Economia dos tokens: gastos explodem, ROI é cobrado

Dossiê de aprofundamento — análise vertical do tema da semana

Vitor Wilher[1]

21 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

Economia dos tokens: gastos explodem, ROI é cobrado

A inflexão do "tokenmaxxing" para o "tokenminimizing"

A semana consolidou uma virada de chave nas grandes corporações americanas no que diz respeito ao consumo de inteligência artificial generativa. A Meta, que até pouco tempo mantinha um leaderboard interno (apelidado de "Claudeonomics") listando os 250 funcionários que mais consumiam tokens — chegando a cogitar incluir a métrica nas revisões de performance —, fez meia-volta e lançou um "AI Gateway" para conter o consumo. Em apenas trinta dias, os funcionários da Meta haviam queimado 60 trilhões de tokens; no mês seguinte, o número saltava para 73 trilhões antes da intervenção. A nova ferramenta impõe orçamentos pré-definidos por funcionário, exige justificativa de projeto e dispara alertas automatizados para picos anormais de gasto (Tech Drops, 15/jun).

O movimento não é isolado. O Uber estourou todo o orçamento anual de IA já no primeiro trimestre; a Microsoft cancelou a licença do Claude para seus funcionários; a Salesforce passou a conectar uso de tokens a resultados de negócio mensuráveis; e o DoorDash implementou controles formais de ROI sobre os investimentos em IA. O padrão é o mesmo: depois de dois anos de adoção orientada por FOMO, o CFO entrou na sala (Tech Drops, 15/jun).

A economia subsidiada dos LLMs e a conta que precisa fechar

Por trás do aperto corporativo está uma descoberta incômoda: o preço atual dos tokens é pesadamente subsidiado pelos próprios laboratórios de IA. Uma estimativa da SemiAnalysis citada pelo Tech Drops indica que uma assinatura mensal de US$ 200 do ChatGPT pode custar até US$ 14.000 à OpenAI, dependendo da intensidade de uso do cliente. É o equivalente ao "cupom de desconto + frete grátis" do modo growth — uma transferência de valor das empresas-modelo para os usuários enquanto a corrida por participação de mercado justifica a queima de caixa. O problema é que esse subsídio tem prazo de validade, e o cálculo de ROI das empresas-cliente terá que ser refeito quando a era da IA barata acabar (Tech Drops, 15/jun).

A OpenAI é o caso emblemático. Documentos de auditoria revelam que a empresa fechou 2025 com prejuízo líquido de US$ 38,6 bilhões — uma quintuplicação em relação aos US$ 5 bilhões de 2024. A receita saltou de US$ 3,7 bi para US$ 13 bi, e o número de usuários ativos mensais passou de 1 bilhão, mas os gastos atingiram US$ 34 bilhões só em 2025: US$ 19,2 bi em P&D, US$ 17,2 bi pagos à Microsoft em treinamento e custo de receita (mais que a própria receita total), além de US$ 41,5 bi de impacto contábil da conversão de non-profit. Mesmo excluindo este último, a perda operacional de US$ 20,9 bi já seria recorde. E o roadmap aponta para ~US$ 600 bi em infraestrutura de IA até 2030 (Tech Drops, 19/jun).

Cobrança por uso como nova fronteira de preço

A resposta dos provedores tem sido migrar do modelo de assinatura plana para precificação variável baseada em consumo, traduzindo no preço final a heterogeneidade do custo marginal. O Copilot Cowork da Microsoft estreou já com créditos escalonados — Plano Leve (100–300 créditos) para entregas simples como updates semanais, Plano Médio (400–700) para briefings com múltiplas fontes, e Plano Pesado (700+) para análises profundas de seis meses de dados. A própria Microsoft se declarou "agnóstica de modelos", permitindo ao usuário escolher entre rodar versões caras (Claude)

ou alternativas mais baratas como o DeepSeek open-source chinês, terceirizando ao cliente a decisão de trade-off custo/qualidade (AiDrop, 18/jun).

Curiosamente, a Anthropic — uma das pioneiras desse modelo via API — fez o movimento oposto e pausou a cobrança por tokens nos seus agentes via SDK, sinalizando que a tarifação granular ainda não convergiu para um equilíbrio (AiDrop, 18/jun).

Capital humano vs. capital de tokens: a tese de Nadella

Satya Nadella publicou um manifesto raro propondo um quadro analítico para essa nova economia: o valor de uma empresa na era da IA se divide entre **capital humano** (conhecimento, julgamento, relacionamentos, reconhecimento de padrões dos funcionários) e **capital de tokens** (a capacidade de IA que a empresa constrói e detém por si própria, independente de provedores). A vantagem competitiva, segundo ele, não está em perseguir o melhor modelo da semana, mas em construir um *learning loop* proprietário que sobreviva à troca de fornecedor. A prova de fogo: trocar o modelo subjacente sem perder a expertise integrada ao sistema (AiDrop, 16/jun).

