

Criando aplicações baseadas em LLMs



Gustavo Pinto



@gustavopinto



gpinto@ufpa.br



gustavopinto.org



ML4SE.substack.com

Antes de Começarmos::

- Consegue ligar a câmera?
- Teremos pausas a cada ~1h:30m
- **Todas aulas teremos exercícios:** na últimas 1h:30m
- Tem dúvidas? Pergunte!
- Aproveite!

Agenda

- Entendendo sobre LLMs
- Conhecendo o langchain
- Testando prompts e engenharia de prompts
- Entendendo de embeddings
- Comparando dados por similaridade
- Conectando com um banco vetorial
- Criando retrievers
- Criando uma interface de chat para o modelo
- Fazendo deploy da aplicação
- ...

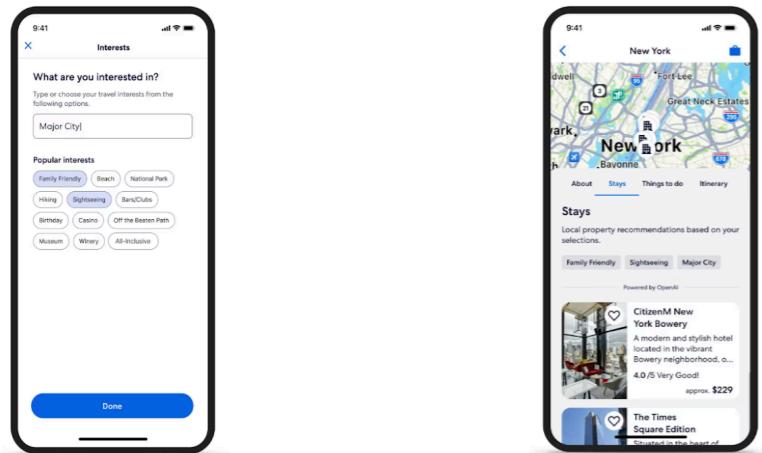
Avaliação

- Frequência
- Exercícios em todas as aulas
- Projeto de disciplina

Motivação

How our AI Travel Tool works

We'll take the hassle out of planning trips so you can focus on enjoying your experience instead of worrying about what to ask and consider. Here's how it works:



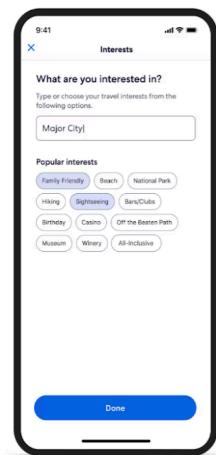
By tapping or entering the relevant keywords, our system can communicate with ChatGPT who will curate everything you need to know for each destination.

Each request will be compared against our internal data and match recommendations to Expedia-based properties and activities.

Motivação

How our AI Travel Tool works

We'll take the hassle out of planning trips so you can focus on enjoying your experience instead of worrying about what to ask and consider. Here's how it works:



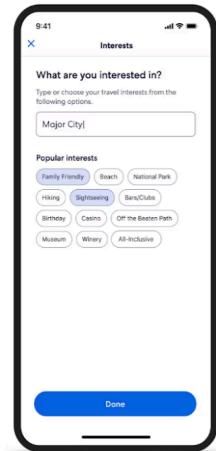
By tapping or entering the relevant keywords, our system can communicate with ChatGPT who will curate everything you need to know for each destination.

The landing page for MayaMD features a large, intricate network graph in the background, symbolizing connectivity and data analysis. At the top, the MayaMD logo is displayed with a sun-like icon. The navigation menu includes "Solutions", "About", and a prominent "Contact Us" button. Below the navigation, the text "AI HEALTH ASSISTANT" is visible. The central message reads: "From symptoms to triage in less than a minute". To the left of this message, there is descriptive text: "Tell Maya your symptoms, answer her questions and get personalized healthcare insights with recommended next steps in under a minute." At the bottom of the page is a blue "Schedule a Demo" button.

Motivação

How our AI Travel Tool works

We'll take the hassle out of planning trips so you can focus on enjoying your experience instead of worrying about what to ask and consider. Here's how it works:



By tapping or entering the relevant keywords, our system can communicate with ChatGPT who will curate everything you need to know for each destination.

mayaMD

Solutions About Contact Us

AI HEALTH ASSISTANT

From symptoms to triage in less than a minute

Tell Maya your symptoms, answer her questions and get personalized healthcare insights with recommended next steps in under a minute.

[Schedule a Demo](#)

AI nutritionist

New version available soon on iPhone and Android



Get answers to all your nutrition related questions

We want everyone to have their own personal nutritionist. With Nutrition.ai, you gain access to a set of features designed to provide you with personalized nutrition guidance, expert support, and a wealth of educational resources:

