

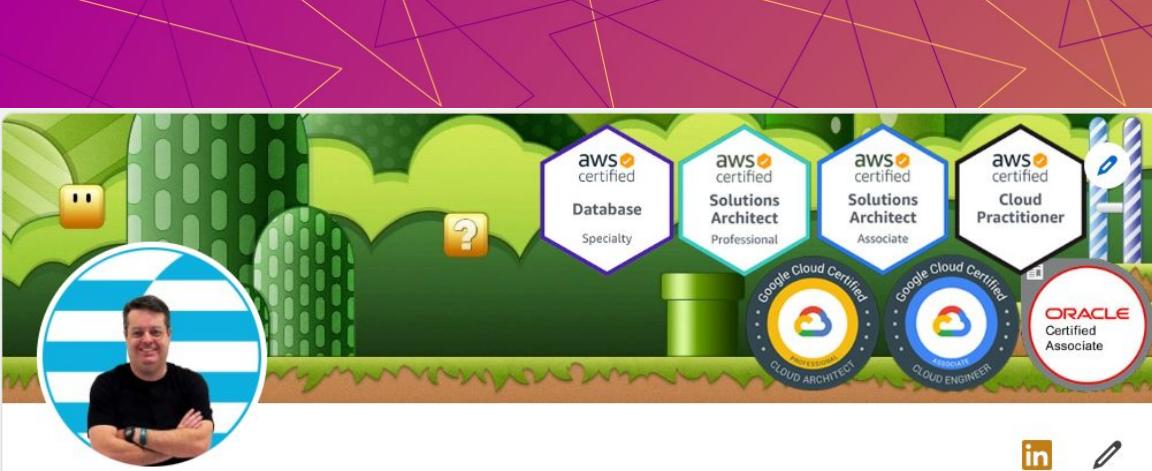
AGILE TREN[⚡]

NORDESTE

O MAIOR EVENTO ÁGIL DO BRASIL
AGORA NO NORDESTE

10 - 13 JUNHO RECIFE

Como as LLM's funcionam



Luis Gustavo (Gus) Amaral ✅

CTO | Head de Tecnologia | Cloud | Analytics | MENSA | AWS (4x) GCP
(2x) Certified Cloud Architect

São Paulo, São Paulo, Brazil · [Contact info](#)

Experience

 Maitha Tech
Full-time · 2 yrs 3 mos

Chief Technology Officer
Jul 2024 - Present · 11 mos

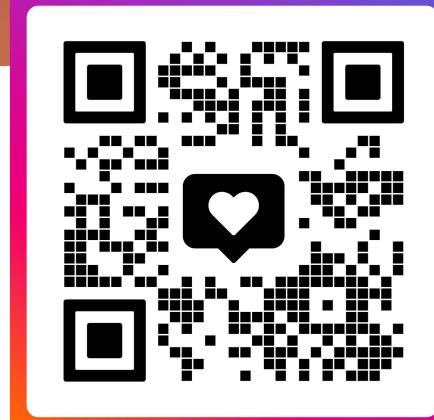
ATUAÇÕES:

- Condução de Due Diligences técnicos de empresas com objetivo de M&A, para clientes externos e i ...see more

