos de que a corrida pela fronteira agora se traduz em capex em escala industrial. A Alphabet anunciou emissão de US$ 80 bilhões em ações para financiar a expansão de IA, com plano de investir entre US$ 180-190 bilhões em capex em 2026 — construção de data centers, compra de chips, treinamento de modelos, geração de energia. A Berkshire Hathaway entrou com US$ 10 bi nessa rodada, sua segunda grande aposta tech após a Apple. O Softbank prometeu US$ 52 bi em data centers na França e ultrapassou a Toyota como maior empresa do Japão. A HPE saltou 28% após resultados trimestrais que mostraram aceleração da demanda por infraestrutura de servidor de IA (TechDrops; AiDrop).

A Dell exemplifica a translação do capex em receita: entregou US$ 43,8 bi de receita no 1T (vs. US$ 35,4 bi esperados), com a divisão de infra crescendo 181% para US$ 29 bi, dos quais US$ 16 bi vieram de servidores otimizados para IA (+757% a/a). A empresa fechou o trimestre com US$ 51,3 bi em backlog de pedidos de IA — receita futura praticamente contratada. As ações dispararam 39% em after-hours, maior salto desde 2018. A margem bruta, porém, caiu 3,3 p.p., revelando o trade-off estrutural: servidores de IA têm margem menor, e o crescimento atual é financiado por compressão de margem (MoneyDrop; TechDrops).

Valuations sem precedente: Anthropic e OpenAI rumo aos trilhões

A Anthropic protocolou confidencialmente o S-1 para IPO nos EUA, com expectativa de valuation perto de US$ 1 trilhão segundo a última rodada — Polymarket aponta US$ 1,8 trilhão. A trajetória de receita anualizada é sem precedentes no capitalismo moderno: US$ 10 mi (2022), US$ 100 mi (2023), US$ 1 bi (2024), US$ 9 bi (2025), US$ 47 bi (maio/2026). No mercado secundário, a empresa já é precificada em US$ 992 bilhões — ultrapassando a OpenAI pela primeira vez. A Série H captou US$ 65 bi a um valuation de US$ 965 bi, com Altimeter, Dragoneer, Greenoaks e Sequoia. Os sete fundadores entraram na lista das 500 pessoas mais ricas do mundo, cada um com ~US$ 8 bi. OpenAI e Anthropic receberam juntas mais de US$ 250 bilhões em funding nos últimos dois anos, enquanto quase metade dos 857 unicórnios não levantou rodada nos últimos três anos e 220+ se tornaram ex-unicórnios — sinal claro de concentração de capital no topo da cadeia (TechDrops).

O movimento Alibaba: da abertura ao fechamento

Decisão sutil mas significativa: Qwen3.7-Max não tem pesos abertos. Junto com Qwen3.6-Max-Preview e Qwen3.6-Plus, segue o caminho fechado; apenas os tiers menores (Qwen3.6-27B e Qwen3.6-35B-A3B) permanecem livres. Simultaneamente, a Alibaba passou a cobrar pelo Qwen Code. A transição vem após mudança de liderança da equipe Qwen e sugere que o objetivo é monetizar os modelos top em vez de maximizar alcance — um espelhamento da estratégia OpenAI/Anthropic em terras chinesas. Andrew Ng manifestou no The Batch lamento explícito por essa virada, dado que a abertura era a contraposição estratégica chinesa frente aos labs americanos (The Batch).

Custos operacionais: capacidade de compute e otimização de inferência

A fronteira não é apenas treino: a inferência domina os custos do ciclo de vida. Servir um LLM de 70B parâmetros consome ~140 GB só para carregar pesos, com cada requisição ativa exigindo seu próprio KV cache em GPU. O curso da DeepLearning.AI com a Red Hat sobre vLLM enfatiza esse gargalo, ensinando quantização, serving eficiente e benchmark de latência/throughput/acurácia — tornando explícito o trade-off entre velocidade, custo e precisão que define a economia operacional dos modelos modernos. O COO da Uber, Andrew Macdonald, sinalizou nessa direção: gastos com tokens estão "cada vez mais difíceis de defender" internamente porque a ligação entre consumo intenso e funcionalidade útil para o consumidor é frouxa (DeepLearning.AI; Data Hackers).