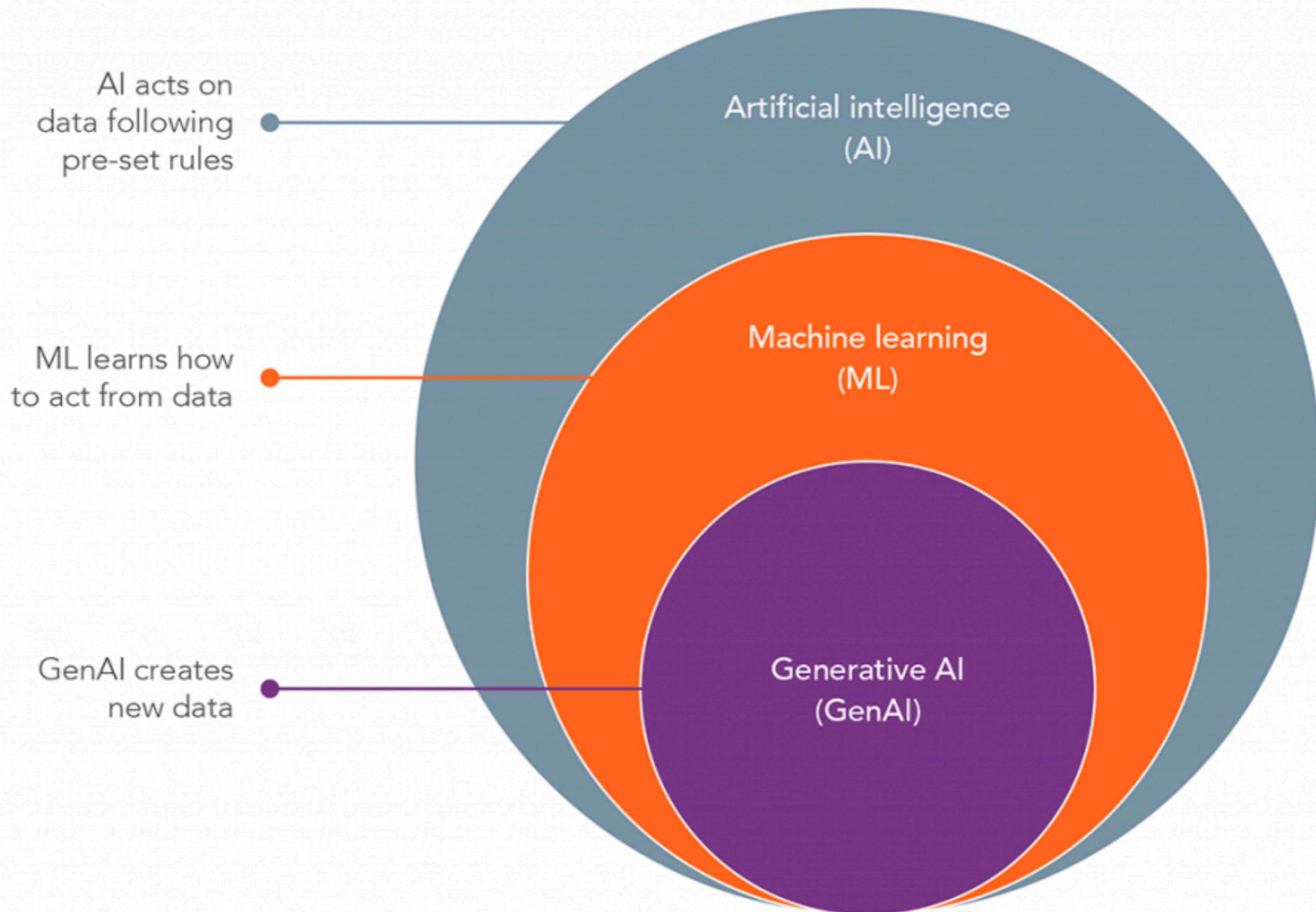
- AI-Powered Chat: Get instant answers to nutrition related questions. It's like having your own personal nutritionist at your fingertips!
- Nutrition Experts: Get specific support around topics



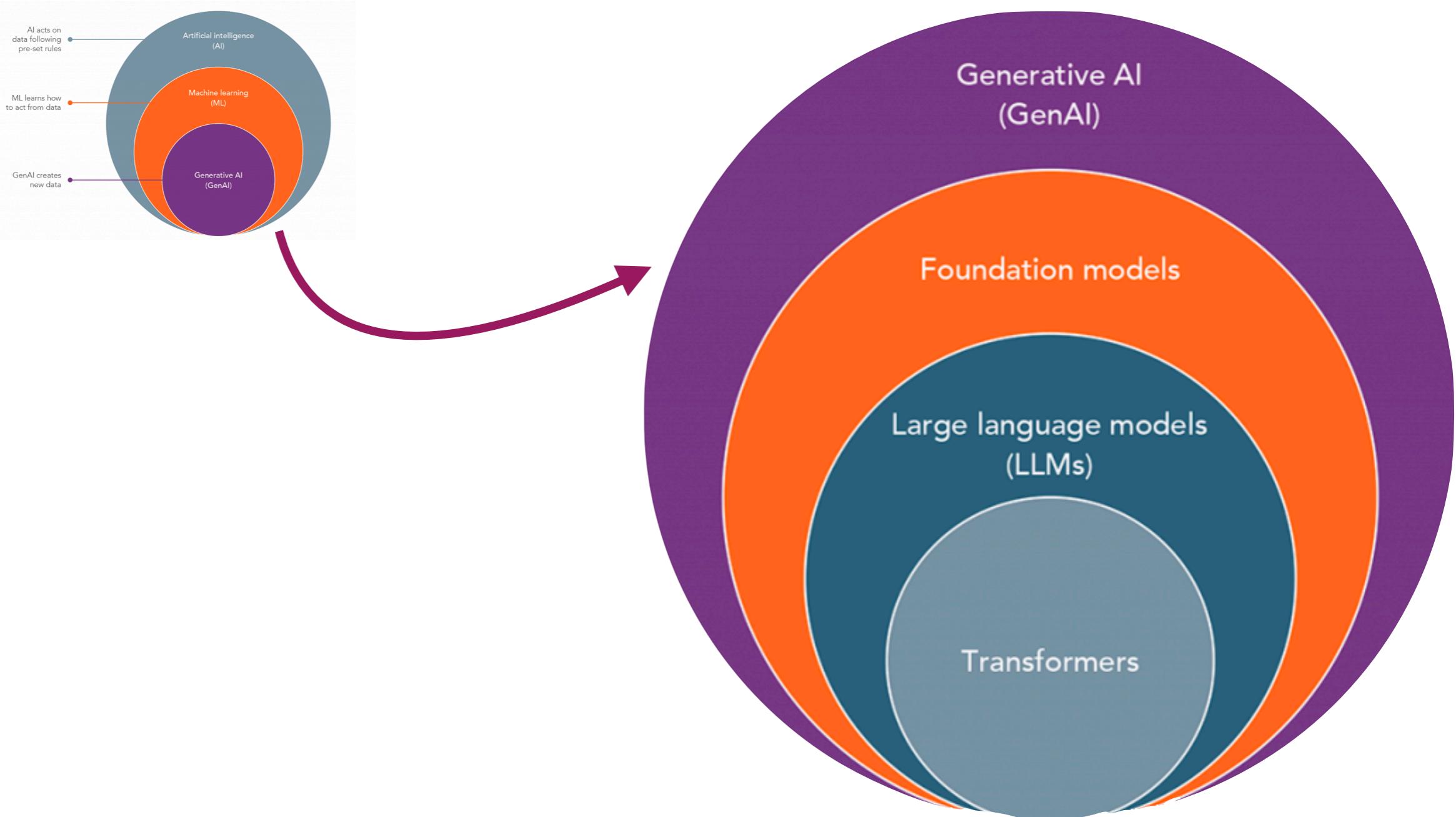
Como criar essas
aplicações baseadas em
LLMs?