O subtexto é geopolítico-corporativo: Nadella alerta que "todas as empresas, em todos os setores, cedem valor a alguns poucos modelos que absorvem tudo o que veem pela frente", e que uma economia de IA administrada por um punhado de modelos devastaria setores inteiros. É uma indireta cifrada à dependência da própria Microsoft em relação à OpenAI, e à concentração de poder nos labs de fronteira (AiDrop, 16/jun).

A "utility tax" das alucinações e o custo embutido de qualidade

Para além do custo nominal por token, há um custo de oportunidade significativo embutido na busca por confiabilidade. Pesquisa do Google quantificou o que os desenvolvedores chamam de *utility tax*: para reduzir a taxa de alucinação de 25% para 5%, os modelos atuais precisam recusar respostas em ~52% dos casos. Em outras palavras, metade das respostas potencialmente corretas é descartada para evitar erros de credibilidade. A proposta da empresa é uma redefinição conceitual — alucinação deixaria de ser "qualquer erro factual" e passaria a ser "informação errada entregue com certeza indevida", abrindo espaço para o conceito de *faithful uncertainty* em que o modelo calibra explicitamente seu grau de confiança. Para chatbots, o ganho é marginal; mas na economia agêntica em escala, o volume de tokens economizados pode ser material (AiDrop, 16/jun).

A consultoria como vítima colateral do paradoxo

A pressão por ROI explícito sobre IA está derrubando o setor que historicamente intermediava a adoção de tecnologia: as grandes consultorias. A Accenture, maior consultoria do mundo em headcount, caiu 18% num único dia e já perde metade do valor no ano. IBM (-15% no ano), Infosys (-37%), Cognizant (-46%) e EPAM (-61%) acompanham o massacre. O fenômeno foi batizado de "Paradoxo do Consultor": as próprias firmas vendem transformação por IA aos clientes, mas essa mesma tecnologia canibaliza suas receitas tradicionais de outsourcing e implementação. Quando o cliente passa a cobrar ROI duro do consumo de tokens, o intermediário humano cobrado por hora vira o primeiro corte (Tech Drops, 19/jun).

A Shadow AI como passivo oculto e a corrida pela governança

O contraponto ao aperto de gastos é o fenômeno da Shadow AI: cerca de **2/3 dos profissionais** usam ferramentas de IA não autorizadas pelas empresas, e 43% inserem e-mails corporativos, 40% notas de reunião, 34% dados de clientes e 31% documentos financeiros sensíveis em modelos públicos. Estudo do Google Workspace com a IDC reforça: 74% dos profissionais usam suas IAs pessoais no trabalho, mas apenas 30% das empresas têm políticas claras sobre o tema. Outro levantamento aponta que funcionários têm acesso a pelo menos um assistente aprovado, mas usam ativamente entre 4 e 8 produtos diferentes — sinal de que a governança formal está significativamente atrás do uso real, criando passivos regulatórios (LGPD/GDPR) e securitários ainda não precificados (AiDrop, 16/jun; Tech Drops, 19/jun).

Números-chave

- **OpenAI**: prejuízo líquido de **US$ 38,6 bi** em 2025 (vs. US$ 5 bi em 2024); receita US$ 13 bi; gastos US$ 34 bi; perda operacional recorrente de US$ 20,9 bi.
- **OpenAI**: P&D US$ 19,2 bi; pagamentos à Microsoft US$ 17,2 bi; queima de US$ 3,7 bi só no 1T26.
- **Infraestrutura de IA prevista**: ~**US$ 600 bi até 2030** (OpenAI).
- **Subsídio implícito**: assinatura ChatGPT de US$ 200/mês pode custar até **US$ 14.000** à OpenAI (SemiAnalysis).
- **Meta**: **60 trilhões de tokens** consumidos em 30 dias; **73 trilhões** no mês seguinte antes do corte.
- **Market share OpenAI**: caiu abaixo de **50%** em março/2026 pela primeira vez desde 2023.
- **Usuários**: ChatGPT 1,1 bi mensais; Gemini 662 mi; Claude 245 mi.
- **Anthropic** já lidera em **% de empresas americanas com assinatura paga de IA**, ultrapassando a OpenAI.
- **Copilot Cowork**: créditos escalonados de 100 a 700+ por tarefa.
- **Utility tax do Google**: reduzir alucinação de 25% para 5% descarta **52% das respostas corretas**.
- **Shadow AI**: 2/3 dos profissionais usam IA não autorizada; 74% usam IA pessoal no trabalho; só 30% das empresas têm políticas claras.
- **Consultorias no ano**: Accenture -50%; EPAM -61%; Cognizant -46%; Infosys -37%; IBM -15%.