Mustafa Suleyman, CEO da Microsoft AI, projetou no Build 2026 que a capacidade computacional de IA cresce 1.000x nos próximos 3 anos, de 5e27 para 5e30 FLOPs — projeção que enquadra a magnitude dos investimentos em capex como pressuposto, não como ambição (TechDrops).

Capacidades agênticas e ciclo de aprimoramento auto-referente

O argumento de venda do Qwen3.7-Max é agêntico: em teste interno, otimizou autonomamente um attention kernel em hardware não visto durante treino, fazendo 1.158 chamadas de ferramenta e 432 avaliações de kernel em 35 horas, gerando código 10x mais rápido que a referência. A Artificial Analysis ainda não validou. O Claude Opus 4.8 traz dynamic workflows que dividem tarefas grandes entre subagentes do Code, executam em paralelo e consolidam resultados — útil para auditorias, mas com custo elevado. Esses dois sinais materializam a preocupação que Andrew Leigh (economista e ministro australiano) articulou em discurso citado pelo Import AI: a governança de "recursive self-improvement" como capacidade pode ser crítica, pois se uma geração de sistemas projeta a próxima, o ator líder pode ampliar sua vantagem rápido demais para escrutínio externo (The Batch; AiDrop; Import AI 459).

Frentes adjacentes: physical AI, biologia, defesa

A fronteira está se ramificando para além de LLMs textuais. A Nvidia lançou o Cosmos 3, primeiro foundation model aberto capaz de raciocinar e gerar em cinco modalidades simultaneamente — texto, imagem, vídeo, som ambiente, ações físicas — mirando robôs humanoides e o mercado de robótica estimado em US$ 200 bi em 2030. O Biohub (Chan-Zuckerberg) liberou ESMFold2 e ESMC (treinado em 2,8 bilhões de sequências), modelo de biologia de proteínas que supera AlphaFold 3 em vários benchmarks e demonstrou hit rates de 36-88% em design de proteínas-binders para alvos de câncer (EGFR, PD-L1, CTLA-4). A OpenAI lançou o programa Rosalind Biodefense baseado em GPT-Rosalind. Em todos os casos, scaling laws clássicas se mantêm: ESMC escalou 2.8B sequências (vs. ~50M do ESM2), e o passe agêntico do ESMFold2 sobe de 49% para 65% com 1.000 samples vs. 1 (The Batch; Import AI; AiDrop).

Números-chave

- PIB nominal da IA nos EUA em 2025: ~US$ 250 bi
- Crescimento real ajustado por qualidade: ~2.600%/ano
- Gasto nominal em compute EUA: US$ 37 bi (2023) → US$ 90 bi (2024) → US$ 219 bi (2025)
- Capex projetado da Alphabet 2026: US$ 180-190 bi
- Emissão de ações da Alphabet: US$ 80 bi
- Investimento de Softbank em DCs França: US$ 52 bi
- Funding combinado OpenAI + Anthropic (2 anos): >US$ 250 bi
- Receita anualizada Anthropic: US$ 10 mi (2022) → US$ 47 bi (mai/2026)
- Valuation Anthropic Série H: US$ 965 bi
- Anthropic mercado secundário: US$ 992 bi (ultrapassa OpenAI)
- Backlog de pedidos de IA da Dell: US$ 51,3 bi
- Receita Dell 1T: US$ 43,8 bi (vs. US$ 35,4 bi esperado)
- Divisão infra Dell: +181% a/a; servidores IA: +757% a/a
- Margem bruta Dell: -3,3 p.p.
- Qwen3.7-Max no AA Intelligence Index: 56,6 (5º-7º)
- Janela de contexto Qwen3.7-Max: 1M tokens entrada / 64K saída
- Velocidade: 208,3 tokens/s
- Preço Qwen3.7-Max: US$ 2,50 / US$ 0,25 / US$ 7,50 por milhão de tokens (in/cached/out)