1.0 que são LLMs?

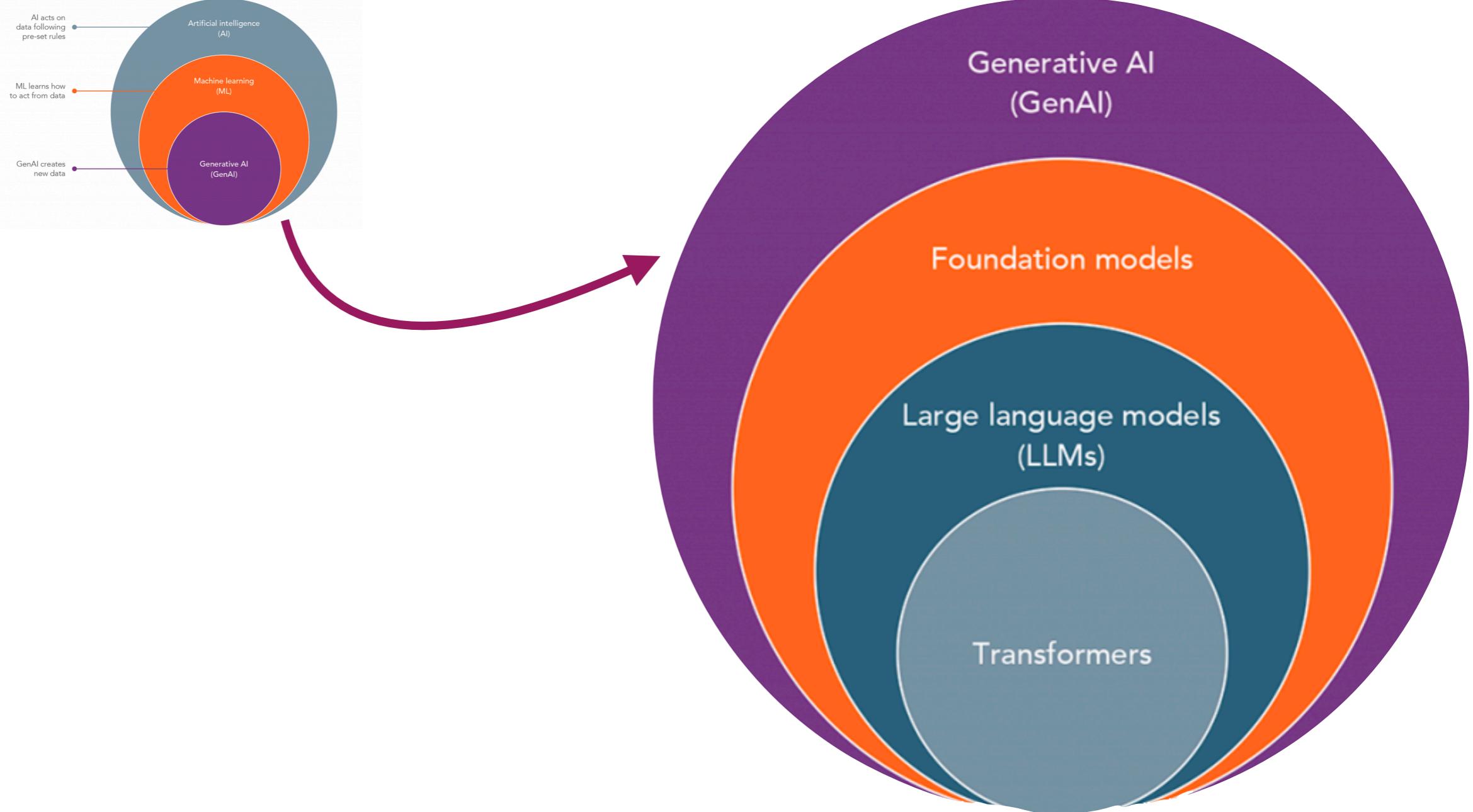


1.0 que são LLMs?