Resolução de problemas, Segurança da informação and +5 skills

Head de Tecnologia

Mar 2023 - Jun 2024 · 1 yr 4 mos

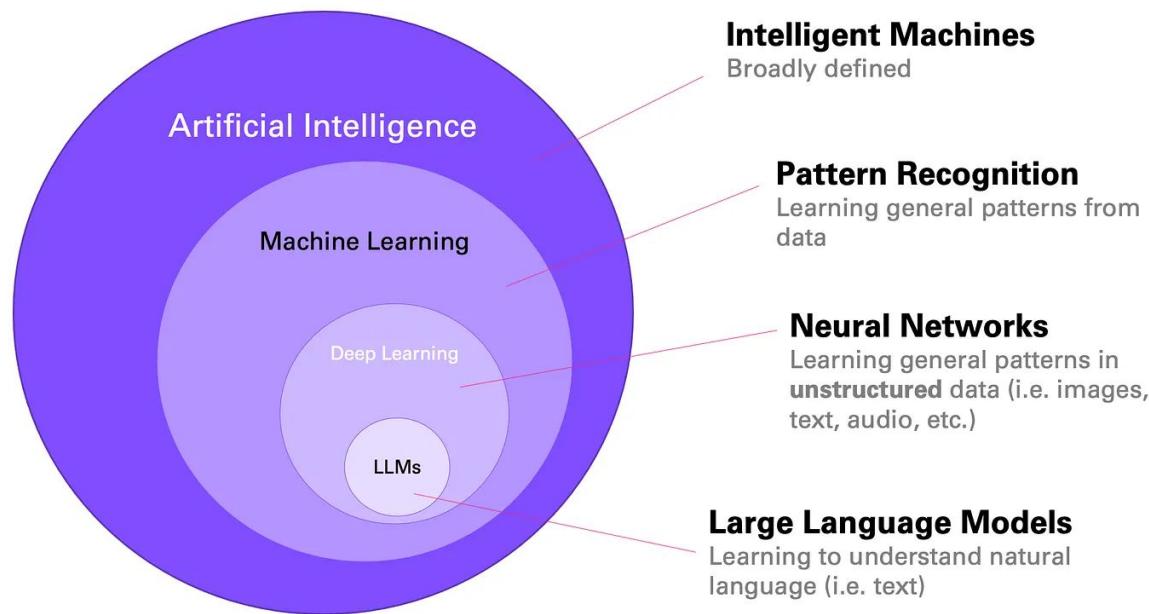


Instagram



O Que São LLMs?

- Modelos de IA treinados para processar e gerar texto.
- Conceito de "grande escala" (enorme quantidade de dados e parâmetros).



LLM 'puras' e Produtos AI

LLMs (como tecnologia):

- São **modelos de base**, geralmente ajustados para tarefas específicas.
- Requerem infraestrutura para serem usadas (como servidores e scripts de aplicação).
- **Não têm conectividade externa** ou acesso dinâmico a dados em tempo real.



Produtos (como ChatGPT):

- Oferecem uma **solução integrada** que vai além da LLM.
- Podem realizar tarefas que envolvem acesso a dados externos (ex.: busca na web).
- Têm uma **interface amigável** e funcionalidades adicionais.

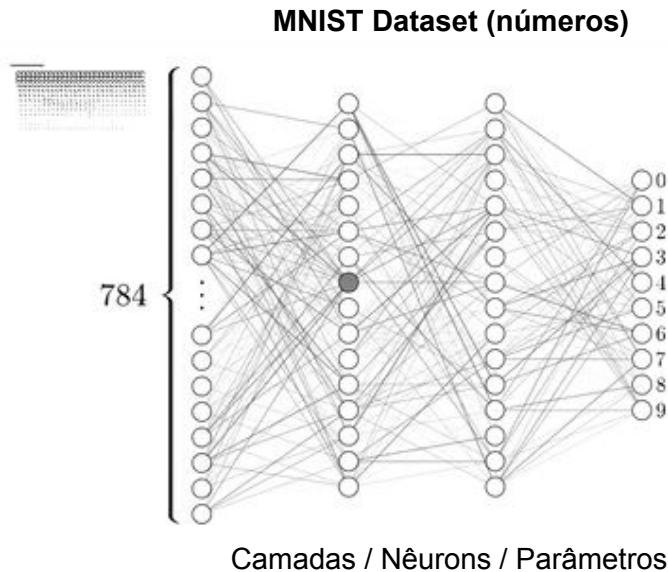


Redes Neurais: a base computacional

São modelos computacionais inspirados no cérebro humano, compostos por “neurônios” artificiais organizados em camadas (entrada, ocultas e saída). Elas aprendem padrões a partir de dados ajustando os pesos das conexões por meio de algoritmos como o **backpropagation**, sendo úteis em tarefas como reconhecimento de imagens, tradução e classificação.

Esses sistemas se destacam por lidar com problemas complexos e não lineares, onde métodos tradicionais falham.

Durante o **treinamento**, a rede otimiza seus parâmetros para minimizar erros, **tornando-se capaz de generalizar** e realizar previsões baseadas em novos dados.

