Artificial Intelligence

marasset.com.br

Junho de 2024

mar asset management

Artificial Intelligence

marasset.com.br

Junho de 2024

aviso legal

As informações aqui contidas são consideradas confiáveis e foram obtidas em fontes consideradas confiáveis. Entretanto, esclarecemos que nós não fazemos nenhuma declaração ou garantia, expressa ou implícita, com respeito à imparcialidade, consistência, precisão, razoabilidade ou integralidade, das informações ou opiniões aqui reportadas. Além disto, não temos nenhuma obrigação de atualizar, modificar ou aditar esse material e tampouco notificar o leitor sobre quaisquer eventos, assuntos aqui declarados ou qualquer opinião, projeção, previsão ou estimativa aqui contempladas que eventualmente mudarem ou se tornarem imprecisas posteriormente.

sumário

AI ao longo das décadas
Machine learning, GPUs, LLMs e ChatGPT
Como os LLMs funcionam?
Impactos de AI na economia
Consumer Applications
Semicondutores
Edge AI
Energia
Próximos passos


AI ao longo das décadas

AI ao Longo das Décadas

1950s

- Turing introduz o conceito do "Teste de Turing" em 1950
- Marvin Minsky e Dean Edmonds desenvolvem a primeira "*artificial neural network*", utilizando 3,000 tubos a vácuo para simular 40 neurônios
- Primeiro uso do termo "Artificial Intelligence", por John McCarthy, em 1956
- Invenção do Perceptron em 1958, primeiro modelo capaz de aprender a partir de bases de dados
- O termo "machine learning" é usado pela primeira vez em 1959, por Arthur Samuel

1960s

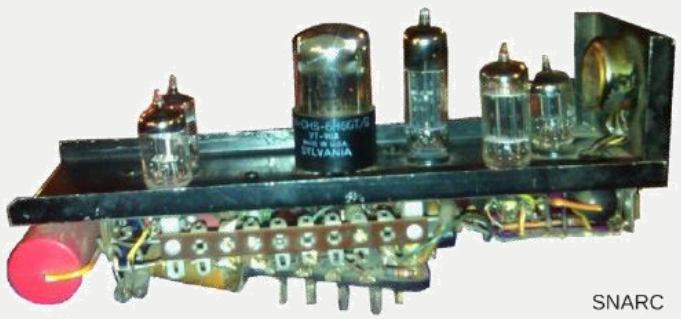
- DARPA financia pesquisa sobre AI no MIT
- Primeiro modelo de linguagem natural (STUDENT), usado para resolver problemas de álgebra (1964)
- Desenvolvimento do primeiro "expert system", voltado para a identificação de moléculas de química orgânica (1965)
- Criação do primeiro *chatbot*, ELIZA, um sucesso na época por ser capaz de conversar com humanos (1966)
- Primeira descrição da ideia de "backpropagation", um elemento fundamental para os avanços posteriores em *deep learning* (1969)
- Em 1968 é lançado "2001: Uma Odisséia no Espaço", com o HAL 9000

1970s

- Um relatório publicado em 1973 trouxe um tom bastante crítico em relação ao estado da pesquisa em AI. Argumentava sobre a falta de avanços concretos, uma falta de visão unificada - com muitos esforços dispersos - e que a tecnologia computacional da época não era suficiente
- Resultou em um longo período de redução nos investimentos em AI, no que ficou conhecido como o primeiro "Inverno AI", durando até o final dos anos 90

AI ao Longo das Décadas

SNARC, a primeira *artificial neural network* (1951)



ELIZA, o primeiro *chatbot* (1966)

HAL 9000, do filme 2001, Uma Odisseia no Espaço (1968)

AI ao Longo das Décadas

1980s

- O foco das pesquisas passou de AI para “Expert Systems”, voltados a tarefas muito mais específicas
- Em 1988 foi publicado o paper que abriu o caminho para o uso de *machine learning* em tradução, um dos principais casos de uso até hoje
- Yann Lecun, Yoshua Bengio e Patrick Haffner mostraram como usar “convolutional neural networks” para reconhecer caracteres escritos à mão

1990s

- O IBM Deep Blue vence o campeão mundial de xadrez, Garry Kasparov, em 1997

2000s

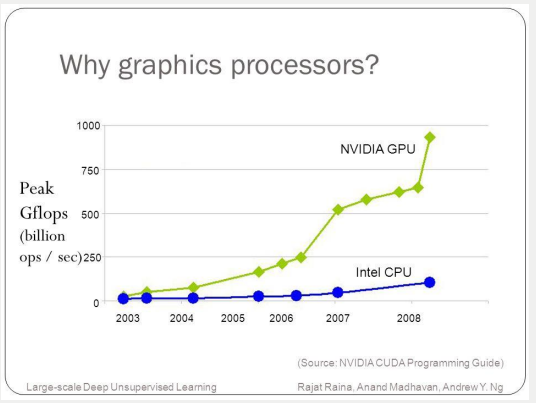
- O primeiro robô humanoide, ASIMO, é desenvolvido em 2002 pela Honda
- Iniciou-se em 2006 o desenvolvimento da base de dados ImageNet, lançada em 2009
- Em 2009, um paper intitulado “Large-Scale Deep Unsupervised Learning Using Graphics Processors” apresenta a ideia e usar GPUs para treinar grandes redes neurais

AI ao Longo das Décadas

Convolutional neural network para identificação de caracteres escritos à mão (1989)

IBM Deep Blue vence Garry Kasparov (1997)

Paper "Large-Scale Deep Unsupervised Learning Using Graphics Processors" (2009)



AI ao Longo das Décadas

2010-2014

- IBM Watson ganha o “Jeopardy!” em 2011.
- Apple lança a Siri em 2011 e Amazon lança a Alexa em 2014
- Em 2012, Hinton, Sustkever e Krizhevsky vencem o ImageNet utilizando um modelo treinado por 2 GPUs, levando a uma nova explosão de interesse e pesquisa em AI
- Em 2013 a DeepMind introduz um modelo de *deep learning* capaz de aprender jogos por meio de tentativa e erro, atingingo nível superior ao de experts humanos
- Em 2014 são inventadas as generative adversarial networks (GANs), capazes de transformar e criar imagens e fotos

2015-2019

- Fundação da OpenAI em 2015
- AlphaGo vence o campeão mundial de Go em 2016
- Publicação do paper “Attention is all you need” pelo Google em 2017, definindo o modelo baseado em *Transformers*, que revolucionou a indústria
- GPT-1 é lançado em 2018, treinado na Wikipedia e com 117 milhões de parâmetros
- GPT-2 é lançado em 2019, com 1,5 bilhão de parâmetros

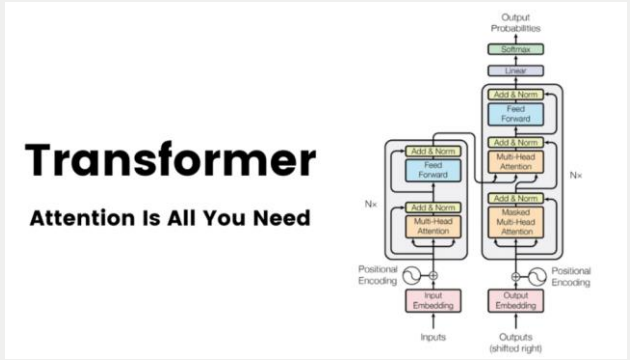
2020-Hoje

- Lançamento do GPT-3 em 2020, com 175 bilhões de parâmetros e do GPT-3.5, uma versão aprimorada, em 2022
- Lançamento do DALL-E, para geração de imagens, em 2021
- ChatGPT é lançado para o público em novembro de 2022, o primeiro grande sucesso B2C de GenAI
- Lançamento de diversos novos modelos, com número de parâmetros crescente (500B a 1,5T): Bard/Gemini do Google, GPT-4 da OpenAI, Claude da Anthropic, LlaMa da Meta, dentre outros
- Em 2024 ocorre o lançamento do Gemini Pro e do GPT-4o, modelos multimodais do Google e OpenAI

AI ao Longo das Décadas

Hinton, Sustkever e Krizhevsky vencem o ImageNet utilizando um modelo treinado por 2 GPUs, levando a uma nova explosão de interesse e pesquisa em AI (2012)

Paper “Attention is All You Need”, introduzindo o modelo de transformers, é publicado pelo Google (2017)

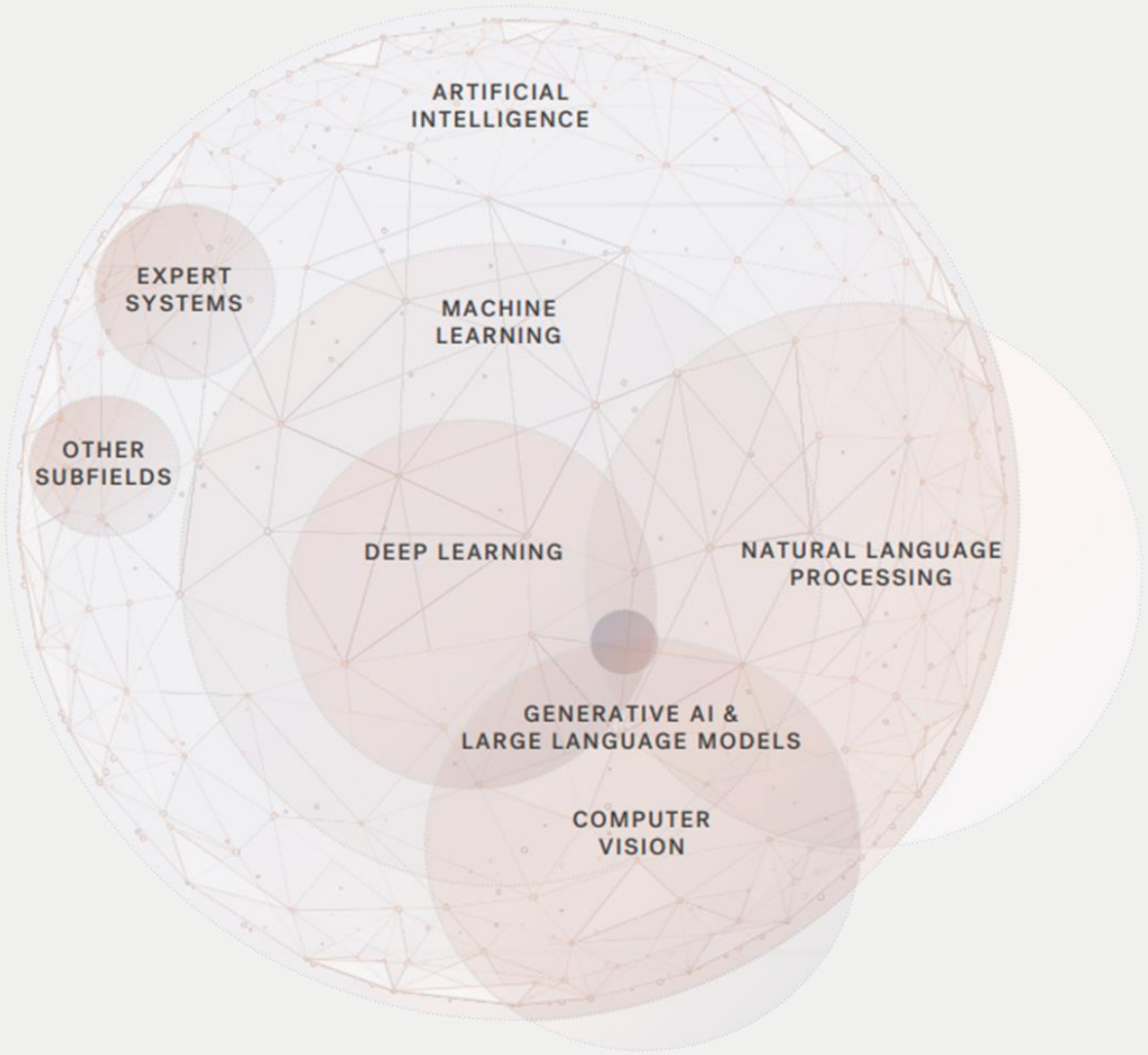


Lançamento do ChatGPT (2022)

ChatGPT


Machine learning, GPUs, LLMs e ChatGPT

AI engloba um amplo conjunto de áreas de pesquisa



- **Machine Learning**: Foco na construção de sistemas que aprendem a partir de dados, sem serem explicitamente programados para tarefas específicas
- **Neural Networks:** Modelos de treinamento de *machine learning* inspirados no funcionamento do cérebro humano. Baseia-se na organização de unidades de processamento simples, os "neurônios", e na relação entre eles
- **Deep Learning**: Uma subárea de *machine learning* envolvendo redes neurais de muitas camadas (*deep neural networks*) que aprendem a partir de grandes quantidades de dados
- **Natural Language Processing:** Um subcampo da IA focado em permitir que computadores entendam, interpretem e gerem a linguagem humana
- **Generative AI & Large Language Models: Tem como objetivo o desenvolvimento de modelos com a capacidade de criar conteúdos novos, como textos, imagens, vídeos ou músicas.**
- **Computer Vision**: Área de estudo que visa permitir que computadores interpretem e tomem decisões com base em dados visuais do mundo real
- **Expert Systems**: Sistemas de AI que emulam as capacidades de tomada de decisão de um especialista humano em um domínio específico

Mas o que é *machine learning*?

- *Machine learning* é um campo da inteligência artificial que permite que computadores **aprendam a partir de dados, sem serem programados explicitamente**. Isso significa que eles podem melhorar seu desempenho em uma tarefa ao longo do tempo, sem a necessidade de intervenção humana

Programação baseada em regras	Machine learning
• Define regras explícitas para o computador seguir • Requer conhecimento humano profundo do problema • Inflexível e difícil de adaptar a novos dados	• Aprende com dados e identifica padrões • Flexível e adaptável a novos dados • Pode lidar com problemas complexos que são difíceis de programar com regras

- *Machine learning* **assemelha-se ao conhecimento adquirido de forma empírica**, baseado em heurísticas e experiências. Exemplos:
  - Entendemos física empiricamente: conseguimos prever a trajetória de uma bola no ar para pegá-la, sem qualquer esforço
  - Somos capazes de aprender uma nova língua sem saber as regras gramaticais, apenas pela exposição à fala e leitura
- **São como o Sistema 1 de Kahneman**: são rápidos, eficientes e capazes de lidar com grandes quantidades de dados. Mas não são capazes de racionalizar ou de tomar decisões complexas, tal como o Sistema 2

Machine learning já está presente em diversas indústrias

Translation

Google Translate

Self-driving vehicles

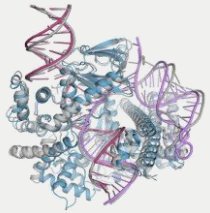
WAYMO

TESLA

Content recommendation

NETFLIX

Drug Discovery



google-deepmind/
alphafold

Credit score

nu

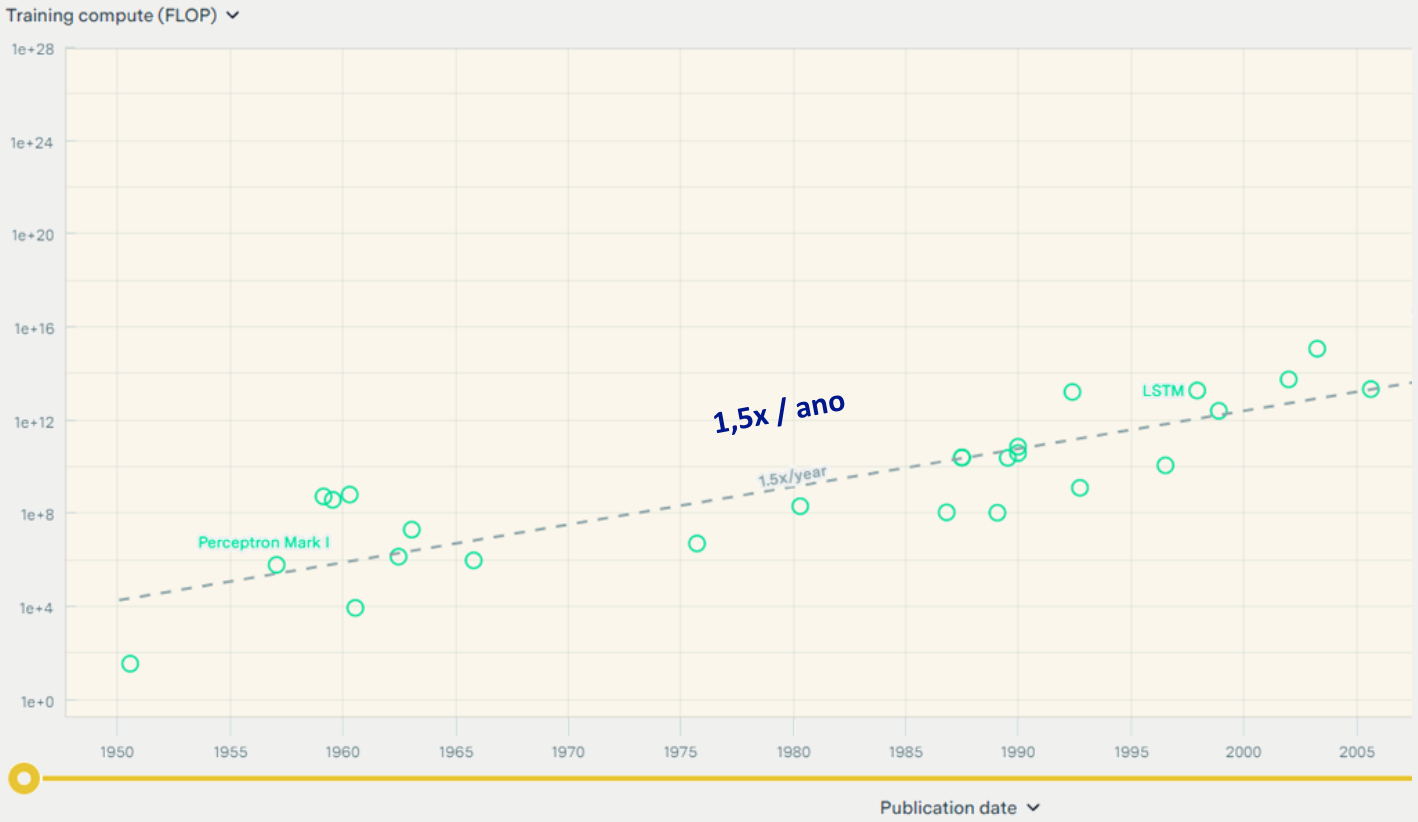
Supply & demand forecasting

SHEIN

TikTok

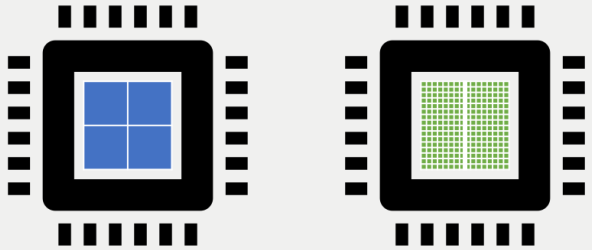
A importância das GPUs para AI

- Um dos principais gargalos para o avanço da pesquisa em AI sempre foi a **quantidade massiva de processamento de dados** necessária para treinamento dos modelos de *machine learning*
- Ao longo das décadas, **a Lei de Moore foi um grande *tailwind*** para os avanços no setor, aumentando a capacidade de processamento em **1,5x por ano**
- **Ainda assim, não foi suficiente**. O setor mostrou avanços consistentes, mas até então nenhum *breakthrough*

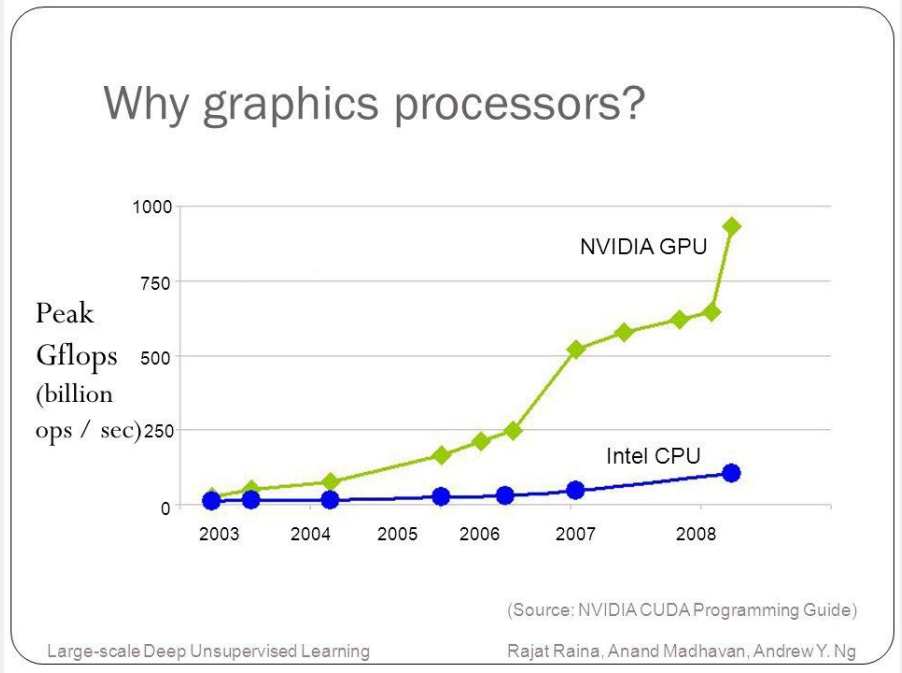


A importância das GPUs para AI

- **As GPUs (Graphic Processing Units) surgiram no final da década de 90, para o uso específico de processamento gráfico**. A ideia chave é simples: CPUs são altamente customizáveis, mas pouco eficientes para lidar com grandes volumes de dados. Processamento gráfico, por outro lado, possui cálculos relativamente simples e "repetitivos", que podem ser realizados em paralelo (ex: cada pixel a ser projetado em uma tela pode ser "calculado" de forma praticamente independente dos demais pixels da mesma imagem)
- Não são todos os problemas computacionais que podem ser resolvidos com o auxílio de GPUs, apenas aqueles paralelizáveis. Ou seja, **problemas que podem ser subdivididos em componentes menores e agrupados posteriormente para chegar à solução**
- Em 2009, com a publicação do paper "Large-Scale Deep Unsupervised Learning Using Graphics Processors", começou a proliferar a **ideia de que o potencial desses chips poderia se estender para além da parte gráfica**

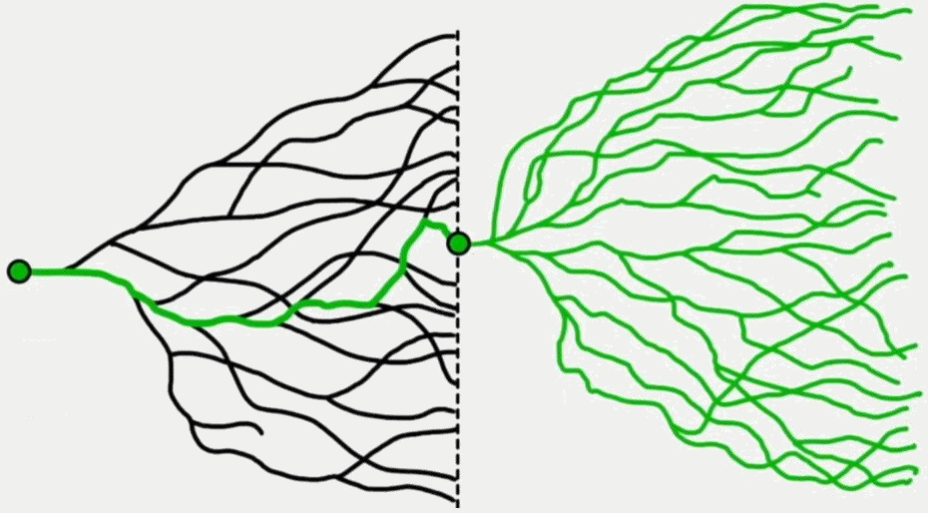


CPU	GPU
Central Processing Unit	Graphics Processing Unit
4-8 Cores	100s or 1000s of Cores
Low Latency	High Throughput
Good for Serial Processing	Good for Parallel Processing
Quickly Process Tasks That Require Interactivity	Breaks Jobs Into Separate Tasks To Process Simultaneously
Traditional Programming Are Written For CPU Sequential Execution	Requires Additional Software To Convert CPU Functions to GPU Functions for Parallel Execution

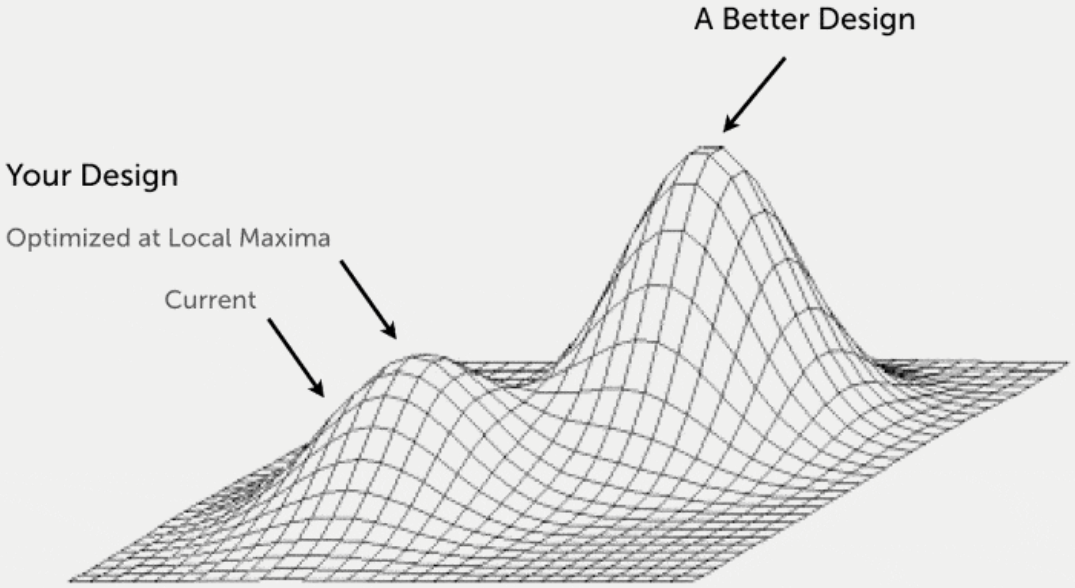


A importância das GPUs para AI

O processamento em paralelo permite testar múltiplos “caminhos” possíveis simultaneamente

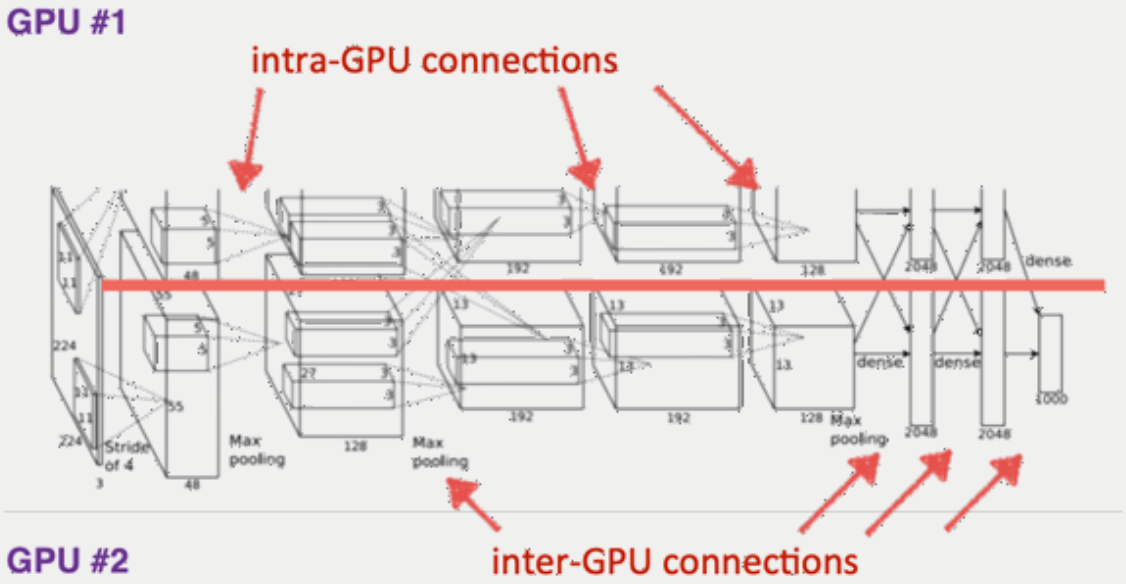


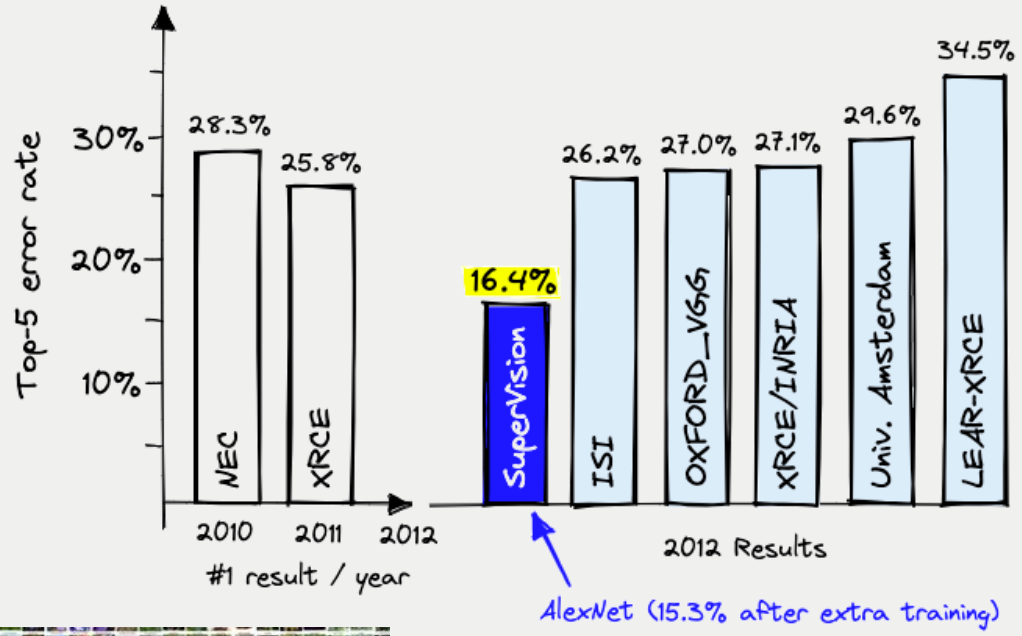
Colocando em termos matemáticos: acelera a busca pelos “máximos globais” de uma função de otimização



A importância das GPUs para AI

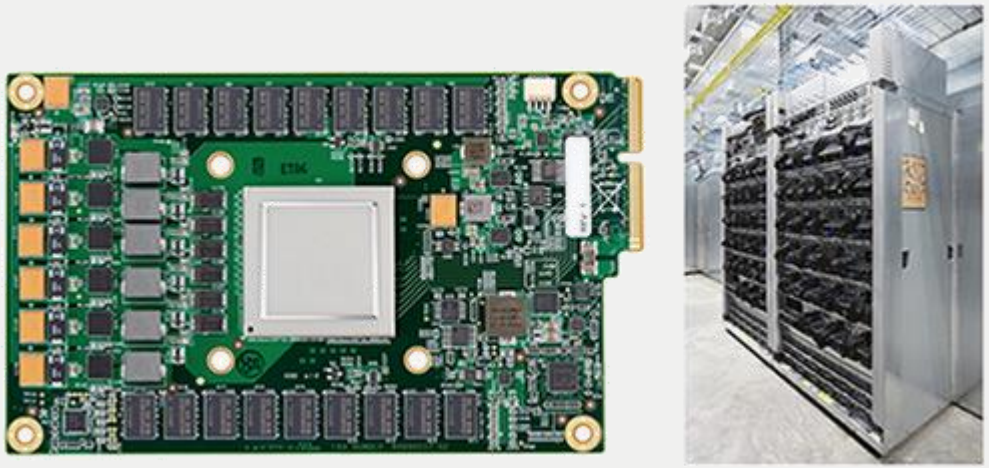
- A vitória do AlexNet na competição ImageNet de 2012 foi um marco para o desenvolvimento de AI
- Foi a primeira vez que o poder das GPUs para treinamento ficou explícito





Placas dedicadas a AI começaram a surgir em 2016

Primeiro TPU (*tensor processing unit*) do Google, lançado em 2016



Primeiro DGX-1 da NVIDIA, entregue à OpenAI em 2016

Attention is All You Need

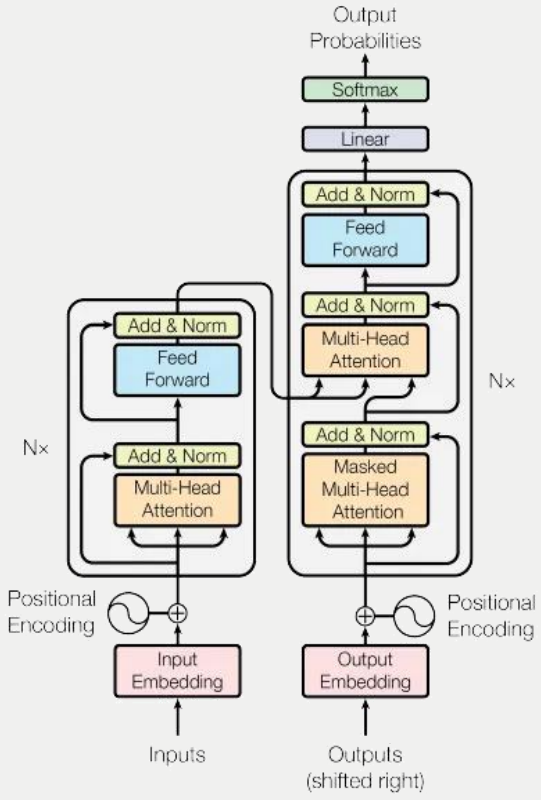
- Em 2017, foi lançado por pesquisadores do Google o paper "Attention is All You Need", que introduziu o modelo de *machine learning* baseado em "**transformers**". Esse paper revolucionou a indústria ao trazer um **framework muito menos intensivo em necessidade de processamento e que permitia paralelização**
- Diferentemente de modelos anteriores, que processavam as palavras de forma sequencial, os *transformers* baseiam-se no mecanismo conhecido como ***self-attention***. Em essência, esse modelo permite que o texto seja lido por completo, mas com cada palavra "prestando atenção" apenas aos valores que surgiram anteriormente

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

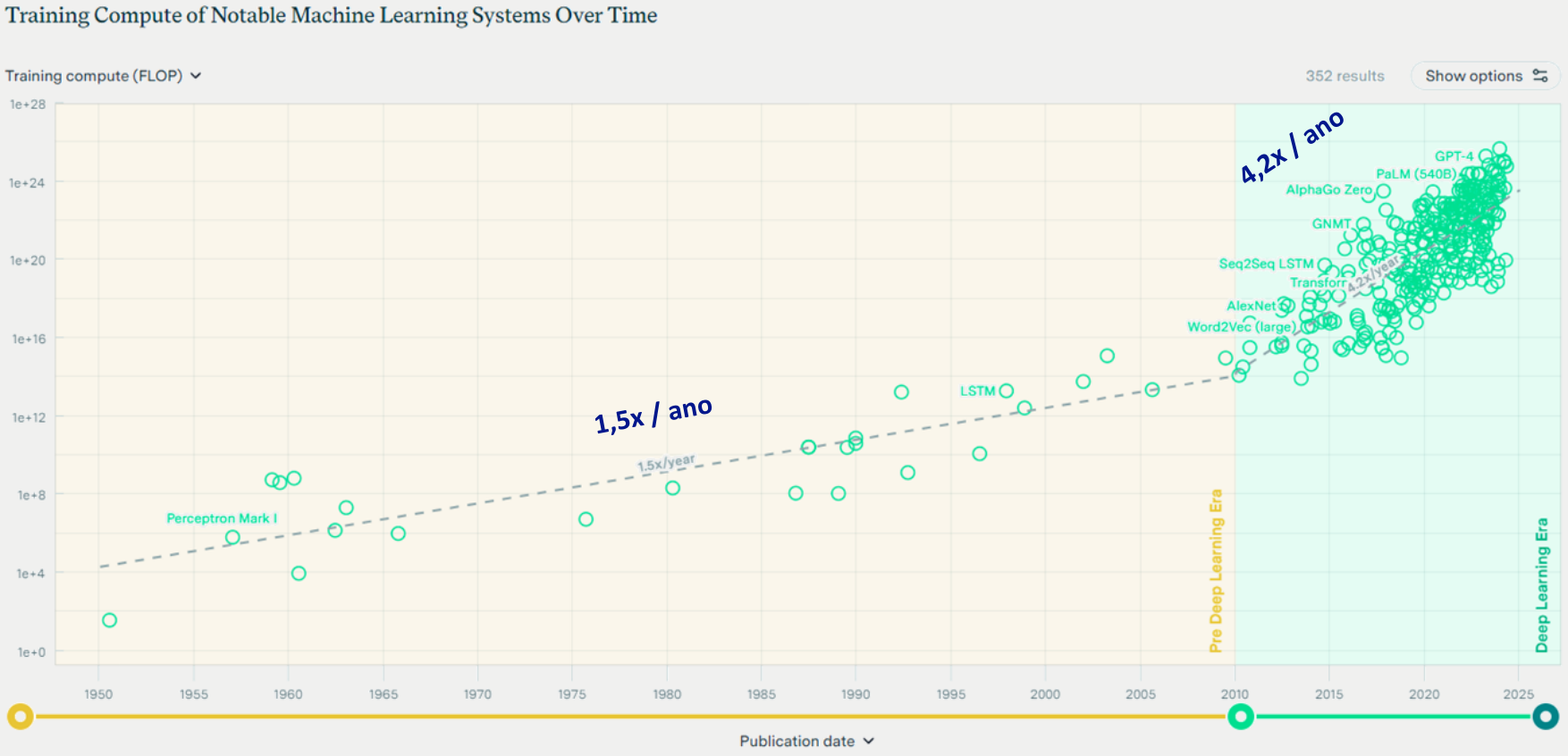
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.