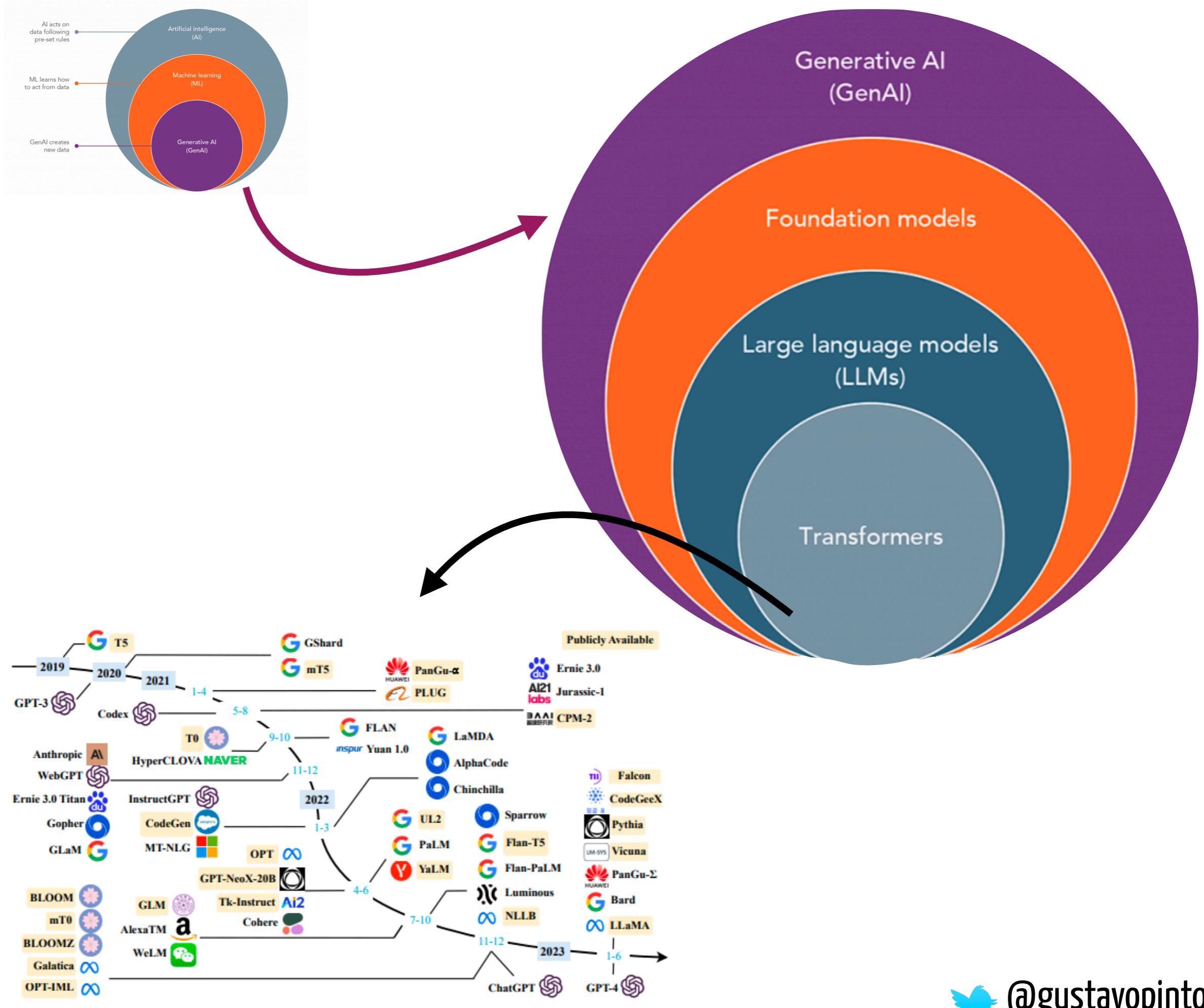


1.0 que são LLMs?

- FMs: Abrangem um escopo mais amplo, incluindo texto, imagem, áudio e dados multimodais. Suas aplicações incluem visão computacional, reconhecimento de fala, e integração multimodal.
- LLMs: Focados exclusivamente em texto e tarefas de PLN. Aplicações predominantemente relacionadas à linguagem natural.



1.0 que são LLMs?