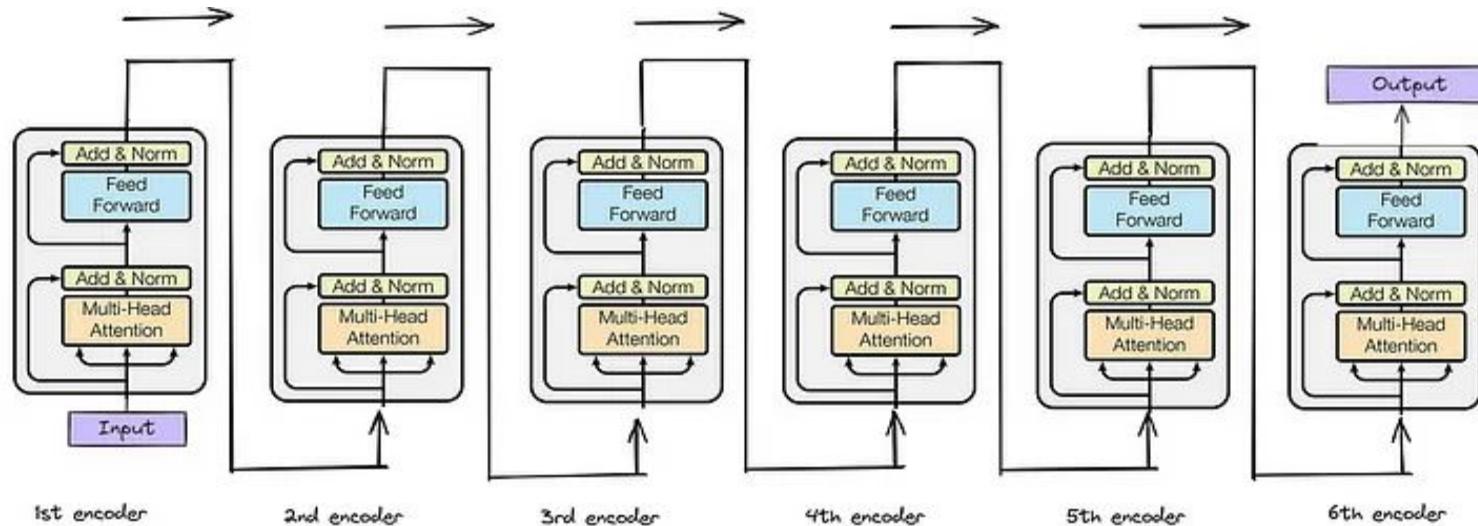


Deep Learning: Redes Neurais Profundas

Deep Neural Networks (DNNs) são redes neurais com **muitas camadas ocultas**, o que permite aprender representações hierárquicas mais complexas. Esse design captura padrões simples nas primeiras camadas e abstratos nas mais profundas, sendo ideal para tarefas como visão computacional e processamento de linguagem natural.

GPT 3: 96 camadas, ~ 12200 “neurônios”, 175.000.000.000 parâmetros (175B!)

Exemplos incluem **CNNs**, para imagens, e modelos baseados no **Transformer**, como o GPT



Datasets: Entra Porco, sai Linguiça

Corpora Gerais para Pré-Treinamento:

- **LAION-2B-en**: Conjunto colossal com 260 bilhões de palavras e 13 bilhões de páginas da web.
- **CCAW (Common Crawl Archive Web)**: Coletado em parceria entre Google AI, Facebook AI e outros.
- **WikiText-103**: 103 GB de textos da Wikipédia.
- **BookCorpus**: Mais de 10 mil obras literárias em

Conjuntos de Dados Específicos por Domínio:

- **PubMed Abstracts**: Resumos de artigos biomédicos.
- **CheXpert**: Radiografias de tórax com laudos médicos.
- **WebText (The Pile)**: Textos e códigos de conversas online.
- **SQuAD**: Conjunto de perguntas e respostas com base na Wikipédia.

Conjuntos para Conversação:

- **OpenSubtitles**: Legendas de filmes em vários idiomas.
- **DailyDialog**: Diálogos informais multi-turno, colhidos de fóruns online.
- **Switchboard**: Transcrições de conversas telefônicas.

Conjuntos de Dados de Código:

- **GitHub**: Repositório com códigos de diversas linguagens.
- **JavaBert**: Códigos em Java para análise e geração automática.
- **Starcoder**: 783 GB de código em 86 linguagens de programação.



LLMDataHub

Datasets for LLM Training

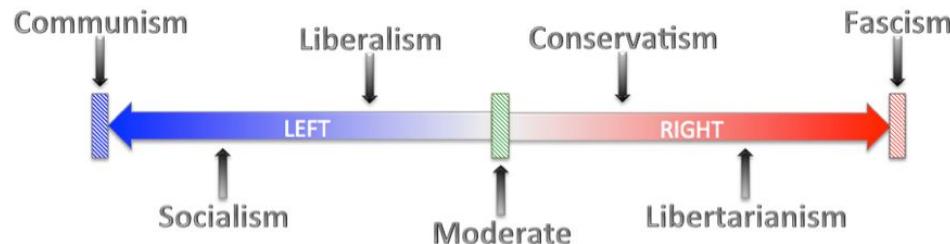
Multidimensionalidade

A **multidimensionalidade** nos *token encodings* dos LLMs (Modelos de Linguagem de Grande Escala) permite representar palavras e conceitos em vetores de alta dimensão, onde relações semânticas são traduzidas em padrões matemáticos. Nesse espaço vetorial, palavras com significados semelhantes ficam próximas, enquanto relações mais complexas — como analogias ou contextos — emergem como direções ou distâncias específicas.