- Preço Claude Opus 4.8: US$ 5 / US$ 25 por milhão (in/out)
- Claude Opus 4.8: 4x menos erros em código vs. 4.7
- GraphWalks 1M tokens: 68,1% (Opus 4.8) vs. 40,3% (Opus 4.7)
- Memória GPU para LLM 70B: ~140 GB
- Projeção crescimento compute IA (Microsoft): 1.000x em 3 anos
- Mercado robótica em 2030 (projeção): US$ 200 bi
- Unicórnios sem nova rodada em 3 anos: ~50% dos 857

Narrativas em disputa

Tese: a fronteira está se concentrando irreversivelmente no topo. Anthropic e OpenAI sozinhas capturaram US$ 250 bi de funding em dois anos, enquanto metade dos unicórnios estagnou. Os hyperscalers (Alphabet, Microsoft, Amazon, Meta) usam balanços para sustentar capex de centenas de bilhões anuais, criando barreira de entrada intransponível. A virada da Alibaba de open para closed reforça que o jogo econômico empurra todos para o mesmo modelo de negócio. **Contraposição: a abertura segue ativa nas tier-2 e em domínios verticais.** GPT-OSS, MiniMax M3 open-source, Cosmos 3 aberto da Nvidia, ESMFold2 do Biohub e GPIC (100M imagens permissivamente licenciadas) mostram que há um ecossistema paralelo de modelos publicamente acessíveis. Pieter Abbeel/Luma também lançaram o Open Physical AI Lab. A fronteira pode estar fechada, mas há "linha de frente" aberta que se atualiza rapidamente.

Tese: o valor da fronteira está em modelos cada vez maiores. Scaling laws continuam segurando — ESMC com 2.8B sequências supera ESM2, Anthropic precisando de US$ 65 bi para "alimentar" próxima geração, Suleyman projetando 1.000x mais compute em 3 anos. **Contraposição: o valor está se deslocando para a camada de aplicação e workflow.** O caso ServiceNow é emblemático: subiu ~40% no mês porque o mercado percebeu que controla os fluxos onde a IA precisa rodar; o "SaaSpocalipse" virou "SaaSalvation". Macdonald (Uber) questiona se mais tokens viram mais valor. O risco é o de uma indústria onde a fronteira dos modelos é commodity e os retornos econômicos vão para quem orquestra workflows.

Tese: o crescimento da IA é genuíno e sustentável. Receitas explosivas (Anthropic, Dell), 800 milhões de horas de podcast no YouTube Premium em abril, casos de uso reais em biologia e defesa. **Contraposição: há sinais de bolha precificada.** Valuations sem precedente sem fluxo de caixa proporcional, Alphabet emitindo US$ 80 bi mesmo com US$ 130 bi de lucro, mercado de previsões avaliando Anthropic em US$ 1,8 trilhão sem ainda saber as margens reais. A própria expressão "SaaSpocalipse" indica reprecificação violenta intra-setor.

Implicações para o Boletim AM

Atividade global e capex. A escala dos compromissos de capex (Alphabet US$ 180-190 bi, Softbank US$ 52 bi na França, Dell com US$ 51 bi de backlog) configura um ciclo de investimento que sustenta crescimento do PIB americano em 2026 mesmo num cenário de juros ainda restritivos. O paper sobre "AI economy" sugere que medidas convencionais subestimam essa contribuição — vale acompanhar revisões de projeções de crescimento dos EUA com lente para o componente "AI compute spending", que já é maior que indústrias inteiras.

Mercados acionários e concentração. O movimento "SaaSpocalipse → SaaSalvation" com ServiceNow +40%/mês, Atlassian +26%/semana, Oracle +17% mostra que a precificação dentro

do setor tech está em violenta rotação. Os múltiplos sugeridos para Anthropic (21-30x receita) são referência para qualquer análise comparativa de tech brasileira. Para o investidor brasileiro, o ponto-chave do Raio-XP citado no MoneyDrop permanece: o Ibov caiu 7,2% em maio enquanto o trade de IA global escalou — saídas estrangeiras de R$ 26 bi em 26 pregões. A divergência se aprofunda enquanto o Brasil não tem exposição direta ao tema.