1. O que São LLMs?

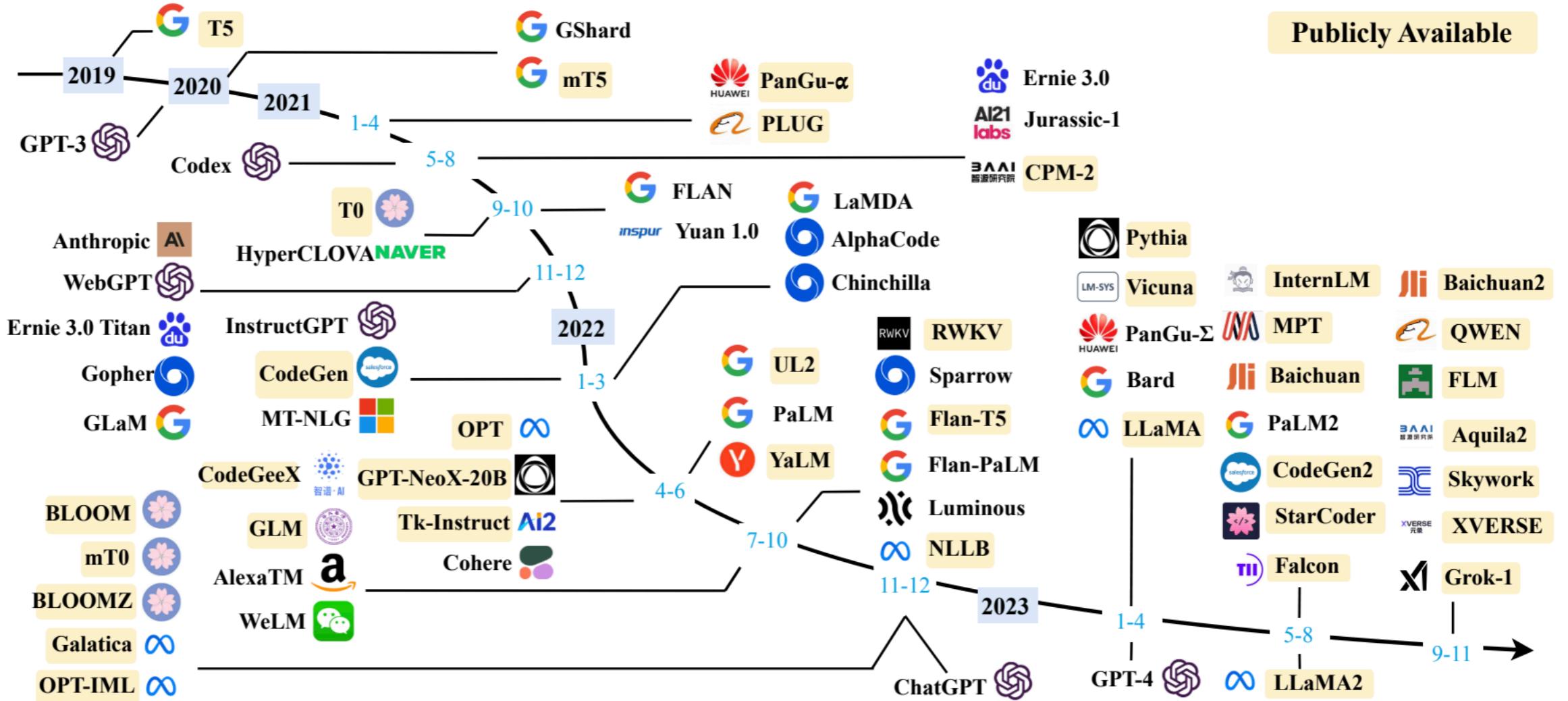
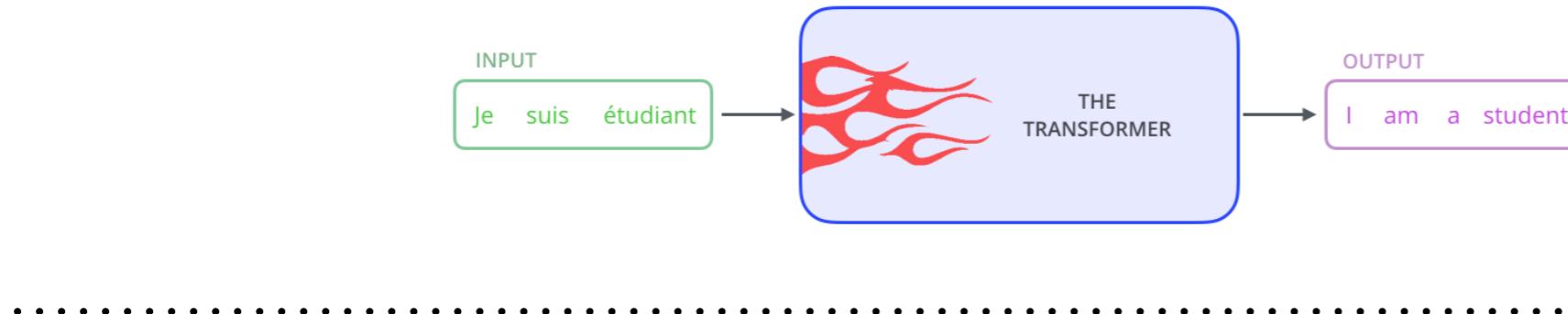


Fig. 3: A timeline of existing large language models (having a size larger than 10B) in recent years. The timeline was established mainly according to the release date (e.g., the submission date to arXiv) of the technical paper for a model. If there was not a corresponding paper, we set the date of a model as the earliest time of its public release or announcement. We mark the LLMs with publicly available model checkpoints in yellow color. Due to the space limit of the figure, we only include the LLMs with publicly reported evaluation results.