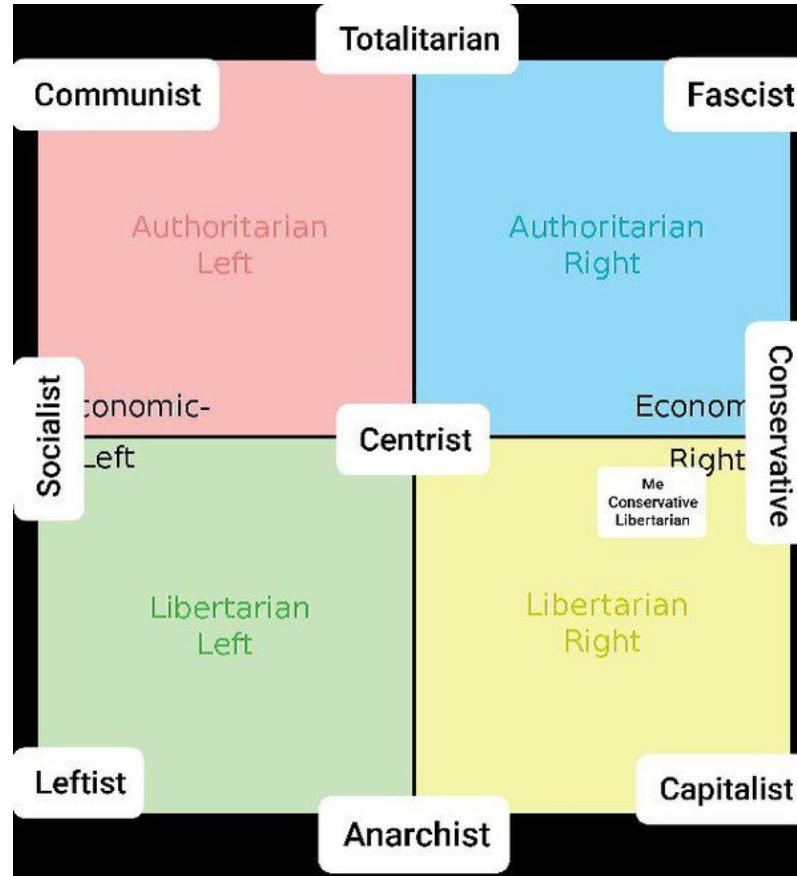
Esse mapeamento vetorial possibilita que os LLMs compreendam **nuances da linguagem** e do conhecimento de forma **quantitativa**.

Assim, semântica, contexto e estrutura gramatical **deixam de ser apenas abstrações humanas** e passam a ser captadas e manipuladas com precisão matemática.

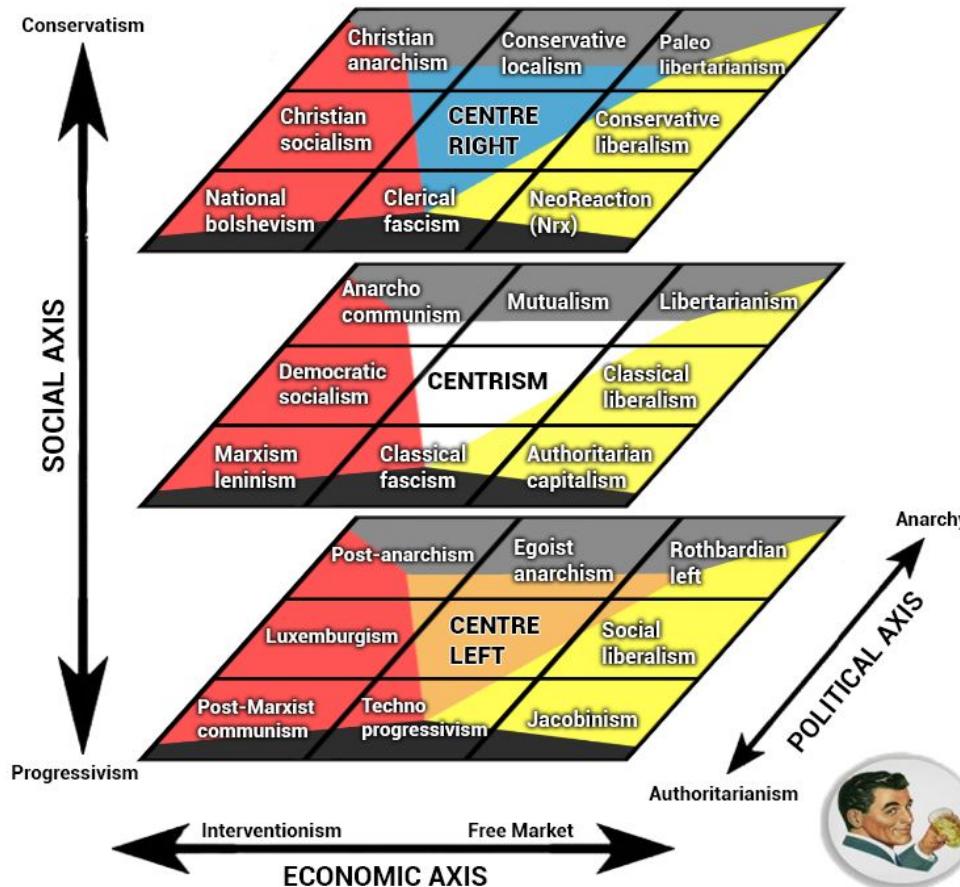
1 Dimensão



2 Dimensões

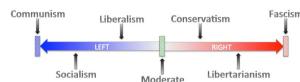


3 Dimensões



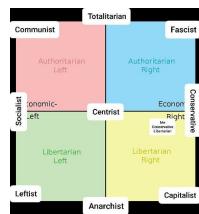
N* Dimensões

Um número maior de dimensões consegue capturar as nuances semânticas, estruturais e de significados da linguagem.