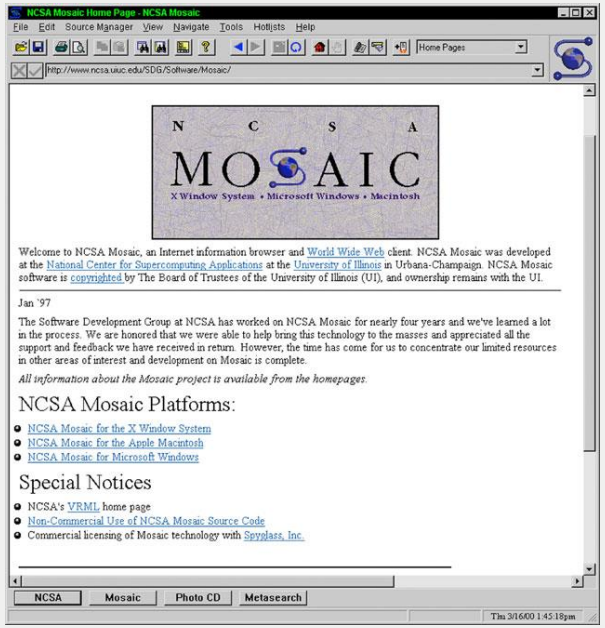
GPUs aceleraram a Lei de Moore

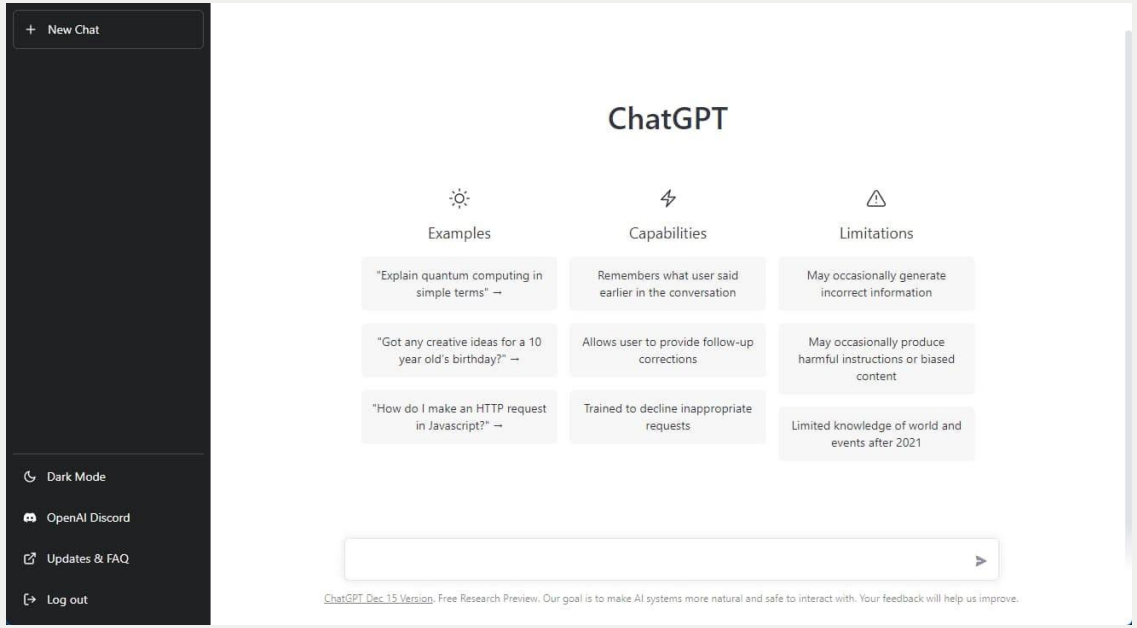
- Com o início da utilização de GPUs, a partir de 2009, e a introdução do modelo baseado em transformers, em 2017, a pesquisa em AI entrou em ebulição
- A evolução da capacidade de treinamento de *neural networks* saiu 1,5x/ano para 4x+/ano, levando a resultados em um espaço de tempo cada vez mais curto e atraindo cada vez mais investimentos



ChatGPT: o "*browser moment*" da Inteligência Artificial

- O lançamento do primeiro *browser* para navegação na web (Mosaic, 1993) foi o grande catalisador para a adoção da internet como produto de massa
- Com sua interface gráfica muito mais amigável e intuitiva, a experiência do usuário (*UX – User Experience*) mudou de forma drástica
- O lançamento do ChatGPT é considerado como o *browser moment* de AI, uma vez que sua interface baseada em linguagem natural permite sua utilização por qualquer pessoa, independente de capacidade técnica

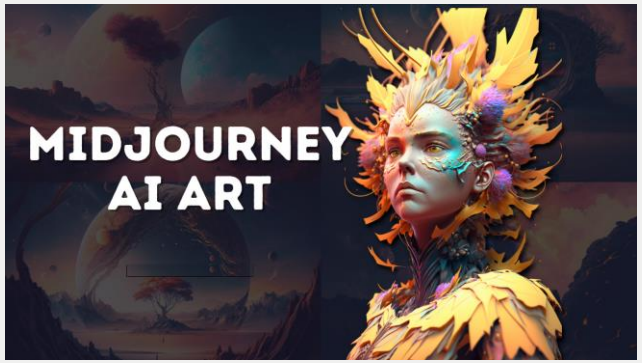




GenAI amplia o mercado endereçável de AI

- LLMs e a *generative AI*, além de trazerem novos casos de uso, **ampliam o mercado endereçável ao permitir que pessoas sem conhecimento técnico possam manipular essa tecnologia**, uma vez que a interface deixa de ser código e passa a ser linguagem natural

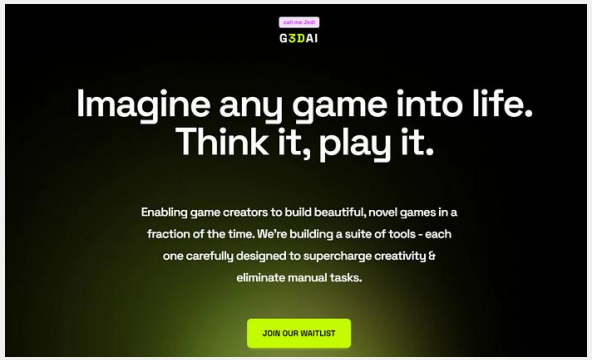
Text to Image



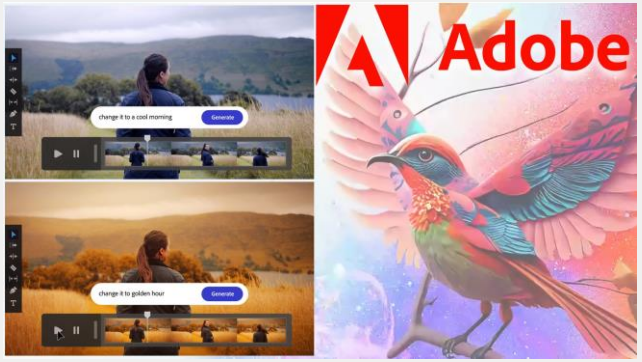
Text to Video

Geração e análise de códigos de programação

Game design

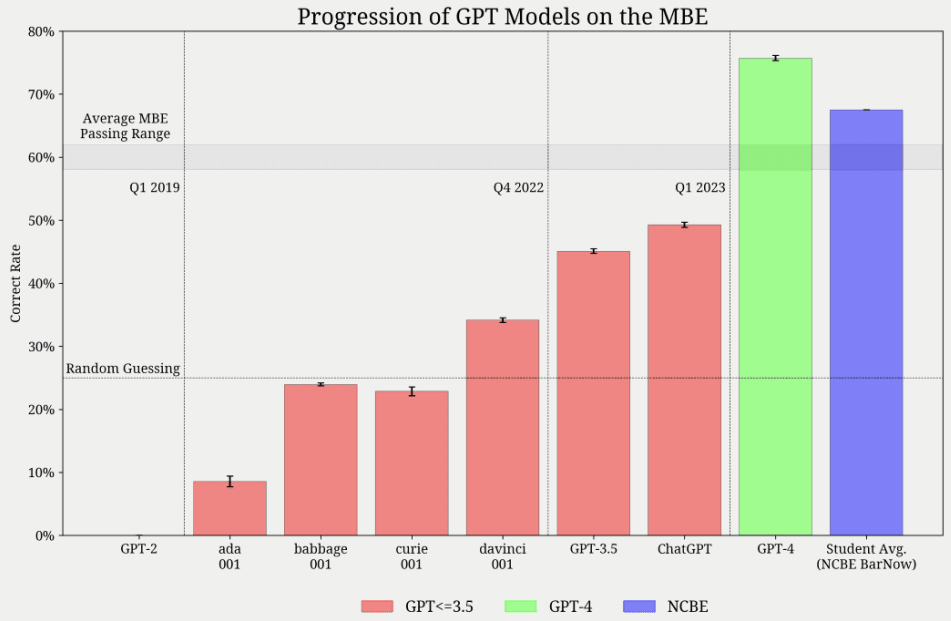


Edição de imagens baseado em texto

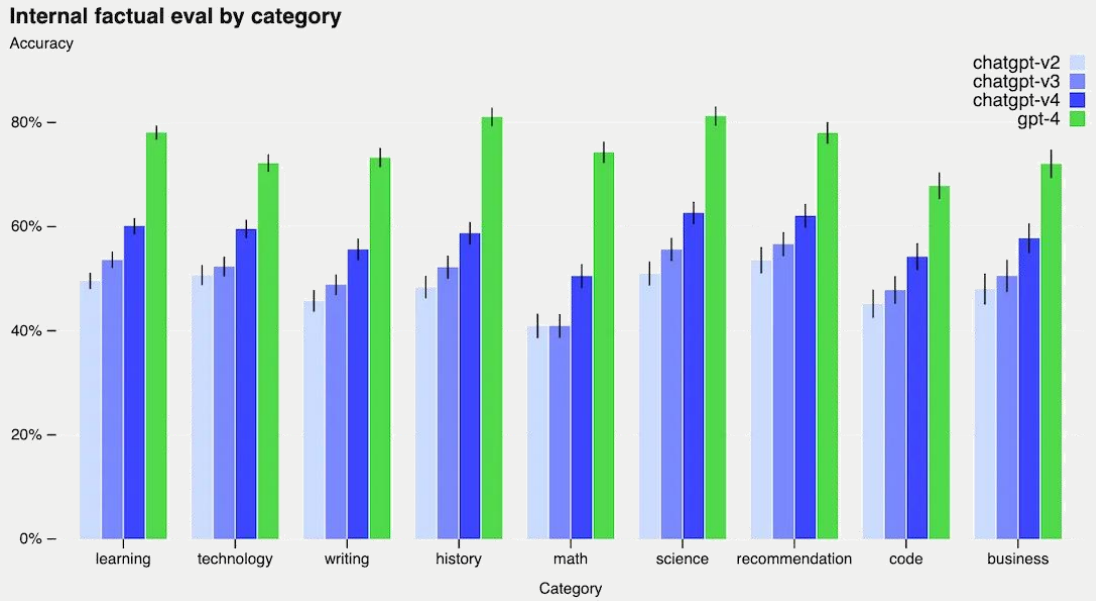


Evolução do ChatGPT

- Desde o lançamento do ChatGPT, em Novembro de 2022, a capacidade dos principais LLMs tem evoluído de forma acelerada
- De modelos treinados apenas com bases de dados antigas, incapazes de fazer contas básicas e extremamente sujeitos a alucinações, hoje temos modelos muito mais capazes. São modelos multimodais, aptos a lidar com diferentes formatos de dados, capazes de acessar informações em tempo real e de interagir com diferentes aplicações



Progression of Recent GPT Models on the Multistate Bar Exam (MBE)



Evolução dos modelos de imagens

"A surreal portrait of a human, the head is a globe, with different landmarks from around the world blending into each other, in the style of a Peter Max's vibrant color palette."

"Simple round logo design of a plant shop "Ali's Plants", pastel color palette "

"Photo of a 6 years old girl smelling a rose in a rose garden"

"Close-up portrait of a young woman with her face partially hidden by a beige sheer fabric, mixture of highlights and shadows, shot on Ektrachrome film"

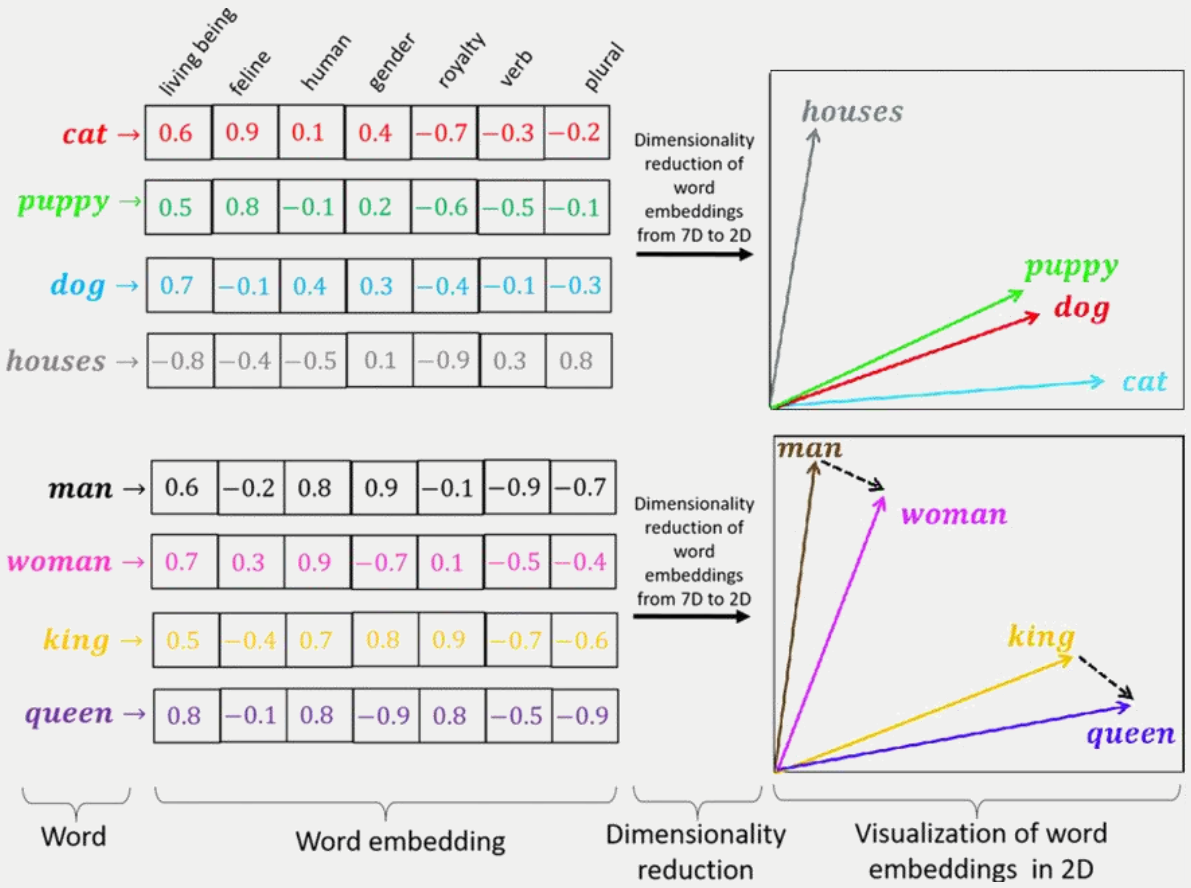

Como os LLMs funcionam?

Como LLMs funcionam?

- Tokens são o elemento básico de um *large language model*. Cada token representa uma palavra, pontuação ou outro elemento fundamental do texto

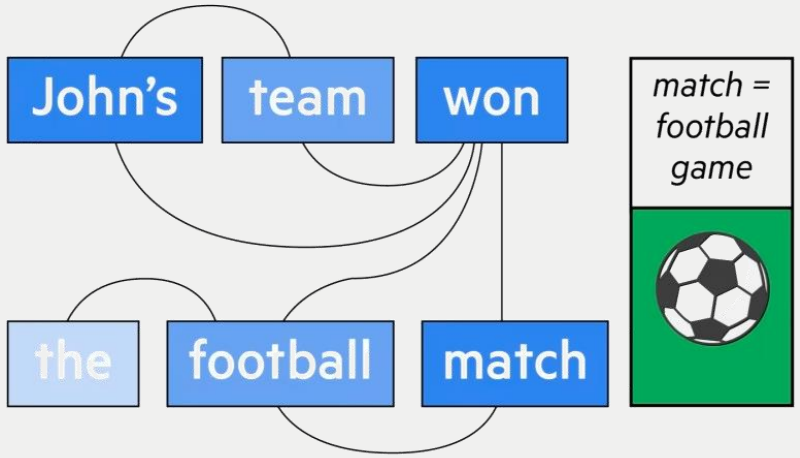
Como LLMs funcionam?

- *Embeddings* transformam os tokens em vetores numéricos. Esses vetores representam o significado da palavra em um espaço multidimensional, capturando nuances como sentido contextual, similaridades com outras palavras e até mesmo informações gramaticais



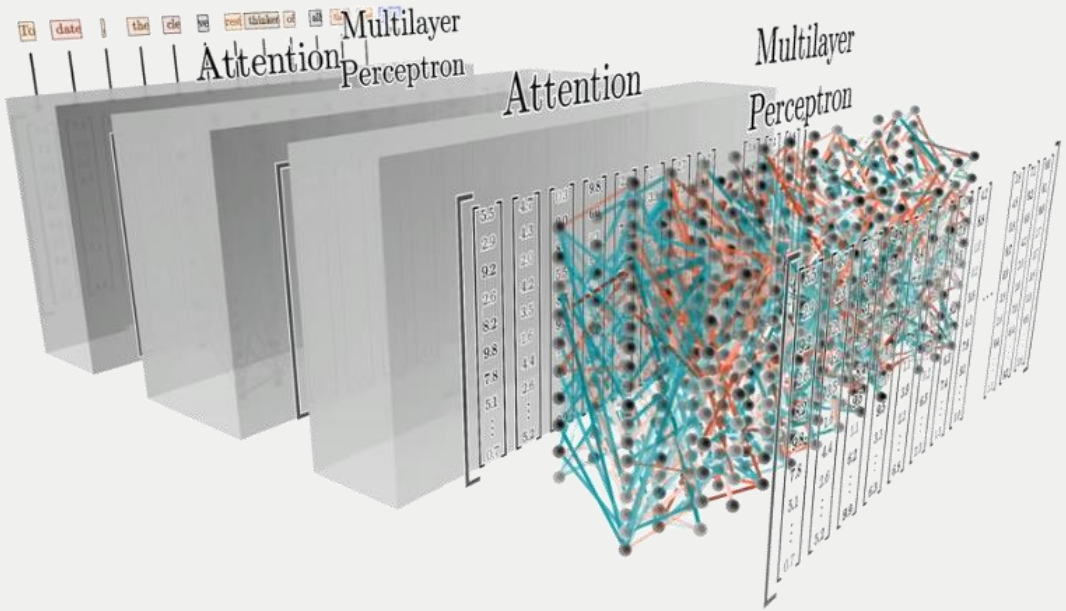
Como LLMs funcionam?

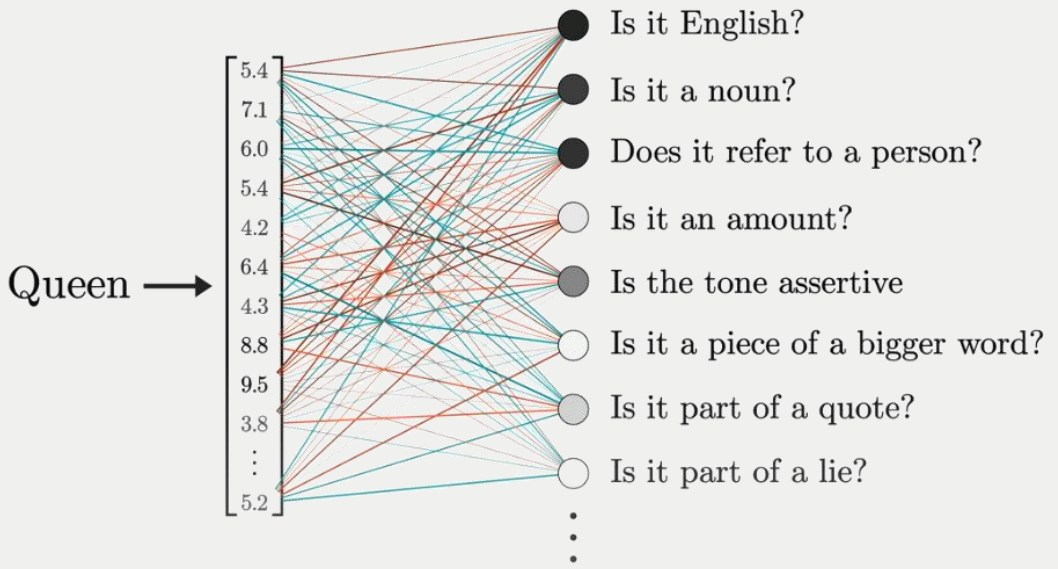
- *Attention layers* são utilizados para identificar as partes do texto mais relevantes para a tarefa em questão. Por meio do mecanismo de *self-attention*, cada token pode "focar" em outros tokens com maior ou menor intensidade, de acordo com o contexto



Como LLMs funcionam?

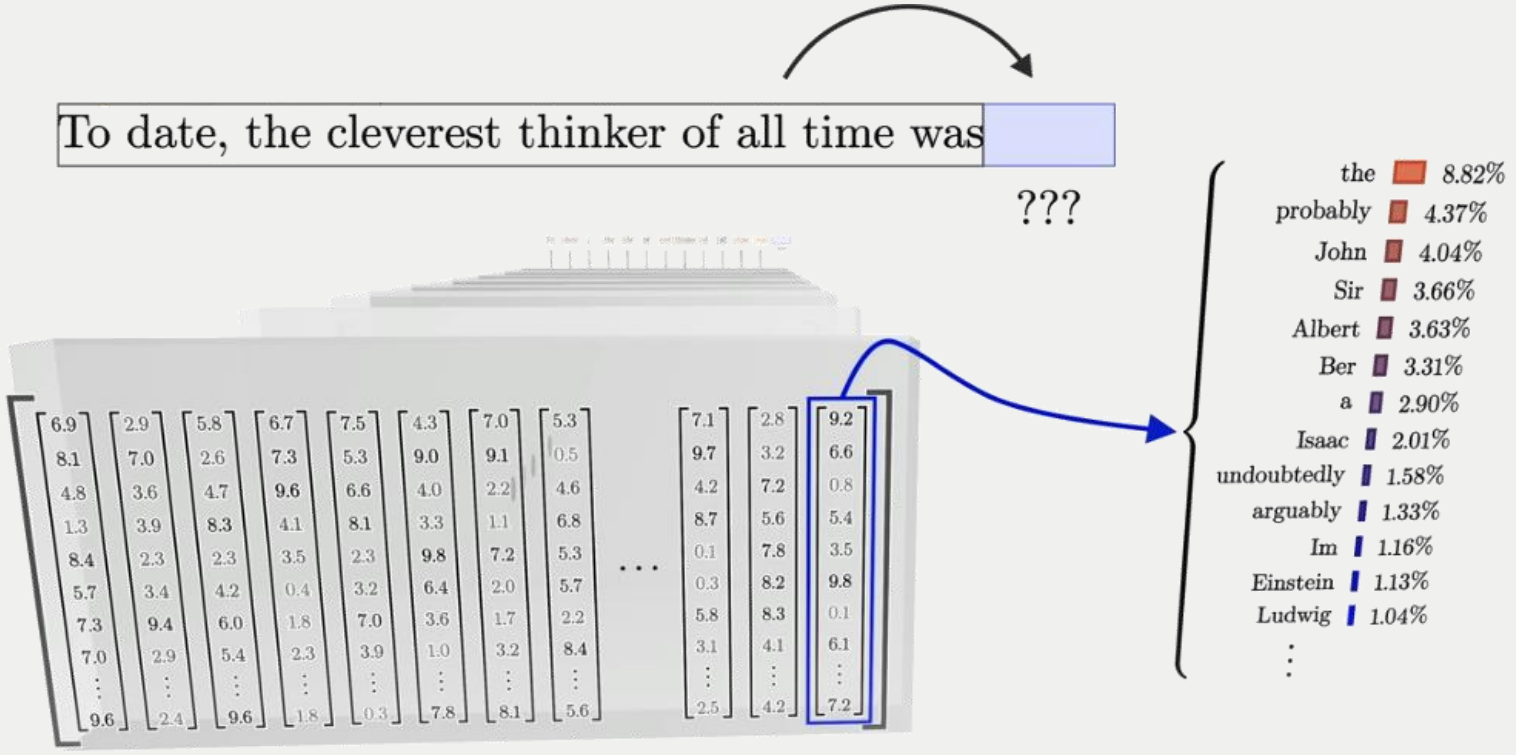
- Os *transformation layers* são compostos por múltiplas camadas de *self-attention* e redes de *feed-forward (perceptron)*, processando os *embeddings* para refinar o entendimento do texto. Esses *transformers* são altamente paralelizáveis, o que os torna eficientes para treinamento e inferência com grandes volumes de dados





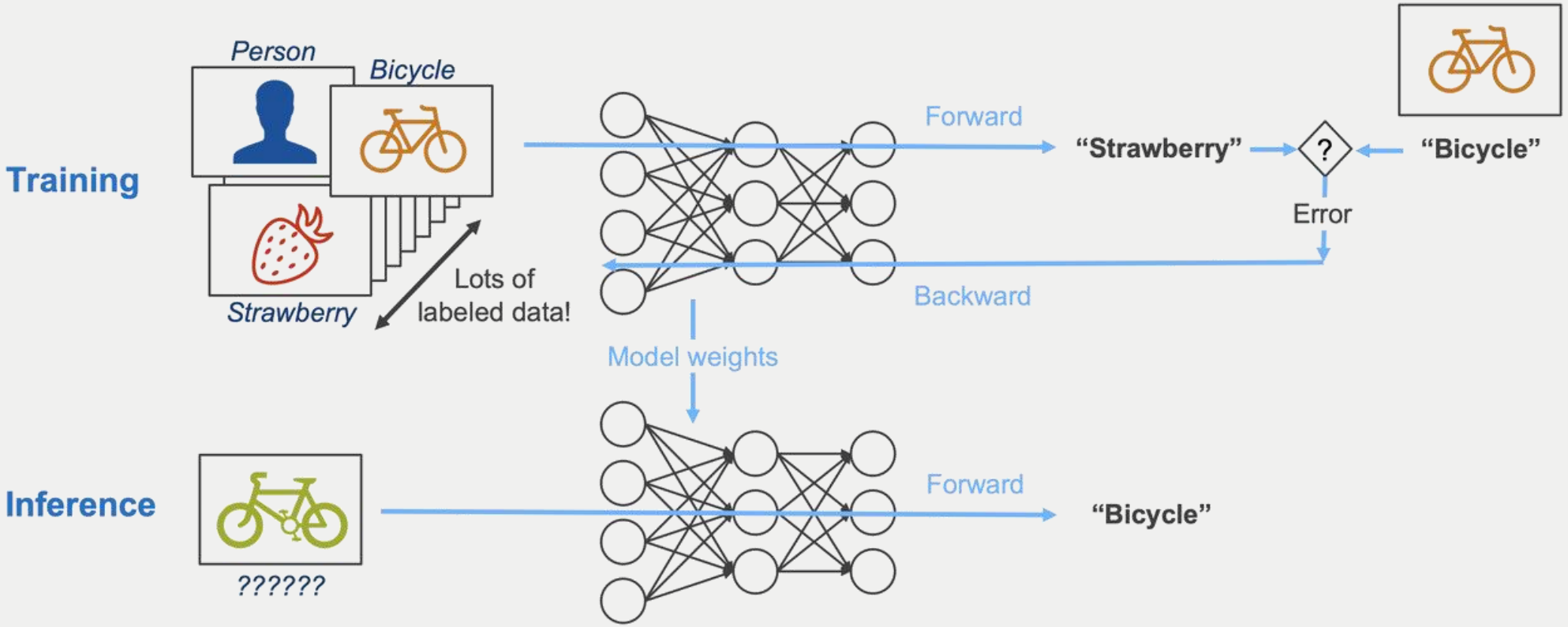
Como LLMs funcionam?