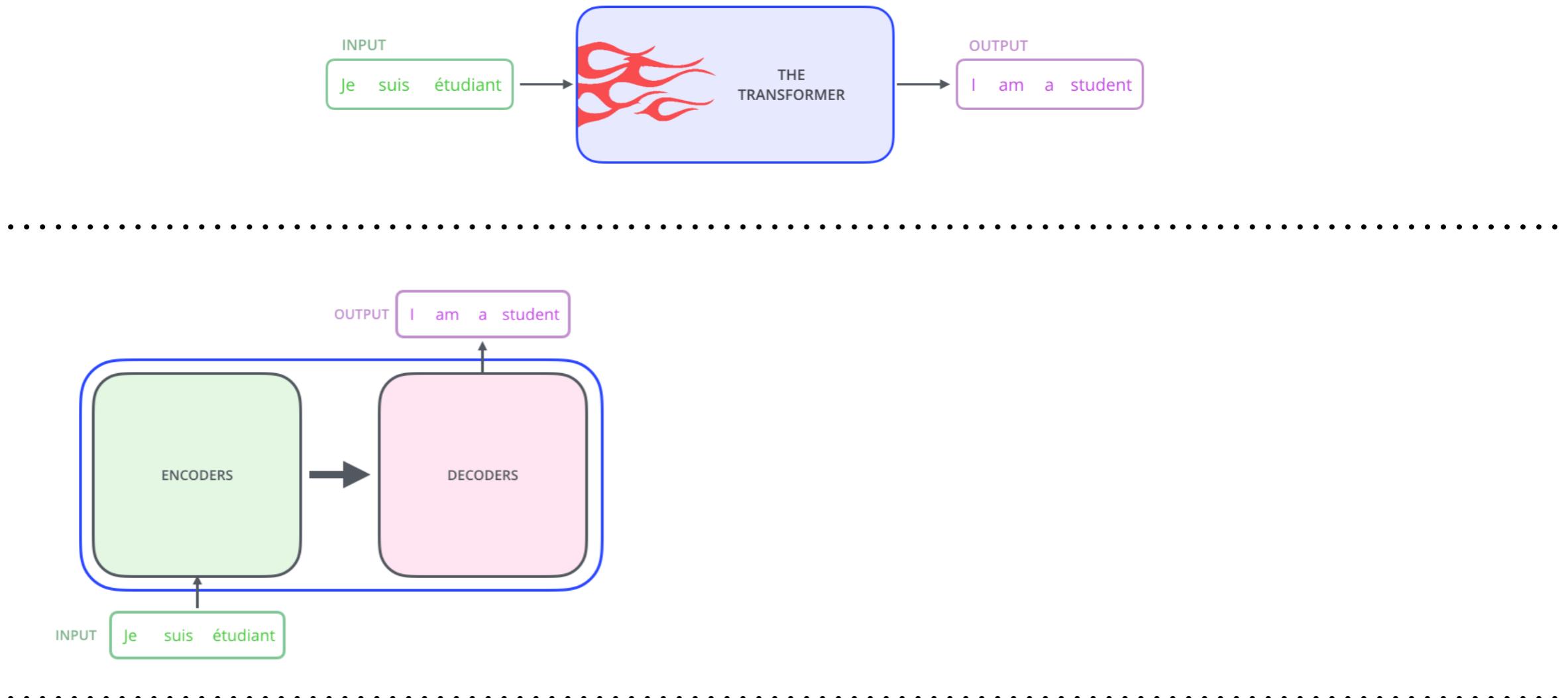
1.0 que são LLMs?



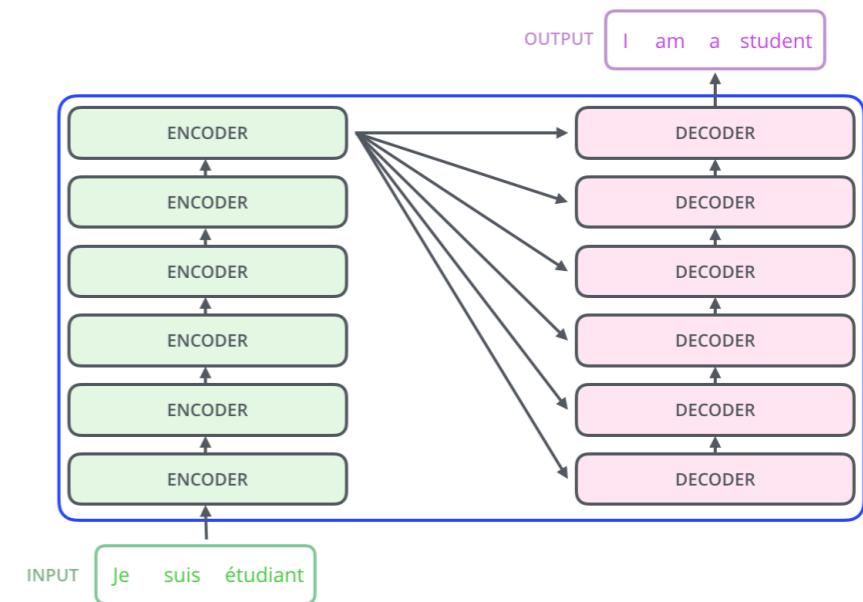
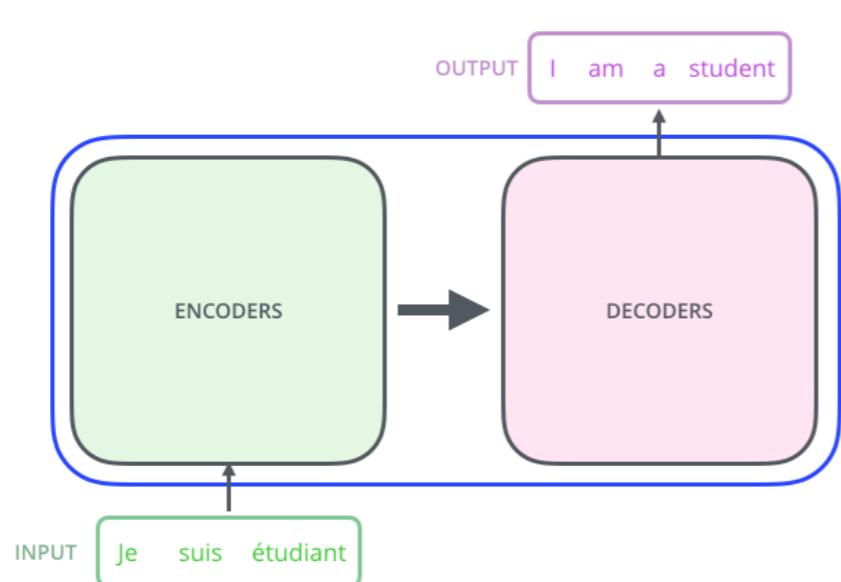
<https://jalammar.github.io/illustrated-transformer/>