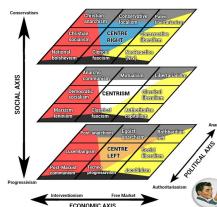
D=1

Ideol1 Ideol2
[-1] [1]



D=2

Ideol1 Ideol2
[-1] [1]
[0.8] [0.6]



D=3

Ideol1 Ideol2
[-1] [1]
[0.8] [0.6]
[-0.95][0.1]



GPT-3

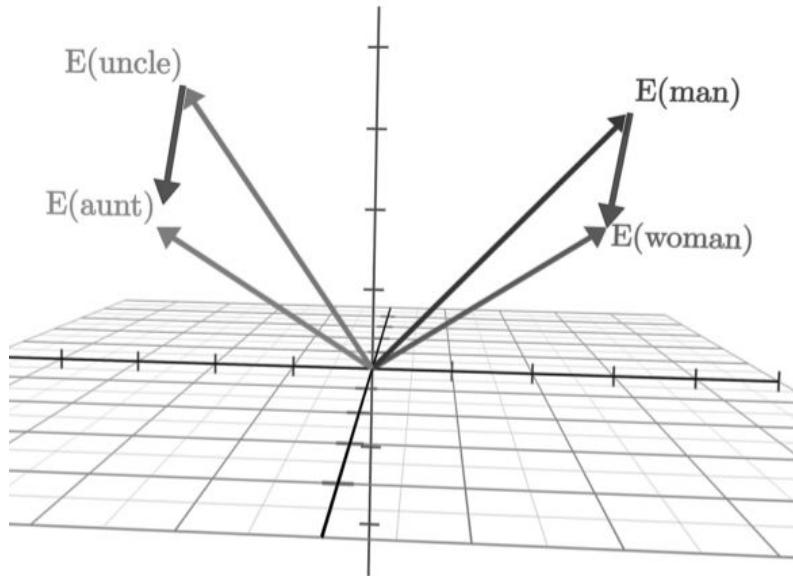
D=1280

Ideol1 Ideol2
[-1] [1]
[0.8] [0.6]
[-0.95][0.1]
[3] [-4]
[2.1] [7]
[0.01] [0.2]
[-0.5] [0.3]

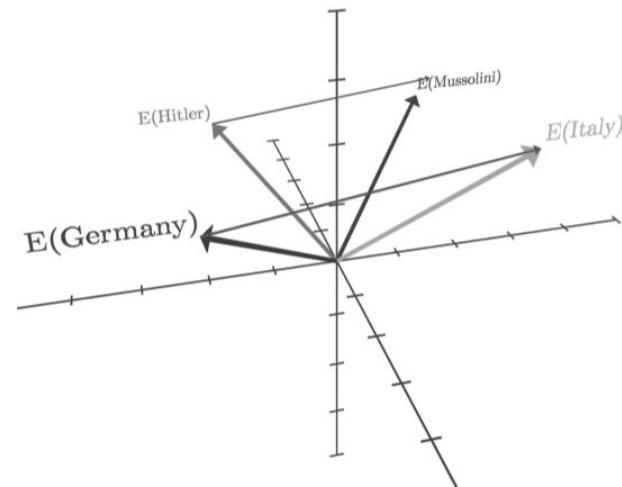
... ∞

Relações N-Dimensionais

$$E(\text{aunt}) - E(\text{uncle}) \approx E(\text{woman}) - E(\text{man})$$



$$E(\text{Hitler}) + E(\text{Italy}) - E(\text{Germany}) \\ \approx E(\text{Mussolini})$$



Predição de Tokens:

I Like



Prompt

Machine Learning: Força Bruta e Escala

O treinamento de modelos de linguagem de grande porte (LLMs) depende de *machine learning*, especialmente redes neurais profundas, que ajustam bilhões de parâmetros processando enormes volumes de dados.

Isso exige cálculos intensivos e força bruta computacional, utilizando GPUs ou TPUs para executar algoritmos de otimização como o gradiente descendente.

