

Inteligência Artificial Generativa

Silvan Ferreira

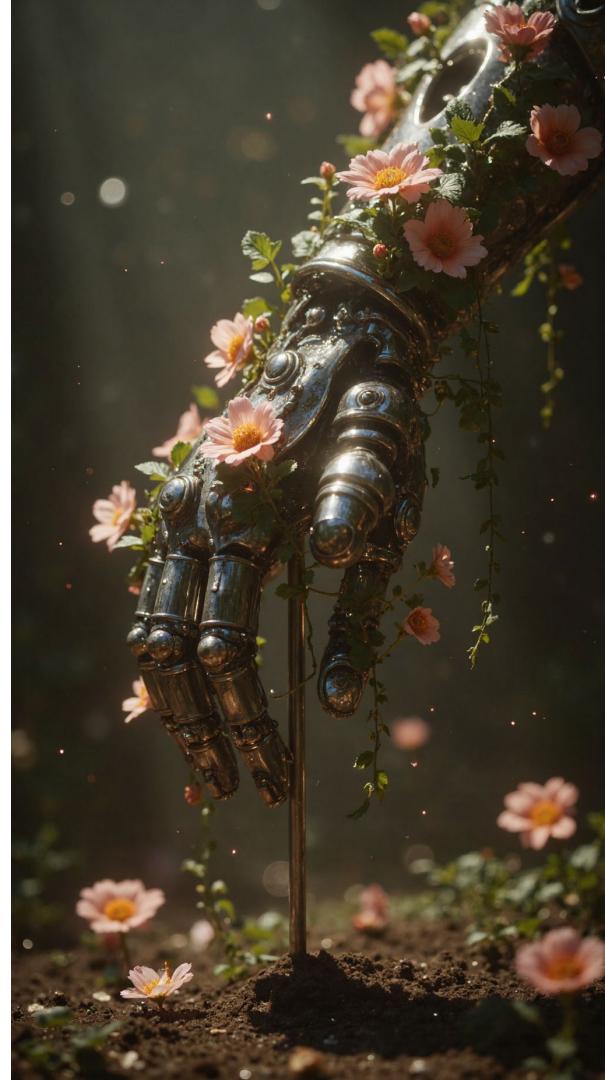
Apresentação

Silvan Ferreira

- Formação
 - Engenharia Elétrica (UFERSA)
 - Mestrado em EEC (UFRN)
 - Doutorado em EEC (UFRN)
- Linhas de Pesquisa
 - IA Neuro-Simbólica
 - Detecção de Objetos Multimodal
- Indústria
 - Tech Lead / Senior Software Engineer (CESAR)
 - Data Scientist (Daitrix)

Agenda da Disciplina

- **Unidade I:**
 - Fundamentos e Matemática
 - Autoregressive Models (ARMs)
 - Modelos de Linguagem (LMs)
- **Unidade II:**
 - Normalizing Flows (NFs)
 - Generative Adversarial Networks (GANs)
 - Variational Autoencoders (VAEs)
- **Unidade III:**
 - Diffusion Models (DDPMs)
 - Energy-Based Models (EBMs)

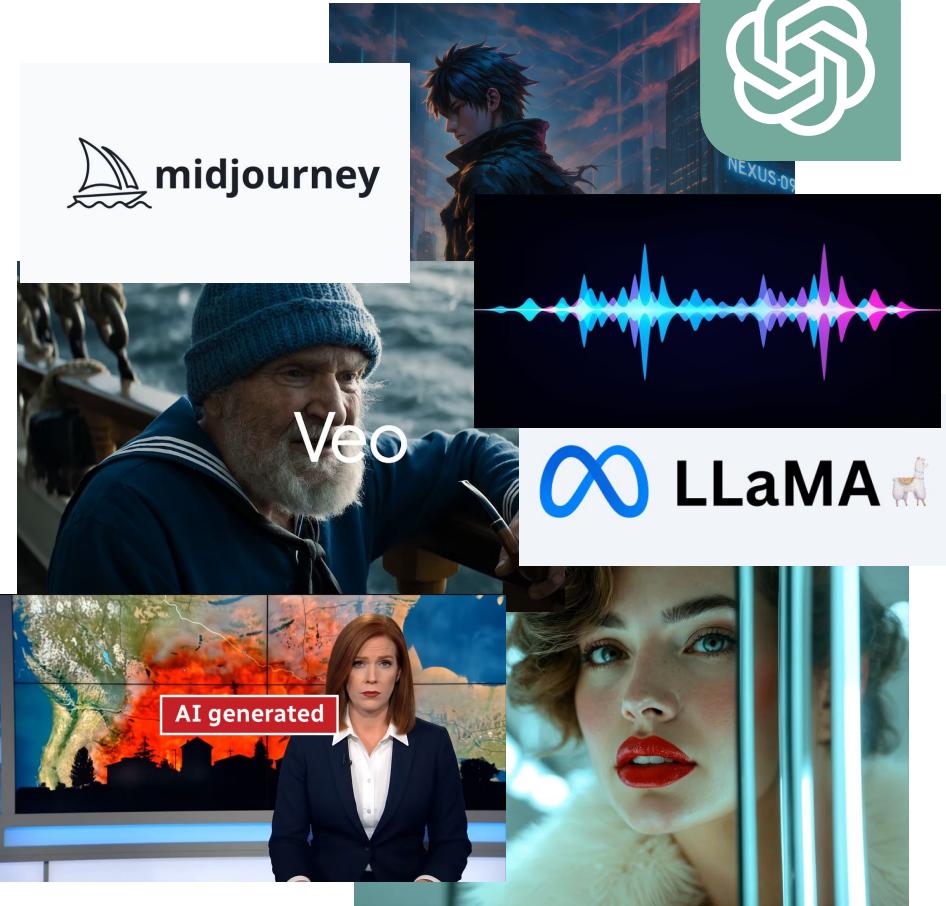


Requisitos

- 1. Probabilidade e Estatística**
- 2. Cálculo Diferencial**
- 3. Álgebra Linear**
- 4. Deep Learning**
- 5. Python**

Introdução

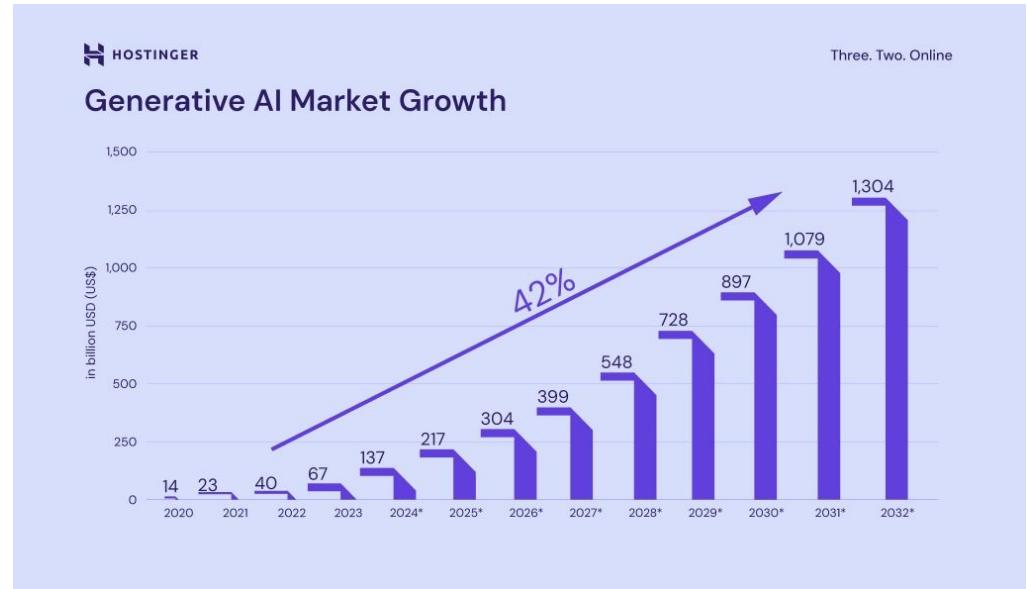
- **Definição:** Modelos capazes de aprender distribuições de dados e gerar novas instâncias com características estatísticas semelhantes
- **Modalidades:** texto, imagens, áudio, vídeo, modelos 3D, código-fonte e dados multimodais
- **Fundamentos:**
 - Modelagem probabilística
 - Aprendizado profundo
 - Otimização de funções ou divergência entre distribuições