 @gustavopinto

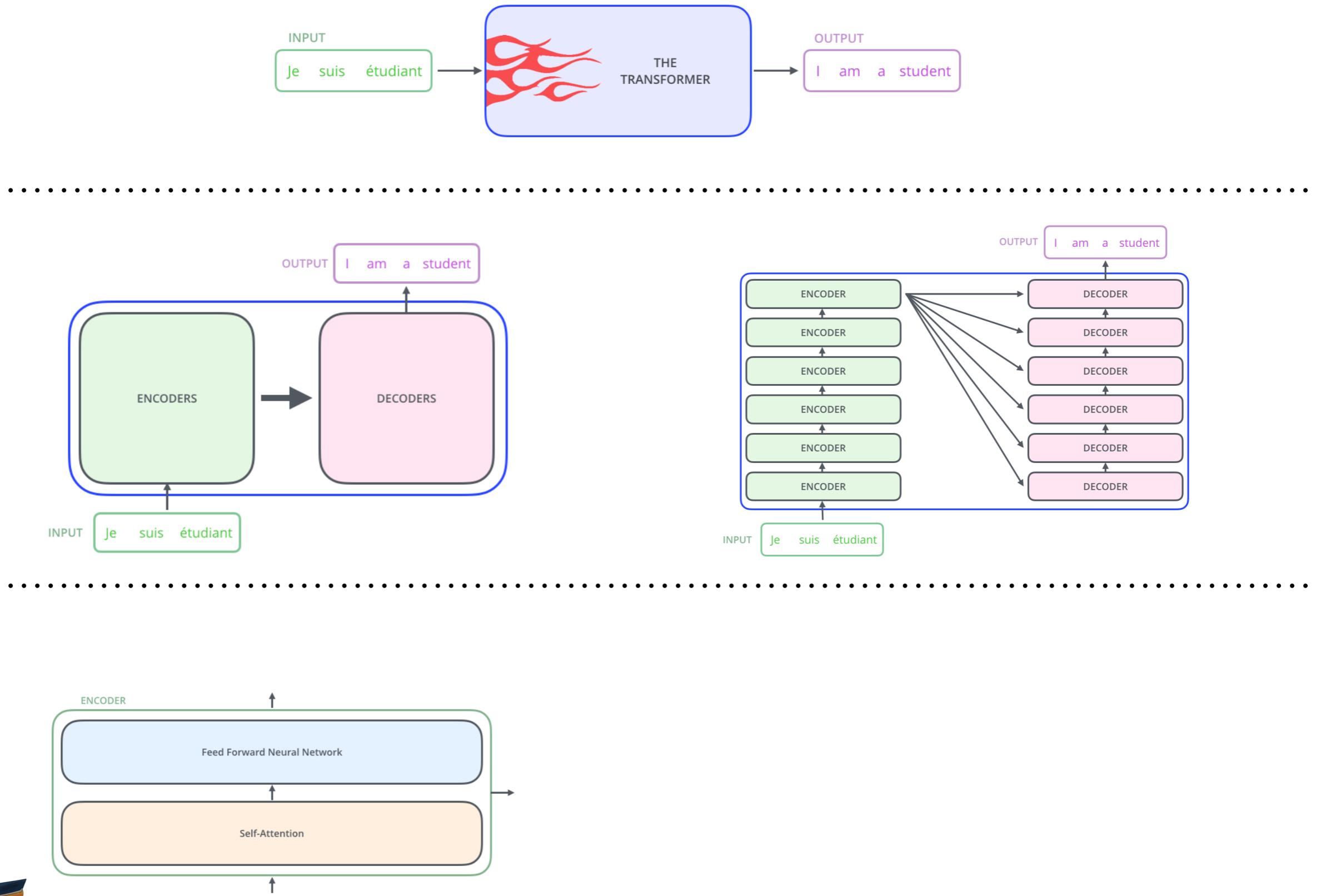
1.0 que são LLMs?



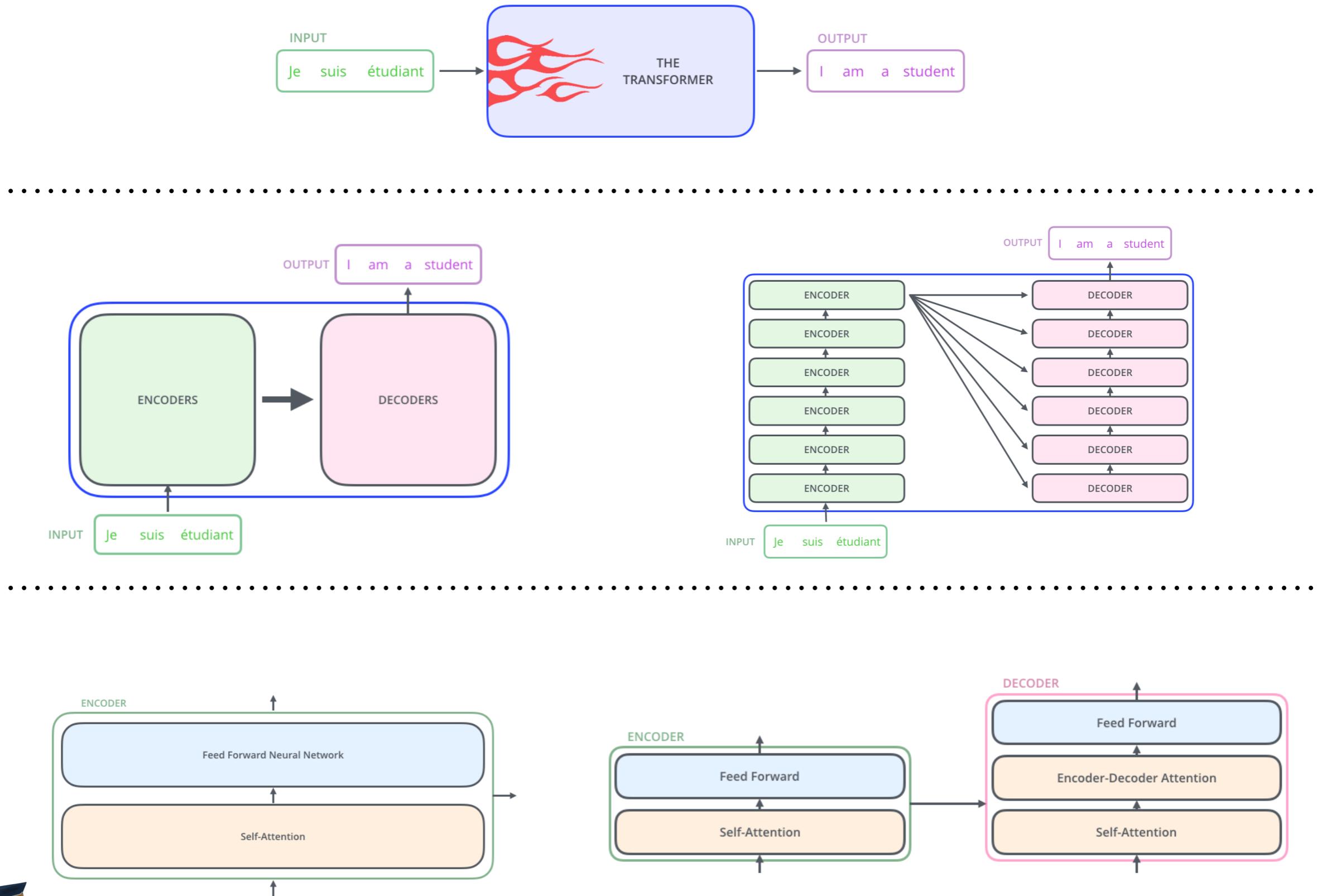
1.0 que são LLMs?