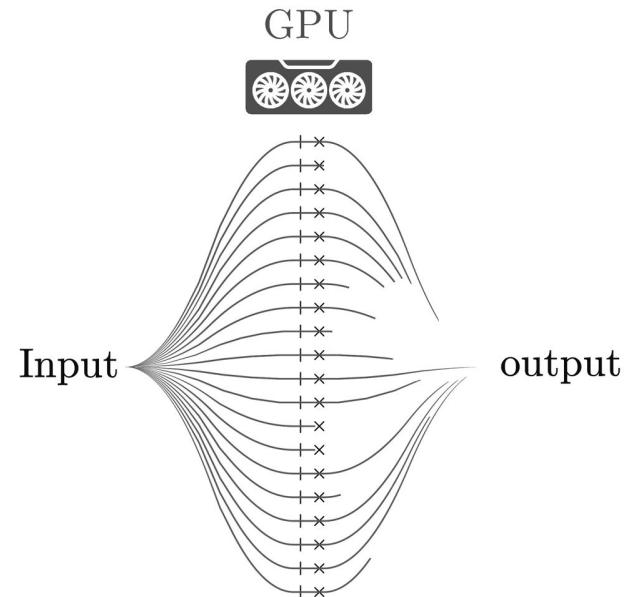
1 Bilhão de Operações por Segundo

$$\begin{aligned}
 4,449.890 + 5,962.869 &= 10,412.759 \\
 3,170.468 \times 2,421.072 &= 7,675,931.034 \\
 1,893.913 + 1,867.295 &= 3,761.209 \\
 4,765.310 \times 898.244 &= 4,280,409.195 \\
 8,767.704 \times 9,233.810 &= 80,959,311.576 \\
 8,981.731 + 9,230.824 &= 18,212.556 \\
 8,378.652 \times 5,245.739 &= 43,952,226.479 \\
 8,494.860 + 8,950.390 &= 17,445.249 \\
 5,174.100 \times 8,144.609 &= 41,587.248.189 \\
 9,638.385 + 6,031.856 &= 15,670.242
 \end{aligned}$$

1 Bilhão de Operações por Segundo

1 Minuto
 1 Hora
 1 Dia
 1 Mês
 1 Ano

100.000.000 Anos para treinar GPT-3.5



Hardware: Força Bruta e Escala



Datacenters Cloud Comuns

- **Hardware:** CPUs, cargas gerais, como hospedagem de sites, bancos de dados e VMs.
- **Rede:** Conectividade padrão com redes de alta velocidade (10–100 Gbps)
- **Armazenamento:** Focado em armazenamento escalável e econômico (HDDs e SSDs), otimizados para leitura/escrita gerais.
- **Energia e Resfriamento:** Consumo energético moderado, com sistemas de resfriamento tradicionais.
- **Escalabilidade:** Escalabilidade horizontal (adicionando mais servidores) para diversos tipos de cargas de trabalho.
- **Uso:** Projetado para cargas genéricas e serviços como VMs, bancos de dados e APIs.

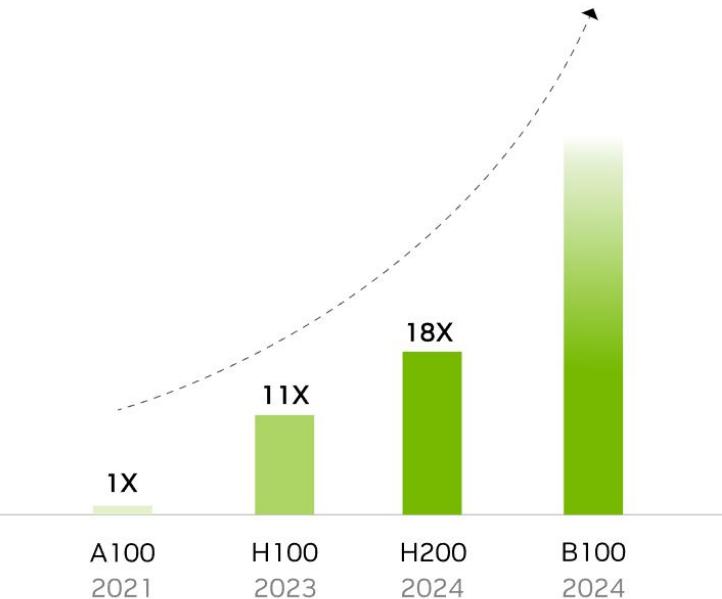


Datacenters de IA

- **Hardware:** Requer GPUs de alto desempenho (NVIDIA A100/H100, AMD MI250, etc.) ou TPUs especializadas para treinamento e inferência de IA.
- **Rede:** Redes ultrarrápidas e de baixa latência (interconexões de 400 Gbps ou mais)
- **Armazenamento:** Necessita de armazenamento SSD NVMe de altíssima velocidade
- **Energia e Resfriamento:** Demanda enorme de energia e sistemas avançados de resfriamento (como líquidos ou imersão) devido ao alto consumo térmico de GPUs.
- **Escalabilidade:** Escalabilidade vertical e horizontal, com clusters de GPUs interligadas para processar modelos massivos.
- **Uso:** Otimizado para treinamento e inferência de IA

Exemplo: NVIDIA H100 / B200

GPT-3 175B Inference Performance



H100



NVLINK



B200



MSRP* STARTING AT

\$46,570

2025 RAM 3500 LIMITED CREW CAB 4X4

<https://indianexpress.com/article/technology/artificial-intelligence/nvidias-blackwell-ai-chip-price-30000-to-40000-us-dollars-9223871/>