 Claude

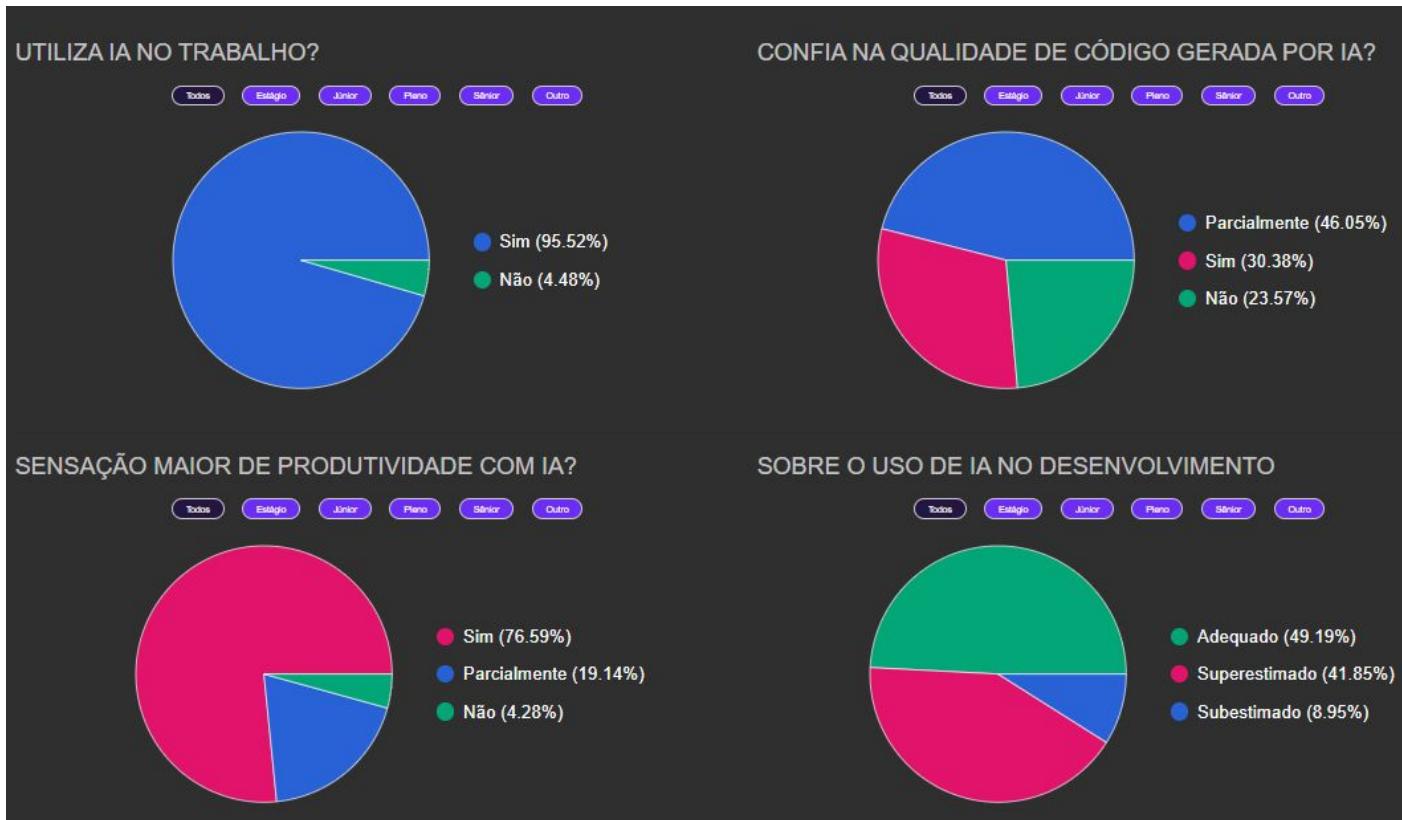
Impacto

- 2023: salto de US\$ 67 bi para US\$ 137 bi (+104%) com adoção em massa de LLMs e ferramentas generativas
- Expansão rápida para marketing, educação, design, engenharia e atendimento
- Novos modelos de negócio: APIs, white-label, soluções verticais, experiências humano-IA
- IA generativa já é camada crítica da economia digital