- O objetivo final de cada “etapa” de um LLM é prever a próxima palavra do texto
- É por meio da “previsão” sequencial, palavra por palavra, que esses modelos são capazes de gerar os textos incrivelmente coerentes dos *chatbots* recentes



Treinamento *vs* Inferência

- É por meio do processo de **treinamento** que os modelos de AI criam suas "compreensões de mundo". Para isso, é necessário uma enorme quantidade de dados e poder computacional
- **Inferência**, por outro lado, é a utilização do "conhecimento" construído no processo de treinamento para realizar interpretações a partir de inputs novos. A necessidade de processamento computacional é menor, podendo ser otimizada para tarefas específicas



A “descoberta” da GenAI

- Grande parte das inovações científicas advém diretamente da engenhosidade humana. Ideias novas sobre como modificar, combinar ou estruturar equipamentos e materiais levam a novas ferramentas. São **invenções**. O motor a combustão, o ar condicionado e o PC enquadram-se nessa categoria

- Outras tantas são resultado da observação de manifestações naturais ou comportamentos emergentes. São **descobertas**. O fogo, as propriedades físico-químicas da água e a mecânica quântica exemplificam essa categoria

- **Diz-se que os modelos de GenAI assemelham-se mais a uma descoberta do que uma invenção**

- A elevada capacidade de criação de conteúdo – que se assemelha, ou em determinados casos ultrapassa, os humanos – não se deu inteiramente a partir de um desenvolvimento intencional. Pelo contrário, **essas capacidades foram uma surpresa até para os próprios pesquisadores, que até hoje não entendem completamente seu comportamento**

- **A lógica computacional por trás das inovações recentes de GenAI não mudou de forma substancial ao longo dos anos. O que mudou, isso sim de forma drástica, foi a quantidade de dados submetida para treinamento associada ao maior número de “neurônios”**. Modelos que, quando alimentados por bases de dados pequenas, mostravam-se quase que completamente inúteis, passaram a demonstrar desempenho impressionante ao serem alimentados por bases de dados cada vez maiores e com mais graus de liberdade para interpretação desses dados (parâmetros)

- A enorme quantidade de dados levou a um **comportamento emergente que poucos previam ou sabem explicar**

A “descoberta” da GenAI

- *“They found that **in certain cases, models could seemingly fail to learn a task and then all of a sudden just get it, as if a lightbulb had switched on**. This wasn’t how deep learning was supposed to work. They called the behavior grokking.”*

- *“**Can we ever be confident that models have stopped learning?** Because maybe we just haven’t trained for long enough.”*

- *“This highlights a remarkable fact about deep learning, the fundamental technology behind today’s AI boom: **for all its runaway success, nobody knows exactly how—or why—it works**.”*

- *“The biggest models are now so complex that **researchers are studying them as if they were strange natural phenomena, carrying out experiments and trying to explain the results**.”*

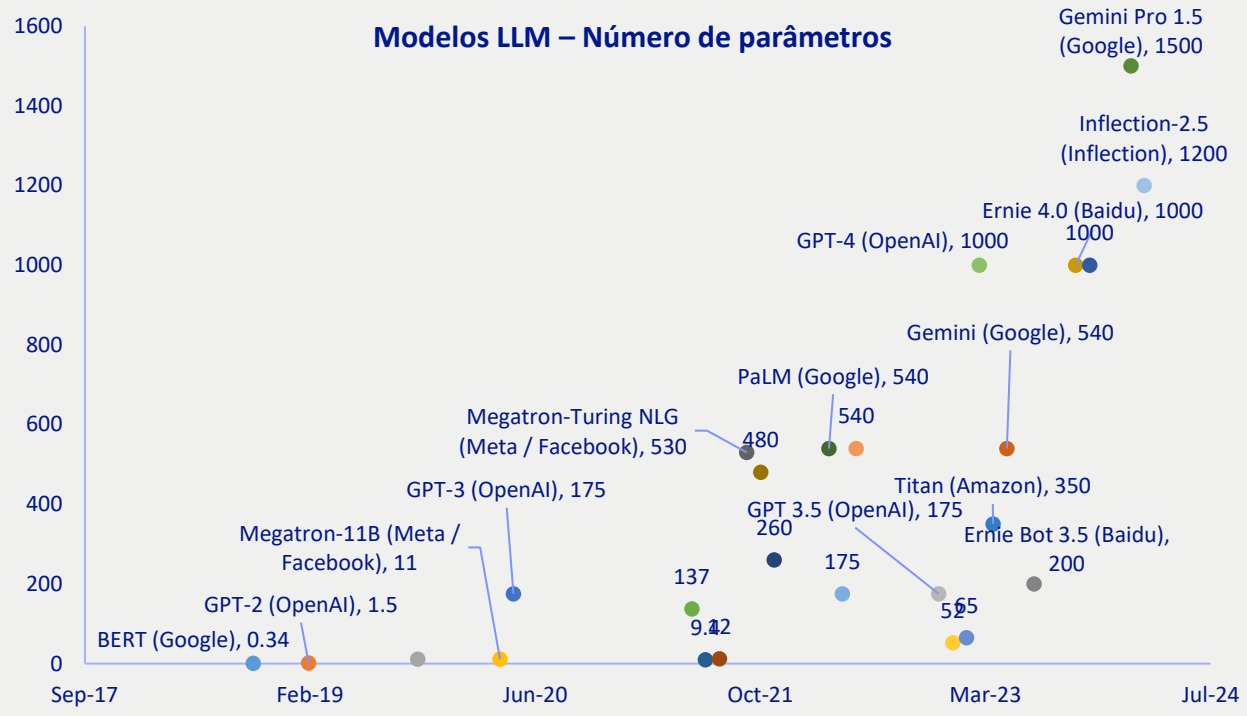
- *“Old code, new tricks: Most of the surprises concern the way models can learn to do things that they have not been shown how to do. Models learn to do a task—spot faces, translate sentences, avoid pedestrians—by training with a specific set of examples. Yet they can generalize, learning to do that task with examples they have not seen before. **Somehow, models do not just memorize patterns they have seen but come up with rules that let them apply those patterns to new cases**.”*

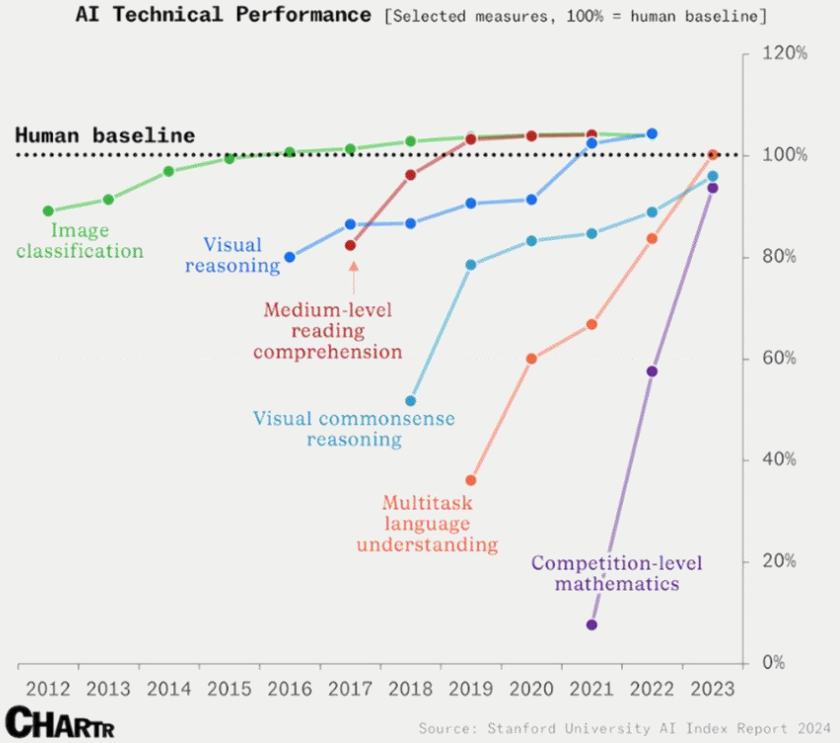
- *“The rapid advances in deep learning over the last 10-plus years **came more from trial and error than from understanding**. Researchers copied what worked for others and tacked on innovations of their own. (…) **And yet for all their success, the recipes are more alchemy than chemistry**.”*

Fonte: https://www.technologyreview.com/2024/03/04/1089403/large-language-models-amazing-but-nobody-knows-why

A “descoberta” da GenAI

- **Maior volume de dados e o maior número de parâmetros leva a modelos cada vez mais capazes**

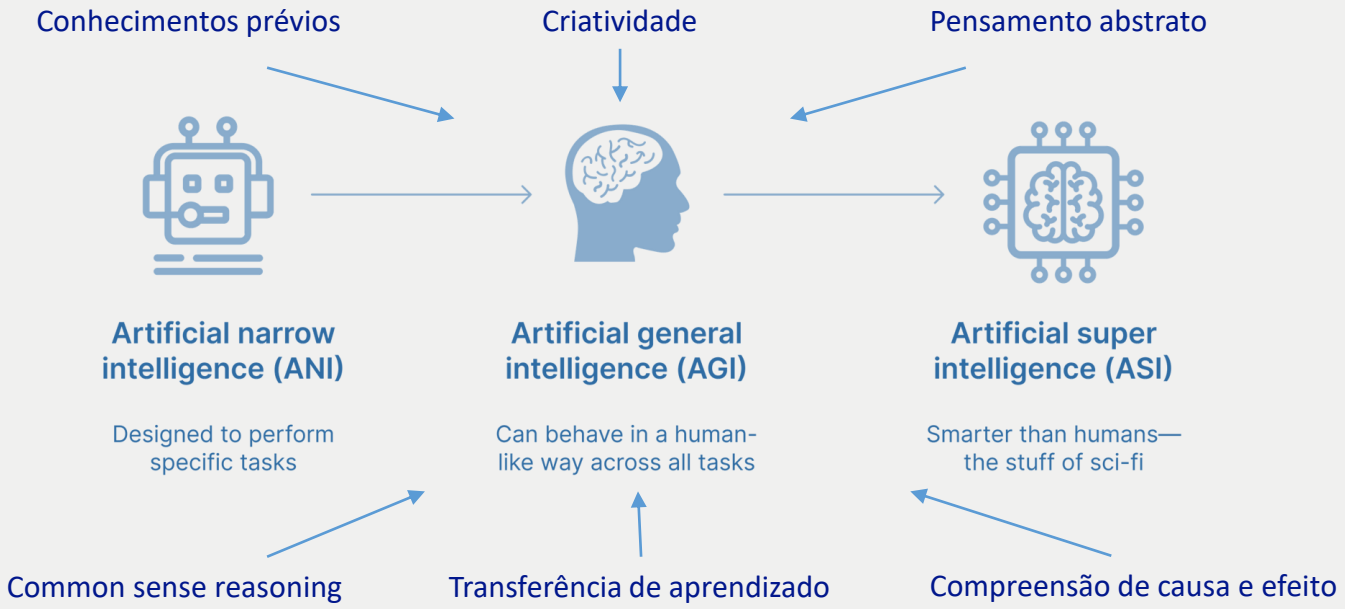




O que a GenAI atual é capaz (ou não) de fazer?

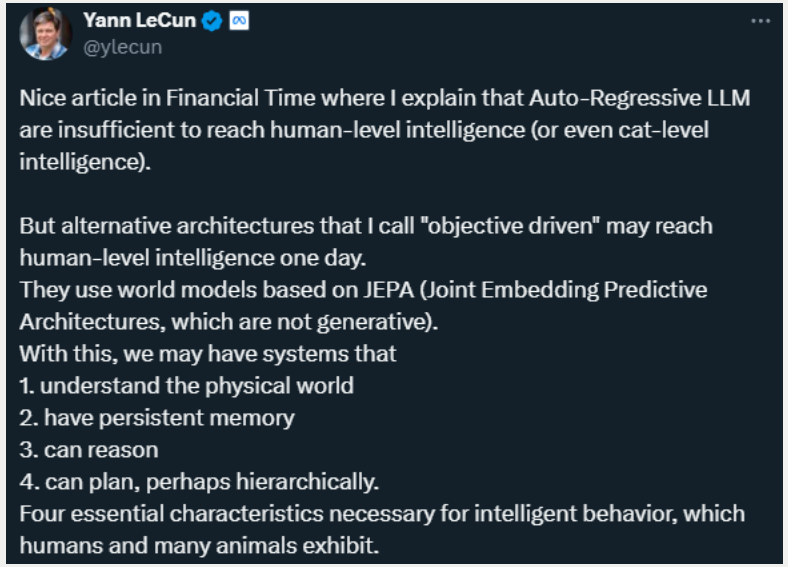
- Os modelos de GenAI mais poderosos já são capazes de criar textos, imagens, vídeos ou música; analisar o conteúdo de imagens e documentos; traduzir textos ou áudio; transcrever diálogos; analisar e gerar códigos de computador; e analisar bases de dados estruturadas, além de muitas outras funções

- Com a evolução exponencial de cada novo modelo ao longo dos últimos meses, muito se pergunta: até onde são capazes de chegar? **Seria a AGI (*artificial general intelligence*) alcançável com os modelos atuais?** Essa é uma pergunta com resposta ainda em aberto, mas para muitos pesquisadores a resposta é **não**



O que a GenAI atual é capaz (ou não) de fazer?

- Yann Lecun, um dos maiores pesquisadores do segmento no mundo e atualmente o Head de AI da Meta, argumenta que não é possível alcançar a AGI baseando-se nos modelos de LLM atuais. Dentre os principais motivos, ele acredita que:
  - LLMs não são capazes de gerar informações factuais, uma vez que as **"alucinações" seriam um problema insolucionável**. Como o modelo é probabilístico, e cada palavra gerada possui uma probabilidade de "erro", **a probabilidade acumulada de que um texto se mostre correto decresce exponencialmente de acordo com o tamanho do texto**
  - LLMs **são incapazes de (i) racionalizar; (ii) planejar; (iii) relembrar informações de forma persistente; e (iv) compreender o mundo físico**. Em comparação, os seres humanos conseguem (i) entender o funcionamento do mundo físico de maneira empírica; (ii) prever as consequências de nossas próprias ações; (iii) executar sequências de raciocínio ("chain of reasoning") com um conjunto ilimitado de passos; e (iv) planejar tarefas complexas decompondo-as em quantas subtarefas menores forem necessárias ("hierarchical planning")
  - **Sistema 1 vs Sistema 2**: *"The type of reasoning that takes place in LLM is very, very primitive, and the reason you can tell is primitive is because **the amount of computation that is spent per token produced is constant**. So, if you ask a question and that question has an answer in a given number of token, the amount of computation devoted to computing that answer can be exactly estimated. (…) And so essentially, **it doesn't matter if the question being asked is simple to answer, complicated to answer, impossible to answer because it's a decidable or something, the amount of computation the system will be able to devote to the answer is constant or is proportional to number of token produced in the answer**. This is not the way we work, the way we reason is that when we're faced with a complex problem or a complex question, we spend more time trying to solve it and answer it because it's more difficult."*



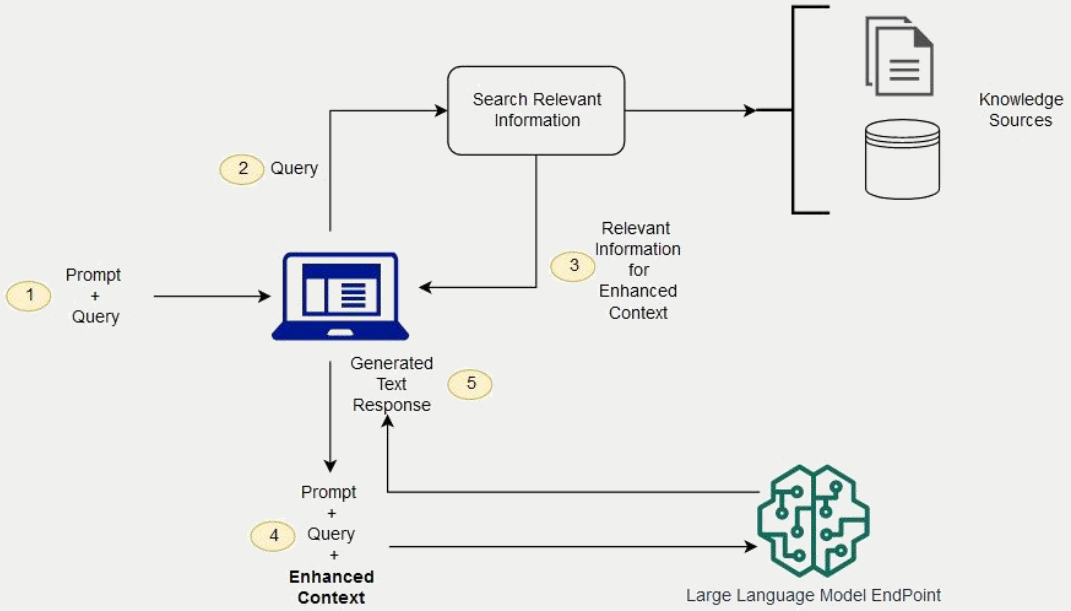
Workarounds

- A comunidade de pesquisadores de AI tem se dedicado a soluções que complementem as capacidades dos LLMs, de forma a mitigar algumas de suas limitações. Com isso, técnicas de *prompting* promissoras tem surgido nos últimos meses

Retrieval-Augmented Generation (RAG)

Ao invés dos modelos basearem-se apenas em suas bases de dados de treinamento, o RAG permite que acessem fontes externas, inclusive em tempo real, para buscar informações relevantes

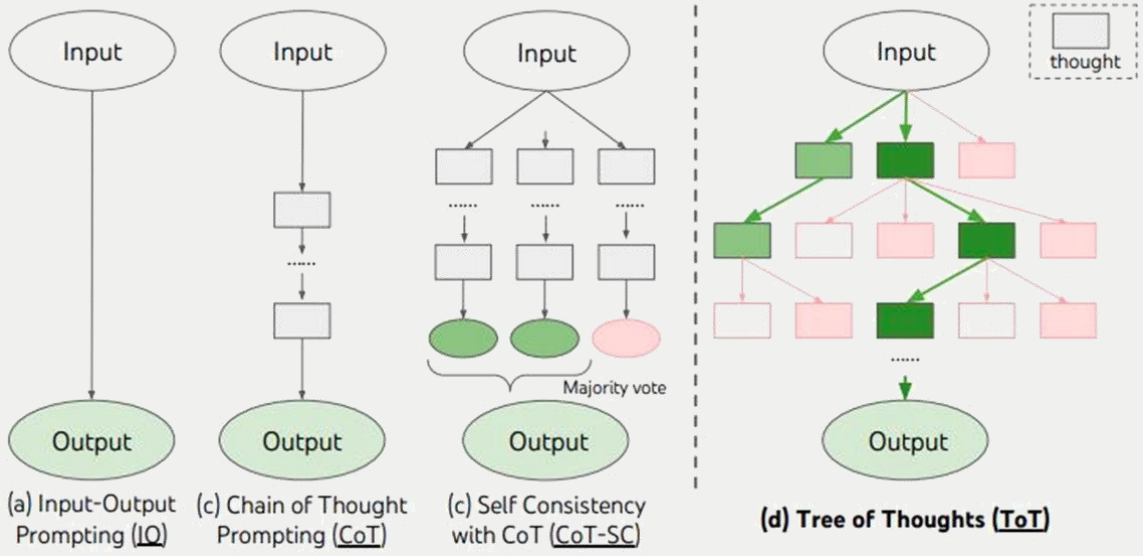
Funciona da seguinte forma: baseado no *prompt* do usuário, o RAG identifica o conjunto dos temas importantes contidos no texto; a partir daí, é realizada uma busca em fontes externas que sejam relevantes para aquele contexto, combinando essas fontes com o *prompt* original para produzir o resultado final



Tree-of-Thoughts

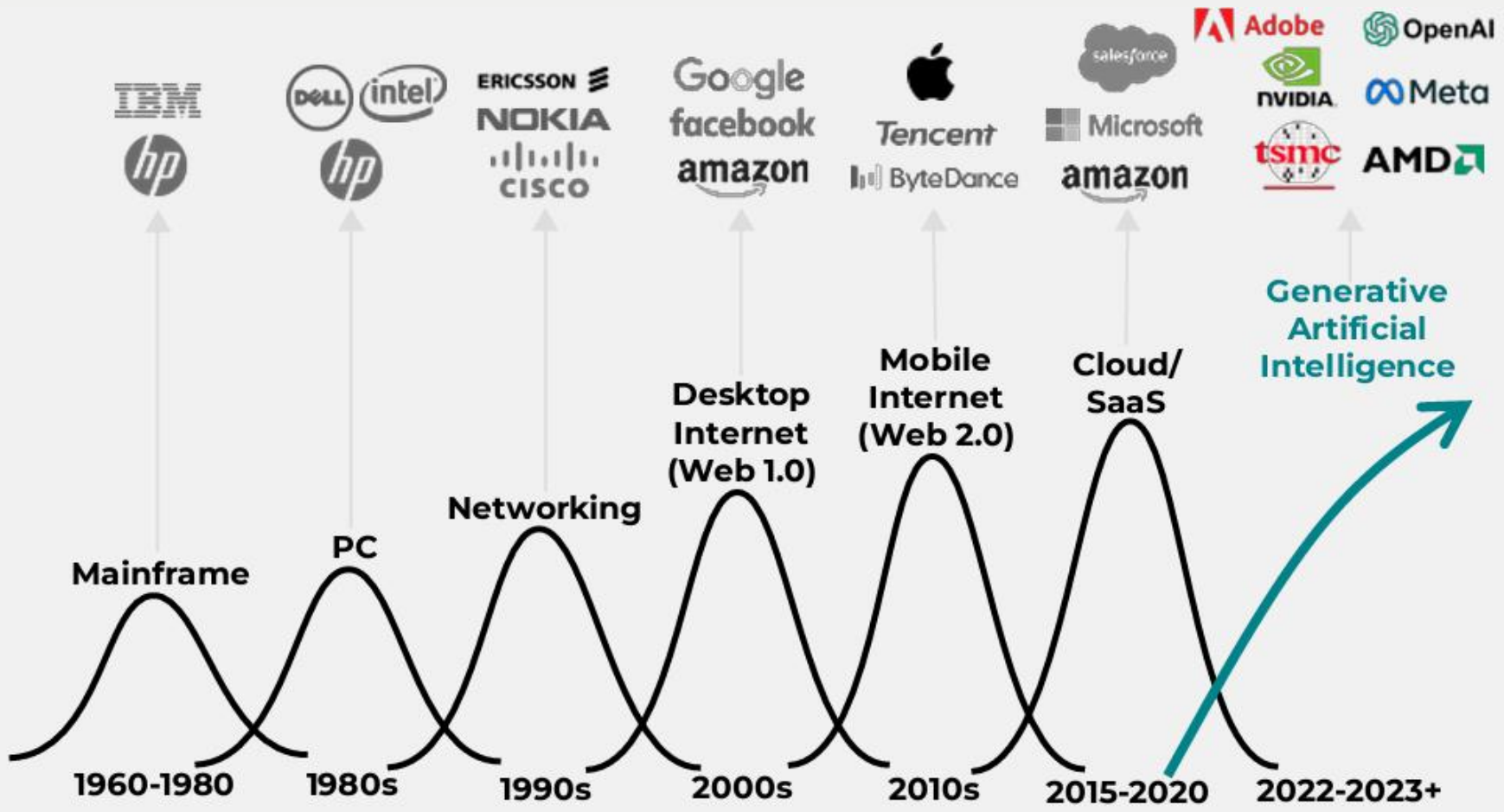
Assim como podemos testar diferentes abordagens ao resolver um problema complicado, o ToT orienta o LLM a considerar várias possibilidades e a se adaptar conforme necessário até chegar à solução mais eficaz

Funciona da seguinte foma: imagine que cada etapa de um processo de tomada de decisão, com suas diferentes ideias e possíveis soluções, seja representada como “ramos” em uma árvore decisória; à medida que o LLM avança através destes ramos, avalia constantemente o seu progresso e decide se continua no caminho atual ou explora um caminho diferente.




Impactos de AI na economia

Estamos em uma nova era tecnológica?



Líderes de tecnologia acreditam que sim

- **Microsoft**: *"AI adoption is like the PC when it became standard issue in early '90s, that's the closest analogy I can come up with."* (CEO Satya Nadella)

- **Alphabet**: *"The AI transition, I think it's a once-in-a-generation kind of opportunity"; AI "impacts the entire breadth of the company"; and "[it] is going to impact every product across every company."* (CEO Sundar Pichai)

- **Amazon**: *"I don't know if any of us have seen a possibility like this in technology in a really long time, for sure since the cloud, perhaps since the internet."* (CEO Andy Jassy)

- **Apple**: *"We believe in the transformative power and promise of AI" and "we see generative AI as a very key opportunity across our products."* (CEO Tim Cook)

- **Meta**: *"There are several ways to build a massive business here"; "we should invest significantly more over the coming years to build even more advanced models and the largest-scale AI services in the world…[We'll] grow our investment envelope meaningfully before we make much revenue from some of these new products."* (CEO Mark Zuckerberg)

Opinião parcial ou não, eles estão investindo de acordo com essa visão

O potencial de novos casos de uso é amplo

Artes e Design

- Criação de imagens, áudio e vídeo
- Designs de logos e marcas
- Projetos de arquitetura e composição de interiores
- Design de vídeo games e personagens virtuais

Geração de conteúdo

- Criação de códigos computacionais
- Criação de textos, artigos e blogs
- Criação e análise de documentos legais
- Peças de marketing e *advertising*
- Posts para *social media*

Chatbots e atendimento ao cliente

- Bots para atendimento ao consumidor
- Assistentes virtuais
- Agentes para reserva de mesas, hotéis, passagens aéreas
- Suporte a vendas e resolução de problemas

Educação e Treinamento

- Materiais personalizados
- "Explicadores" pessoais
- Provas e testes customizados
- Treinamento baseado em simulações
- Laboratórios virtuais

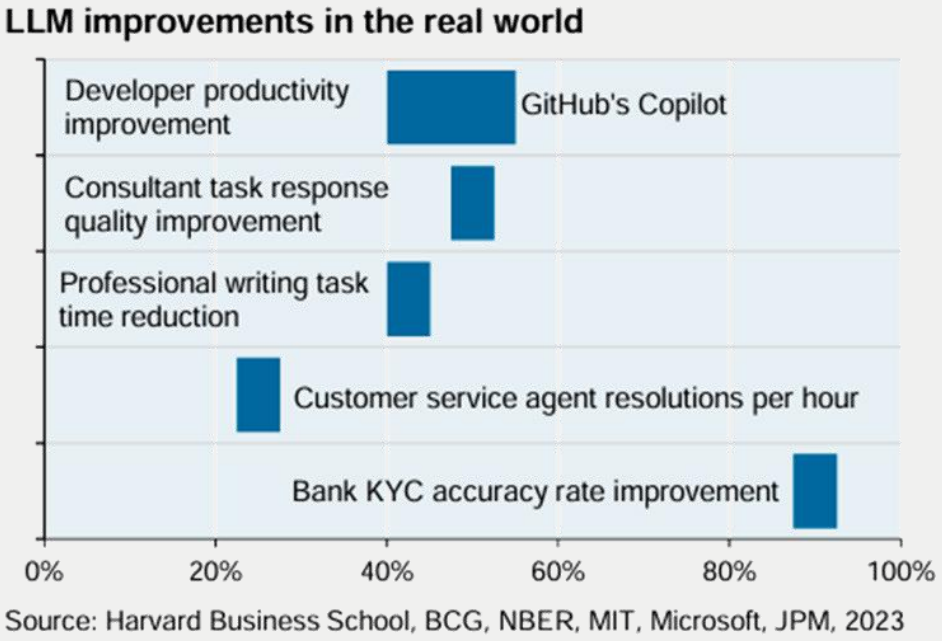
Healthcare

- Análise de imagens médicas
- Auxílio a diagnósticos
- Tratamentos personalizados
- Pesquisa de novos medicamentos

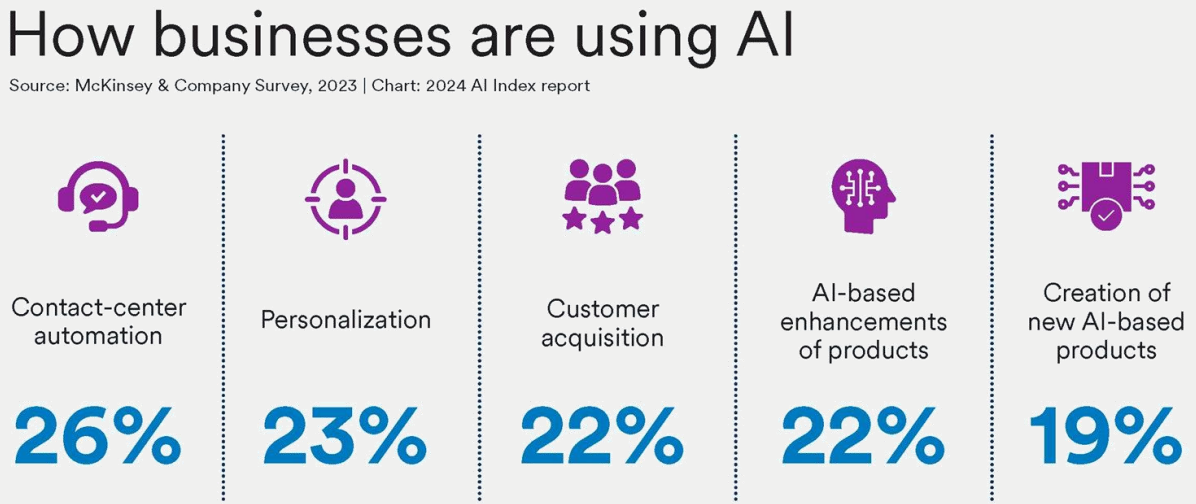
Indústria e Manufatura

- Design de projetos
- Pesquisa de materiais
- Otimização de processos
- Controle de qualidade

Mas apesar da *hype*, a adoção concreta ainda é limitada

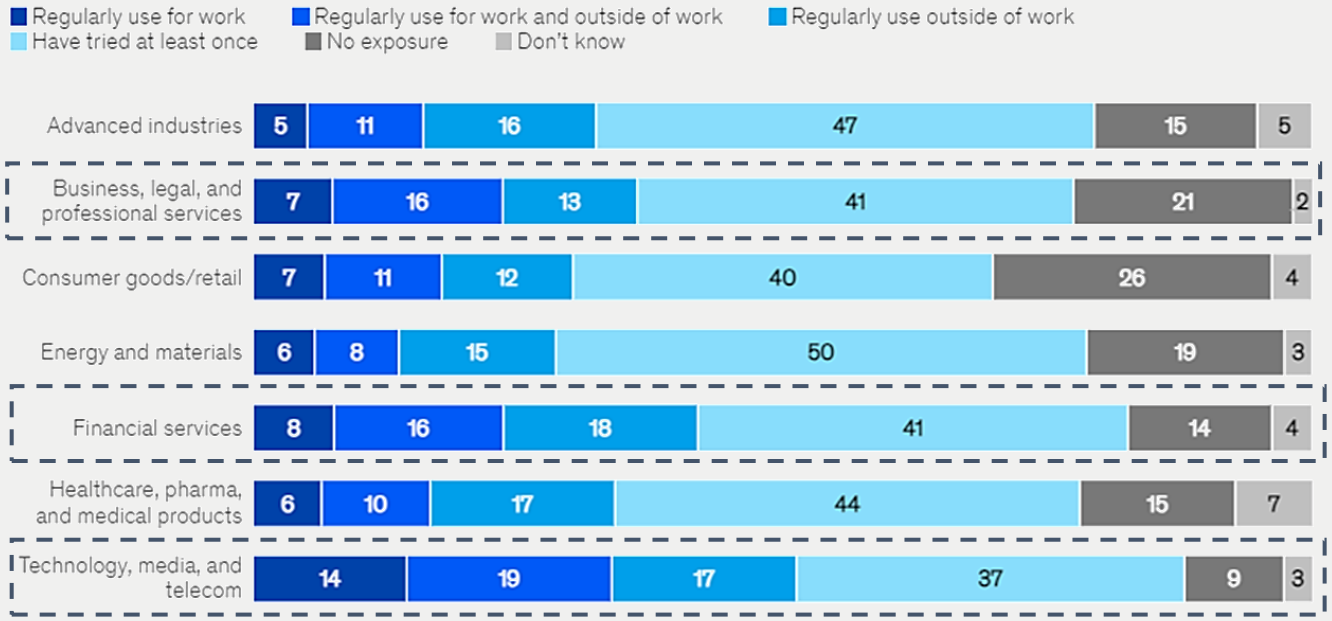


Source: Harvard Business School, BCG, NBER, MIT, Microsoft, JPM, 2023



"Knowledge professions" tem sido as mais impactadas

- As indústrias com maior taxa de adoção são TMT, Serviços Financeiros e Consultoria / Advocacia

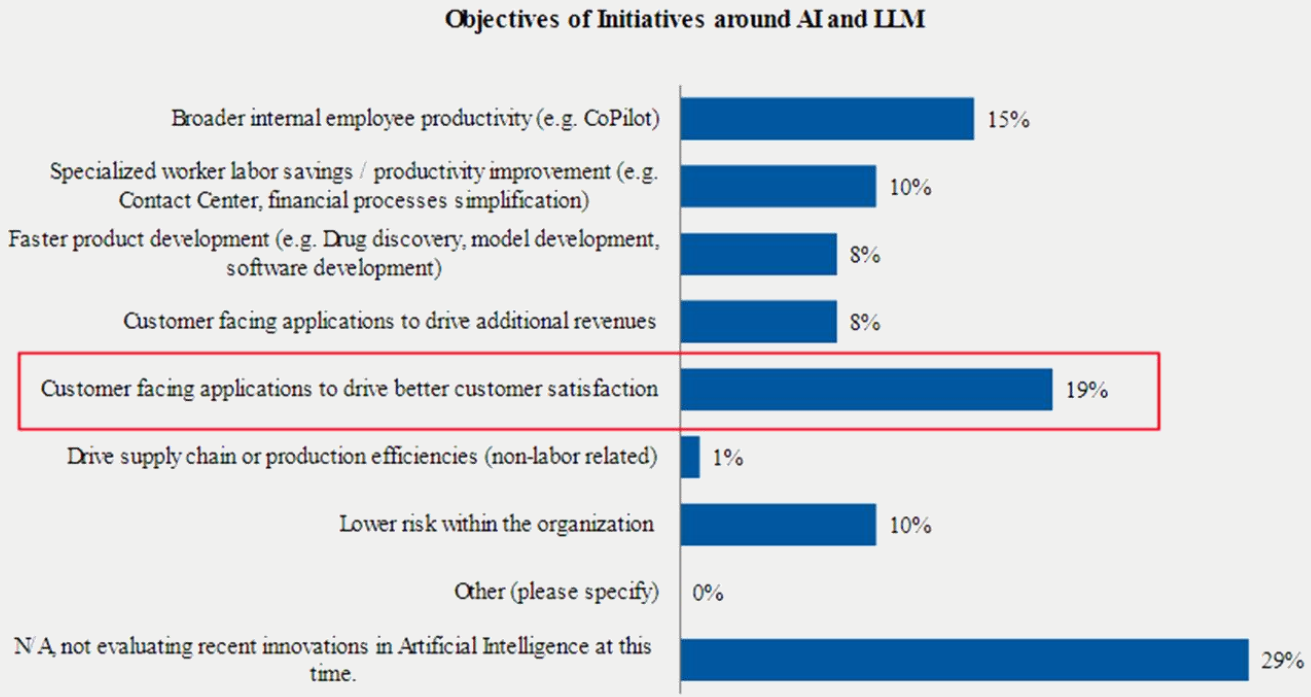


Sam Altman: *"If you asked the 10-year-old version of me, who used to spend a lot of time daydreaming about AI, what was going to happen, my pretty confident prediction would have been that first we're gonna have robots, and they're going to perform all physical labor. Then we're going to have systems that can do basic cognitive labor. A really long way after that, maybe we'll have systems that can do complex stuff like proving mathematical theorems. Finally we will have AI that can create new things and make art and write and do these deeply human things. That was a terrible prediction—it's going exactly the other direction."*

E atividades voltadas para o atendimento a clientes

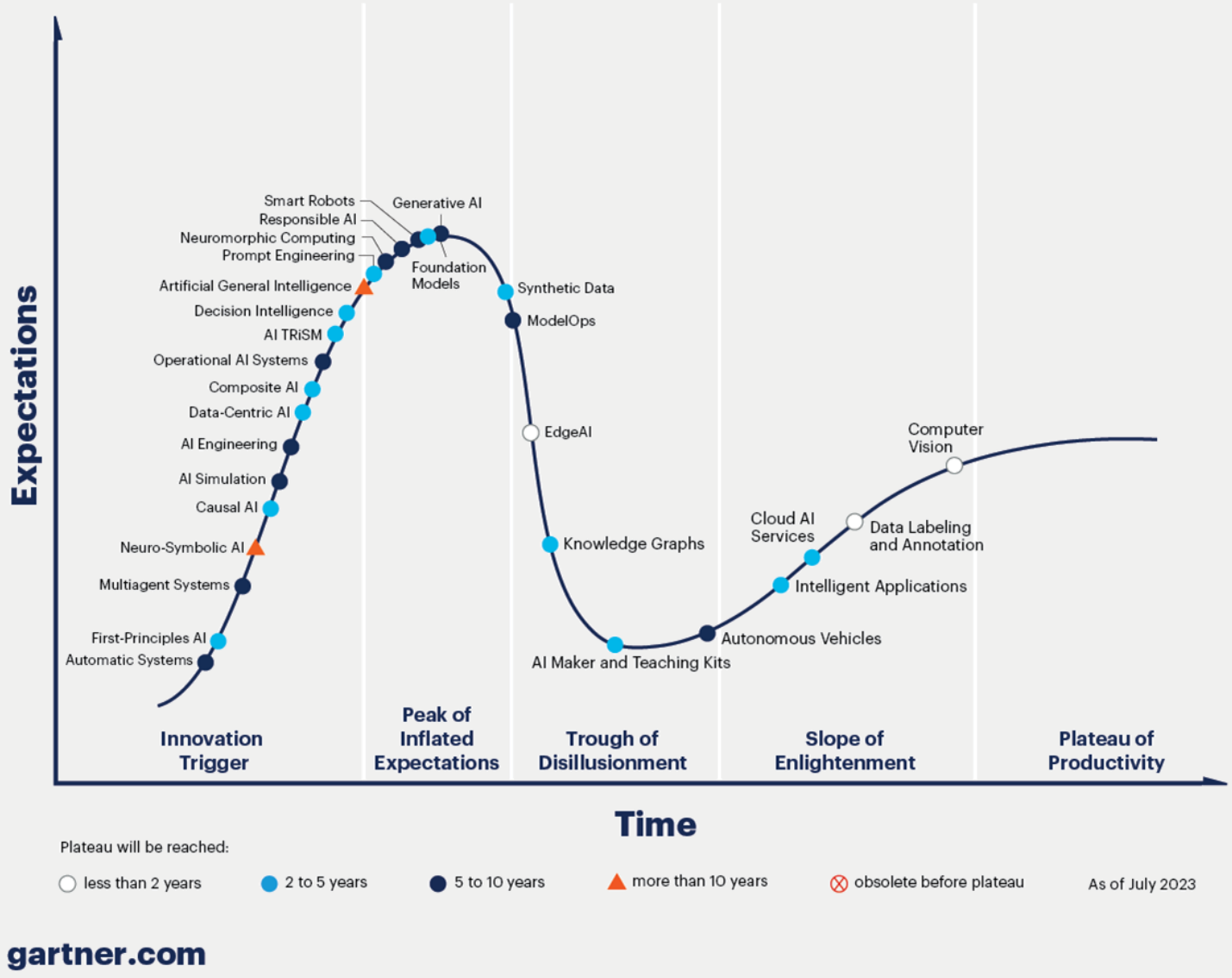
- Pesquisas com CIOs (*Chief Information Officers*) indicam que a maior parte dos investimentos tem sido direcionada a aplicações de relacionamento com o consumidor, além de ferramentas de ganho de produtividade (CoPilot)

Exhibit 6: Our 3Q23 CIO Survey Indicates that Front Office Use Case Is The Top Objective of AI/LLM Initiatives



Source: Morgan Stanley Research, AlphaWise, n=100 (US and EU data)

Onde estamos no ciclo?