Top Players: Estimativa de Capacidade

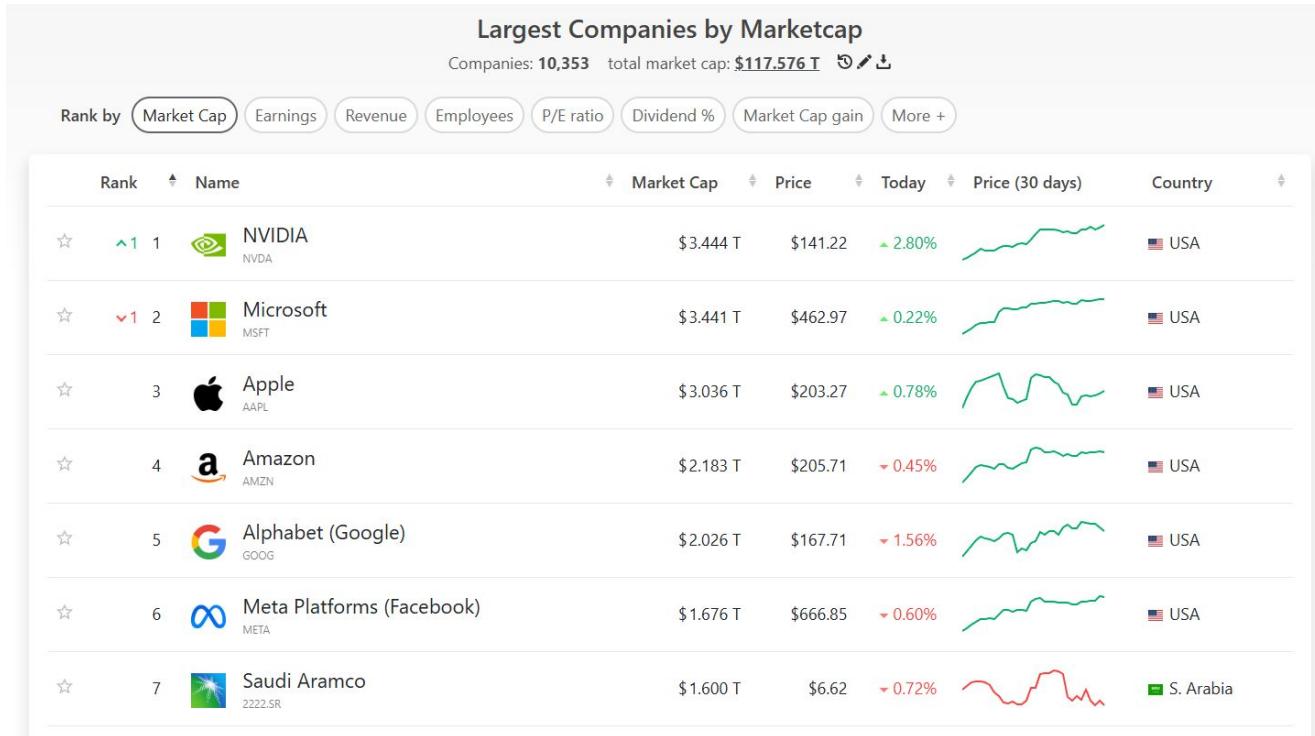
Summary of estimated chip counts [11]

	2024 YE (H100 equivalent)	2025 (GB200)	2025YE (H100 equivalent)
MSFT	750k-900k	800k-1m	2.5m-3.1m
GOOG	1m-1.5m	400k	3.5m-4.2m
META	550k -650k	650k-800k	1.9m-2.5m
AMZN	250k-400k	360k	1.3m-1.6m
XAI	~100k	200k-400k	550k-1m

NVIDIA: Em Ascensão



NVIDIA: Distorções





Key Points:

Multidimensionalidade: Entra Porco, sai Linguiça

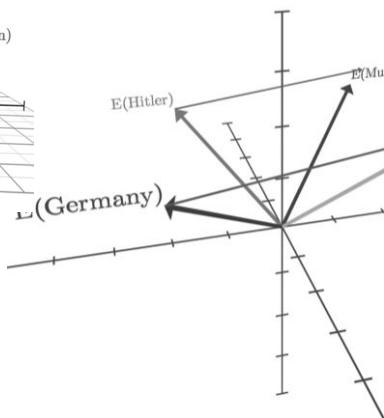
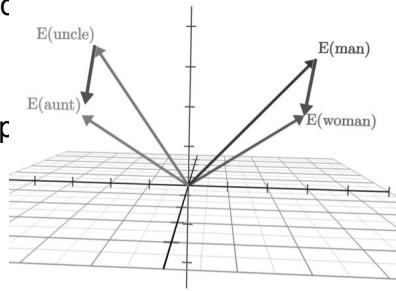
$$E(\text{Hitler}) + E(\text{Italy}) - E(\text{Germany})$$