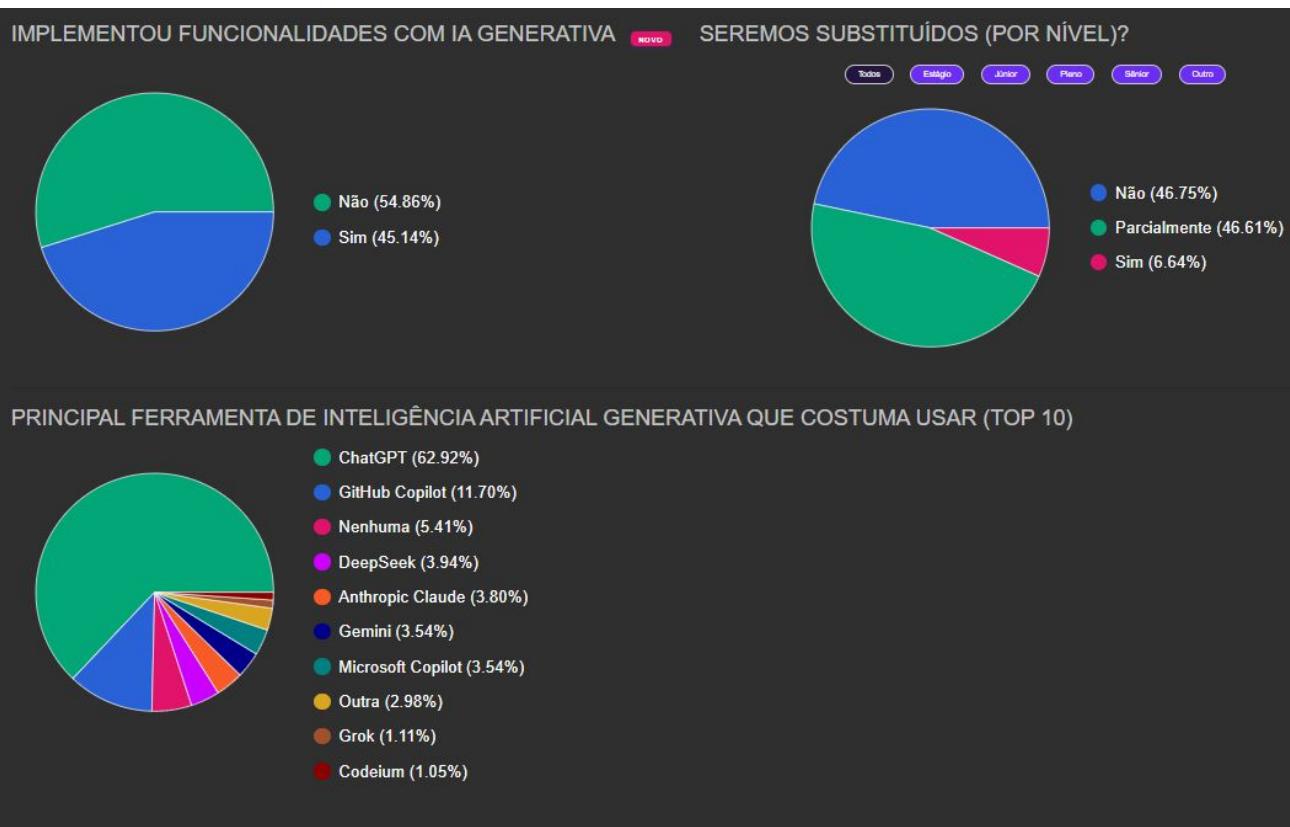


Impacto

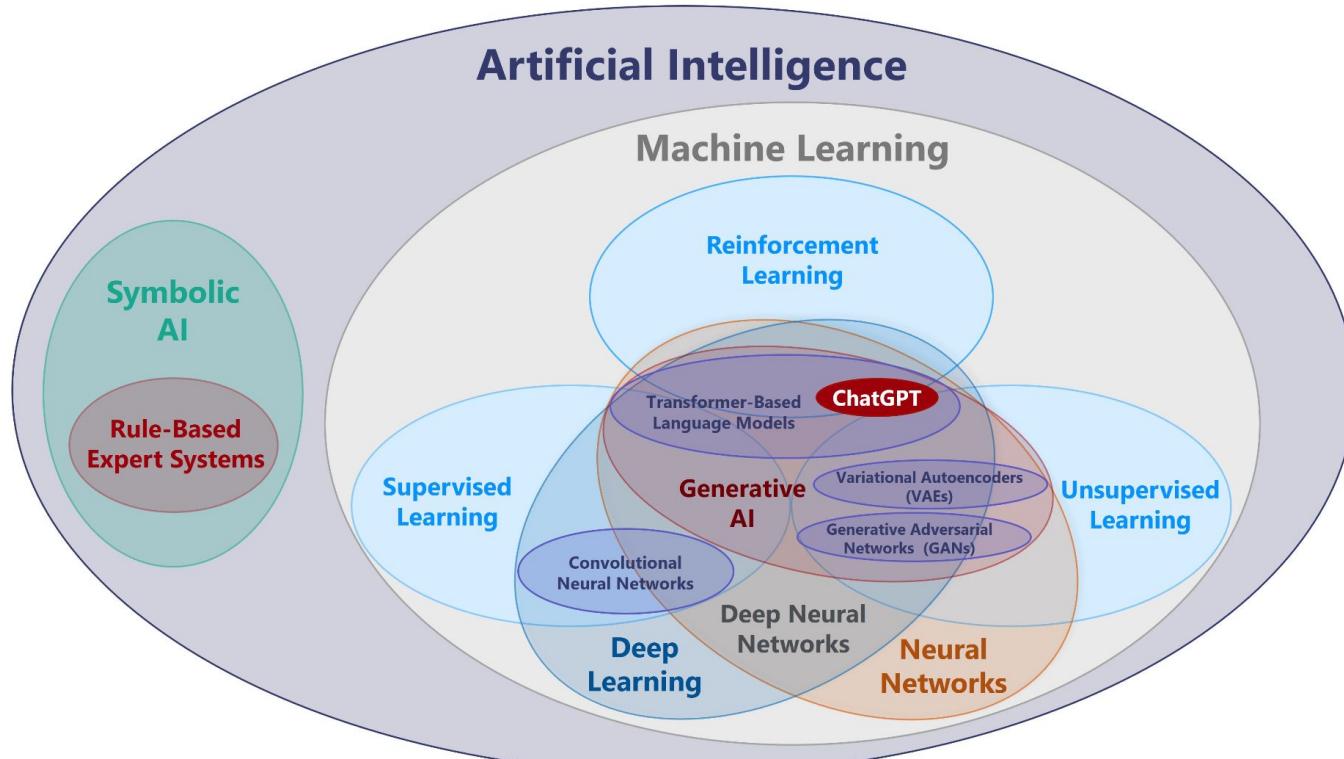
Código Fonte TV



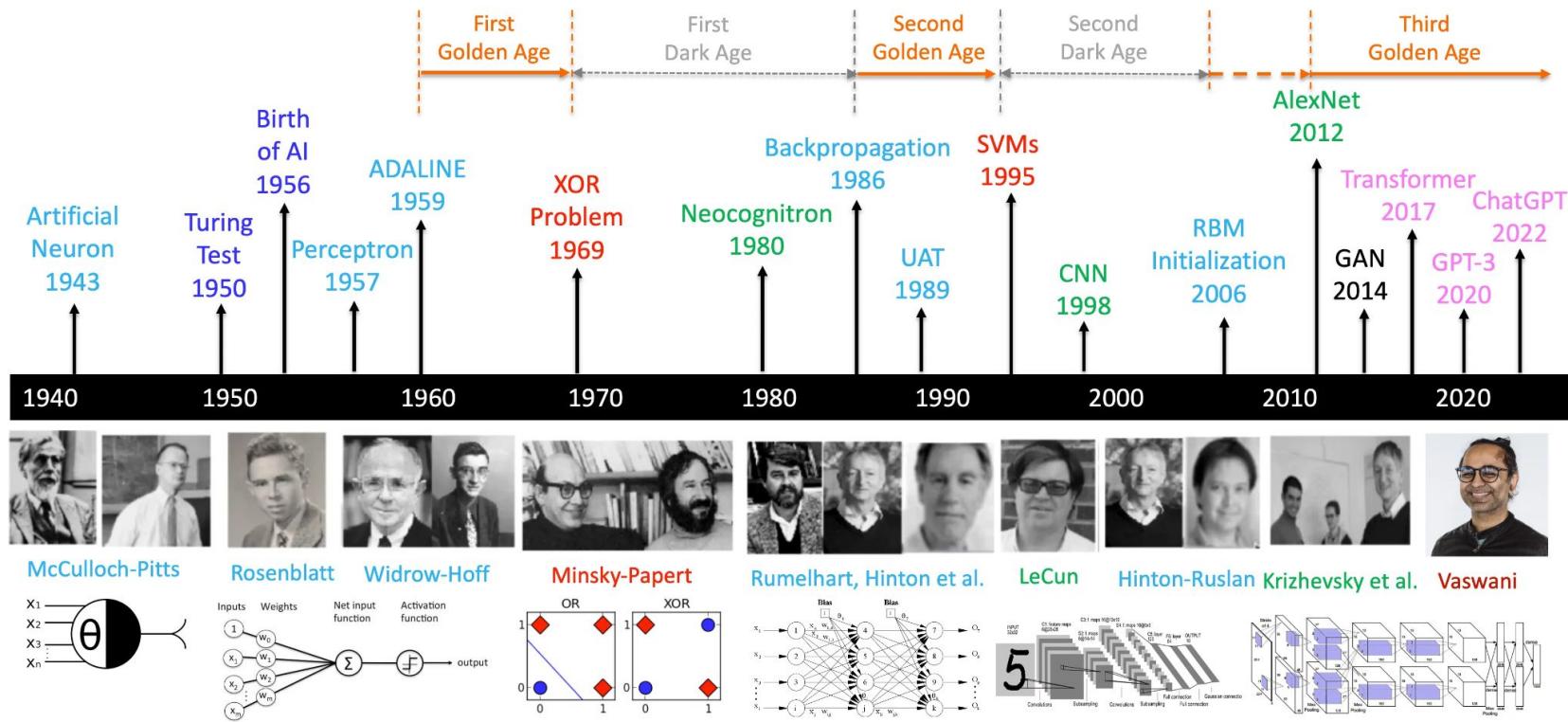
Impacto



Introdução



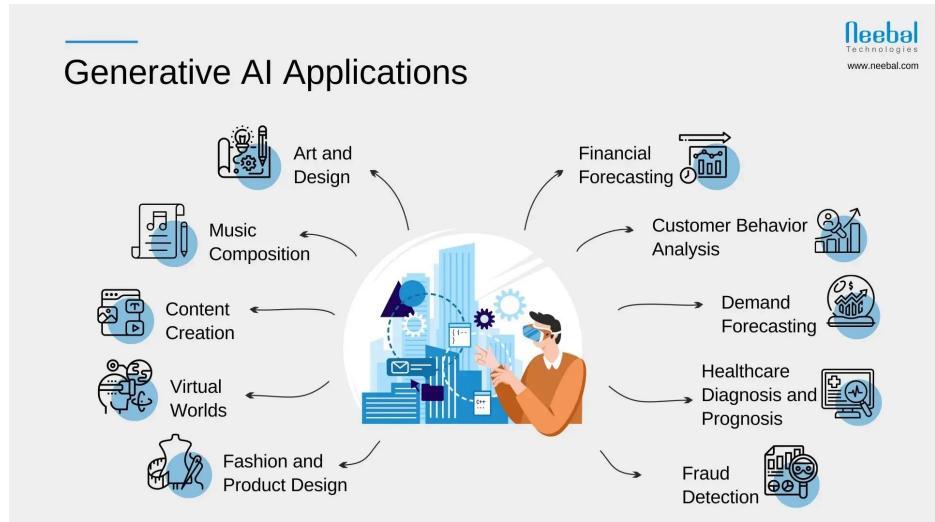
Linha do Tempo



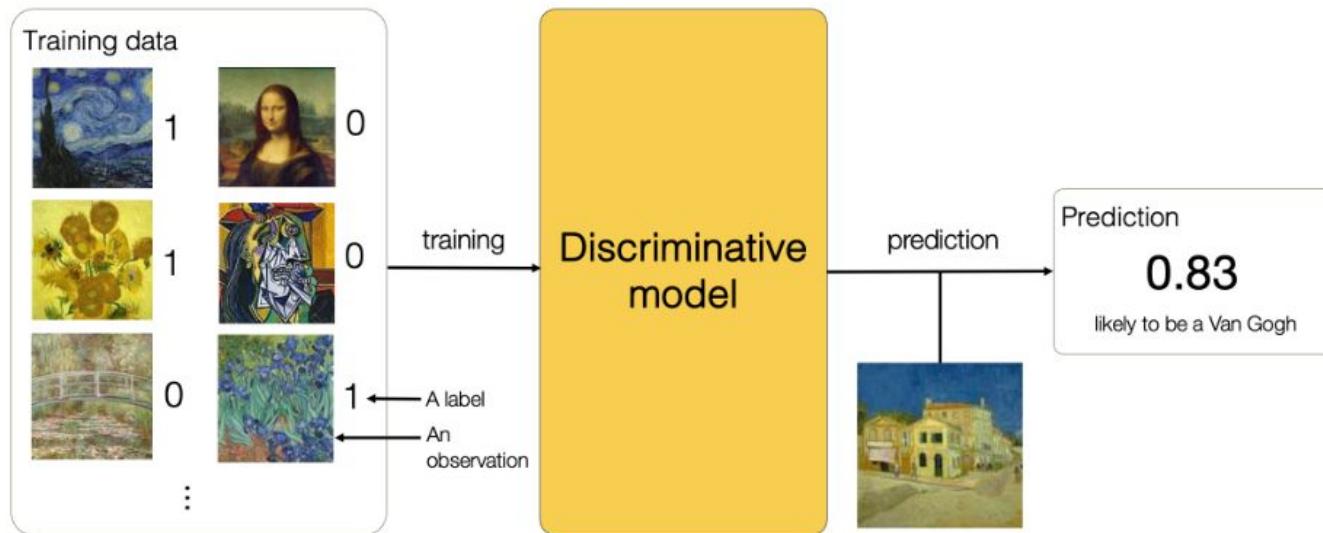
IA Generativa

O que é?