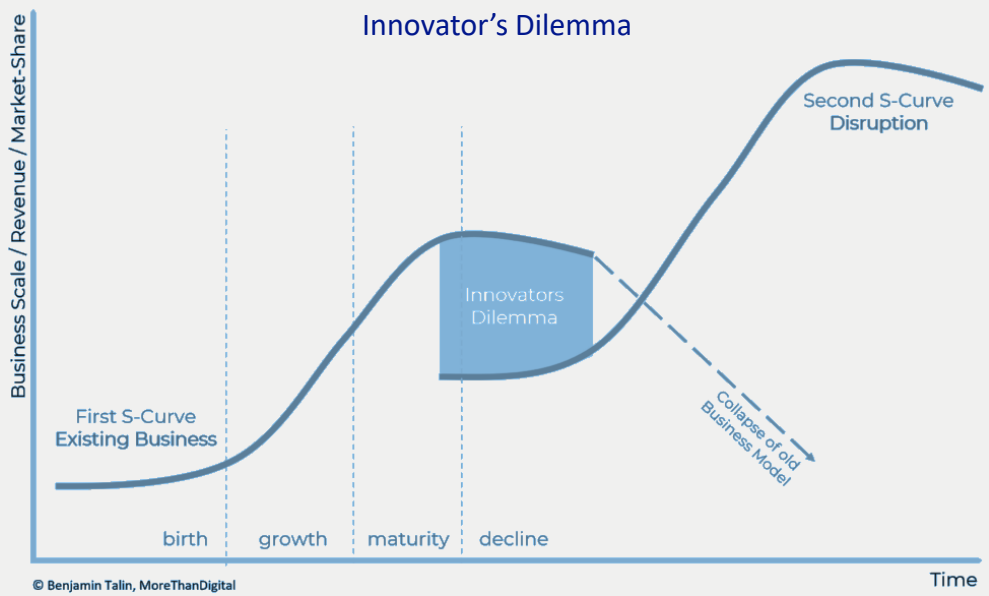


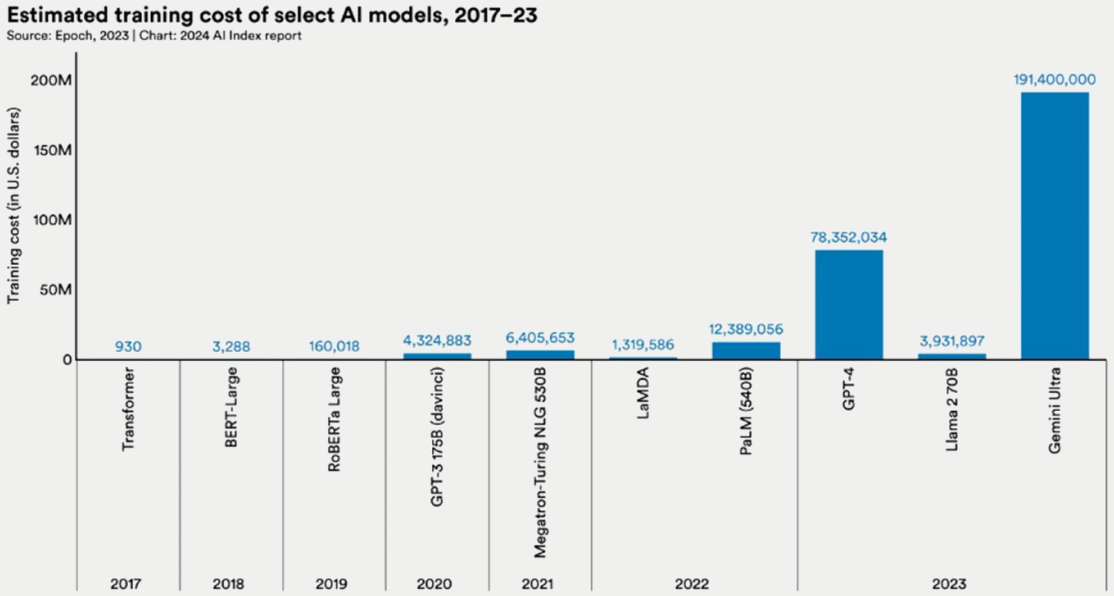
- Análise da Gartner coloca Generative AI no "pico das expectativas infladas"
- Estimam um prazo para atingimento do plateau de produtividade entre 5 e 10 anos
- AGI, apesar das expectativas cada vez maiores, ainda estaria a pelo menos 10 anos de ser atingida
- *Edge AI* já é uma realidade, mas ainda deve passar por um "vale de desilusão" antes de atingir maturidade
- Outras tecnologias adjacentes a AI, como *Computer Vision* e *Cloud AI* já estariam entrando nessa nova fase

- **Innovation Trigger**: A potential breakthrough, such as early proof-of-concept or media interest, generates publicity.
- **Peak of Inflated Expectations**: Early success stories, often accompanied by failures, lead to high expectations.
- **Trough of Disillusionment**: Interest wanes as experiments and implementations fail to meet expectations, and the technology's limitations become apparent.
- **Slope of Enlightenment**: The innovation's maturity increases, leading to a second rise in expectations and real value.
- **Plateau of Productivity**: The innovation is widely adopted and businesses can benefit from it.

Incumbentes ou *challengers*?

- De acordo com o "Innovator's Dilemma", do Clayton Christensen, os incumbentes tem dificuldade em lidar com inovações disruptivas. Em geral, costuma haver maior incentivo econômico na preservação de seus negócios legados, ainda que por tempo limitado, do que em colocar essa linha em risco para perseguir uma nova estratégia ainda incerta
- No caso de AI, no entanto, essa dinâmica não parece válida. Apesar de as GPUs ficarem cada vez mais eficientes, o tamanho dos modelos avança de forma ainda mais acelerada. Com isso, o custo para treinamento e inferência dos LLMs também cresce exponencialmente. Necessidade de escala e capacidade de investimento favorecem os incumbentes, em especial os *cloud providers*, que diluem os custos em ativo fixo por meio do aluguel da infraestrutura para terceiros
- Além disso, acesso a dados e a um canal de distribuição robusto se tornam diferenciais competitivos ainda mais relevantes, permitindo a criação de soluções customizadas, com maior valor agregado, e com custo de distribuição praticamente zero

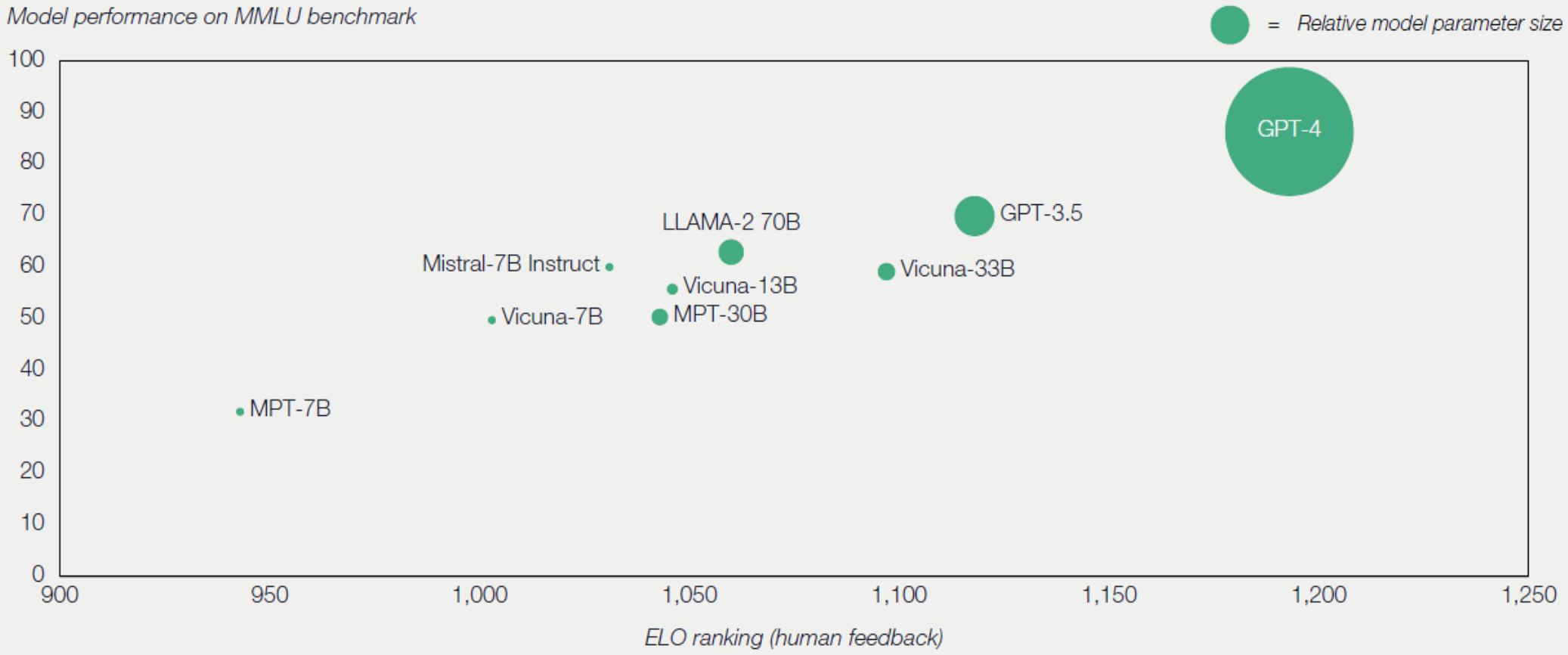




Na corrida da AI, ninguém quer ficar para trás

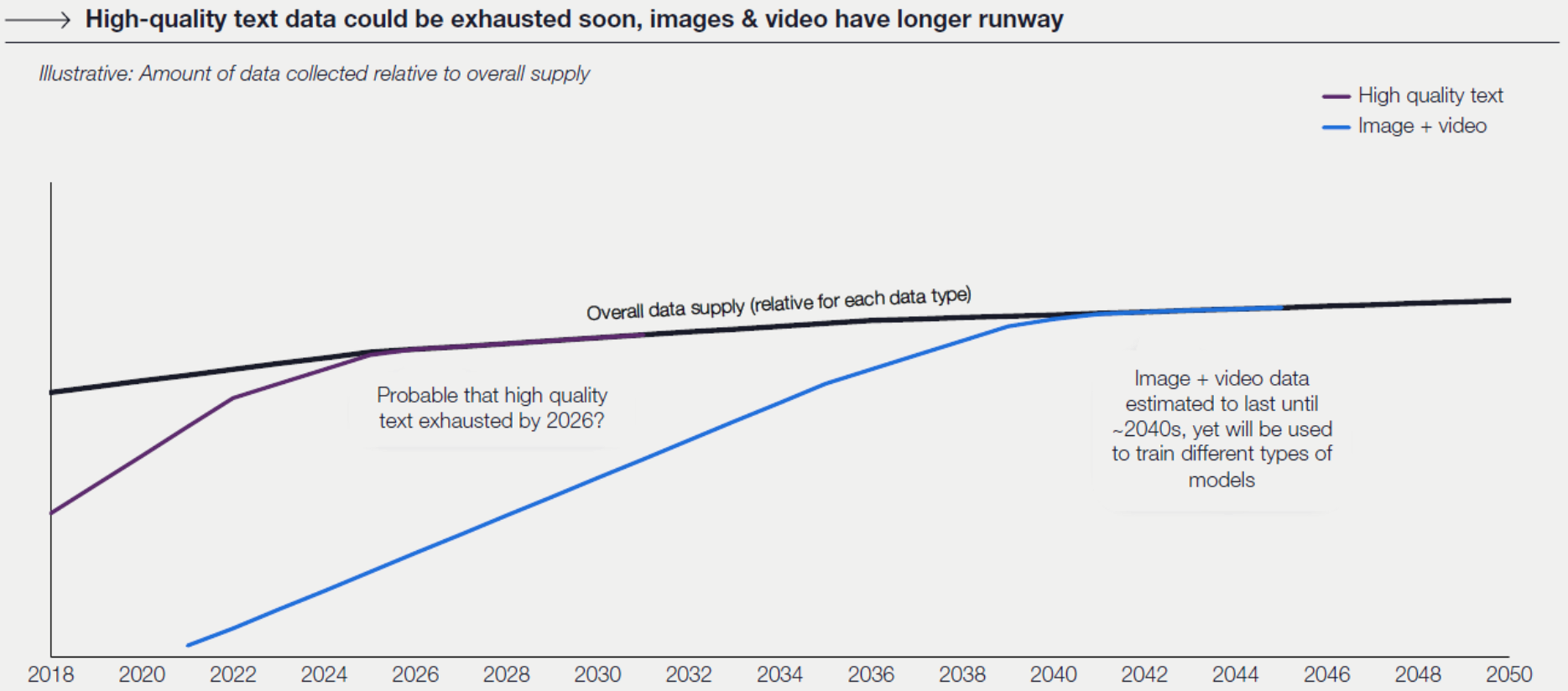
- **Apesar das incertezas em relação à real demanda, os incentivos são grandes para a manutenção de investimentos massivos**, especialmente pelas *big techs*
  - Os modelos tem apresentado melhoras constantes conforme aumenta o tamanho da base de dados e o número de parâmetros de treinamento. **Enquanto essa dinâmica persistir, a busca pelo próximo grande *breakthrough* deverá continuar**
  - **Essa é uma corrida que ninguém sabe onde fica a linha de chegada** ("Quão melhores os LLMs podem ficar?"; "É possível atingir AGI?"). Nesse caso, **a disputa deixa de ser em relação a uma métrica tangível e passa a ser relativa**: enquanto os concorrentes estiverem investindo, o incentivo é de investir mais
  - As grandes empresas de tecnologia possuem geração de caixa suficiente para "pagar para ver". Afinal, estão sob risco tanto o *business* legado quanto a nova onda tecnológica. **O custo potencial de ficar para trás é maior do que o custo de investir em excesso**
- "*This idea of standing up a data center instantaneously is so valuable and getting this thing called time to train is so valuable. The reason for that is because* ***the next company who reaches the next major plateau gets to announce a groundbreaking AI. And the second one after that gets to announce something that's 0.3% better. And so the question is, do you want to be repeatedly the company delivering groundbreaking AI or the company delivering 0.3% better?*** *And that's the reason why this race, as in all technology races, the race is so important. And you're seeing this race across multiple companies because this is so vital to have technology leadership*" – Jensen Huang
- "*Expectations about the future benefits of AI don't need to be universally held to induce tremendous investment, so long as some of the key actors believe them. In the near term, those key actors are likely to be large corporations (particularly in tech) and some of the world's large pools of capital*"– Bridgwater

Maior volume de dados / parâmetros melhora o modelo



O que pode parar essa tendência?

- Modelos cada vez maiores devem absorver todo o conteúdo já produzido pela humanidade ao longo dos próximos anos
- A partir desse ponto, a maior parte dos novos dados usados para treinamento deverá vir de conteúdos gerados pela própria GenAI, o que pode levar a uma deterioração relevante na qualidade



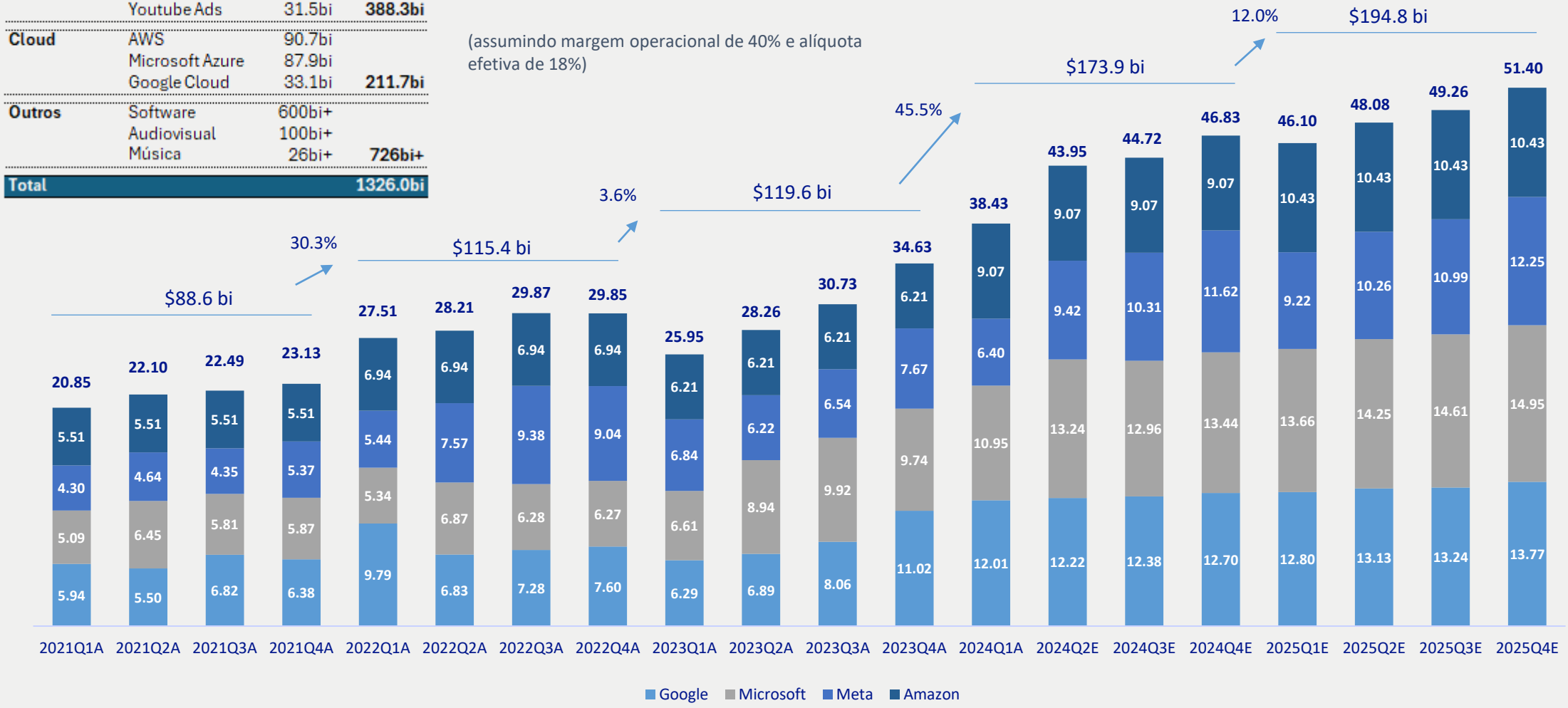
Estimativa da Capex das *Hyperscalers*

Referências de mercados potenciais

Ads	Google Search	175.0bi	
	Meta	134.9bi	
	Amazon Ads	46.9bi	
	Youtube Ads	31.5bi	**388.3bi**
Cloud	AWS	90.7bi	
	Microsoft Azure	87.9bi	
	Google Cloud	33.1bi	**211.7bi**
Outros	Software	600bi+	
	Audiovisual	100bi+	
	Música	26bi+	**726bi+**
Total			**1326.0bi**

Para cada **$100 bi em capex** serão necessários cerca **$60 bi em receita adicional** para um ROIC marginal de 20%. Otimismo?

(assumindo margem operacional de 40% e alíquota efetiva de 18%)

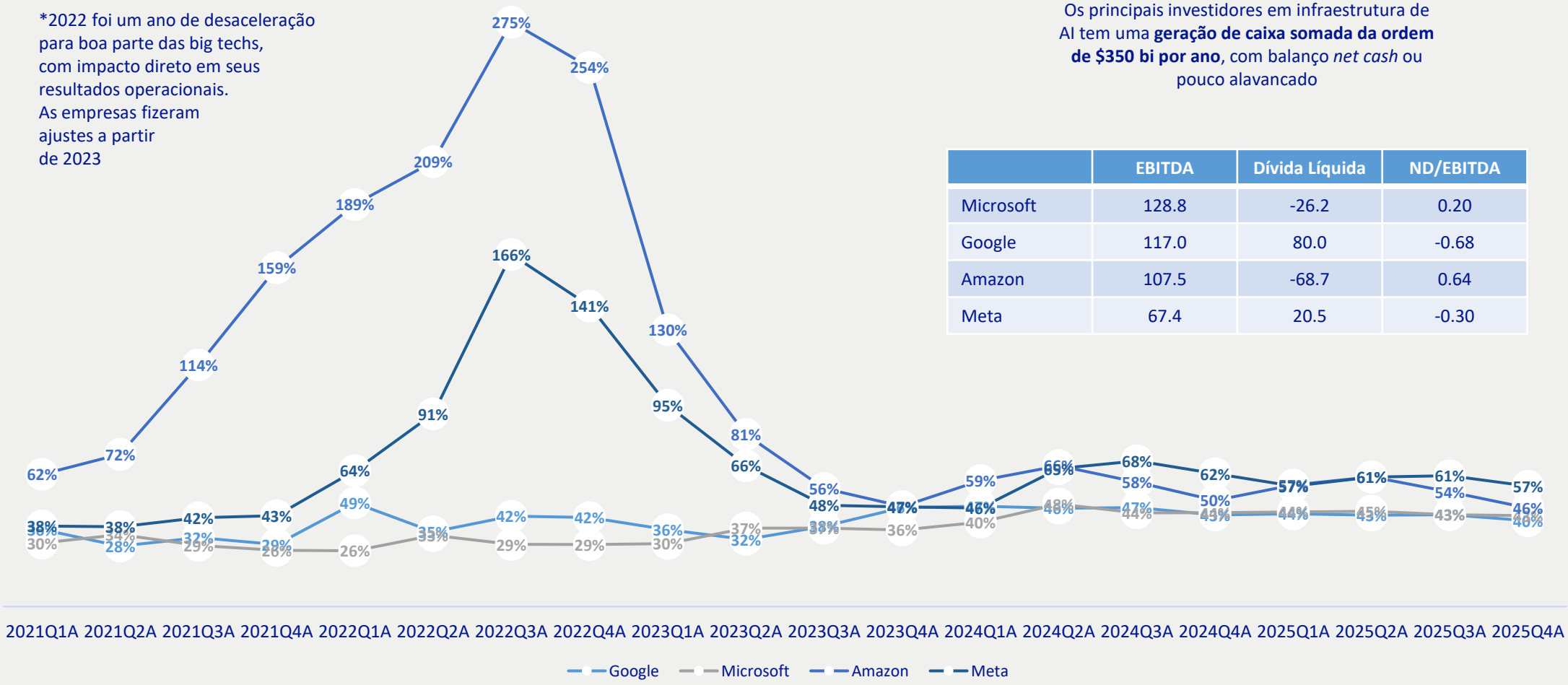


Fonte: Bloomberg (Jun-24)

Balanço e *cashflow* para investir não são um problema

Capex % Operating Income

*2022 foi um ano de desaceleração para boa parte das big techs, com impacto direto em seus resultados operacionais. As empresas fizeram ajustes a partir de 2023



Os principais investidores em infraestrutura de AI tem uma **geração de caixa somada da ordem de $350 bi por ano**, com balanço *net cash* ou pouco alavancado

	EBITDA	Dívida Líquida	ND/EBITDA
Microsoft	128.8	-26.2	0.20
Google	117.0	80.0	-0.68
Amazon	107.5	-68.7	0.64
Meta	67.4	20.5	-0.30

mar asset management

Dados financeiros

Empresa	Mkt Cap	ND/ EBITDA	P/E 2024	P/E 2025	P/FCF 2024	P/FCF 2025	Rev growth CAGR L5Y	Rev growth CAGR 2024	Rev growth CAGR 2025	Free Cash Flow 2023	Free Cash Flow 2024	Free Cash Flow 2025	FCF margin 2021	FCF margin 2022	FCF margin 2023	5yr ROE	5yr ROIC	TSR 1Y	TSR 5Y
Microsoft	3,334,285	0.17 x	37.95	33.51	47.78	42.27	13.9%	14.6%	14.9%	59,475	69,779	78,888	28.1%	28.7%	28.2%	42.0%	26.7%	30.3%	242.1%
Google	2,200,081	-0.49 x	23.31	20.72	27.25	23.84	17.6%	12.7%	11.9%	69,495	80,742	92,281	22.6%	23.3%	24.0%	23.4%	20.1%	43.3%	220.5%
Amazon	1,916,232	0.44 x	36.97	27.91	31.17	24.53	19.8%	11.1%	11.3%	32,217	61,470	78,108	5.6%	9.6%	11.0%	18.6%	8.9%	46.4%	92.4%
Meta	1,291,898	-0.20 x	24.41	21.44	29.37	25.76	19.3%	17.8%	15.2%	43,847	43,990	50,155	32.5%	27.7%	28.0%	26.1%	21.1%	79.8%	169.5%
Apple	3,352,337	-0.41 x	33.06	30.88	32.38	29.16	7.6%	1.0%	3.7%	99,584	103,521	114,983	26.0%	26.7%	27.9%	124.8%	39.3%	15.5%	345.1%
Oracle	386,331	2.50 x	24.53	21.78	32.88	41.48	6.0%	8.4%	9.5%	8,470	11,748	9,313	16.0%	20.5%	14.7%	n/a	15.2%	11.7%	184.1%
Alibaba	179,568	-3.08 x	8.78	9.40	10.03	9.15	20.1%	(0.2%)	4.1%	21,176	17,897	19,619	17.6%	14.9%	15.0%	13.0%	6.0%	(17.7%)	n/a
Baidu	32,557	-2.20 x	8.94	8.48	12.08	9.58	5.6%	3.7%	5.2%	3,255	2,695	3,397	18.9%	15.1%	17.8%	7.4%	3.5%	(36.9%)	n/a
Tencent	459,104	-0.83 x	18.17	16.17	19.42	15.94	14.3%	9.0%	9.5%	25,795	23,637	28,806	33.1%	27.8%	30.8%	16.7%	14.5%	5.5%	21.8%

Fonte: Bloomberg (17/06/24)

O potencial é grande, mas como monetizar?

Aplicações voltadas para os usuários finais são hoje as de maior incerteza

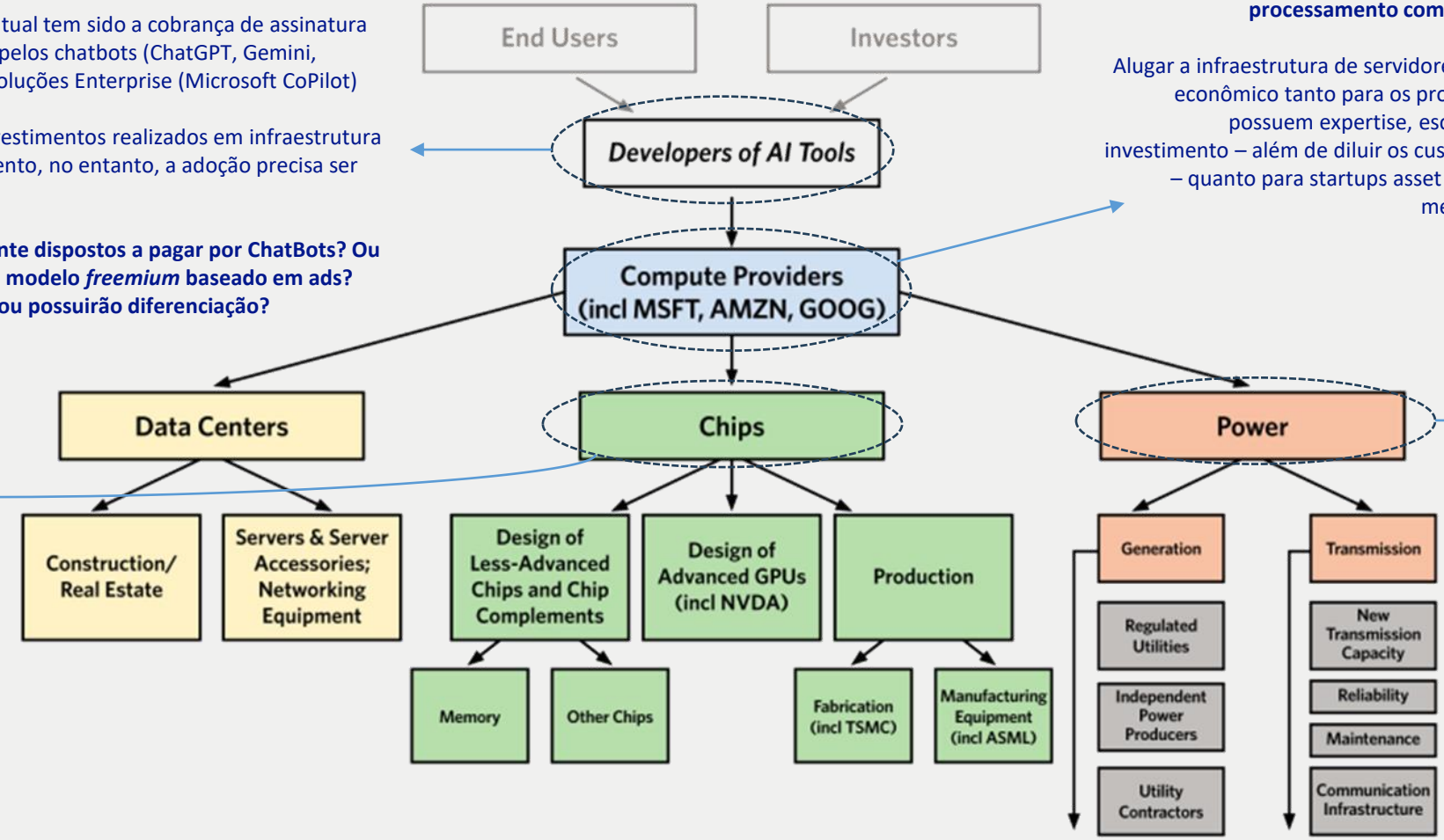
A forma de monetização atual tem sido a cobrança de assinatura mensal dos usuários, seja pelos chatbots (ChatGPT, Gemini, Perplexity, etc), seja por soluções Enterprise (Microsoft CoPilot)

Para justificar todos os investimentos realizados em infraestrutura e em custo de processamento, no entanto, a adoção precisa ser massiva

Os usuários estão realmente dispostos a pagar por ChatBots? Ou deverá caminhar para um modelo *freemium* baseado em ads? LLMs serão commodities ou possuirão diferenciação?