-
- **Slide 8: A Arquitetura Transformer**

- Introdução ao Transformer, a base dos LLMs modernos
- Componentes principais:
 - **Mecanismo de atenção** (self-attention).
 - Embeddings e representação vetorial de palavras.
- Diagrama simplificado de um Transformer.

$$E(\text{aunt}) - E(\text{uncle}) \approx E(\text{woman}) - E(\text{man})$$

$$\approx E(\text{Mussolini})$$



Olá

7. Futuro dos LLMs

- **Slide 17: Tendências**

- Modelos mais eficientes e sustentáveis (menores com resultados semelhantes aos grandes).
- Personalização de LLMs para aplicações específicas.
- Integração com outras áreas, como visão computacional.

- **Slide 18: O Que Esperar?**

- Previsões:
 - LLMs mais acessíveis e economicamente viáveis.
 - Maior regulamentação e foco em ética.



Olá

8. Conclusão

- **Slide 19: Resumo**
 - Recapitação dos principais pontos:
 1. O que são LLMs.
 2. Como funcionam.
 3. Desafios e impacto.
- **Slide 20: Mensagem Final**
 - "LLMs não são caixas mágicas, mas uma combinação de ciência, engenharia e investimento. Entender seus bastidores é essencial para aproveitarmos seu potencial de maneira ética e sustentável."
- **Slide 21: Perguntas?**
 - Convide o público a fazer perguntas ou compartilhar reflexões.

Olá

Dicas Adicionais

1. Use Visuais de Qualidade:

- Diagramas claros para explicar redes neurais e o mecanismo de atenção.
- Fotos ou gráficos para ilustrar datacenters, GPUs, e consumo de energia.

2. Simplifique o Complexo:

- Evite jargões excessivos, mas mantenha a profundidade técnica.

3. Engaje o PÚBLICO:

- Faça perguntas retóricas ou use curiosidades para capturar a atenção.

4. Referências e Fontes:

- No último slide, inclua referências confiáveis para quem quiser se aprofundar no tema.

Olá

Conceito de Multidimensionalidade

- 1D, 2D, 3D: Um Começo Intuitivo
 - Comece com algo familiar para o público:
 - 1 Dimensão (1D): Um ponto em uma linha.
 - 2 Dimensões (2D): Um ponto em um plano (ex.: um quadrado ou gráfico cartesiano).
 - 3 Dimensões (3D): Um ponto no espaço (ex.: um cubo ou esfera).
- Mais de 3 Dimensões?
 - Explique que, embora não possamos visualizar diretamente dimensões acima de 3D, a matemática permite representá-las.
 - Nos LLMs, palavras, frases e contextos são representados como **vetores em espaços multidimensionais** (geralmente com **milhares ou milhões de dimensões**).

Evolução Visual da Multidimensionalidade

Slide 1: 1 Dimensão

- Um ponto em uma linha reta.

Olá

4. Os Bastidores Computacionais

- **Slide 9: Treinamento de LLMs**
 - Processo de treinamento:
 - Ajuste de bilhões (ou trilhões) de parâmetros.
 - Ciclos de retropropagação (backpropagation).
 - Visual: gráfico mostrando o "loop" de treinamento.
- **Slide 10: Hardware Necessário**
 - Datacenters massivos, GPUs de última geração, memórias rápidas.
 - Exemplo visual: infraestrutura de um datacenter.
 - Dados: consumo de energia e custo estimado (ex.: GPT-3 custou milhões para treinar).
- **Slide 11: Paralelismo**
 - Como o paralelismo acelera o treinamento:
 - Divisão de dados em múltiplos processadores.
 - Técnicas como **pipeline parallelism** e **model parallelism**.
 - Diagrama para ilustrar o conceito.

Olá

5. Desafios e Limitações

- **Slide 12: Custos Computacionais**
 - Alto consumo de energia elétrica e impacto ambiental.
 - Gráficos comparando a energia usada por diferentes modelos.
- **Slide 13: Limitações Técnicas**
 - Dependência de grandes quantidades de dados rotulados.
 - Problemas de viés nos dados.
 - Exemplos de erros comuns de LLMs (alucinações, respostas incorretas).
- **Slide 14: Ética e Sustentabilidade**
 - Discussão sobre os desafios éticos no uso de LLMs:
 - Privacidade de dados.
 - Viés algorítmico.
 - Sustentabilidade ambiental.





AGILE
TREND
NORDESTE



AGILE
TREND
NORDESTE

AGILE TREND

NORDESTE

O MAIOR EVENTO ÁGIL DO BRASIL
AGORA NO NORDESTE

10 - 13 JUNHO

RECIFE

AGILE TREND

NORDESTE

O MAIOR EVENTO ÁGIL DO BRASIL
AGORA NO NORDESTE

10 - 13 JUNHO

RECIFE