Narrativas em disputa

Tese da demanda inelástica (OpenAI, Anthropic e investidores em data centers): o consumo de IA crescerá tão rápido que justificará a queima atual de bilhões; a curva de aprendizado e os ganhos de produtividade tornarão o ROI evidente nos próximos 24-36 meses. SpaceX e Anthropic já firmaram acordos para data centers no espaço, sinalizando que a aposta em escala continua.

Tese do ROI cobrado (Meta, Microsoft, Uber, Salesforce, DoorDash): o subsídio dos labs distorce o cálculo econômico; quando o preço dos tokens convergir para o custo real, boa parte dos casos de uso atuais deixará de fazer sentido. A reação é instalar gateways, dashboards e governança de consumo *antes* da normalização dos preços.

Tese de Nadella (capital humano + capital de tokens): o valor não está em comprar o melhor modelo, mas em construir loops de aprendizado proprietários que sobrevivam à troca de fornecedor — uma defesa contra a captura de valor pelos labs de fronteira.

Contraponto da Shadow AI: enquanto CFOs apertam, funcionários driblam — sinalizando que a demanda real por IA na ponta é maior do que as políticas oficiais admitem, e que a "racionalização" pode ser parcialmente artificial.

Implicações para o Boletim AM

O tema toca diretamente três vetores macro relevantes para a semana. **Primeiro, mercados de capitais**: o prejuízo de US$ 38,6 bi da OpenAI, somado à formalização do acrônimo MANGOS (Meta, Anthropic, Nvidia, Google, OpenAI, SpaceX) e ao IPO da SpaceX precificado a 95x receita (fechando o primeiro dia a 112x), reforça que a capitalização do setor segue dissociada de fundamentos correntes — risco de revisão abrupta caso a tese de demanda inelástica seja desafiada. O caso da gestora da Flórida que perdeu US$ 50 bi em aposta errada em IA é canário relevante.

Segundo, ciclo de capex e implicações para juros: os ~US$ 600 bi em infraestrutura de IA projetados pela OpenAI até 2030, combinados com a captação de US$ 25 bi em dívida da Nvidia (primeira grande emissão desde 2021) e os US$ 7 bi da DeepSeek a valuation de US$ 50 bi, sinalizam que o ciclo de investimento em IA pressionará demanda por capital, semicondutores e energia — variável relevante para inflação de bens duráveis e para a curva de juros americana, especialmente no contexto da estreia hawkish de Warsh no Fed mantendo juros em 3,50%-3,75%.

Terceiro, produtividade corporativa e revisão de tese: o movimento de "tokenminimizing" sugere que o ganho de produtividade via IA — usado por bancos centrais e analistas para justificar projeções otimistas de PIB potencial — pode estar superestimado no curto prazo. Se Meta, Microsoft e Uber estão *cortando* uso de IA por ROI insuficiente, a narrativa do choque positivo de oferta via IA precisa ser qualificada. Para o Brasil, vale acompanhar o anúncio dos US$ 550 mi do Rio AI City e o modelo Rio 3.5 397B como sinal de que mesmo orçamentos públicos modestos (R$ 500 mil de custo de desenvolvimento) já participam dessa cadeia — abrindo flanco para discussão sobre IA soberana e dependência cambial de provedores de fronteira.

Fontes (16 newsletters)

- **Daily papers of 19 Jun 2026** — daily_papers_digest@notifications.huggingface.co
- **Midjourney Spa** — tech-drops-newsletter@mail.beehiiv.com
- **Testing Mythos and Fable, Moving Beyond SWE-bench, Nvidia's Open Contender** — thebatch@deeplearning.ai
- **Daily papers of 18 Jun 2026** — daily_papers_digest@notifications.huggingface.co
- **Durma enquanto eles trabalham** — aidrop@mail.beehiiv.com
- **MANGOS: o kit de IA em Wall Street** — moneydrop@mail.beehiiv.com
- **Daily papers of 17 Jun 2026** — daily_papers_digest@notifications.huggingface.co
- **Voice + AI agents: fast, reliable, and easier than you think** — hello@deeplearning.ai
- **Fin de uma era** — tech-drops-newsletter@mail.beehiiv.com
- **Daily papers of 16 Jun 2026** — daily_papers_digest@notifications.huggingface.co

- **Shadow AI e os riscos da rebeldia** — aidrop@mail.beehiiv.com
- **SpaceX: IPO de fé ou de fundamentos?** — moneydrop@mail.beehiiv.com
- **Daily papers of 15 Jun 2026** — daily_papers_digest@notifications.huggingface.co
- **O que a Salesforce aprendeu após lançar mais de 20 mil agentes de IA?** — newsletter@mail.datahackers.com.br
- **Claude perdeu o visto** — tech-drops-newsletter@mail.beehiiv.com
- **EUA barra a melhor IA de código do mundo** — aiwhisperbr@mail.beehiiv.com