Até aqui, os desenvolvedores de chips vem se mostrando os grandes vencedores dessa onda de AI

Com o choque de demanda repentino, associado a um capacidade restrita, as empresas líderes tem crescido faturamento e margens de forma exponencial



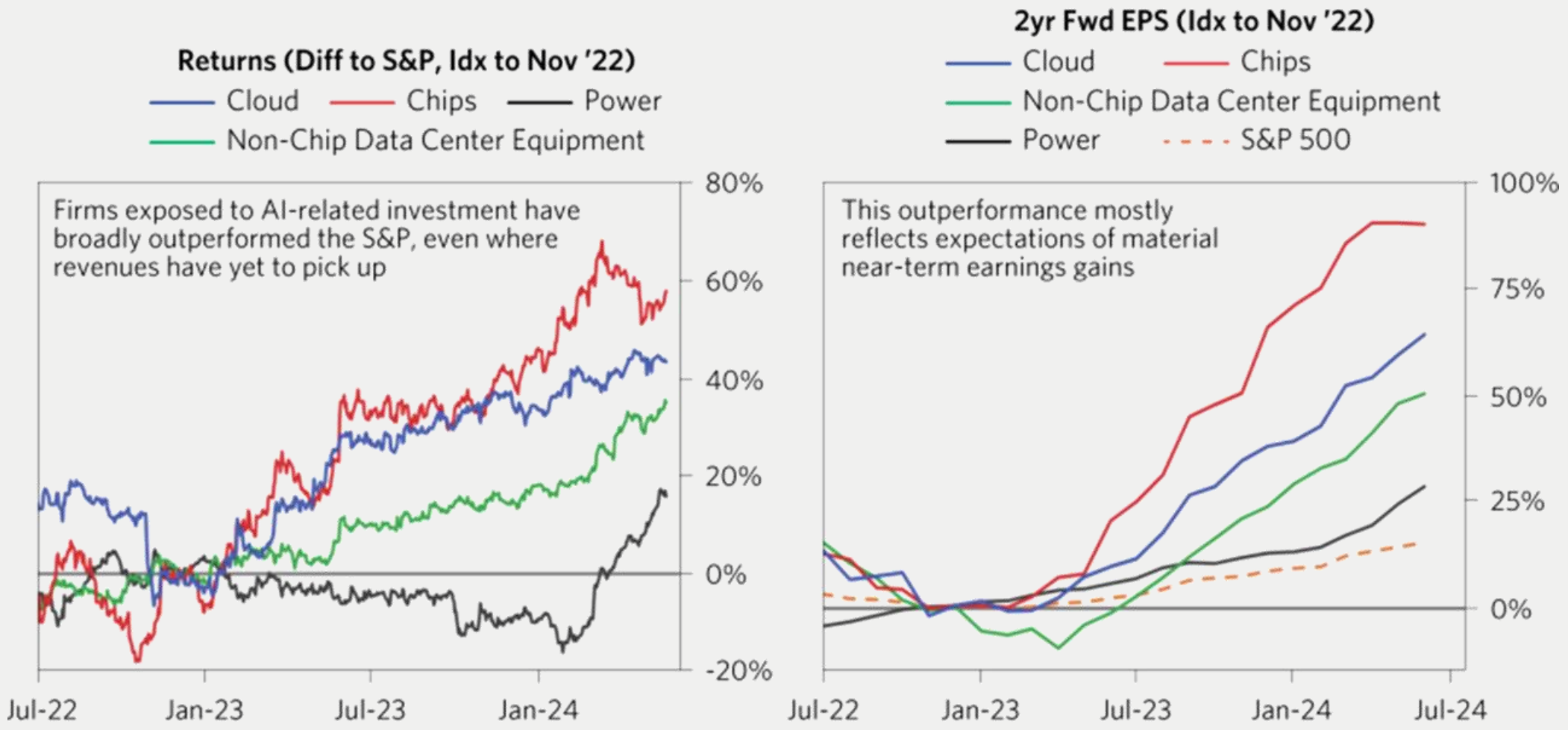
A avenida de monetização pelos *cloud providers* é mais direta: **com o avanço dos aplicações baseadas em AI a demanda por processamento computacional deverá acelerar**

Alugar a infraestrutura de servidores, CPUs e GPUs faz sentido econômico tanto para os provedores desse serviço, que possuem expertise, escala e capital para realizar o investimento – além de diluir os custos para seus usos próprios – quanto para startups asset light buscando escalar com menor necessidade de capital

Data centers para AI possuem demanda energética elevada, o que tem levado a revisões crescentes na expectativa por oferta de geração e transmissão de energia

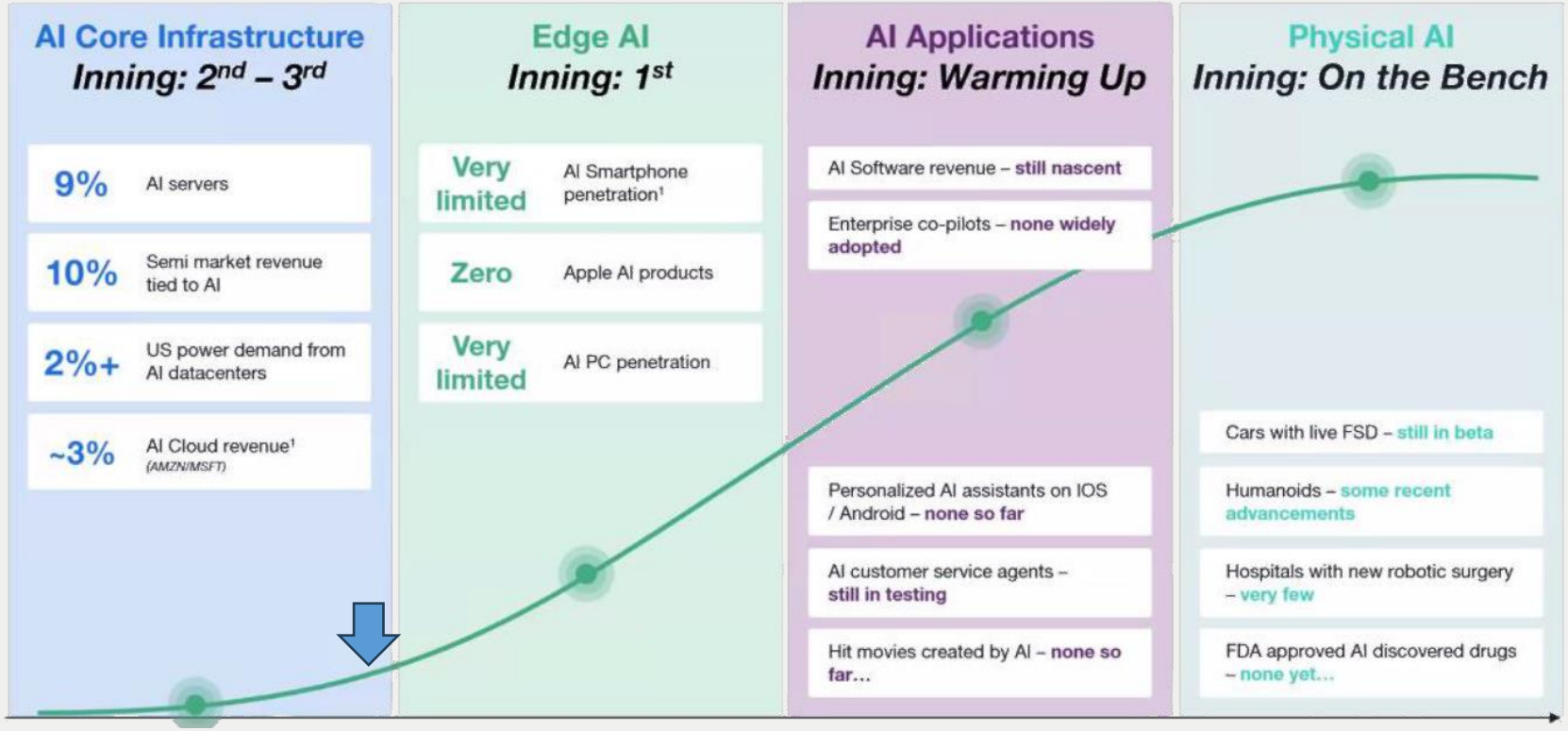
Depois de quase uma década de demanda de energia estagnada nos EUA, as projeções voltaram a mostrar crescimento

Na corrida do ouro, quem ganha é o vendedor de pá



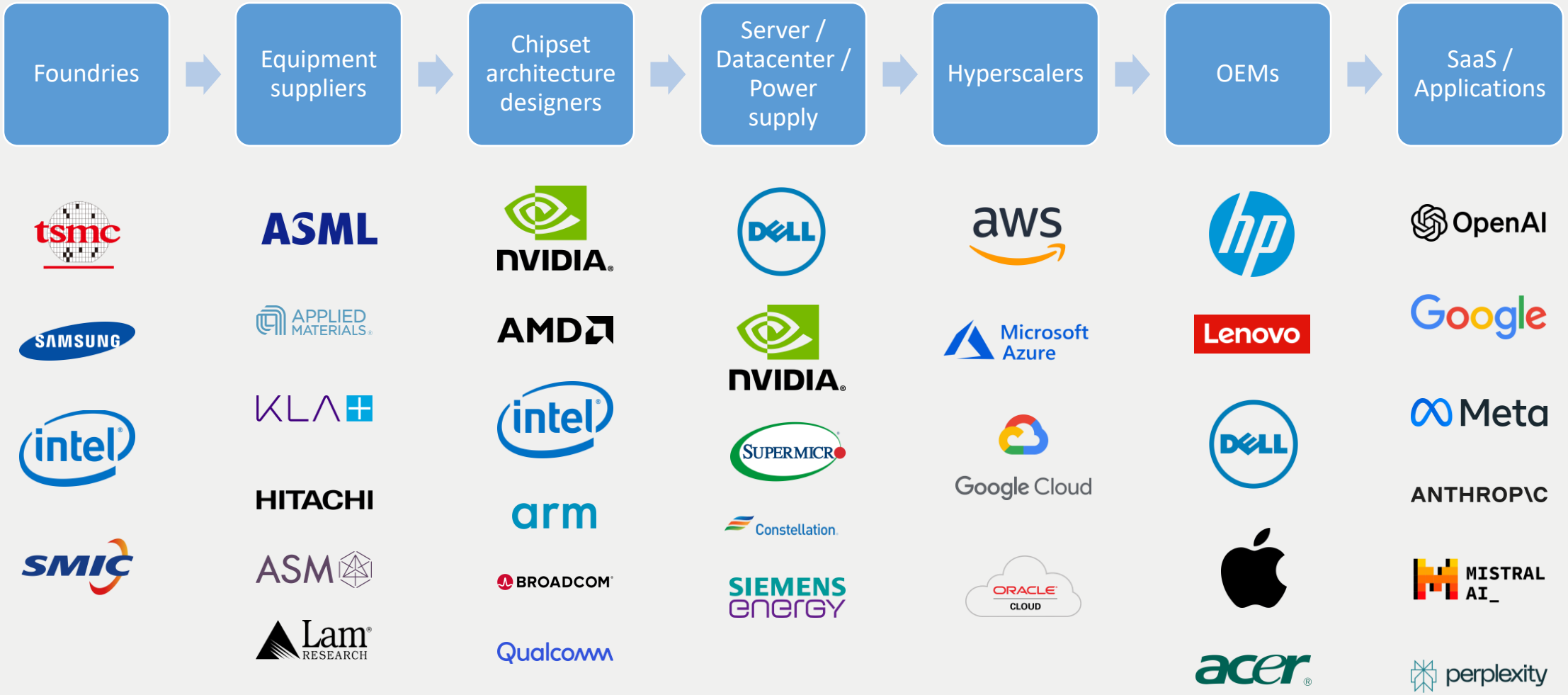
Ainda estamos no começo da jornada

- Análise do Coatue indica que estamos entre as fases 1 e 2 da curva de adoção



mar asset management

Principais *Players*

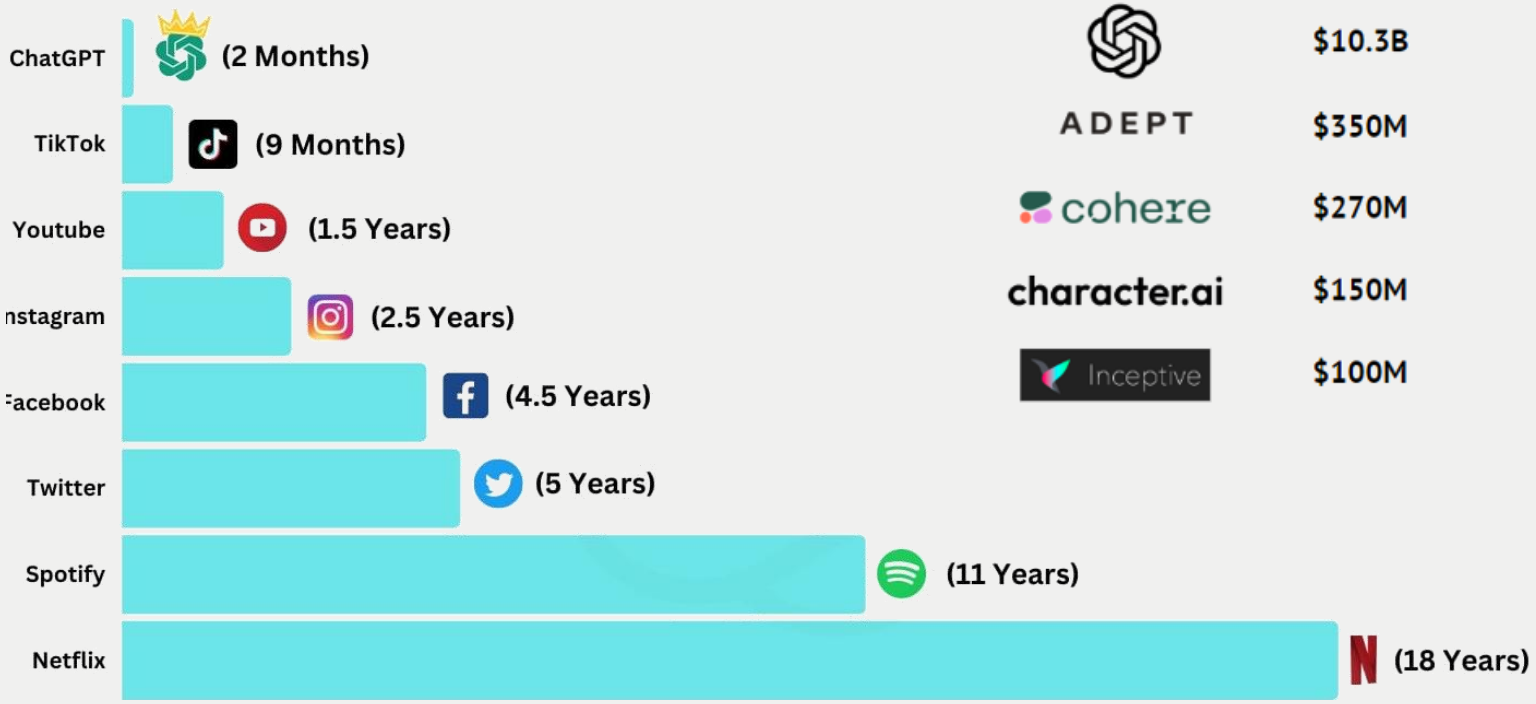



Consumer Applications

Chatbots são o *killer app* de AI?

- O ChatGPT foi o aplicativo mais rápido a atingir 100 milhões de usuários, chegando à marca em apenas 2 meses
- Desde então, diversos novos chatbots surgiram no mercado, atraindo mais de $25 bilhões em investimento de *venture capital* e de empresas de tecnologia apenas em 2023

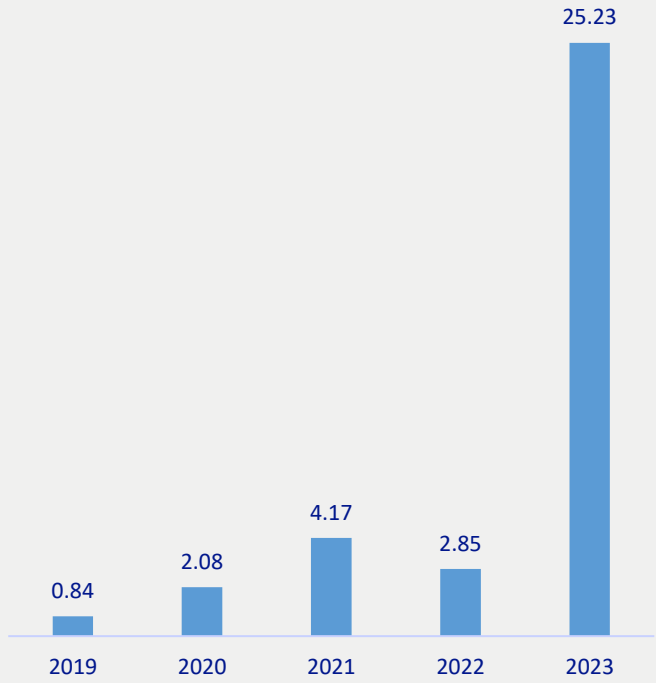
Tempo para atingir 100 milhões de usuários



Capital raised in 2023 alone

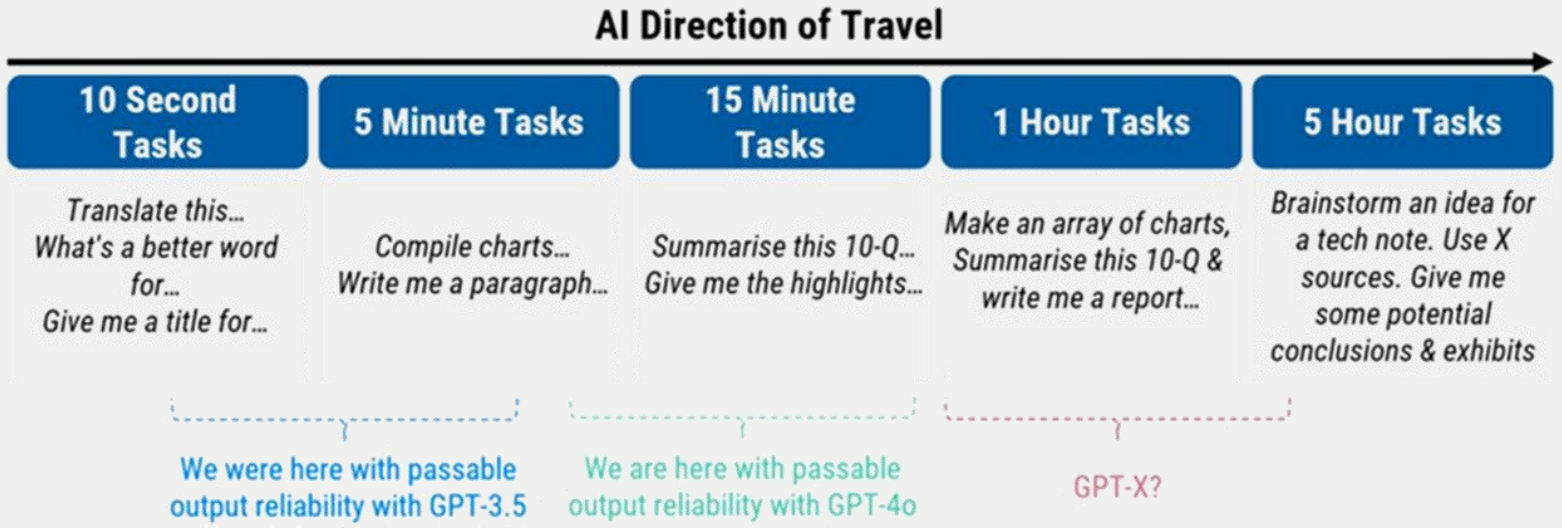
	$10.3B
ADEPT	$350M
cohere	$270M
character.ai	$150M
Inceptive	$100M

Investimentos privados em Generative AI



Como monetizar?

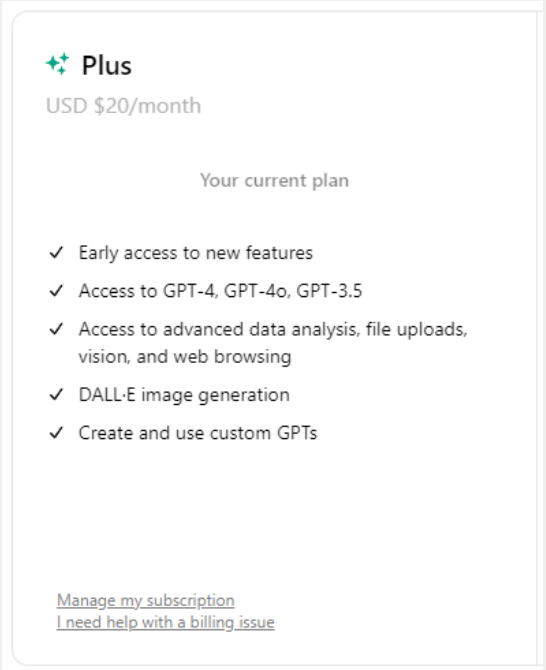
- Até aqui, o modelo adotado tem sido o *freemium*: uma versão gratuita com capacidades limitadas associada a uma versão paga completa
- Ainda estamos em estágios bastante iniciais tanto da capacidade desses produtos quanto de adoção, mas ficam as perguntas:
  - Os usuários estão realmente dispostos a pagar por chatbots?
  - LLMs serão commodities ou possuirão diferenciação?
  - Conseguirão escapar do modelo baseado em Ads?
- **A grande aposta é que os modelos ficarão cada vez mais capazes ao longo do tempo**



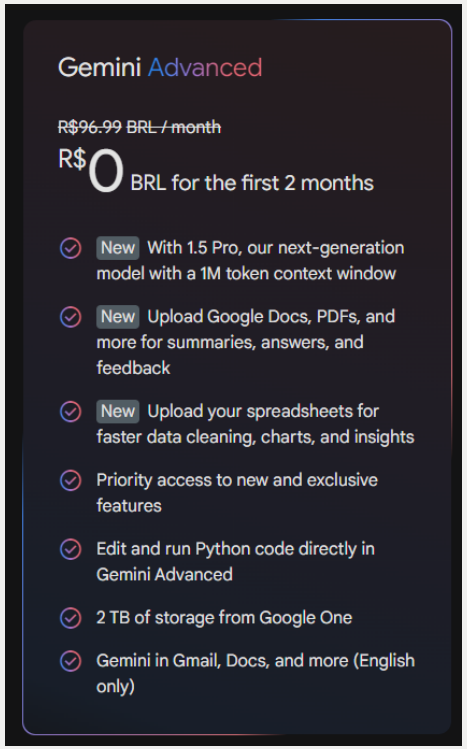
Precificação

- Precificação para uso pessoal em $20 por mês e para uso profissional em $25-30 por mês por usuário
- Deverá haver uma tendência ao *bundling*, especialmente nas licenças empresariais

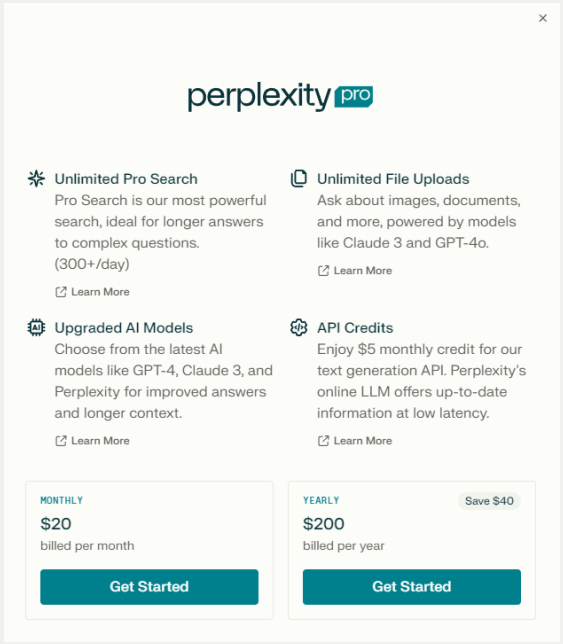
ChatGPT



Gemini



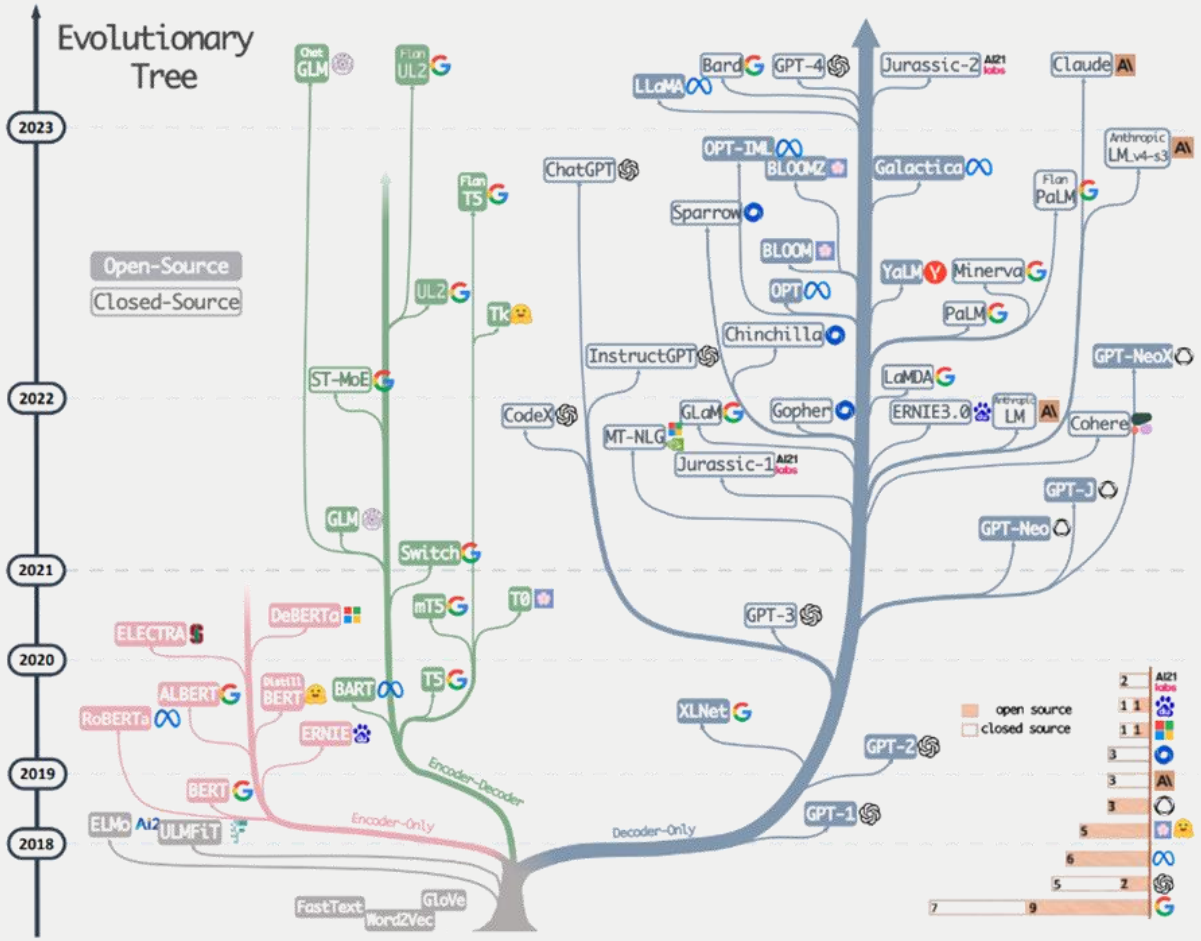
perplexity



Claude

Opções variadas, mas parece haver pouca diferenciação

- Os modelos líderes possuem ratings muito próximos no *Chatbot Arena*, mas há uma clara vantagem dos modelos proprietários. A versão *open source* mais bem posicionada é o Llama-3, da Meta

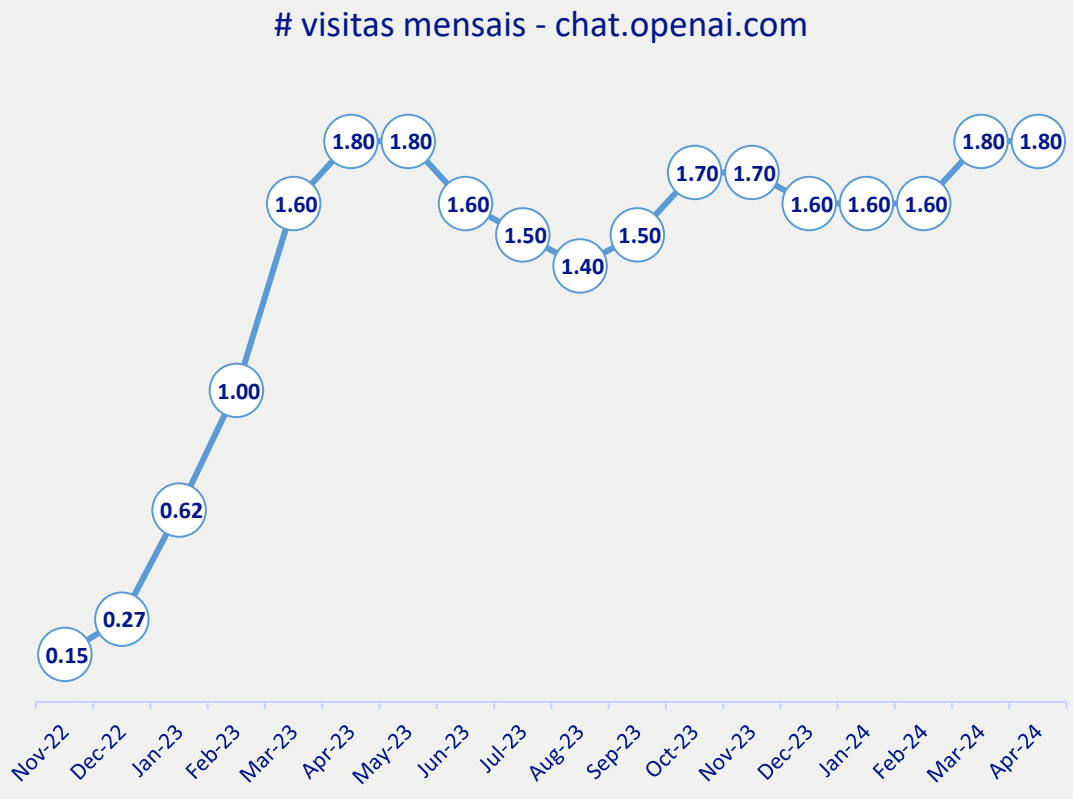


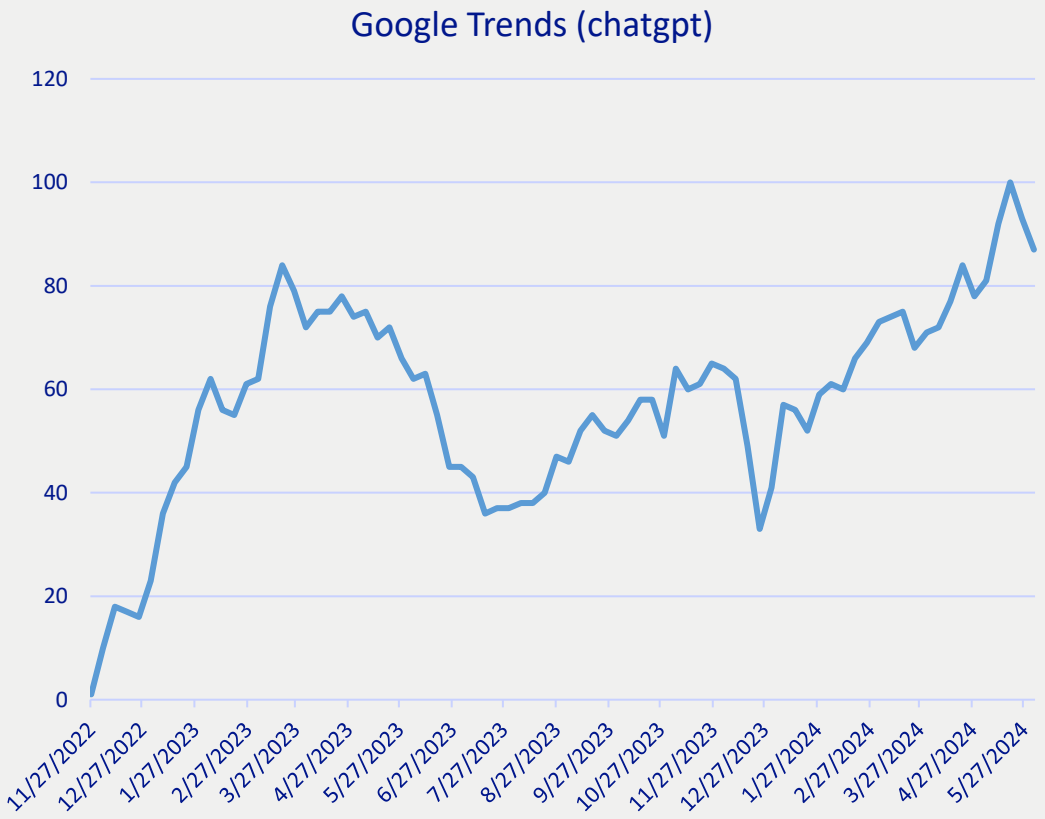
Ranking do Chatbot Arena

Rank	Model	Arena Elo	Votes	Organization	License
1	GPT-4o-2024-05-13	1287	34985	OpenAI	Proprietary
2	Gemini-Advanced-0514	1267	29838	Google	Proprietary
2	Gemini-1.5.Pro-API-0514	1266	28170	Google	Proprietary
4	Gemini-1.5-Pro-API-0409	1258	55731	Google	Proprietary
4	GPT-4-Turbo-2024-04-09	1256	61122	OpenAI	Proprietary
6	GPT-4-1106-preview	1251	80987	OpenAI	Proprietary
6	Claude. 3..0pus.	1249	126356	Anthropic	Proprietary
6	GPT-4-0125-preview	1246	74232	OpenAI	Proprietary
9	Yi-Large preview	1239	36412	01 AI	Proprietary
10	Gemini-1.5-Flash-API-0514	1232	26409	Google	Proprietary
11	Bard (Gemini Pro)	1208	11853	Google	Proprietary
11	Llama-3-70b-Instruct	1208	127901	Meta	Open Source
12	Claude..3..Sonnet	1202	98168	Anthropic	Proprietary
12	Reka-Core-20240501	1200	44097	Reka AI	Proprietary

Chatbots

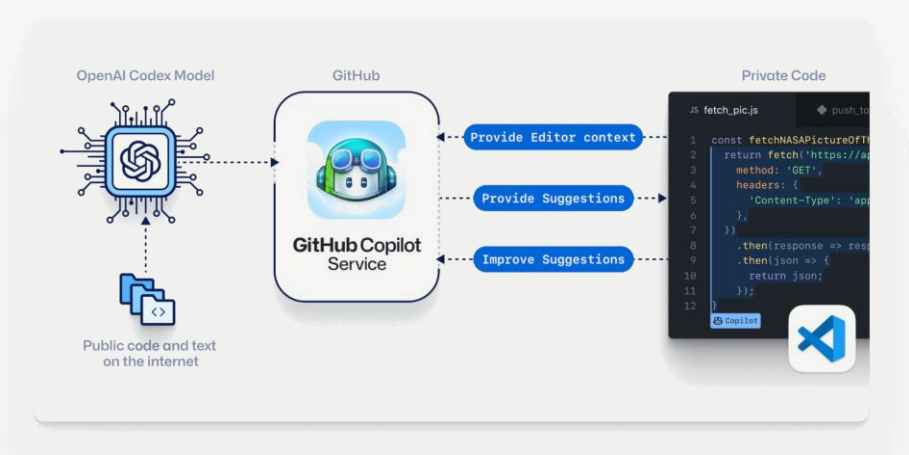
- A taxa de visitas mensais do chatbot líder, o ChatGPT, está estável nos últimos 12 meses, com 1,8 bilhão de visitas mensais

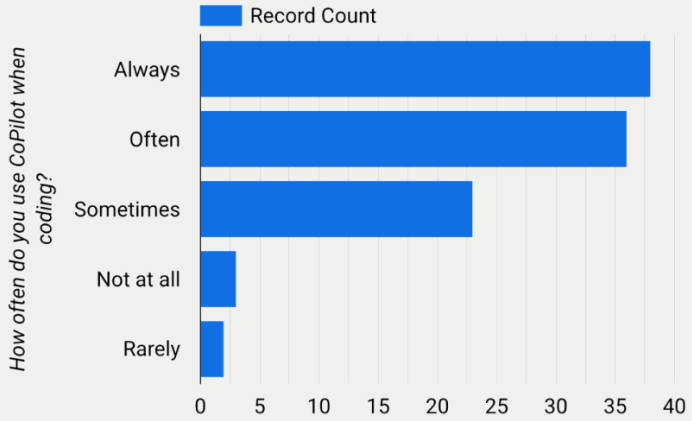




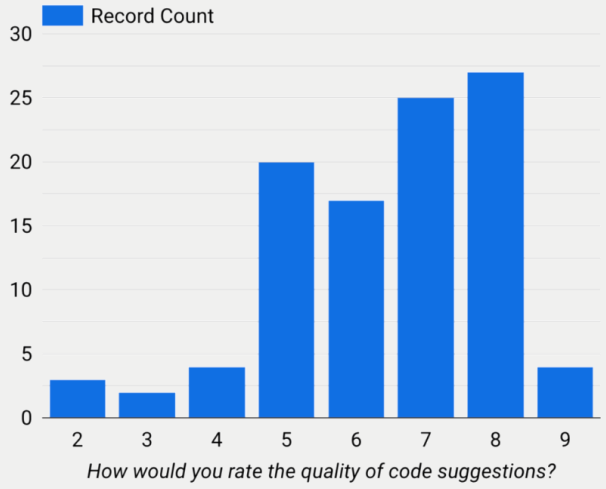
Assistentes de programação lideram em adoção

- Um dos casos de uso mais bem sucedidos até aqui são os assistentes de programação, com destaque para GitHub Copilot