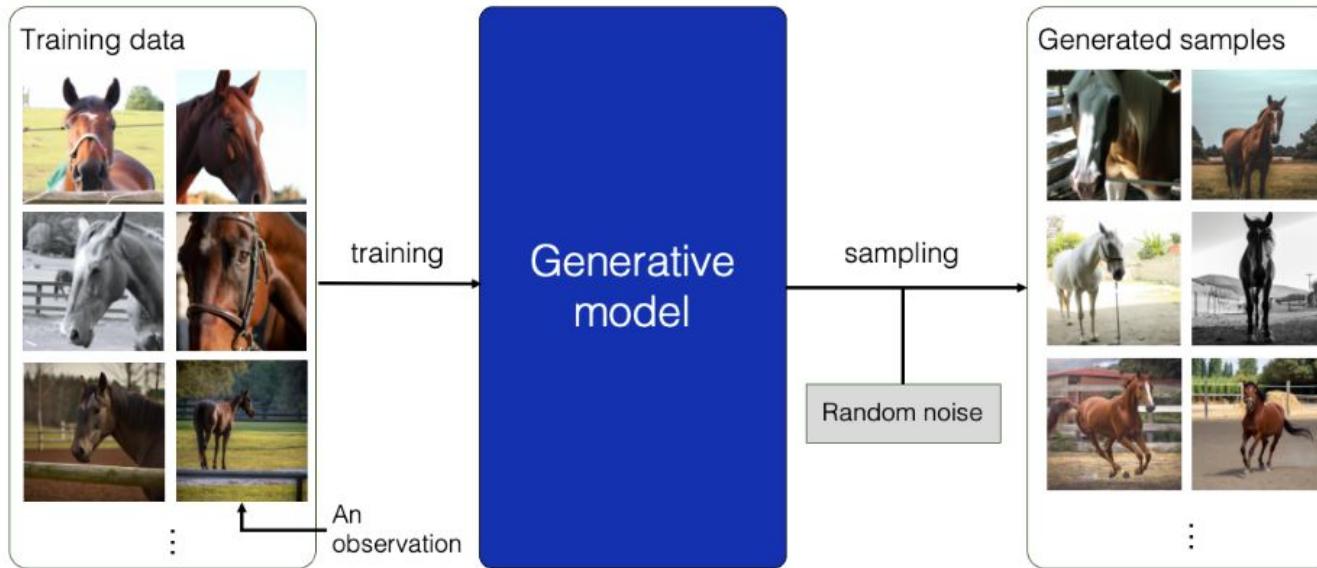
- Área da IA focada na criação de novos dados a partir de padrões aprendidos
- Baseia-se em aprender a distribuição de probabilidade dos dados originais
- Produz resultados originais, mas coerentes com os exemplos de treino



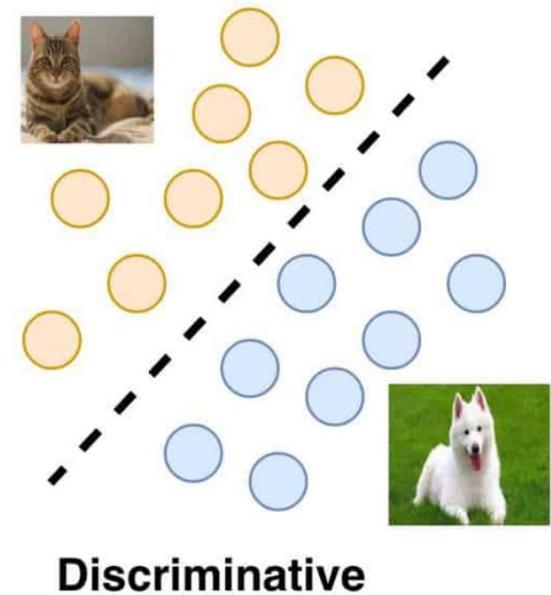
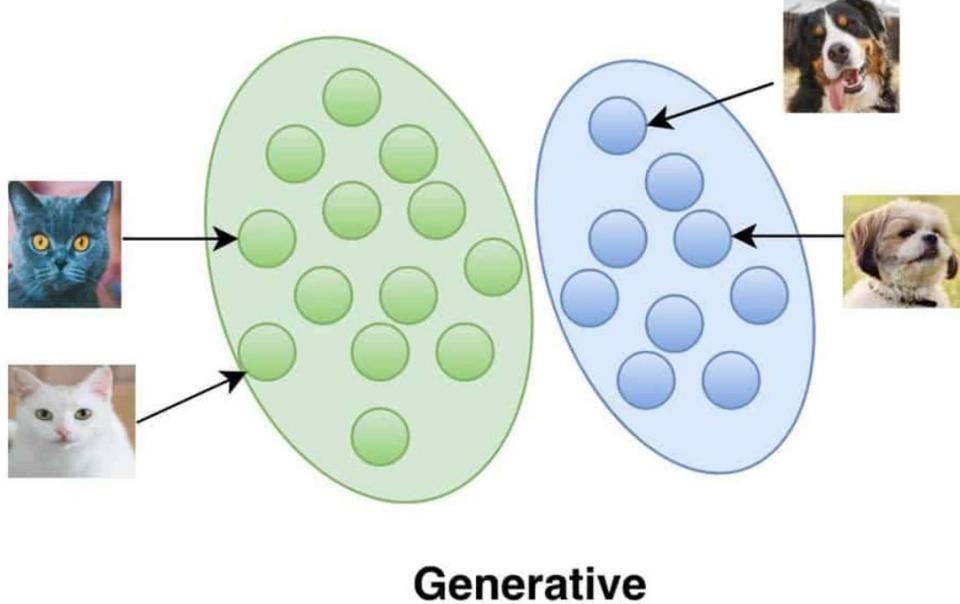
Modelos Discriminativos



Modelos Generativos



Generativa vs Discriminativa



Generativa vs Discriminativa

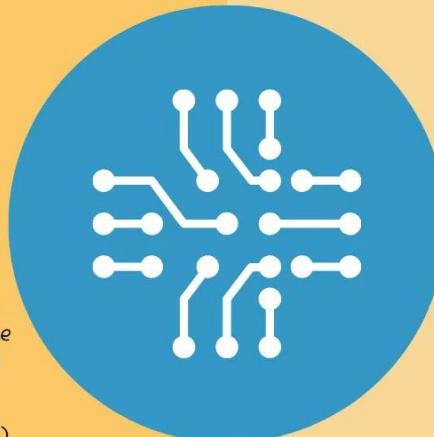
Generative AI

Typically trained on huge language models to perform almost any task

Objective of the model is to create entirely new content using the data the model has been trained on

Input and output are very flexible, often requiring prompt engineering to determine the best input to get the output needed

Sometimes makes things up (hallucinates), which requires a human to confirm the output's accuracy



Discriminative AI

Can be trained on narrow models to perform very specific tasks

Objective of the model is to make a decision based on data the model has been trained on

Input is typically fixed schema

Having a human reviewer in the loop to moderate low confidence decisions and retrain the model with new annotated data can help improve model performance

Histórico: Primeiros Algoritmos

ELIZA:

- Um dos primeiros chatbots da história, criado em 1966 por Joseph Weizenbaum
- Baseado em regras simples de correspondência de padrões e substituição de texto
- Simulava um psicoterapeuta rogeriano, respondendo com perguntas abertas
- Não entendia o significado das palavras, apenas manipulava texto superficialmente
- “Estou me sentindo *” → “Por que você está se sentindo {x}?”

```
Welcome to
      EEEEEE  LL      IIII    ZZZZZZ  AAAAAA
      EE     LL      II      ZZ  AA  AA
      EEEE   LL      II      ZZZ  AAAAAAA
      EE     LL      II      ZZ  AA  AA
      EEEEEEE LLLLLL  IIII  ZZZZZZ  AA  AA

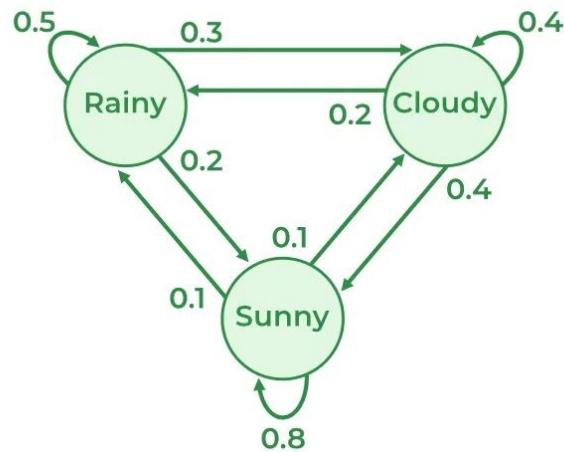
Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU: Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU: They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU: Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU: He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU: It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

Histórico: Primeiros Algoritmos

Cadeias de Markov para Geração:

- Modelo probabilístico que prevê o próximo elemento com base no estado atual
- Assume que o futuro depende apenas do presente, não de toda a sequência passada (propriedade de Markov)
- Para geração de texto, calcula a probabilidade de cada palavra dada a palavra anterior ou anteriores
- Aprende uma matriz de transições entre tokens a partir de um corpus



Histórico: Deep Generative Models

Autoencoder Variacional:

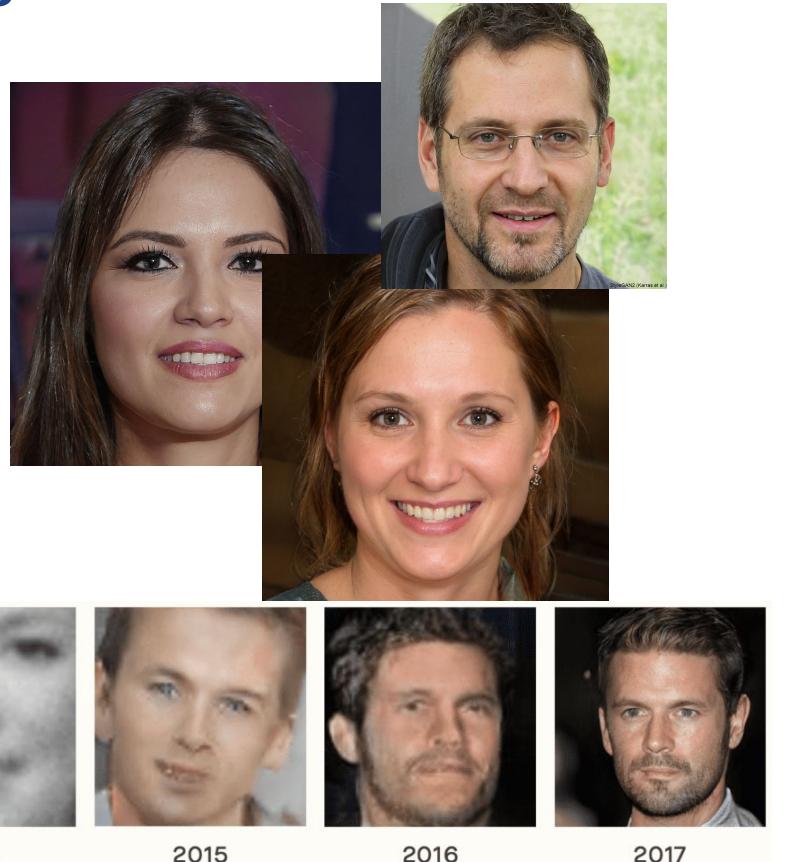
- Introduzido em 2013 para unir autoencoders e modelos probabilísticos
- Aprende uma distribuição latente que permite gerar novas amostras realistas
- Vetor latente pode ser manipulado para controlar características do conteúdo gerado
- Popular em aplicações que exigem variação controlada, como edição de imagens e síntese de dados



Histórico: Deep Generative Models

Generative Adversarial Networks:

- Propostos em 2014 como abordagem adversarial para geração de dados
- Consiste em um gerador e um discriminador treinados em competição
- Geração pode ser controlada condicionando o gerador a rótulos ou vetores específicos
- Amplamente usados para criar imagens realistas, deepfakes e aumentar datasets



Histórico: Deep Generative Models



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

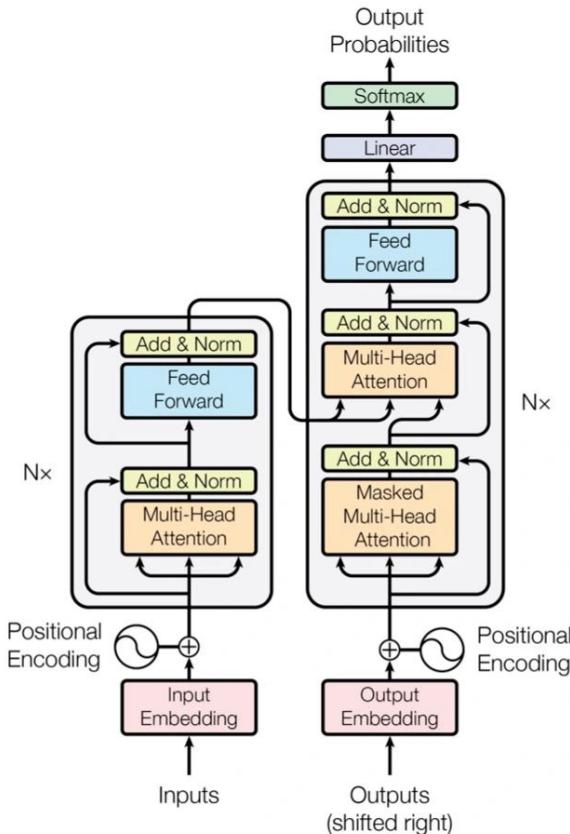
Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukasz.kaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

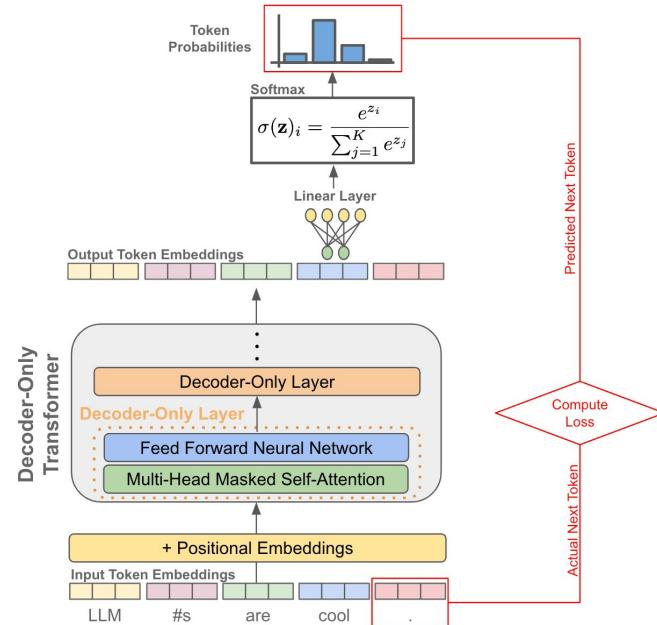
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, 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.



Histórico: Deep Generative Models

Generative Pre-Trained Transformer:

- Introduzido pela OpenAI em 2018, baseado na arquitetura Transformer
- Modelo de linguagem autoregressivo treinado para prever o próximo token em grandes corpus
- Escalonamento de parâmetros e dados levou a avanços significativos em coerência e versatilidade
- Capaz de realizar múltiplas tarefas sem ajuste específico, apenas via prompting



Histórico: AI Boom

- Em 2019, é lançado o GPT-2
 - 1.5B de parâmetros
 - Treinado em 40GB de texto
- Em 2020, o GPT-3
 - 175B de parâmetros
 - Treinado em 570GB de texto

Language Models are Unsupervised Multitask Learners

Alec Radford ^{* 1} Jeffrey Wu ^{* 1} Rewon Child ¹ David Luan ¹ Dario Amodei ^{** 1} Ilya Sutskever ^{** 1}