Política monetária. Indiretamente, a explosão do capex em IA absorve poupança global e pode sustentar a estrutura a termo dos juros americanos em níveis mais altos por mais tempo — a Berkshire alocando US$ 10 bi e Buffett apostando em capex de IA é um sinal de que o "muro de dinheiro" para a fronteira continua. Para o Brasil, isso reforça o cenário de prêmio de risco persistente e dificulta a normalização do diferencial de juros.

Câmbio. O fluxo direcionado a hubs de IA (EUA, China, agora França via Softbank) tende a manter o dólar firme contra moedas de mercados emergentes sem narrativa tech relevante. Real continua sem âncora de fluxo estrutural ligada a essa onda.

Risco regulatório e geopolítico. A ordem executiva americana citada pelo Andrew Ng e o discurso de Andrew Leigh na Austrália sobre "precificar risco de extinção" sinalizam que a janela de regulação está se abrindo. Para o Boletim, vale registrar que o ambiente regulatório frontier-AI hoje é o terceiro vetor (junto a tarifas e tensões em commodities) que pode reprecificar abruptamente o setor — e por arrasto, índices americanos onde Mag7 pesa.

Produtividade e mercado de trabalho. Korinek e coautores alertam para "mismeasurement of labor-tax-base shock". Mesmo no Brasil, onde a adoção é mais lenta, vale começar a tratar IA como variável de produtividade total dos fatores nas projeções de médio prazo — especialmente em serviços modernos e financeiro, segmentos onde o uso já é mensurável.

Fontes (20 newsletters)

- **Daily papers of 5 Jun 2026** — daily_papers_digest@notifications.huggingface.co
- **O segredo sujo do podcasting:Ninguém quer seus números reais (inclusive eu)** — nao_responder@castnews.com.br
- **Qwen3.7-Max Challenges Google for Third Place, AI Saves Whales, Fine-Tuning Breaks Copyright Alignment** — thebatch@deeplearning.ai
- **Daily papers of 4 Jun 2026** — daily_papers_digest@notifications.huggingface.co
- **YouTube supera 1 bilhão de usuários mensais de podcast** — nao_responder@castnews.com.br
- **Daily papers of 3 Jun 2026** — daily_papers_digest@notifications.huggingface.co
- **Stop wasting GPU memory** — hello@deeplearning.ai
- **Eu estava errado, e não custa admitir: o podcast de IA invadiu o Brasil** — nao_responder@castnews.com.br
- **A loja de tudo ads** — tech-drops-newsletter@mail.beehiiv.com
- **Daily papers of 2 Jun 2026** — daily_papers_digest@notifications.huggingface.co
- **Nvidia: Let's Get Physical** — aidrop@mail.beehiiv.com
- **Correios: a bagunça nova e abagunça antiga** — moneydrop@mail.beehiiv.com
- **Acast diz que marcas erram alvo na publicidade em podcasts** — nao_responder@castnews.com.br
- **Daily papers of 1 Jun 2026** — daily_papers_digest@notifications.huggingface.co

- **Import AI 459: AI oversight is difficult; scaling laws for protein folding models; and pricing the extinction risk...** — importai@substack.com
- **Amazon transforma a Alexa em geradora de podcasts por IA** — nao_responder@castnews.com.br
- **Claude Opus 4.8 chegou e todo mundo está elogiando** — aiwhisperbr@mail.beehiiv.com
- **Até o Papa está preocupado com aInteligência Artificial...** — newsletter@mail.datahackers.com.br
- **SaaSpocalipse → SaaSalvation** — tech-drops-newsletter@mail.beehiiv.com
- **Empresas de podcast criam aliança para padronizar amensuração de anúncios** — nao_responder@castnews.com.br