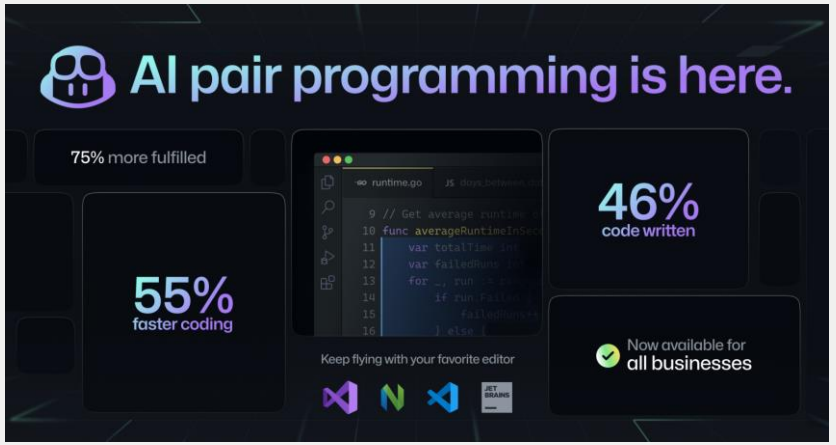




Adevinta Copilot Survey – November 2023 (Activity)

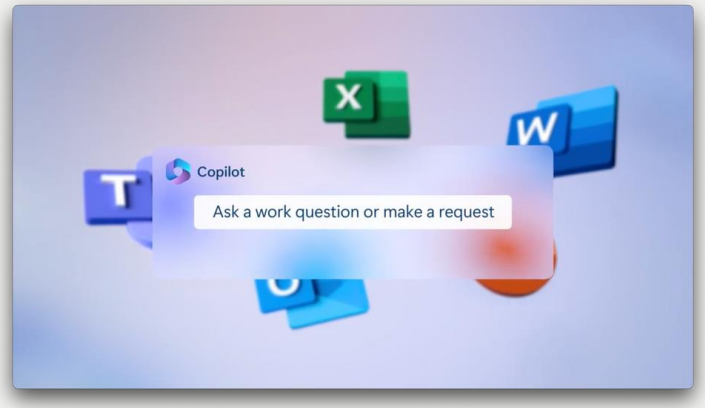


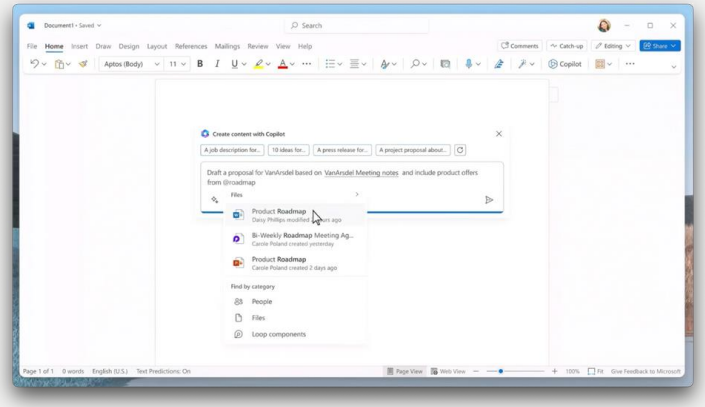
Adevinta Copilot Survey – November 2023 (Suggestion Quality)



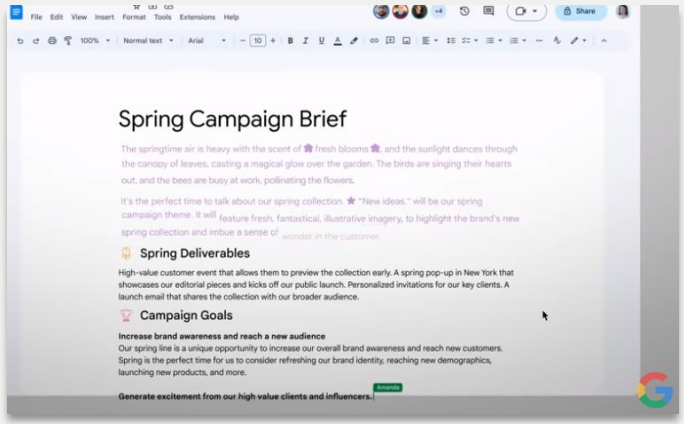
Soluções de produtividade são outra grande aposta

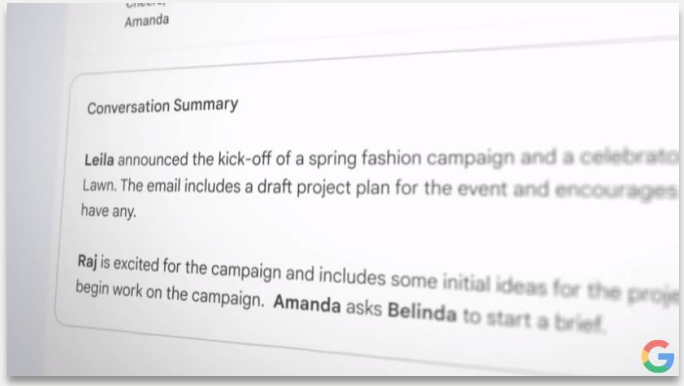
Microsoft 365 Copilot



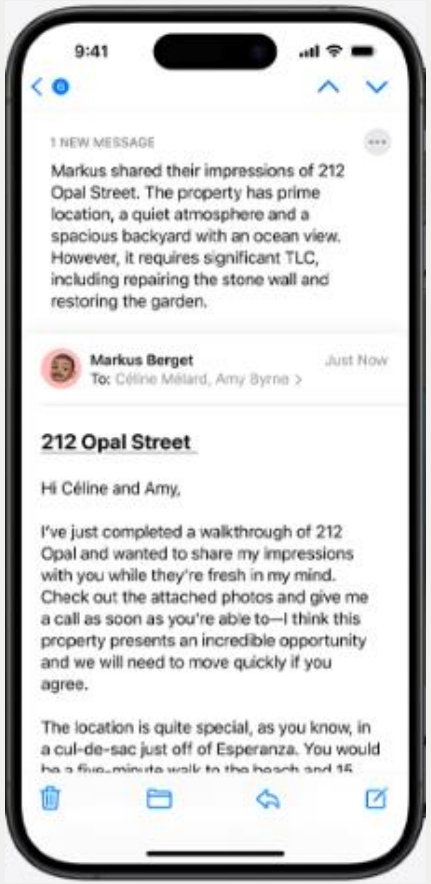


Google Workspace

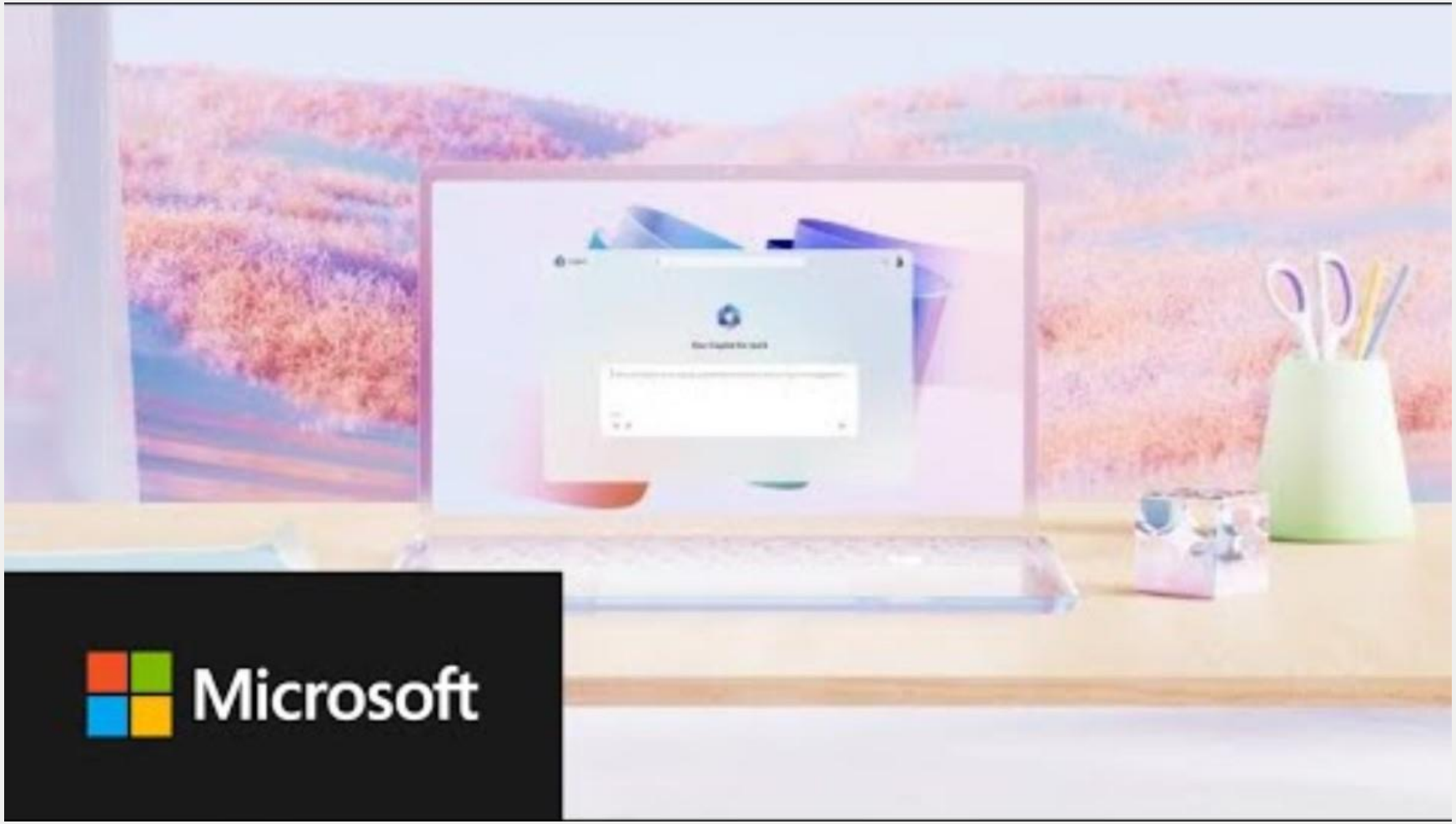




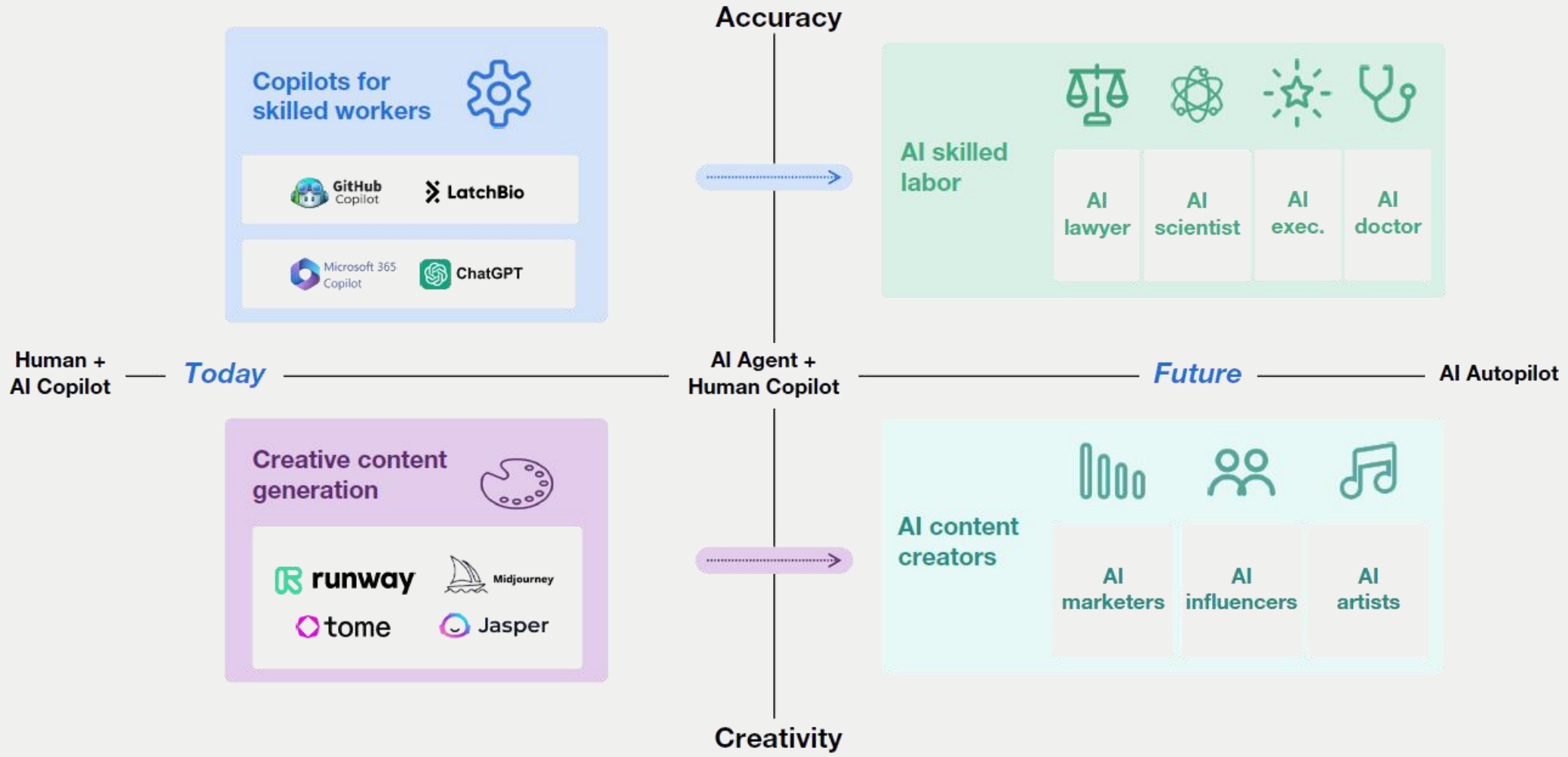
Apple Intelligence



Exemplo: Microsoft 365 Copilot



Visão de Futuro

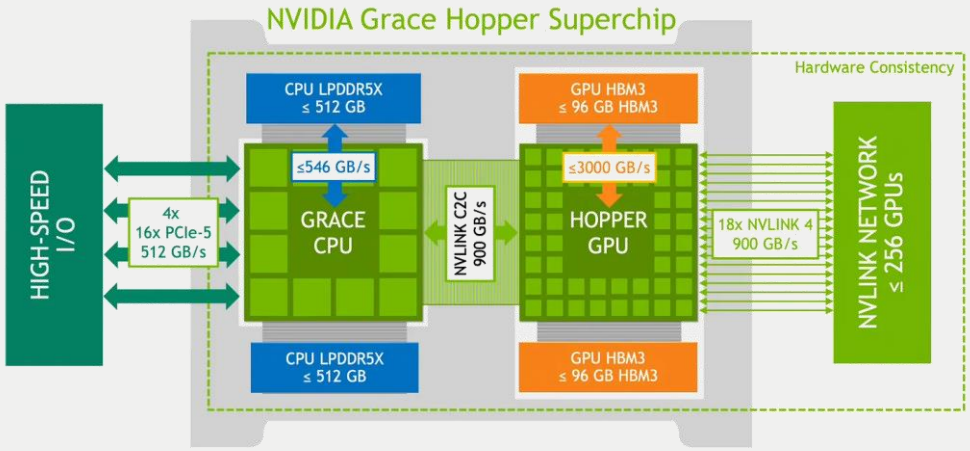


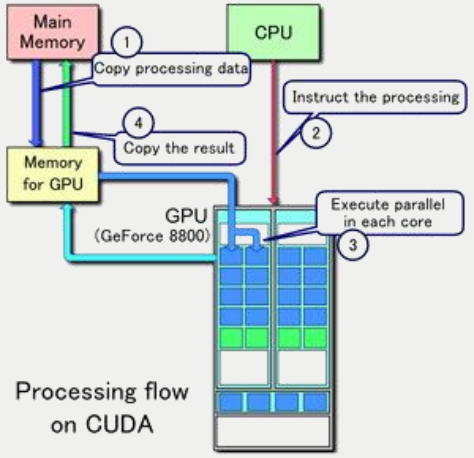

Semicondutores

O momento Nvidia

No lugar certo, na hora certa

- Os resultados recentes da Nvidia são consequência de uma combinação poderosa de **vantagens competitivas**:
  - ***First to market***: A empresa foi uma das primeiras a perceber o potencial das GPUs para AI
  - **Produtos *top notch***: Líder de mercado em GPUs, com produtos de melhor qualidade e esteira consistente de lançamentos
  - **Solução integrada**: A companhia entrega não apenas GPUs, mas uma solução completa de *accelerated computing*, incluindo CPU, GPU, memória e *networking*
  - **CUDA**: Plataforma integrada de software para desenvolvimento de aplicações com processamento em paralelo nas GPUs da Nvidia
  - **Efeito rede**: A maior parte de suas placas gráficas é compatível entre si, permitindo que um mesmo código escrito em CUDA funcione em diferentes versões de seus hardwares. Adicionalmente, aprender a desenvolver código em CUDA – o que não é uma tarefa fácil – estimula os desenvolvedores a permanecer no ecossistema Nvidia
- *"Our AI technology leadership is reinforced by our **large and expanding ecosystem in a virtuous cycle**. **Our computing platforms are available from virtually every major server maker and cloud service provider**, as well as on our own AI supercomputers. **There are over 4.7 million developers worldwide using CUDA** and our other software tools to help deploy our technology in our target markets."*

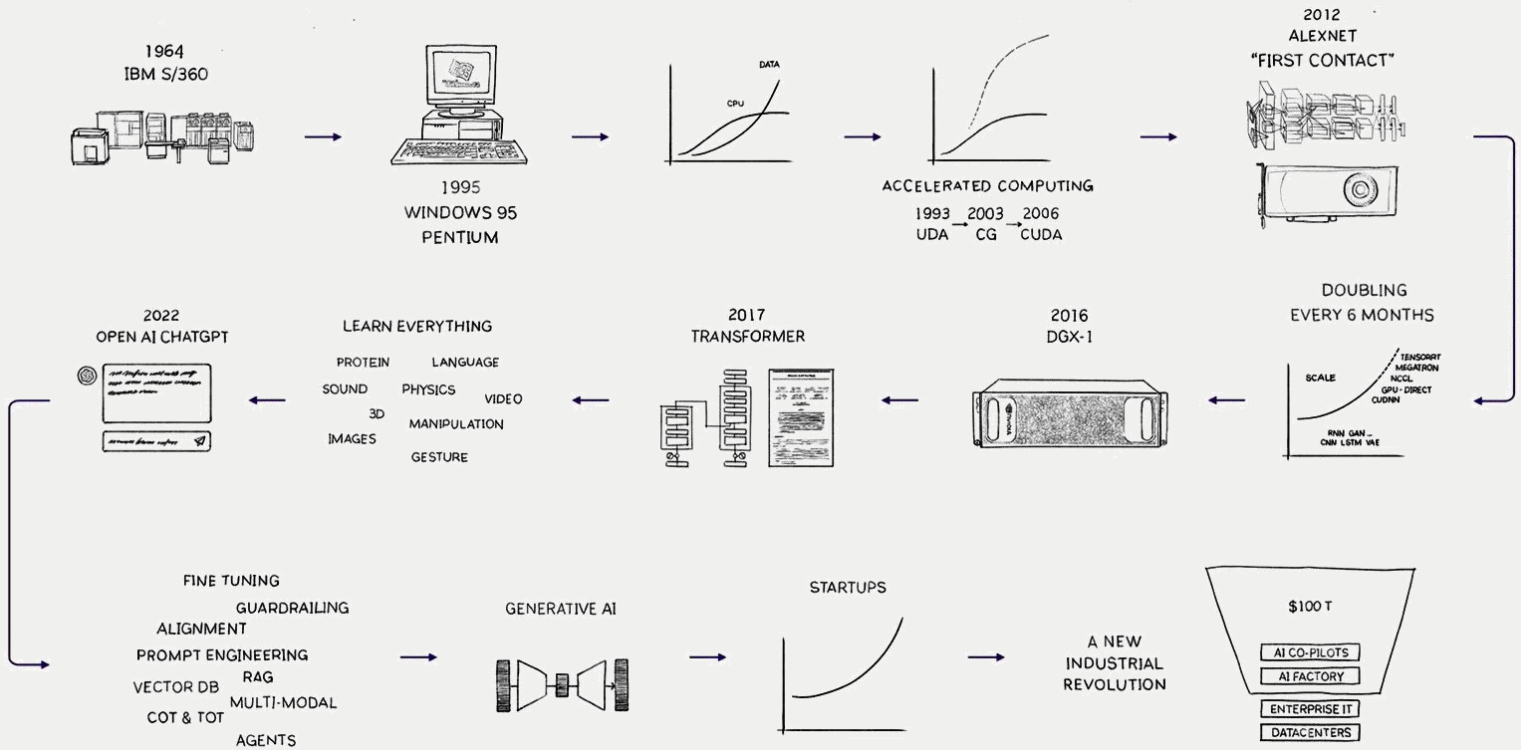




Pioneirismo em AI

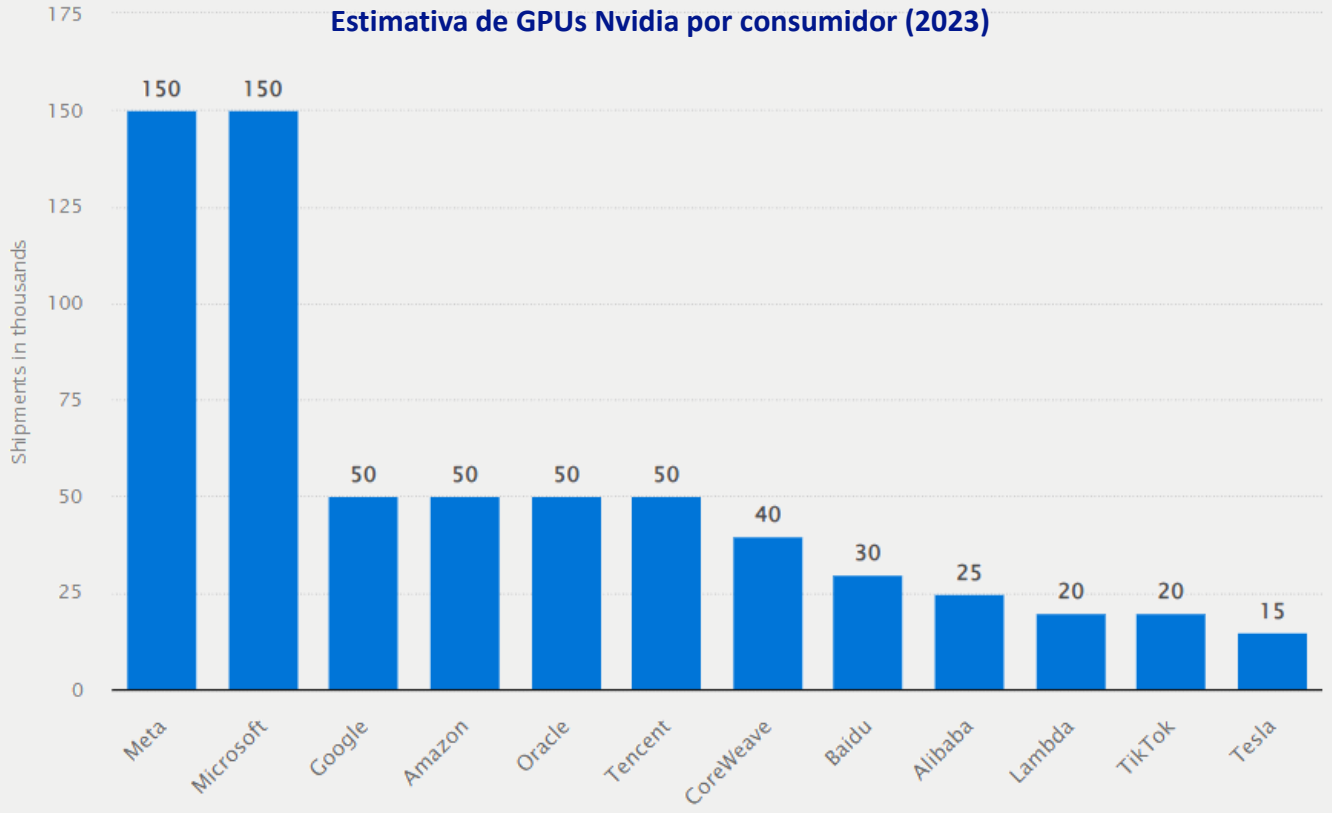
- Impressiona a presciência de ser uma das primeiras companhias a realizar investimentos intensivos em AI. Ainda em 2017, 5 anos antes do lançamento do ChatGPT, o CEO Jensen Huang disse ao MIT Tech Review: "***Software is eating the world, but AI is going to eat software***."
- Hoje, essa visão começa a se materializar: "*It is our job to create computing technology such that nobody has to program. And that the programming language is human.* ***Everybody in the world is now a programmer****. This is the miracle of artificial intelligence.*"

A tese Nvidia

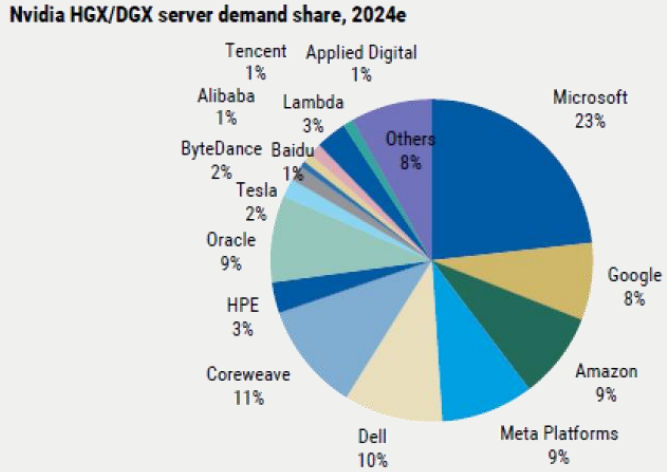


“He who controls NVIDIAs, controls the world”

- GPUs da Nvidia são hoje um dos produtos mais demandados do mundo. Parafraseando o livro Duna, de Frank Herbert: “He who controls NVIDIAs, controls the world”
- Jensen Huang: "*The demand for both Hopper and Blackwell platforms will outstrip supply well into next year. The complexity of these chips is a significant challenge for us in keeping up with the demand.*"

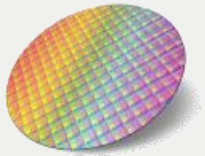
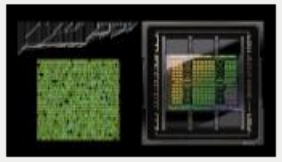


"People think that Nvidia GPUs is like a chip. But the Nvidia Hopper GPU is 35,000 parts. It weighs 70 pounds," Huang said. "These things are really complicated...people call it an AI supercomputer for good reason. If you ever looking look in the back of the data center, the systems, the cabling system is mind boggling. It is the most dense, complex cabling system for networking the world's ever seen.“ – Jensen Huang



Source: Morgan Stanley Research (E) estimates

Nvidia é apenas um elo de uma cadeia muito mais ampla

Semi manufacturing inputs		Semi design and manufacturing		Device assembly
Materials	Capital equip.	IC design/embedded software	Fabrication, pkg. & testing	Assembly
GlobalWafers Co., Ltd. 環球晶圓股份有限公司 Unimicron 欣興電子 JSR	Lam Research ASML	NVIDIA intel AMD arm BROADCOM Qualcomm	tsmc	FOXCONN wistron
				
Ex: Silicon wafer	*Ex: EUV Lithography*	*Ex: NVIDIA chip design*	*Ex: 3nm chip*	*End product: NVIDIA H100*

Com um ciclo de investimento que pode chegar a $1T+

	Mainframe	PC + Internet	Smartphone	Public Cloud	AI
Time Period	Pre 1990	1990 - 2006	2007 - 2014	2014 - 2022	2022 onwards
Incremental Semi Spend	~$50B	~$100B	~$150B	~$250B	~$500B+ ?
Total Semi Spend	~$50B	~$200B	~$350B	~$600B	~$1.5tn+ ?
Key HW players	IBM	intel, DELL, Qualcomm, BROADCOM, Google, Microsoft, Lam RESEARCH, AMD, Microsoft, hp, arm, SKYWORKS, amazon, APPLIED MATERIALS, cerebras, NVIDIA, tsmc, ASML, SAMSUNG			

Potenciais vencedores em diversos segmentos da cadeia

- **GPUs generalistas**: placas voltadas para aplicações gerais de *parallel computing*. São muito mais flexíveis, mas menos eficientes para casos de uso específicos

- **GPUs para aplicações específicas** (ASICs – *Application Specific Integrated* Circuits; que podem ser TPUs – *Tensor Processing Units,* NPUs – *Neural Processing Units* ou similares): Chips construídos para otimizar a eficiência computacional de aplicações específicas em larga escala

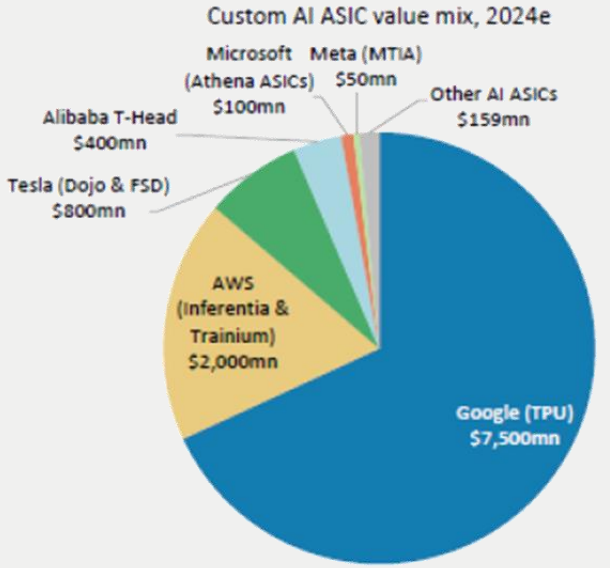
- **CPUs, GPUs e NPUs voltados para *edge AI***: chips dedicados ao processamento de *machine learning* "on edge", i.e., diretamente no device dos usuários finais (smartphones e PCs), ao invés de servidores remotos

- **Fabricação e *packaging* dos chips**: Os efetivos fabricantes dos semicondutores. Etapa mais complexa e capital intensiva da cadeia

Custom silicon deve ganhar relevância

- Conforme a demanda por recursos computacionais migre de treinamento para inferência, os chips construídos para aplicações específicas deverão ganhar relevância
- O Google é hoje a *big tech* mais avançada nesse tipo de design, desenvolvendo seus chips próprios em conjunto com a Broadcom desde 2016. Mais recentemente, diversas empresas tem buscado soluções proprietárias

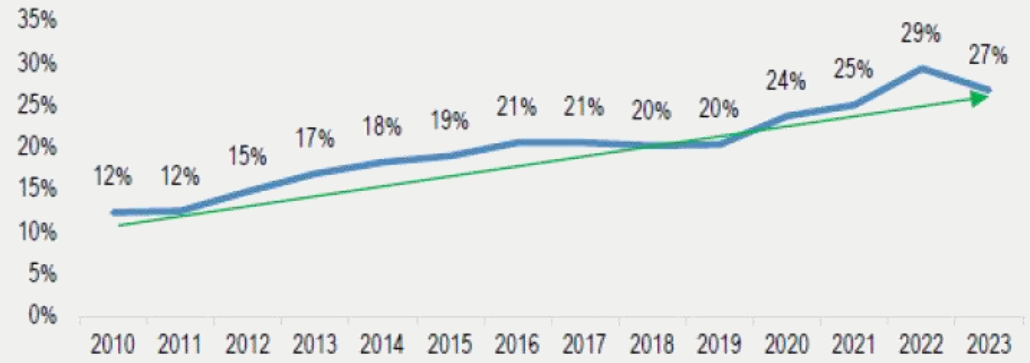
Company	First official announcement for AI ASIC	AI ASIC project	ASIC partners	ASIC's benefits vs. merchant solutions
Google	2016	TPU	Broadcom	2-3x greater energy efficiency
AWS	2018	Inferentia, Trainium	Alchip, Marvell	50% greater performance per watt
Tesla	2018	D1 (Dojo supercomputer), FSD (on-car AD/ADAS)	Alchip, Samsung	33% of cost savings
Microsoft	2023	Maia 100 AI Accelerator, Cobalt 100 CPU	GUC	40% performance improvement
Meta	2023	MTIA, MSVP	N/A	2x performance



TSMC, um dos maiores *moats* da indústria

- O processo de fabricação de um chip é uma das atividades industriais mais complexas do mundo. Ao longo do tempo, a TSMC foi aprimorando sua expertise técnica por meio de muito investimento em R&D e parcerias sólidas com as principais empresas de design de chips e fabricantes de equipamentos, o que permitiu que ganhassem escala e fossem estimulados a estar sempre no *edge* da inovação
- Ela foi inovadora no conceito de **"foundry" independente**. Ou seja, atuando apenas como fabricante dos chips, e não no design. Dessa forma, ela não concorre com seus clientes (como fazem Intel e Samsung), o que permite que seja a escolha preferencial da maioria das *fabless companies* (Nvidia, AMD, Broadcom, Qualcomm, etc)
- Apesar de a Lei de Moore ter sido conjecturada por Gordon Moore, co-fundador da Intel, pode-se dizer que o fato dessa "Lei" ter sido mantido válida ao longo dos últimos anos deve-se em boa parte à TSMC
- Morris Chang, CEO da TSMC: "***The semiconductor business is like a treadmill that speeds up all the time. If you can't keep up, you fall off***"

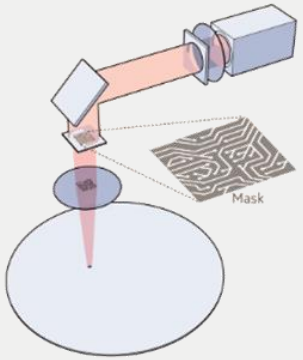
Market share de semis (ex-memory) em *foundries*



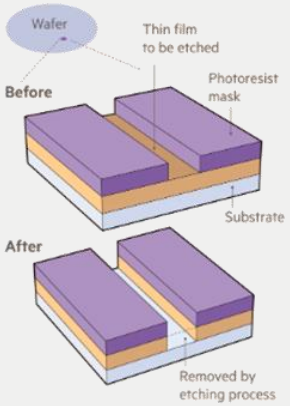
Market share da TSMC dentre *foundries*

Fontes: Quartr, Gartner

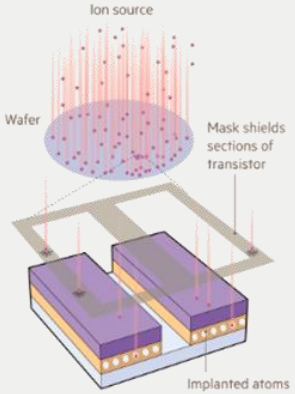
Necessidade de escala para R&D consolidou o mercado