Abstract

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. Our largest model, GPT-2, is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting but still underfits WebText. Samples from the model reflect these improvements and contain coherent paragraphs of text. These findings suggest a promising path towards building language processing systems which learn to perform tasks from their naturally occurring demonstrations.

competent generalists. We would like to move towards more general systems which can perform many tasks – eventually without the need to manually create and label a training dataset for each one.

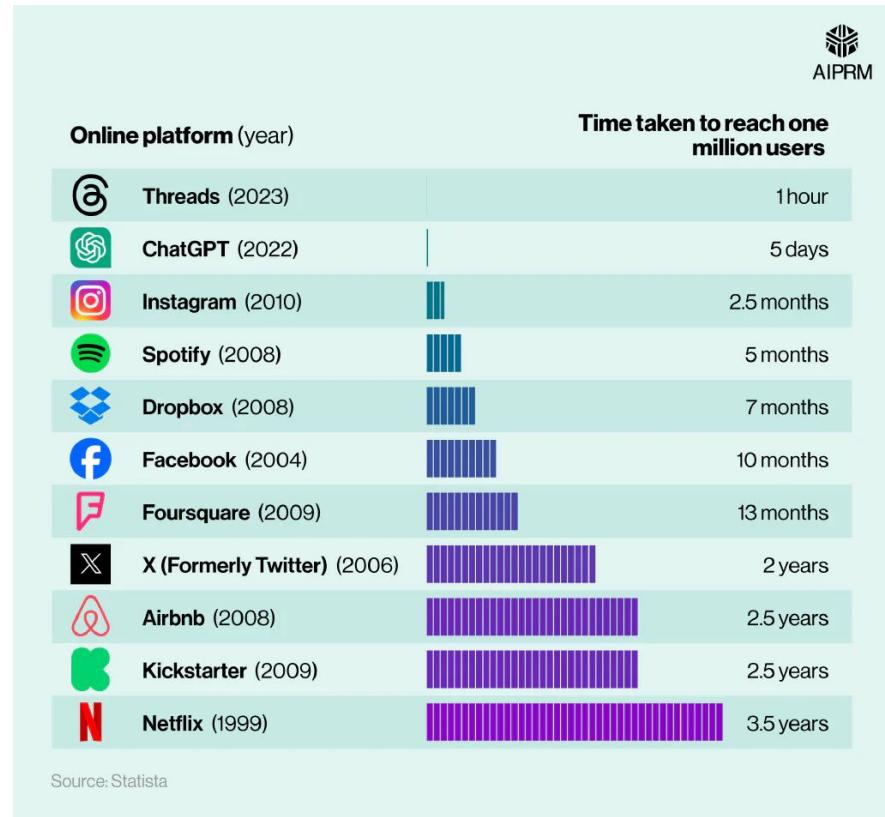
The dominant approach to creating ML systems is to collect a dataset of training examples demonstrating correct behavior for a desired task, train a system to imitate these behaviors, and then test its performance on independent and identically distributed (IID) held-out examples. This has served well to make progress on narrow experts. But the often erratic behavior of captioning models (Lake et al., 2017), reading comprehension systems (Jia & Liang, 2017), and image classifiers (Alcorn et al., 2018) on the diversity and variety of possible inputs highlights some of the shortcomings of this approach.

Our suspicion is that the prevalence of single task training on single domain datasets is a major contributor to the lack of generalization observed in current systems. Progress towards robust systems with current architectures is likely to require training and measuring performance on a wide range of domains and tasks. Recently, several benchmarks have been proposed such as GLUE (Wang et al., 2018) and decaNLP (McCann et al., 2018) to begin studying this.

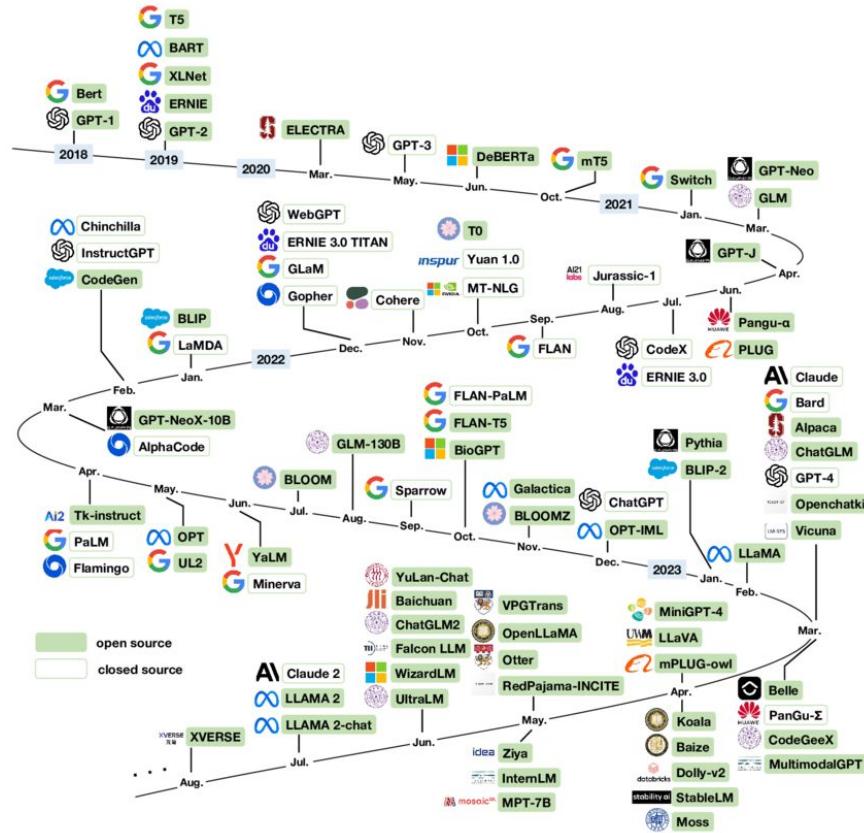
Multitask learning (Caruana, 1997) is a promising framework for improving general performance. However, multitask training in NLP is still nascent. Recent work reports modest performance improvements (Yogatama et al., 2019) and the two most ambitious efforts to date have

Histórico: AI Boom

- No final de 2022, é lançado o ChatGPT
- GPT-3 ajustado por RLHF (InstructGPT)
 - Humanos ranqueiam múltiplas respostas geradas
 - É treinado um modelo de recompensa para prever essas preferências
 - Aplica-se PPO para ajustar o GPT-3



Histórico: AI Boom



Histórico: AI Boom

- **Objetivo:** Avaliar 4 sistemas (ELIZA, GPT-4o, LLaMa-3.1-405B e GPT-4.5) em dois testes de Turing randomizados, controlados e pré-registrados.
- **Metodologia:**
 - Conversas de 5 minutos com um humano e um sistema simultaneamente.
 - Participantes deveriam identificar qual parceiro era humano.
- **Resultados principais:**
 - GPT-4.5: identificado como humano 73% das vezes.
 - LLaMa-3.1-405B: 56%, sem diferença significativa em relação ao humano real.
 - GPT-4o e ELIZA: desempenho abaixo do acaso (21% e 23%, respectivamente).

Large Language Models Pass the Turing Test

Cameron R. Jones

Department of Cognitive Science
UC San Diego
San Diego, CA 92119
cameron@ucsd.edu

Benjamin K. Bergen

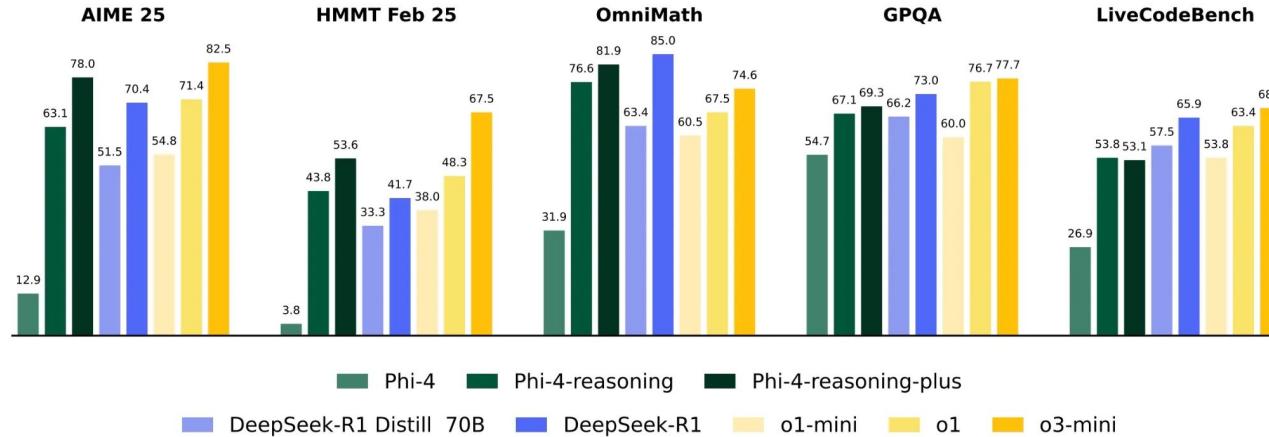
Department of Cognitive Science
UC San Diego
San Diego, CA 92119
bkbergen@ucsd.edu

Abstract