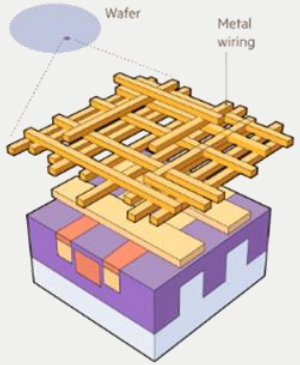
Ultraviolet light, projected through a stencil, transfers tiny patterns onto the wafer



Many thin films of materials are added and etched away using these intricate patterns as a guide

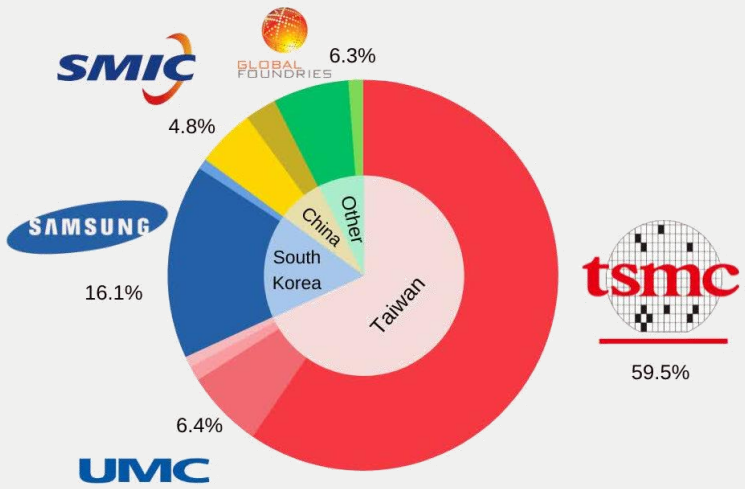


The water is hit with ions, or charged atoms, to make areas more conductive or insulating



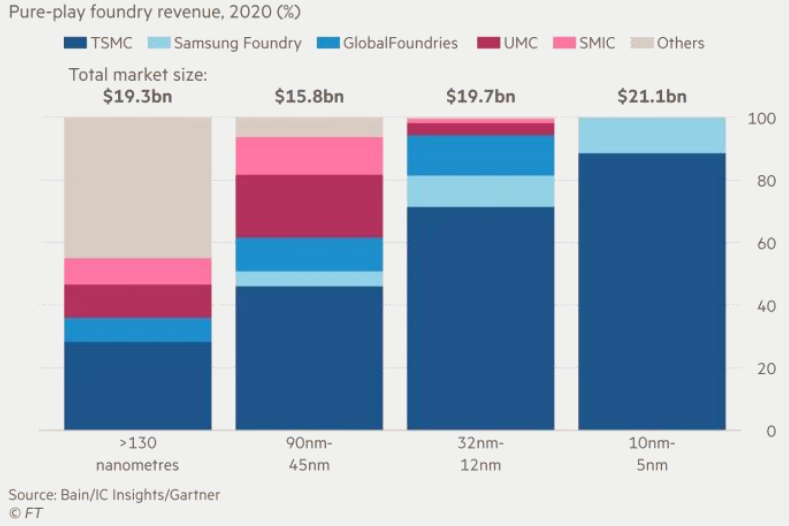
Hundreds of stages of layering build up the chip's components. Metal wiring completes the circuit

Foundries market share

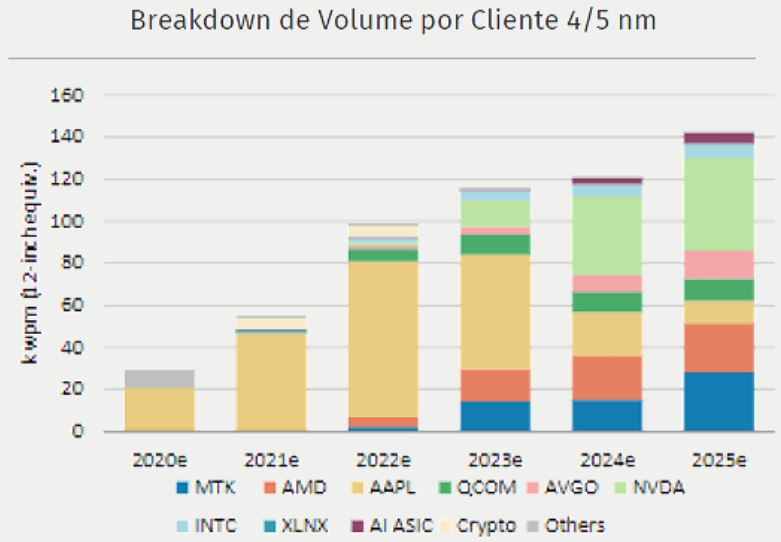


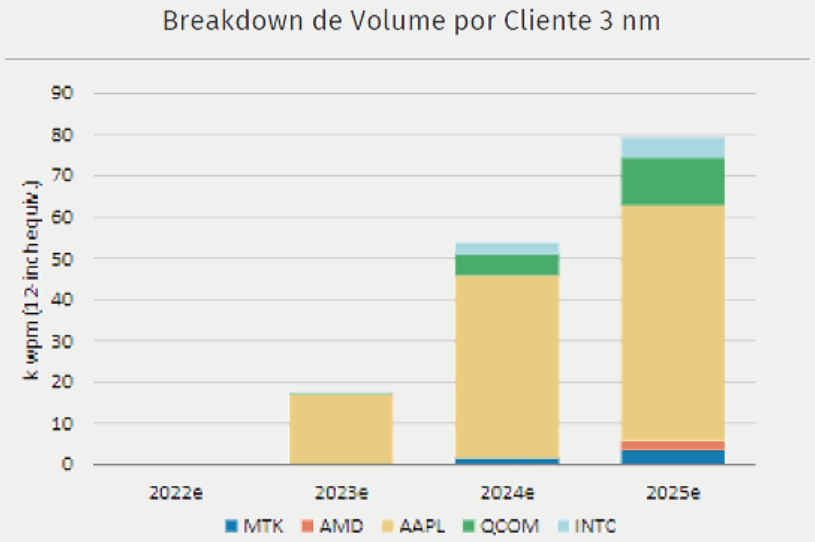
Empresas capazes de produzir cada *node size*

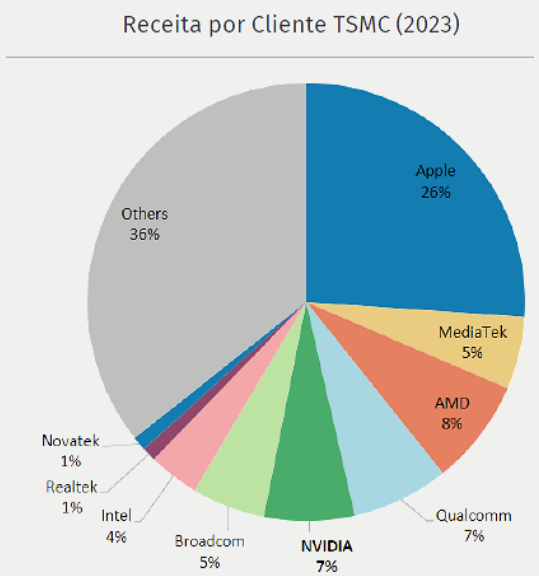
Panasonic							
STM							
HLMC							
UMC							
IBM		UMC					
SMIC	IBM	SMIC					
GF	GF	GF	SMIC				
Samsung	Samsung	Samsung	Samsung	Samsung	Samsung	Samsung	Samsung
TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC
Intel	Intel	Intel	Intel	Intel	Intel	Intel	Intel
32nm/28nm	**22nm/20nm**	**16nm/14nm**	**10nm**	**7nm**	**5nm**	**3nm**	**2nm**

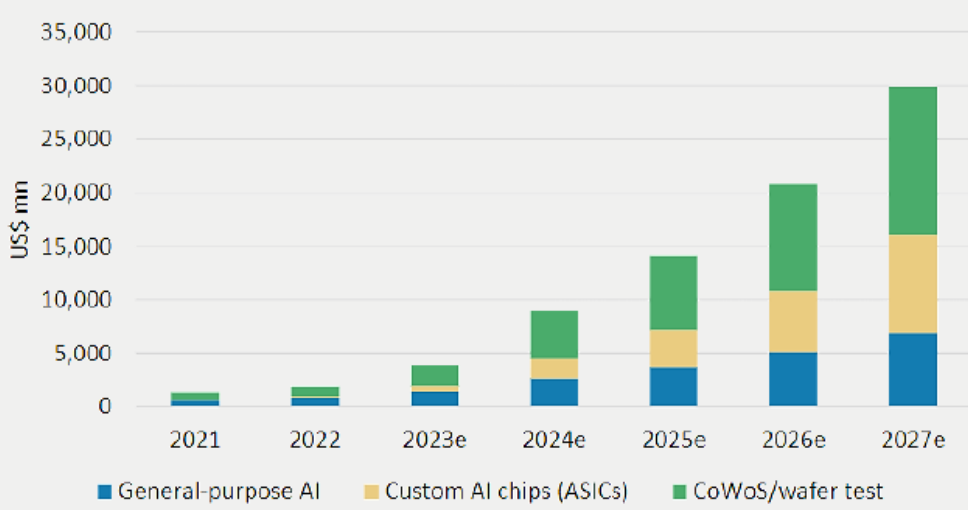


E deve ganhar *share* com maior demanda por *leading edge*



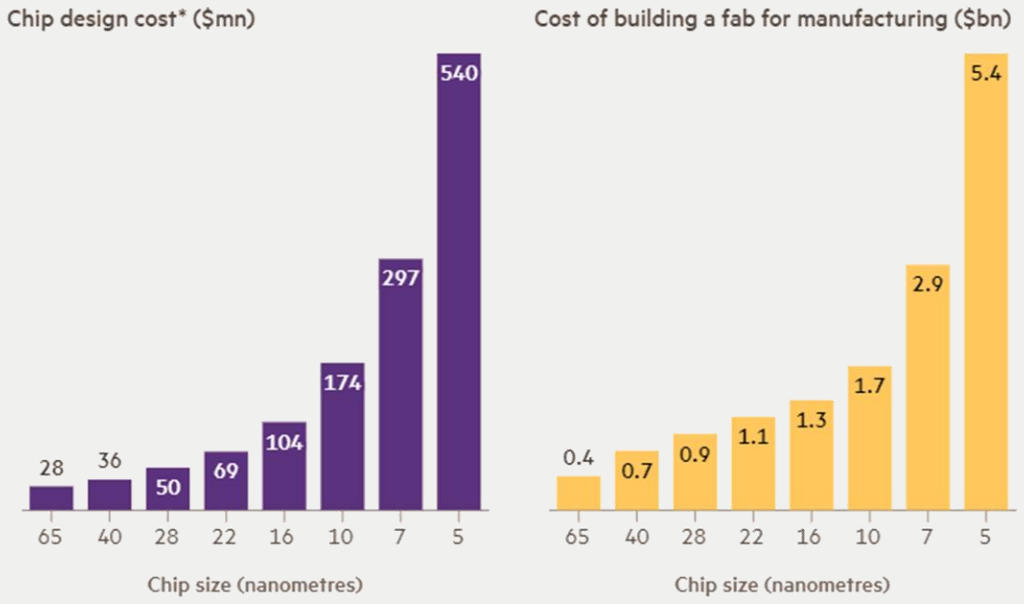






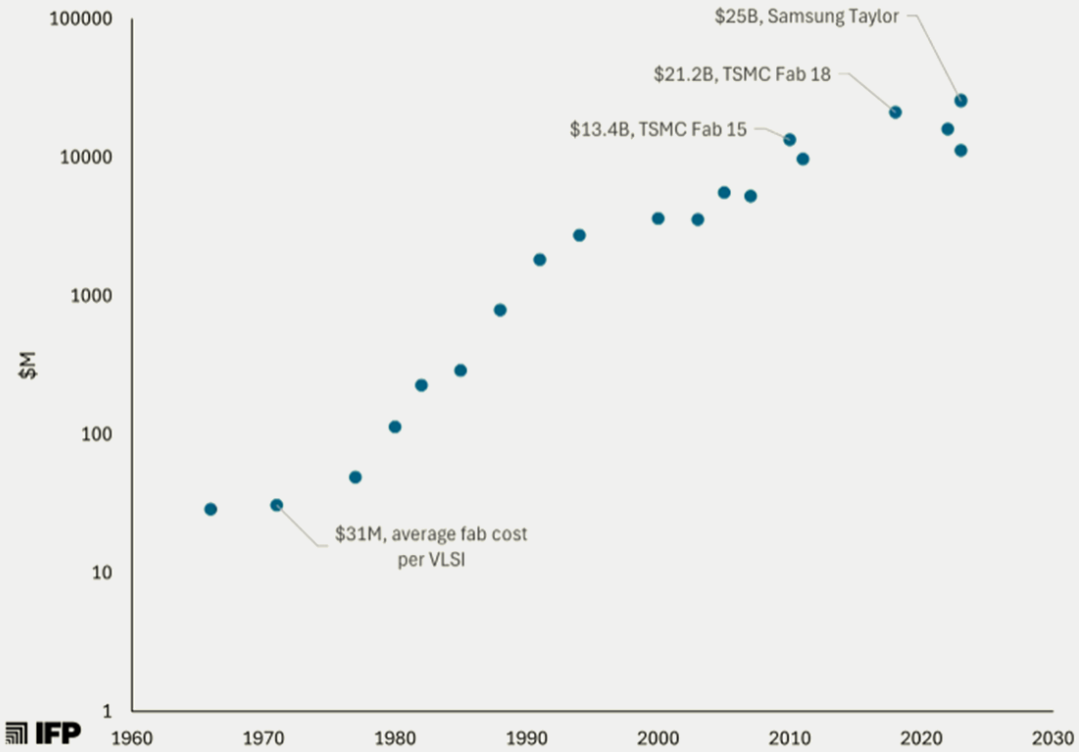
O diferencial competitivo da escala fica cada vez maior

"These enormous costs are ultimately due to the same factor that has steadily driven down the cost of semiconductors: Moore's Law, the observation that the number of components on an integrated circuit tends to double every two years. ***(There is a Moore's Second Law, also known as Rock's Law, which posits that the cost of a semiconductor fab doubles every four years****.) The smaller semiconductor components get, the more difficult it is to create the conditions to manufacture them."*



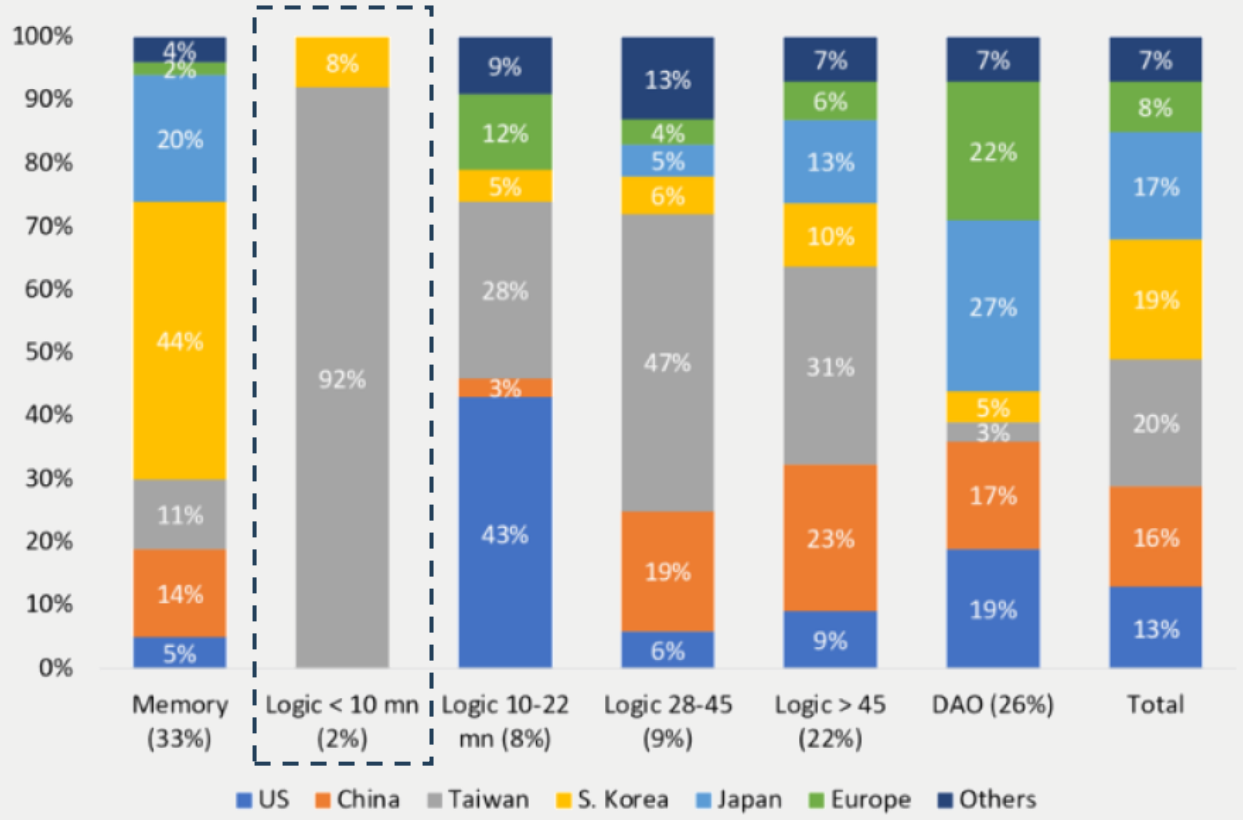
Sources: IBS, McKinsey • Data covers 2006-2019 chip developments. *Major components include IP qualification, architecture, verification, physical, software, prototyping and validation

Custo de uma fábrica de semicondutores ($bi)



Importância fundamental na geopolítica

- **Mais de 90% dos chips topo de linha são produzidos pela TSMC, em Taiwan**
- São os chips presentes nos *devices* tecnológicos mais avançados, como *smartphones*, servidores, PCs, GPUs, automóveis e máquinas industriais avançadas, dentre muitos outros
- Isso torna a TSMC um **foco de tensão geopolítica** relevante, uma vez que tanto o ocidente quanto o oriente dependem de seus chips, e **nenhum dos blocos tem capacidade tecnológica para desenvolvê-los de forma independente hoje**
- Por esse motivo, o Governo americano tem direcionado subsídios fiscais para a construção de fábricas da TSMC nos EUA por meio do CHIPS Act



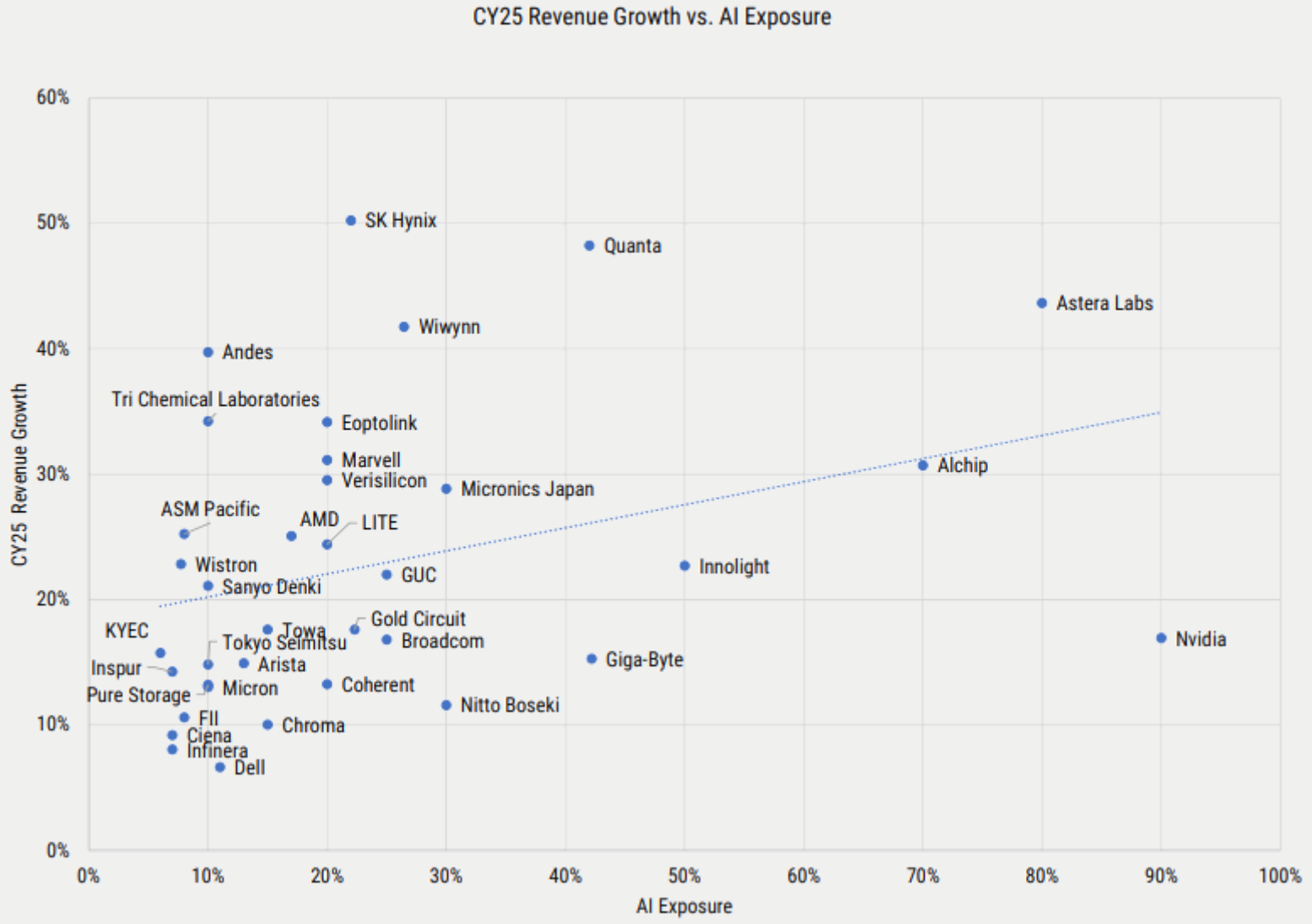
mar asset management

Dados financeiros

Empresa	Mkt Cap	ND/EBITDA	P/E 2024	P/E 2025	P/FCF 2024	P/FCF 2025	Rev growth CAGR L5Y	Rev growth CAGR 2024	Rev growth CAGR 2025	Free Cash Flow 2023	Free Cash Flow 2024	Free Cash Flow 2025	FCF margin 2021	FCF margin 2022	FCF margin 2023	5yr ROE	5yr ROIC	TSR 1Y	TSR 5Y
Nvidia	3,252,856	-0.26 x	106.14	49.07	125.47	52.43	39.1%	(3.0%)	39.4%	3,808	25,925	62,036	6.3%	43.9%	52.4%	45.3%	32.3%	209.0%	3,365.0%
AMD	257,139	-0.29 x	45.01	29.60	60.29	34.03	28.5%	12.4%	19.3%	1,121	4,265	7,556	4.9%	16.7%	23.4%	21.4%	26.7%	32.9%	423.4%
Intel	130,985	1.50 x	28.70	15.74	(12.25)	(27.36)	(5.2%)	4.7%	7.3%	(14,279)	(10,696)	(4,787)	(26.3%)	(18.8%)	(7.7%)	15.3%	10.6%	(15.1%)	(26.3%)
Broadcom	856,843	1.71 x	36.91	29.22	37.21	28.29	11.4%	42.9%	28.6%	17,633	23,025	30,290	49.2%	45.0%	51.1%	35.5%	14.1%	103.8%	634.8%
Qualcomm	244,058	0.10 x	21.79	19.59	22.33	19.87	9.6%	6.9%	8.8%	9,849	10,929	12,281	27.5%	28.5%	29.0%	90.2%	29.4%	79.5%	236.5%
Marvell	62,193	2.02 x	47.40	51.15	59.57	63.07	14.0%	(0.1%)	(1.0%)	1,083	1,044	986	19.7%	19.0%	18.3%	(2.6%)	(0.5%)	20.4%	210.2%
Mediatek	68,179	-1.01 x	22.64	20.34	27.63	21.51	12.7%	19.0%	17.1%	4,843	2,467	3,169	36.2%	15.5%	17.2%	18.2%	14.7%	111.6%	520.5%
Arm	169,870	-1.39 x	135.25	103.48	352.29	133.33	n/a	(1.6%)	10.7%	675	482	1,274	20.9%	15.2%	32.2%	n/a	n/a	n/a	n/a
TSMC	737,867	-0.40 x	23.64	18.67	33.37	25.57	15.9%	26.5%	23.6%	10,322	22,110	28,853	15.5%	26.2%	28.3%	29.3%	22.6%	59.5%	336.5%
SMIC	26,842	3.82 x	61.86	39.02	(6.77)	(26.44)	13.5%	730.3%	207.7%	(588)	(3,966)	(1,015)	(67.5%)	(54.8%)	(12.3%)	6.4%	2.4%	(16.3%)	n/a
ASML	415,789	-0.05 x	53.15	33.57	59.86	37.58	20.3%	(1.2%)	14.4%	3,529	6,946	11,065	11.9%	23.8%	28.6%	45.2%	34.2%	45.7%	469.4%
Applied Materials	200,821	-0.17 x	28.77	25.46	31.37	29.20	9.7%	1.5%	6.2%	7,594	6,401	6,877	28.6%	23.8%	23.0%	46.5%	32.4%	71.9%	486.0%
Lam Research	139,584	-0.12 x	35.63	30.09	31.57	31.85	9.5%	(14.8%)	0.2%	4,677	4,422	4,382	26.8%	29.8%	25.0%	59.1%	31.5%	70.4%	507.3%
Foxconn	75,313	-0.90 x	21.32	17.70	208.24	38.58	2.8%	16.6%	17.1%	2,844	362	1,952	4.3%	0.5%	2.2%	18.1%	12.2%	16.0%	159.6%
Wistron	9,530	1.16 x	16.90	12.95	18.82	(412.15)	(0.5%)	15.9%	16.8%	1,089	506	(23)	4.1%	1.6%	(0.1%)	11.6%	6.8%	38.7%	514.4%
Kla	113,418	0.52 x	35.84	30.06	36.73	31.47	21.1%	(9.8%)	3.2%	3,328	3,088	3,604	31.7%	32.6%	32.2%	99.3%	32.2%	79.1%	687.3%
Super Micro	53,428	-0.09 x	36.41	24.37	(29.35)	66.25	16.2%	109.7%	82.8%	627	(1,821)	807	8.8%	(12.2%)	3.4%	19.6%	14.5%	266.0%	4,416.3%

Fonte: Bloomberg (17/06/24)

Crescimento vs Exposição a AI

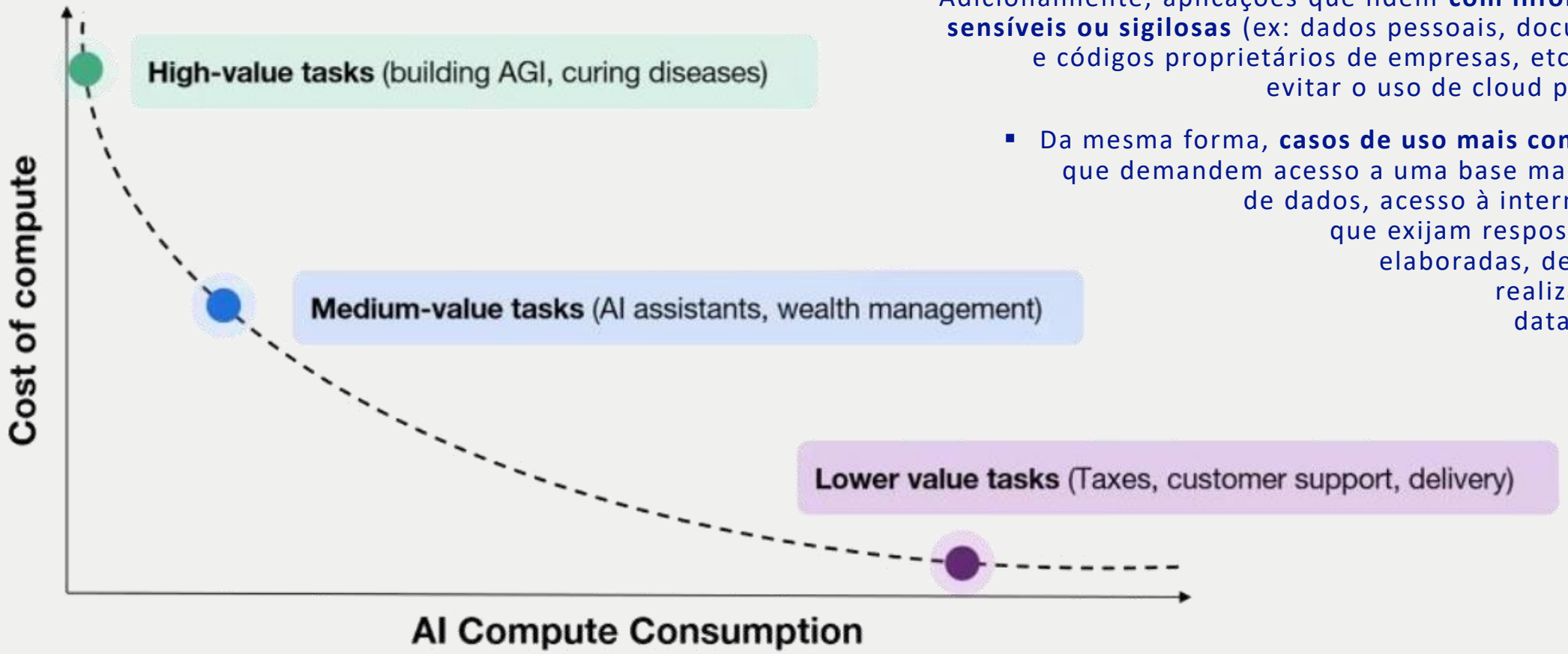


Fonte: Morgan Stanley


Edge AI

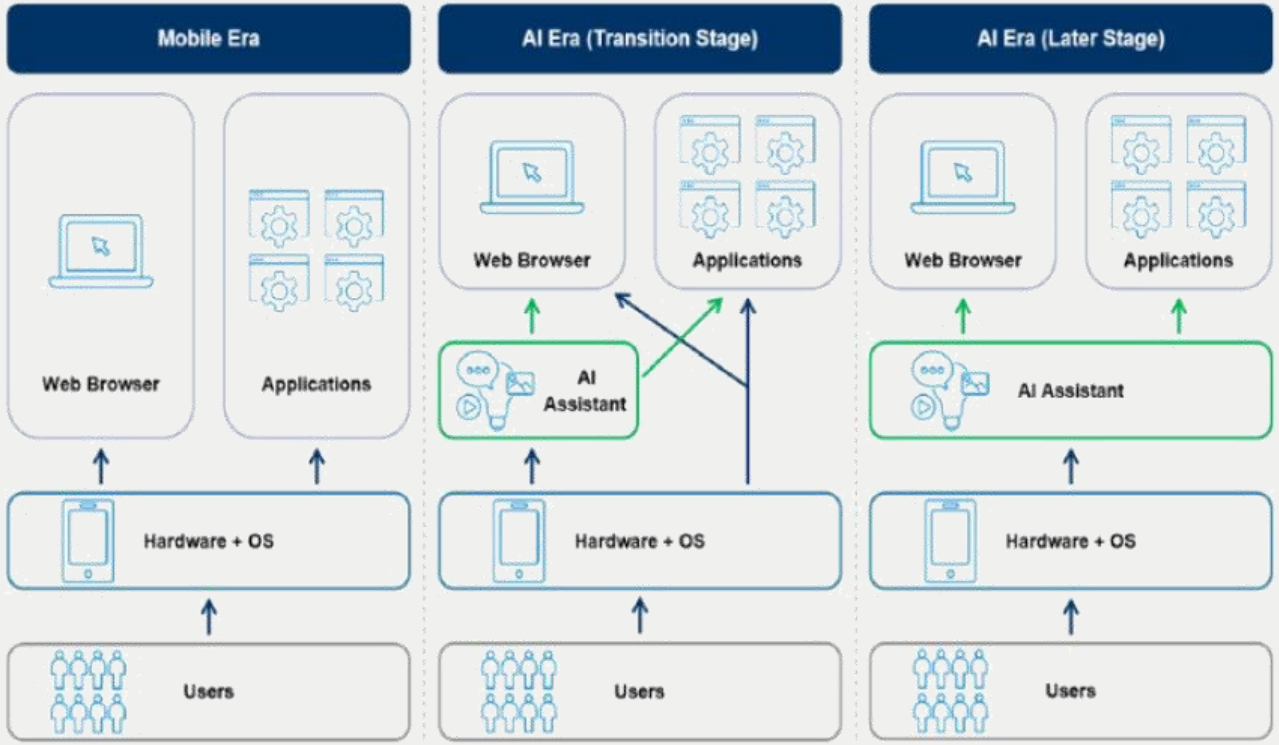
Nem todo processamento de AI se dará em *data centers*

- **Diferentes casos de uso levarão a diferentes necessidades de processamento, latência e privacidade**
  - Aplicações que demandem **resposta em tempo real e que possuam escopo majoritariamente local** tendem a ser processadas diretamente nos devices
  - Adicionalmente, aplicações que lidem **com informações sensíveis ou sigilosas** (ex: dados pessoais, documentos e códigos proprietários de empresas, etc) devem evitar o uso de cloud providers
  - Da mesma forma, **casos de uso mais complexos**, que demandem acesso a uma base mais ampla de dados, acesso à internet e/ou que exijam respostas mais elaboradas, devem ser realizados em data centers



Smartphones seguem como principal interface

- Apesar das inovações da *generative* AI, **a forma como interagimos com essa tecnologia não parece estar sob o risco de disrupção no curto ou médio prazos**. Com isso, smartphones e PCs deverão continuar sendo a principal interface de interação entre os usuários e AI
- Mas **é possível que os "agentes pessoais" baseados em AI ganhem cada vez mais autonomia ao longo do tempo**, se tornando a ponte entre nossos aparelhos e as aplicações que hoje usamos diretamente, **tornando-se assim o grande arbitrador do fluxo de informações e, potencialmente, de nossas escolhas**

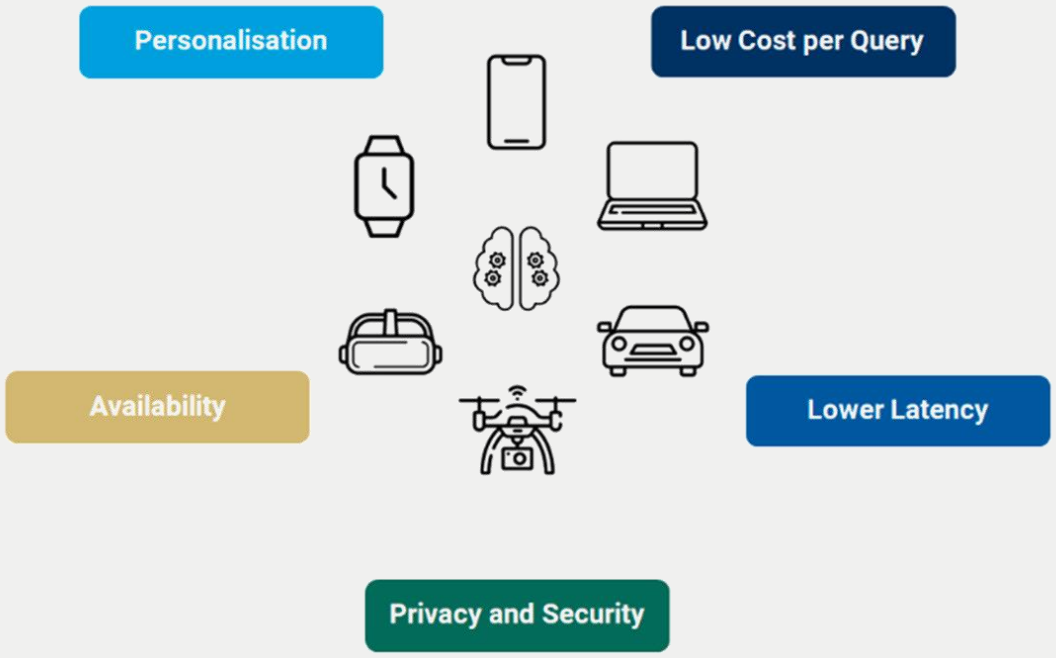


Data centers vs Edge AI

Data center	Edge AI
• Treinamento dos LLMs • Inferência de *queries* mais complexas • LLMs	• "Assistentes" AI personalizados com contexto local • Inferência de *queries* mais simples • SLMs
• Custo elevado • Intensivo em consumo de energia • Centralizado • Maior risco em relação à privacidade de dados	• Mais barato • Eficiente em energia • Descentralizado • Menor risco em relação à privacidade de dados
• Alta latência	• Baixa latência
• Servidor / *datacenter* AI	• PC / smartphone AI

Vantagens e desafios do *Edge AI*

Vantagens



Desafios

Consumo de energia / bateria

- O tempo de duração da bateria é um dos principais fatores de escolha de um PC ou smartphone
- Chips e modelos eficientes no consumo de energia e/ou inovação na eficiência das baterias em si é essencial para maior adoção

Poder de processamento / memória

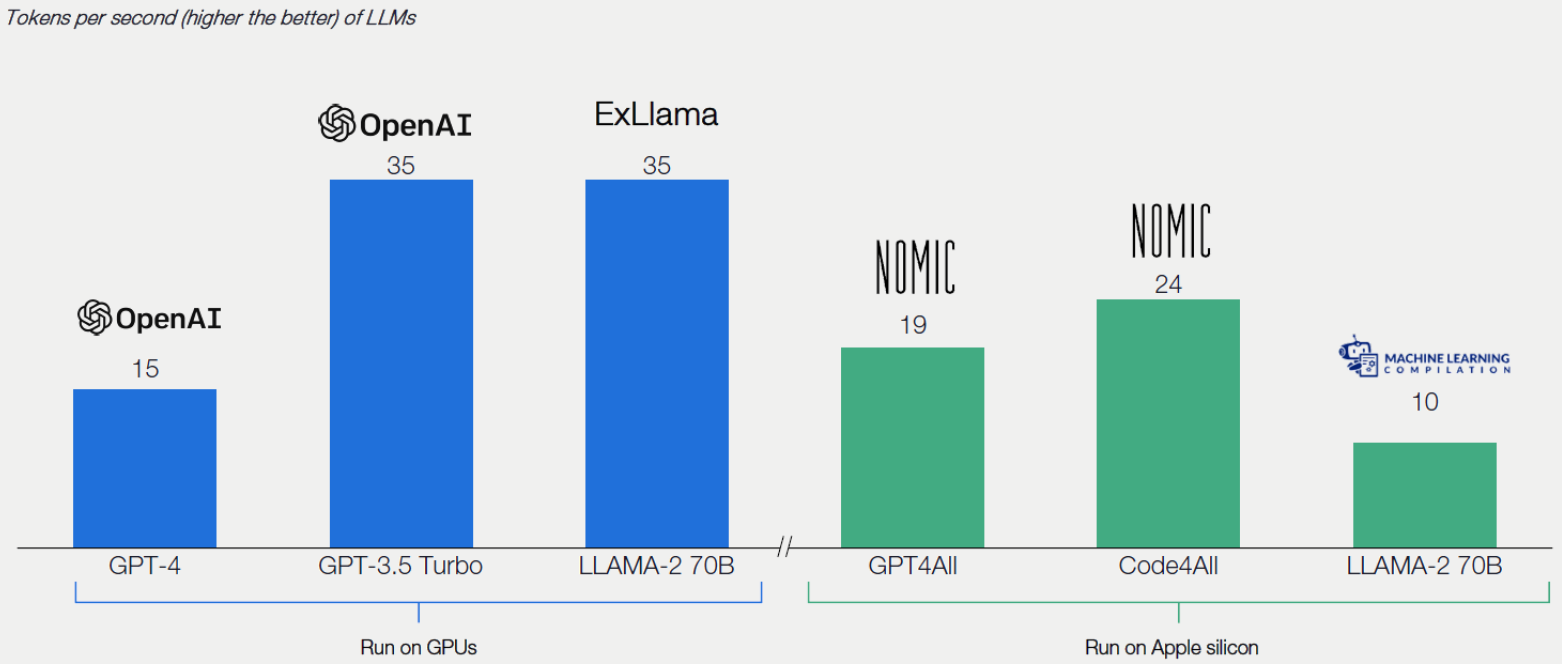
- Modelos de *machine learning* demandam capacidade de processamento e espaço em memória muito superior às principais aplicações que utilizamos hoje
- Chips dedicados a AI (ex: NPUs) e aparelhos com maior volume de memória dedicada 8+GB serão necessários

Forma / tamanho

- Outra demanda importante para PCs e smartphone é o tamanho e o peso
- Conciliar a maior necessidade de hardware com esses requisitos seguirá um desafio importante para *edge devices*

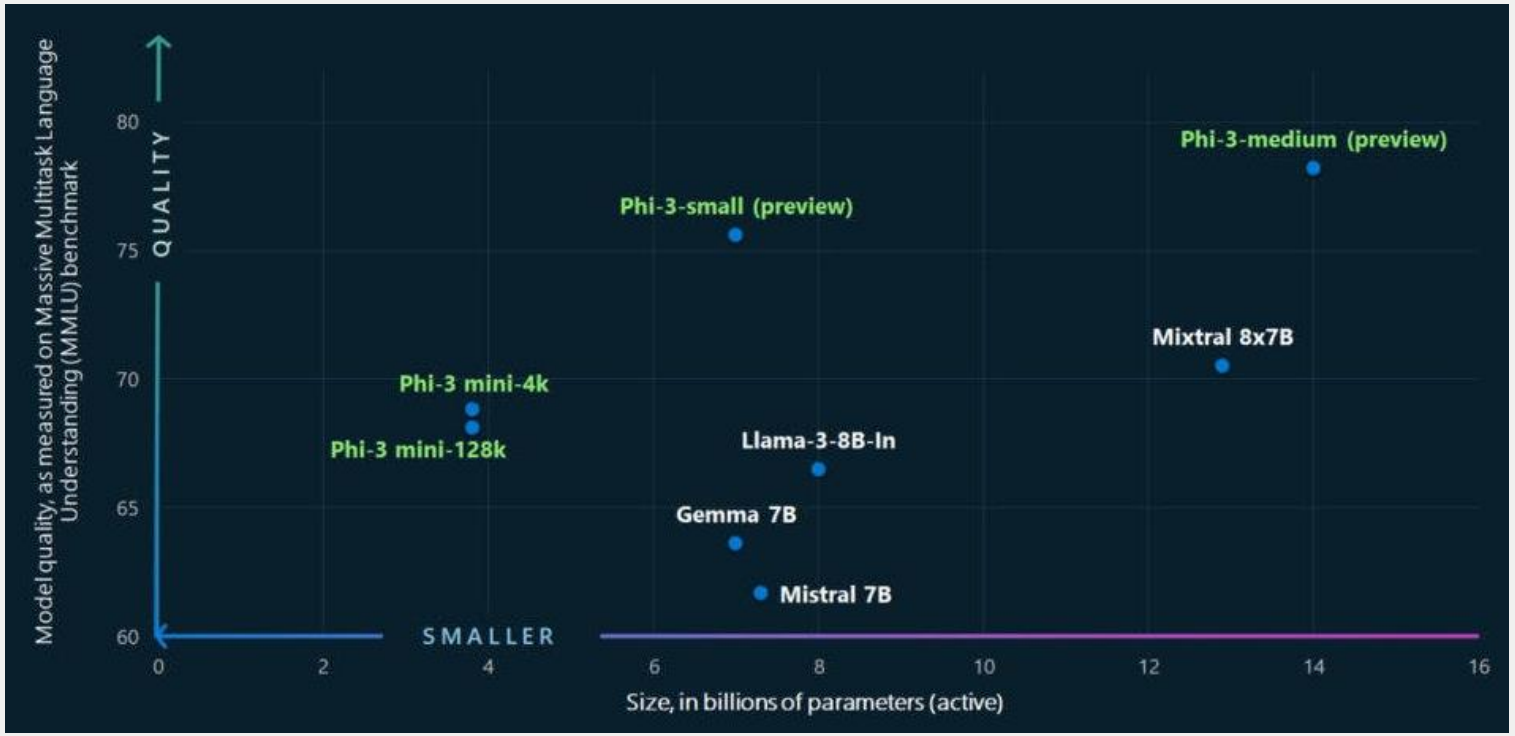
Modelos menores cada vez mais competitivos

- Os **modelos adaptados para execução em chips *on edge* estão cada vez mais competitivos**, tanto em termos de velocidade quanto de qualidade das respostas
- Pesquisadores estão sendo capazes de criar modelos baseados em um número muito inferior de parâmetros do que os LLMs, e ainda assim entregar performance robusta. Para isso, estão utilizando bases de dados menores, mas de altíssima qualidade, tendo passado por filtros e classificações rigorosas. Em essência, estão descobrindo que **a qualidade importa tanto quanto a quantidade dos dados utilizados para treinamento**



Small Language Models (SLMs)

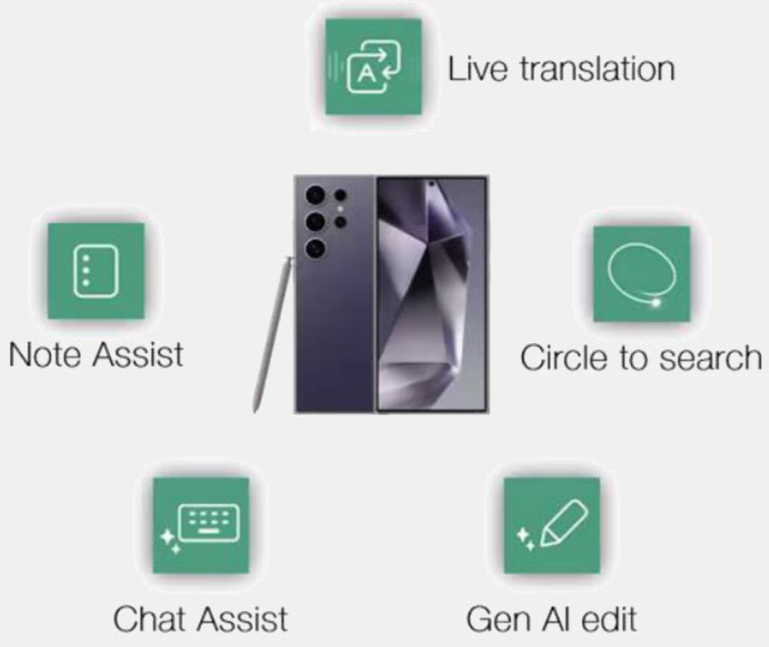
- Os *small language models* (SMLs) são modelos de *machine learning* treinados em um **conjunto de dados muito menor, mais específico e, muitas vezes, de qualidade superior** ao de um LLM. Possui muito menos parâmetros e uma arquitetura mais simples
- São normalmente **implementados para a realização de tarefas específicas** (responder perguntas em *call-centers*, resumir ligações de *leads* de vendas ou redigir e-mails de marketing)
- Podem ser mais eficientes computacionalmente e mais rápidos do que LLMs por conta do tamanho reduzido e da base de dados mais direcionada e de maior qualidade



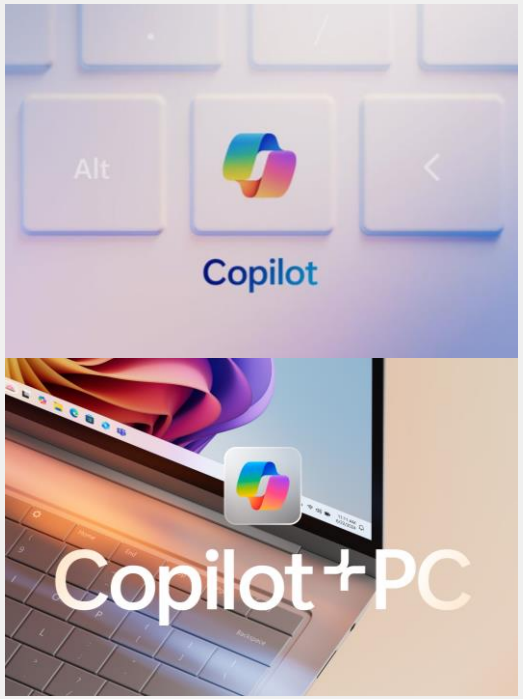
Edge AI já é uma realidade

- Com avanços tecnológicos acelerando ao longo dos últimos anos, os desafios tem sido sequencialmente superados. Hoje a *edge* AI já é uma realidade

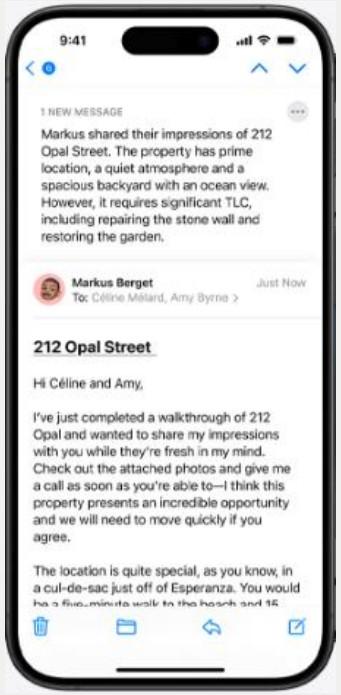
Android



Windows CoPilot+ PC



Apple Intelligence



mar asset management

Dados financeiros

Empresa	Mkt Cap	ND/EBITDA	P/E 2024	P/E 2025	P/FCF 2024	P/FCF 2025	Rev growth CAGR L5Y	Rev growth CAGR 2024	Rev growth CAGR 2025	Free Cash Flow 2023	Free Cash Flow 2024	Free Cash Flow 2025	FCF margin 2021	FCF margin 2022	FCF margin 2023	5yr ROE	5yr ROIC	TSR 1Y	TSR 5Y
Microsoft	3,334,285	0.17 x	37.95	33.51	47.78	42.27	13.9%	14.6%	14.9%	59,475	69,779	78,888	28.1%	28.7%	28.2%	42.0%	26.7%	30.3%	242.1%
Google	2,200,081	-0.49 x	23.31	20.72	27.25	23.84	17.6%	12.7%	11.9%	69,495	80,742	92,281	22.6%	23.3%	24.0%	23.4%	20.1%	43.3%	220.5%
Meta	1,291,898	-0.20 x	24.41	21.44	29.37	25.76	19.3%	17.8%	15.2%	43,847	43,990	50,155	32.5%	27.7%	28.0%	26.1%	21.1%	79.8%	169.5%
Apple	3,352,337	-0.41 x	33.06	30.88	32.38	29.16	7.6%	1.0%	3.7%	99,584	103,521	114,983	26.0%	26.7%	27.9%	124.8%	39.3%	15.5%	345.1%
Dell	102,310	1.88 x	20.93	18.34	21.42	19.24	(0.6%)	9.2%	8.4%	562	4,777	5,317	0.6%	4.9%	5.1%	n/a	12.6%	177.5%	431.1%
HP	35,174	1.56 x	10.41	10.25	11.49	10.18	(1.7%)	(0.2%)	2.0%	2,962	3,061	3,454	5.5%	5.7%	6.2%	n/a	79.5%	18.6%	102.7%
Lenovo	16,993	-30.93 x	144.19	100.86	82.74	133.91	2.2%	(1.3%)	4.0%	216	205	127	3.0%	2.9%	1.6%	27.6%	20.5%	31.6%	128.2%
Acer	4,755	-2.81 x	24.75	22.01	(41.15)	(64.64)	(0.1%)	8.3%	6.3%	384	(116)	(74)	5.2%	(1.4%)	(0.9%)	8.9%	6.8%	56.3%	238.4%

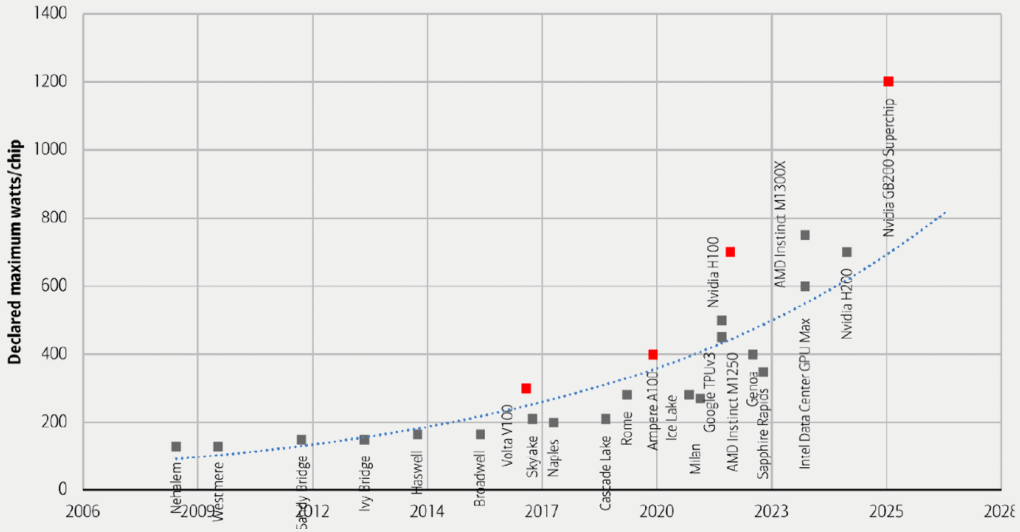
Fonte: Bloomberg (17/06/24)

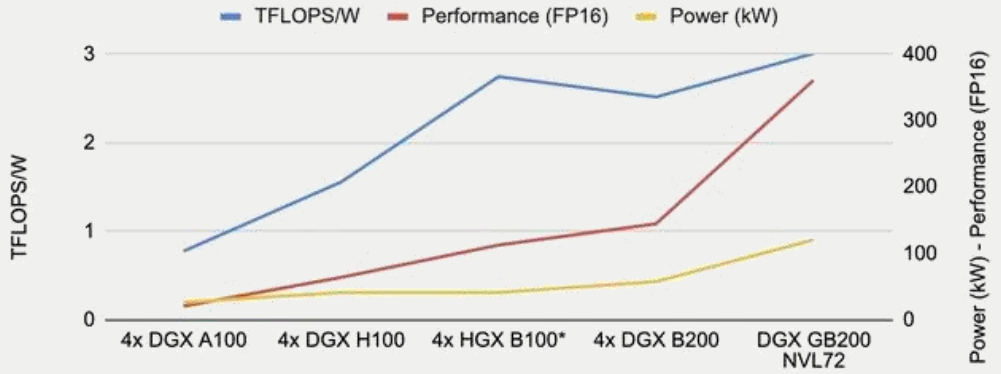
Energia

GPUs mais eficientes, mas modelos ainda maiores

- À medida que aumentam os investimentos em *data centers* para treinamento e inferência, aumenta proporcionalmente a demanda por energia elétrica para alimentar toda essa base instalada de servidores
- Apesar de as GPUs e os servidores de *accelerated computing* terem evoluído exponencialmente em termos de capacidade de processamento, o tamanho dos modelos de *machine learning* tem crescido ainda mais rápido. Ou seja, **o custo por unidade de processamento computacional cai, mas os *data centers* ficam cada vez mais *power hungry***
- A expectativa é de uma aceleração na construção de novos *data centers*, com uma pressão sobre toda a cadeia de geração e transmissão de energia, o que talvez represente um gargalo ainda maior do que o acesso aos chips

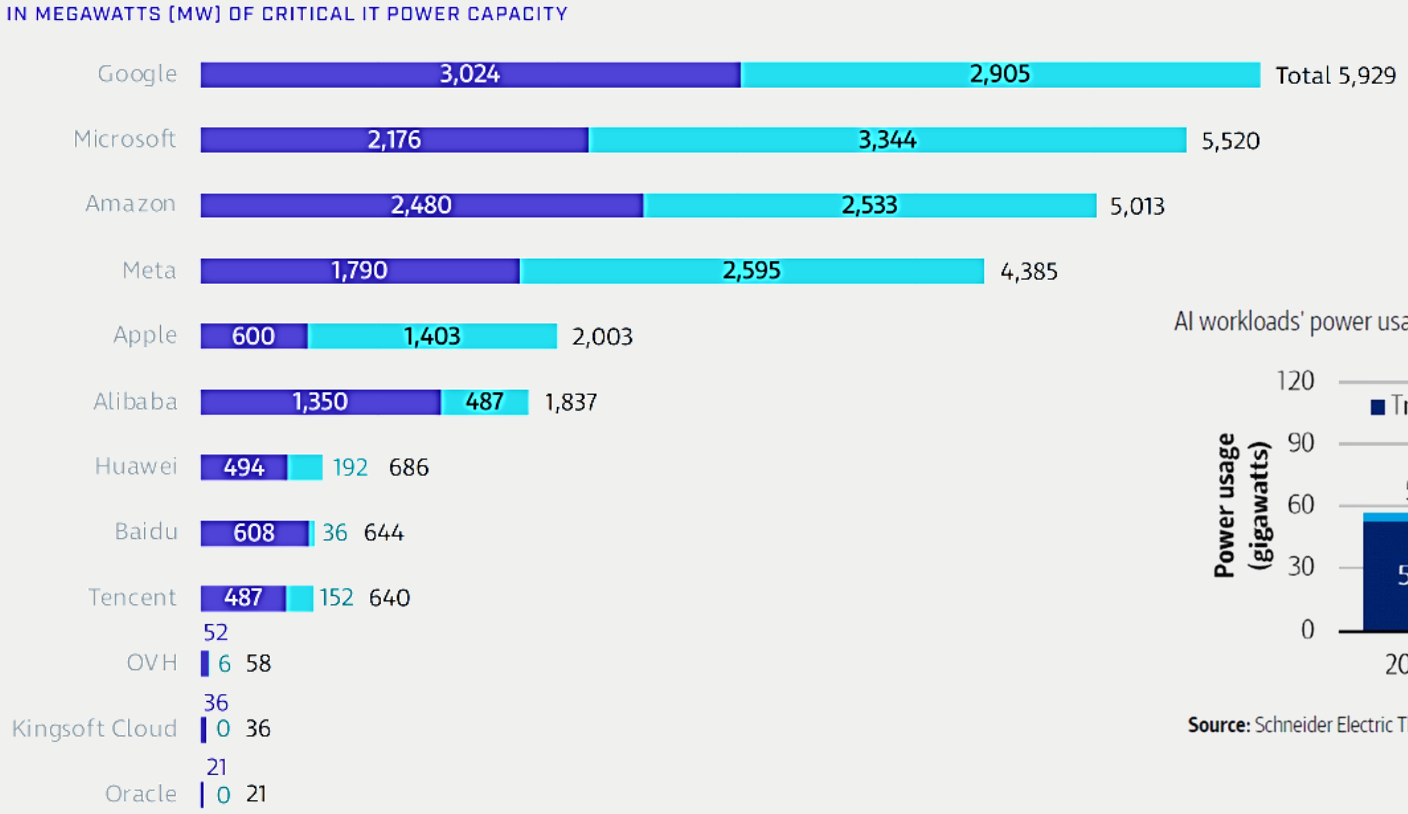
Potência declarada (W) / GPU

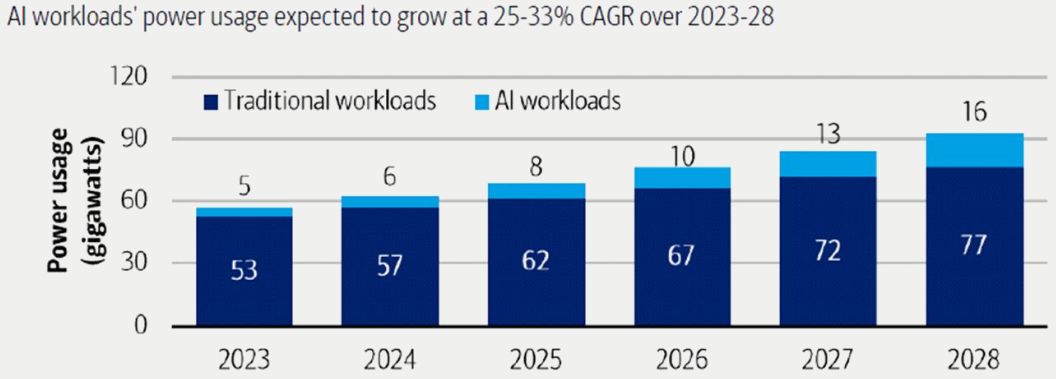




GPT-5 *Training* Analysis	Hopper GPUs - Indicative Range of Compute Increase	
GPT-5 Training Parameters (indicative range: 5-10x GPT-4)	8,800,000,000,000	17,600,000,000,000
vs. GPT-4 Parameters	1,760,000,000,000	1,760,000,000,000
GPT-5 Tokens for Training (indicative range: 5-10x GPT-4)	65,000,000,000,000	130,000,000,000,000
vs. GPT-4 Tokens	13,000,000,000,000	13,000,000,000,000
Training teraFLOPs required for GPT-4	137,280,000,000,000	137,280,000,000,000
Training teraFLOPs required for GPT-5	3,432,000,000,000,000	13,728,000,000,000,000
Multiple of Training Compute Increase: GPT-5 vs GPT-4	**25**	**100**
OpenAI Chips Used	200,000	300,000
Training exaFLOPs Available from OpenAI Chips Used (FP8)	791,600,000	1,187,400,000
Days Required to Train GPT-5	**201**	**134**

Infraestrutura das *big techs* cada vez mais *power hungry*





Source: Schneider Electric The AI Disruption: Challenges and Guidance for Data Center Design

Análises da Structure Research e da Schneider Electric indicam um crescimento de servidores voltados para AI da ordem de **30% a.a. pelos próximos 5 anos**

Schneider Electric estimate	2023	2028
Total data center workload	54 GW	90 GW
AI workload	4.3 GW	13.5-20 GW
AI workload (% of total)	8%	15-20%
AI workload (Training vs Inference)	20% Training, 80% Inference	15% Training, 85% Inference
AI workload (Central vs Edge)	95% Central, 5% Edge	50% Central, 50% Edge

Contribuindo para o aumento da demanda por energia

Historical CAGR (2013-23)	**0.4%**
Data centers	0.5%
Industrial growth	0.4%
EV adoption	0.5%
Building electrification	1.0%
Forecast CAGR (2023-30E)	**2.8%**

O consumo de energia no mercado americano cresceu apenas **0.4% a.a nos últimos 10 anos**

Com expectativa de maior eletrificação (EVs, prédios e indústrias), além dos maiores investimentos em *data centers*, a perspectiva é de uma **aceleração para 2.8% a.a.**

Source: BofA Global Research, US Energy Information Administration

Exagero ou realidade?

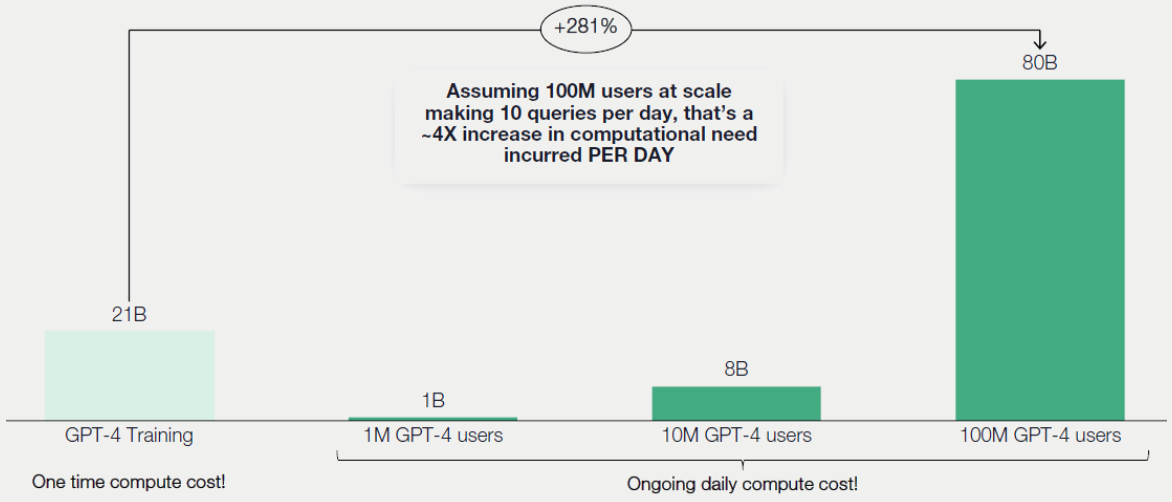
- Olhando para trás, apesar de todo o sucesso da internet e dos servidores de computação em nuvem, ganhos de eficiência permitiram que o **consumo de energia por *data center* crescesse a taxas muito inferiores à efetiva demanda computacional**. Será esse o caso ou prevalecerá o "**Paradoxo de Jevons**"?

Indicator	2015	2022	Change
Internet users	3 billion	5.3 billion	78%
Internet traffic	0.6 ZB	4.4 ZB	600%
Data center workload	180 million	800 million	340%
Data center energy use	200 TWh	240-340 TWh	20-70%

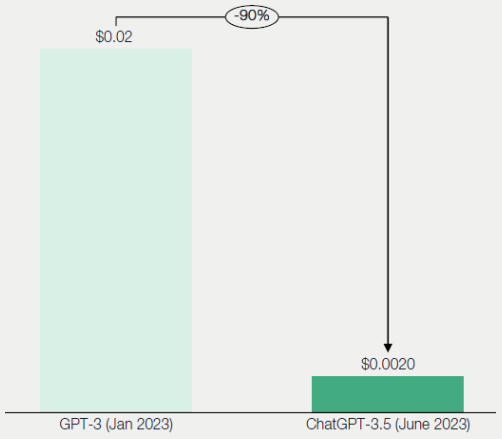
Table: Brian Potter • Source: IEA • Get the data • Created with Datawrapper

*O Paradoxo de Jevons diz que ganhos de eficiência podem ampliar o mercado endereçável e aumentar o consumo total do recurso, ao invés de reduzi-lo

Petaflops de computação (B)



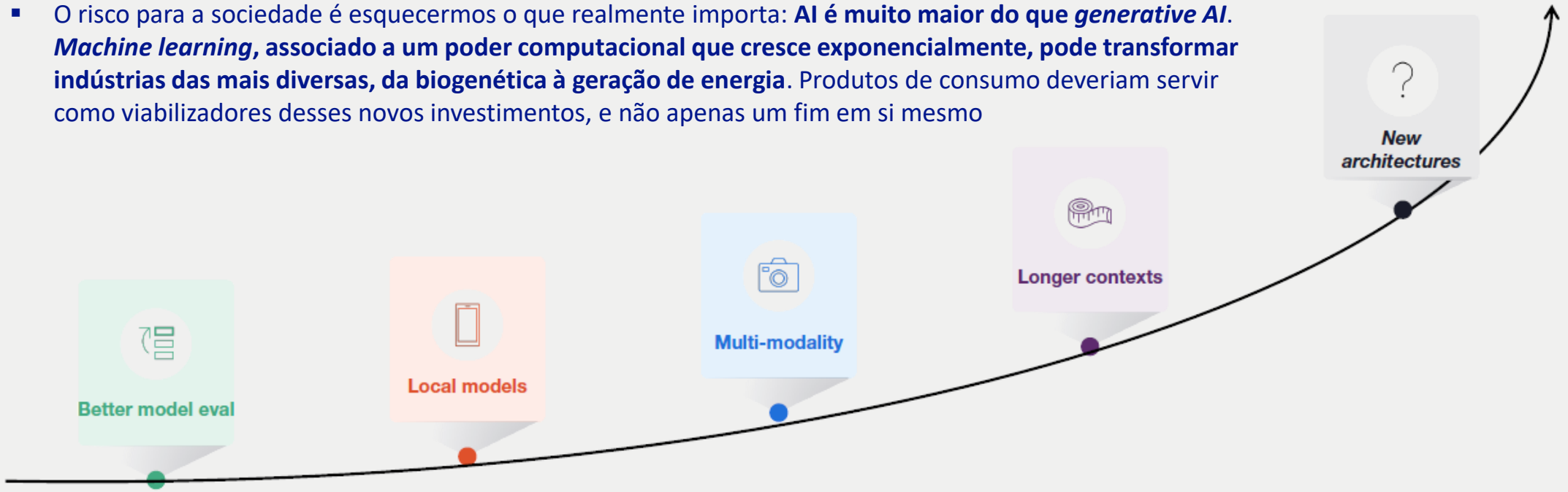
Custo por 1k tokens



Próximos passos

Próximos Passos

- Depois de décadas de pesquisas e avanços em *machine learning*, pode-se dizer que pela primeira vez entramos em um verdadeiro "**AI Summer**", quando AI efetivamente penetrou a consciência popular por meio do ChatGPT, atraindo um volume massivo de investimentos
- Apesar dos avanços, ainda há desafios diversos que precisam ser superados para que essa tecnologia alcance todo seu potencial. Em especial, lidar com as limitações dos LLMs atuais em direção a modelos mais confiáveis e eficientes
- **Ainda é cedo para um veredito se trata-se de uma bolha ou verdadeira revolução (*provavelmente os dois*),** mas a certeza é que tanto as grandes empresas quanto startups acelerarão investimentos em busca de um espaço de destaque no ecossistema desse novo paradigma tecnológico
- O risco para a sociedade é esquecermos o que realmente importa: **AI é muito maior do que *generative AI*. *Machine learning*, associado a um poder computacional que cresce exponencialmente, pode transformar indústrias das mais diversas, da biogenética à geração de energia**. Produtos de consumo deveriam servir como viabilizadores desses novos investimentos, e não apenas um fim em si mesmo



mar asset management

Relação com Investidores:

Igor Galvão

55 21 99462 3359

igalvao@marasset.com.br

rio de janeiro – rj • av. ataulfo de paiva 1351, 3º andar, leblon • 22440 034

marasset.com.br