We evaluated 4 systems (ELIZA, GPT-4o, LLaMa-3.1-405B, and GPT-4.5) in two randomised, controlled, and pre-registered Turing tests on independent populations. Participants had 5 minute conversations simultaneously with another human participant and one of these systems before judging which conversational partner they thought was human. When prompted to adopt a humanlike persona, GPT-4.5 was judged to be the human 73% of the time: significantly more often than interrogators selected the real human participant. LLaMa-3.1, with the same prompt, was judged to be the human 56% of the time—not significantly more or less often than the humans they were being compared to—while baseline models (ELIZA and GPT-4o) achieved win rates significantly below chance (23% and 21% respectively). The results constitute the first empirical evidence that any artificial system passes a standard three-party Turing test. The results have implications for debates about what kind of intelligence is exhibited by Large Language Models (LLMs), and the social and economic impacts these systems are likely to have.

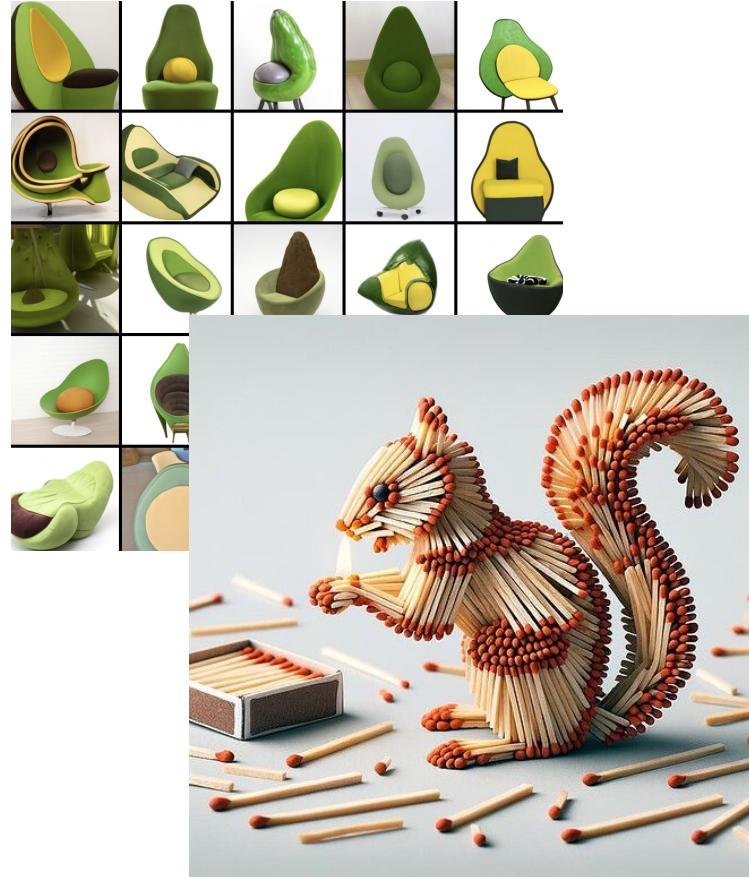
Histórico: AI Boom

- 2024-2025: Modelos especializados em raciocínio, como DeepSeek-R1 e OpenAI o1



Histórico: AI Boom

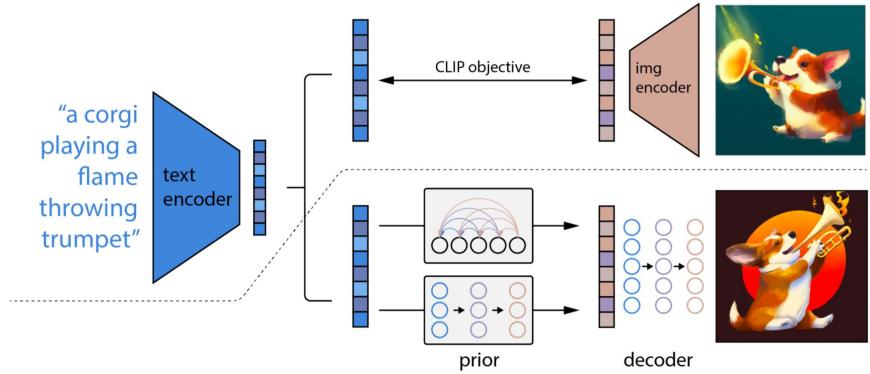
- 2021: OpenAI apresenta um modelo Transformer treinado para mapear texto em imagem
- 2022: Melhora de resolução e fidelidade via diffusion models condicionados por CLIP embeddings
- Impacto:
 - Democratização da criação visual
 - Viralização nas redes sociais, ampliando interesse público e acadêmico em IA generativa
 - Contribuiu para a popularização do conceito de *prompt engineering*



Histórico: AI Boom

Modelos de Difusão:

- Ganharam destaque entre 2020 e 2022 com avanços como DDPM e Stable Diffusion
- Funcionam revertendo um processo gradual de adição de ruído até reconstruir dados detalhados
- Oferecem maior controle de geração com condicionamentos, como descrições textuais (text-to-image)
- Midjourney, lançado em 2022, popularizou o uso de IA generativa para arte de forma acessível via Discord



Histórico: AI Boom

Geração de Vídeos:

- Evoluíram rapidamente entre 2022 e 2025, combinando avanços em difusão, transformers e geração condicional
- Exemplos recentes: [Sora \(OpenAI\)](#) e [Veo 3 \(Google DeepMind\)](#)
- Desafios históricos incluíam consistência temporal, distorções e falta de continuidade entre quadros
- Condicionamento quadro a quadro, atenção temporal e predição em múltiplas escalas para melhorar a estabilidade
- [Will Smith comendo espaguete](#)

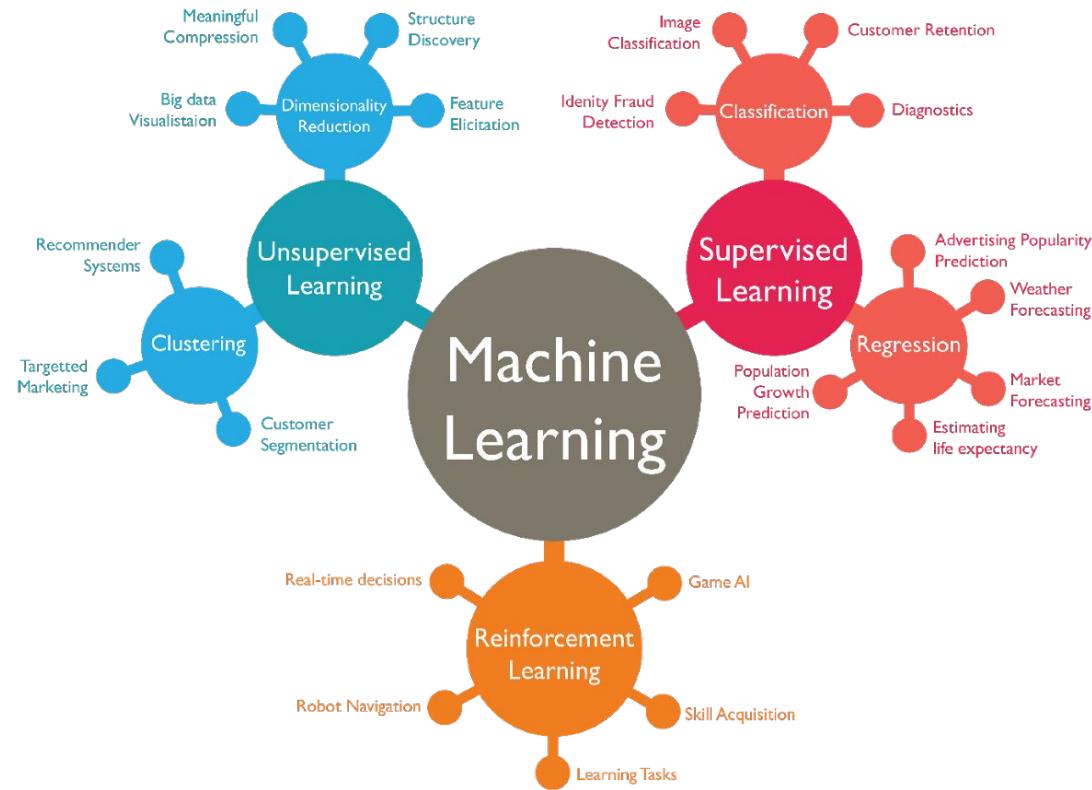


Histórico: AI Boom



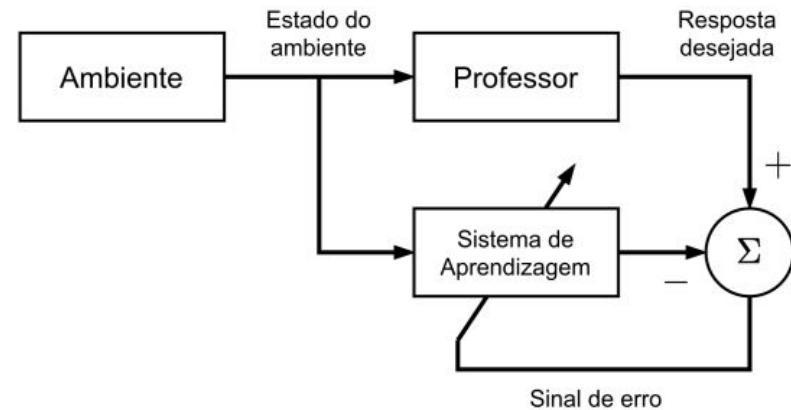
Aprendizado

Tipos de Aprendizado



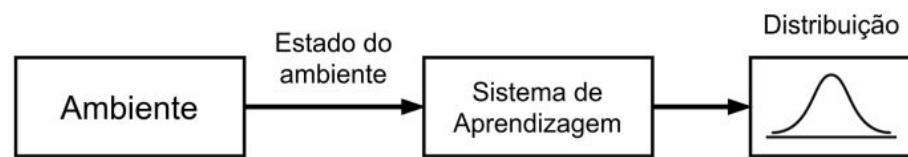
Aprendizado Supervisionado

- Dados possuem correspondência clara entre entrada e saída
- Cada exemplo contém vetor de entrada x_i e vetor de saída y_i
- O modelo aprende com a orientação de um “professor” ou supervisor
- Ajuste de parâmetros é feito com base no sinal de erro



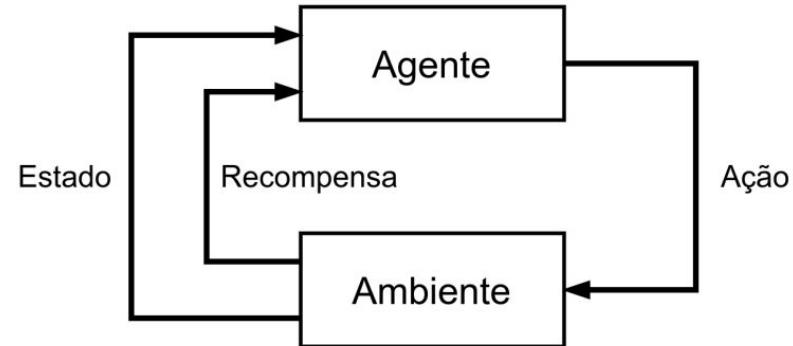
Aprendizado Não-Supervisionado

- Não existe a figura do professor nem correspondência entrada-saída
- O modelo não recebe sinal de erro supervisionado
- Objetivo é aprender a distribuição de probabilidades dos dados
- Aprende regularidades estatísticas do conjunto de treinamento



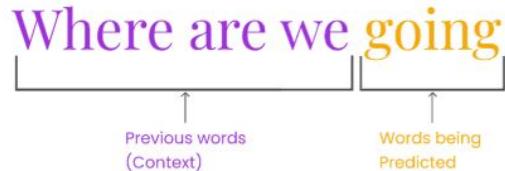
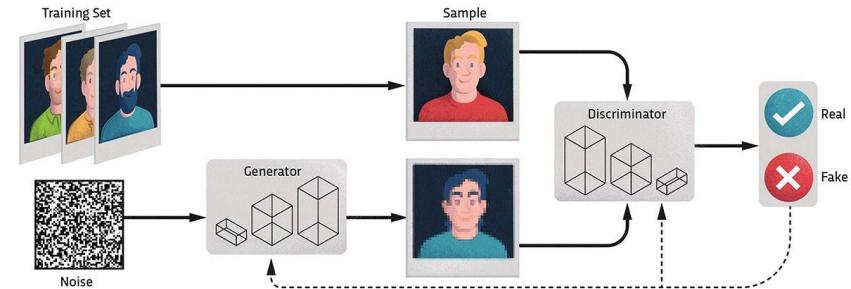
Aprendizado por Reforço

- O agente interage com o ambiente executando ações
- Aprende qual sequência de ações maximiza a recompensa acumulada
- Baseia-se no estado atual do ambiente e do próprio agente
- Experiência é adquirida diretamente pelas interações

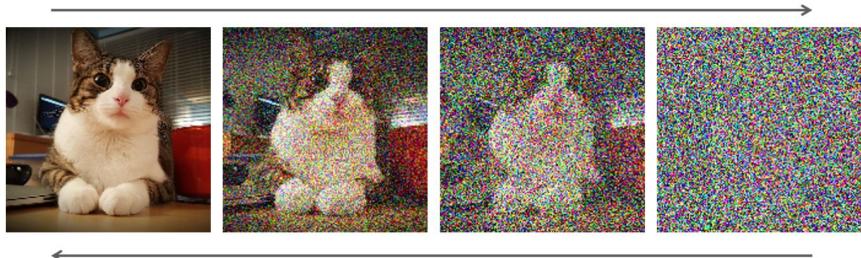


Aprendizado Auto-Supervisionado

- É necessário uma grande quantidade de dados para treinar modelos de IA Generativa
- Rotular dados é custoso e exige muita intervenção humana
- SSL: Paradigma de aprendizado onde o próprio dado fornece o rótulo



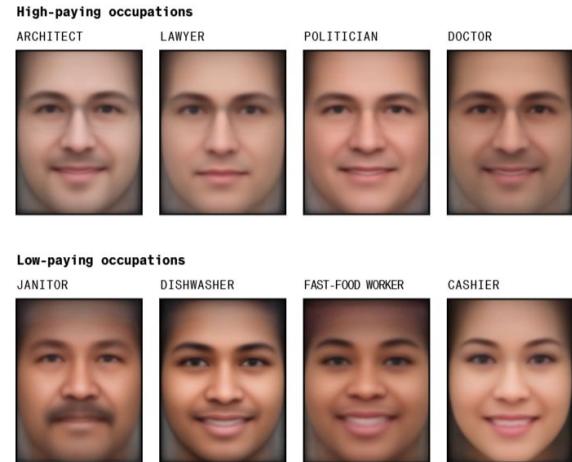
$$P(S) = P(\text{Where}) \times P(\text{are} | \text{Where}) \times P(\text{we} | \text{Where are}) \times P(\text{going} | \text{Where are we})$$



Riscos

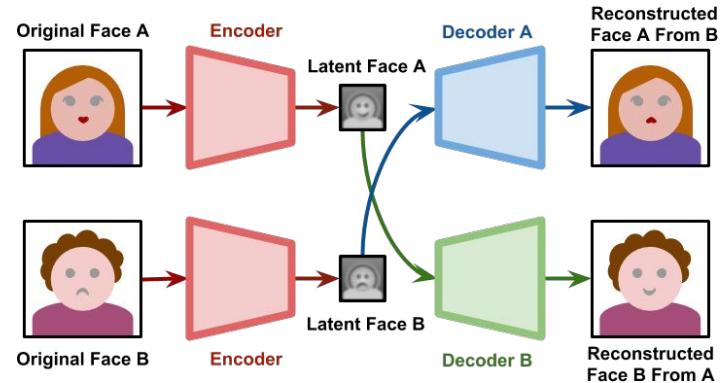
Vieses

- Modelos generativos aprendem padrões e vieses presentes nos dados de treino
- Isso pode reforçar estereótipos culturais, de gênero, raça ou classe social
- Discussões:
 - É possível criar um modelo totalmente livre de vieses?
 - Quem decide quais vieses devem ser removidos ou preservados?
 - Como equilibrar neutralidade e representatividade nos resultados gerados?
 - A mitigação de vieses pode introduzir novos vieses não intencionais?



Desinformação

- Modelos gerativos podem criar conteúdos falsos convincentes, como textos, imagens, áudios e vídeos
- Risco de amplificação de narrativas falsas em redes sociais e ambientes políticos
- Discussões:
 - Quem deve ser responsabilizado pela criação e disseminação de deepfakes nocivos?
 - A detecção automática de desinformação conseguirá acompanhar o ritmo da geração?
 - Marcas d'água digitais seriam suficientes para conter a propagação de conteúdo falso?
 - A população está preparada para lidar criticamente com mídias hiper-realistas?



Copyright

- IA generativa pode reproduzir estilos ou elementos protegidos por direitos autorais sem permissão
- Quem detém o copyright de obras criadas por IA?
- Discussões:
 - A IA generativa deve ter liberdade para aprender com qualquer obra existente?
 - É ético treinar modelos com material protegido por copyright sem autorização?
 - Como equilibrar inovação tecnológica e proteção dos direitos dos criadores?



Impacto no Trabalho

- A IA generativa pode substituir algumas funções e criar novas oportunidades de trabalho
- Setores criativos, de mídia e marketing estão entre os mais vulneráveis à automação criativa
- Discussões:
 - A IA generativa vai substituir empregos ou criar novas oportunidades?
 - Como preparar profissionais para trabalhar junto com sistemas generativos?
 - É necessário criar políticas públicas para mitigar impactos negativos no mercado de trabalho?

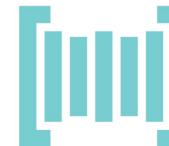


Por onde ficar por dentro?

- [arXiv: pré-prints de pesquisa em IA e deep learning](#)
- [Papers with Code: artigos com código e benchmarks comparativos](#)
- [Kaggle: competições, datasets e notebooks colaborativos](#)
- [Hugging Face: modelos pré-treinados, datasets e documentação prática](#)
- YouTube: canais técnicos e conferências gravadas
- Repositórios no GitHub: implementação e experimentação de modelos



kaggle



Papers With
Code

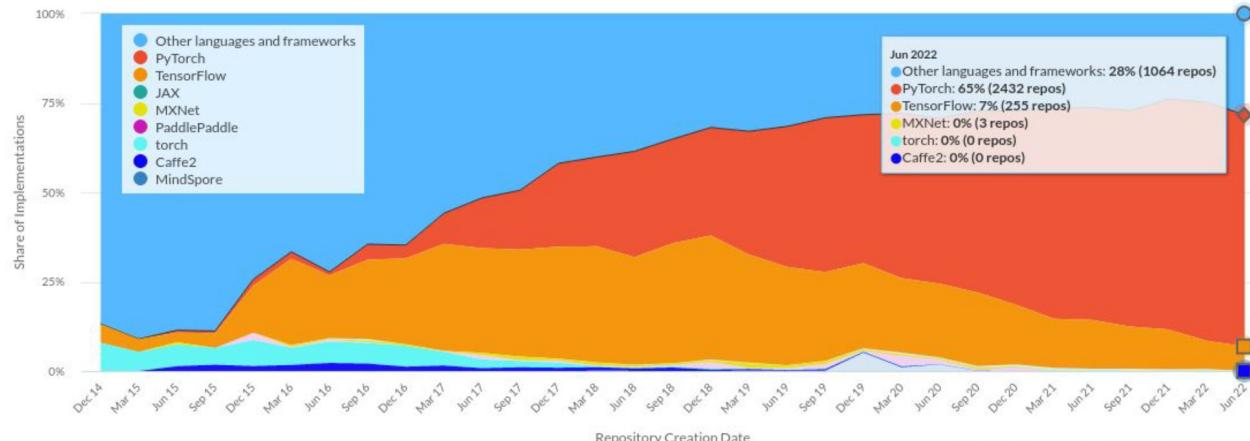


Hugging Face

Frameworks



Paper Implementations grouped by framework



Links Úteis

- [Esta pessoa não existe](#)
- [GANPaint](#)
- [Stable Diffusion no Colab](#)