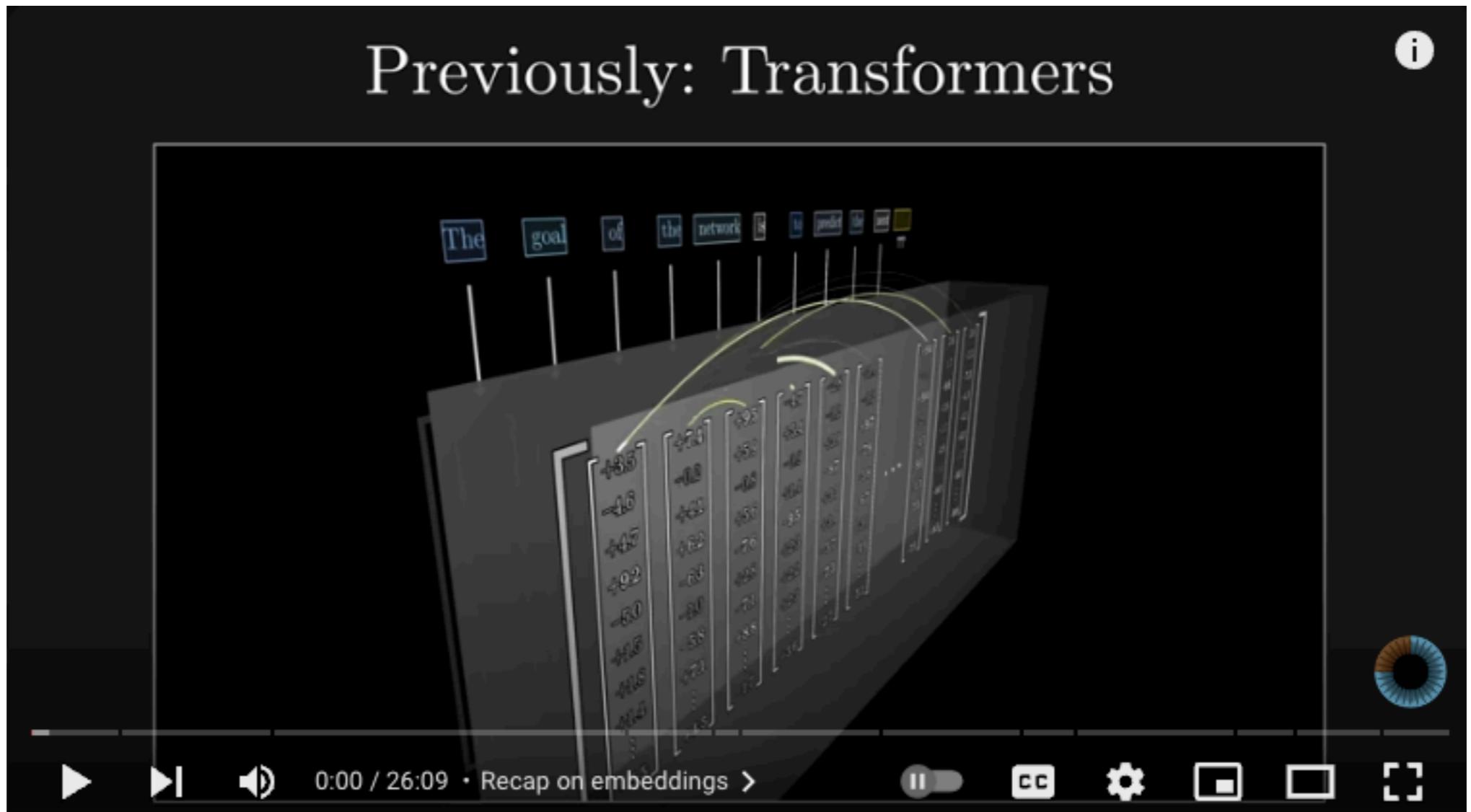
1.0 que são LLMs?



1.0 que são LLMs?



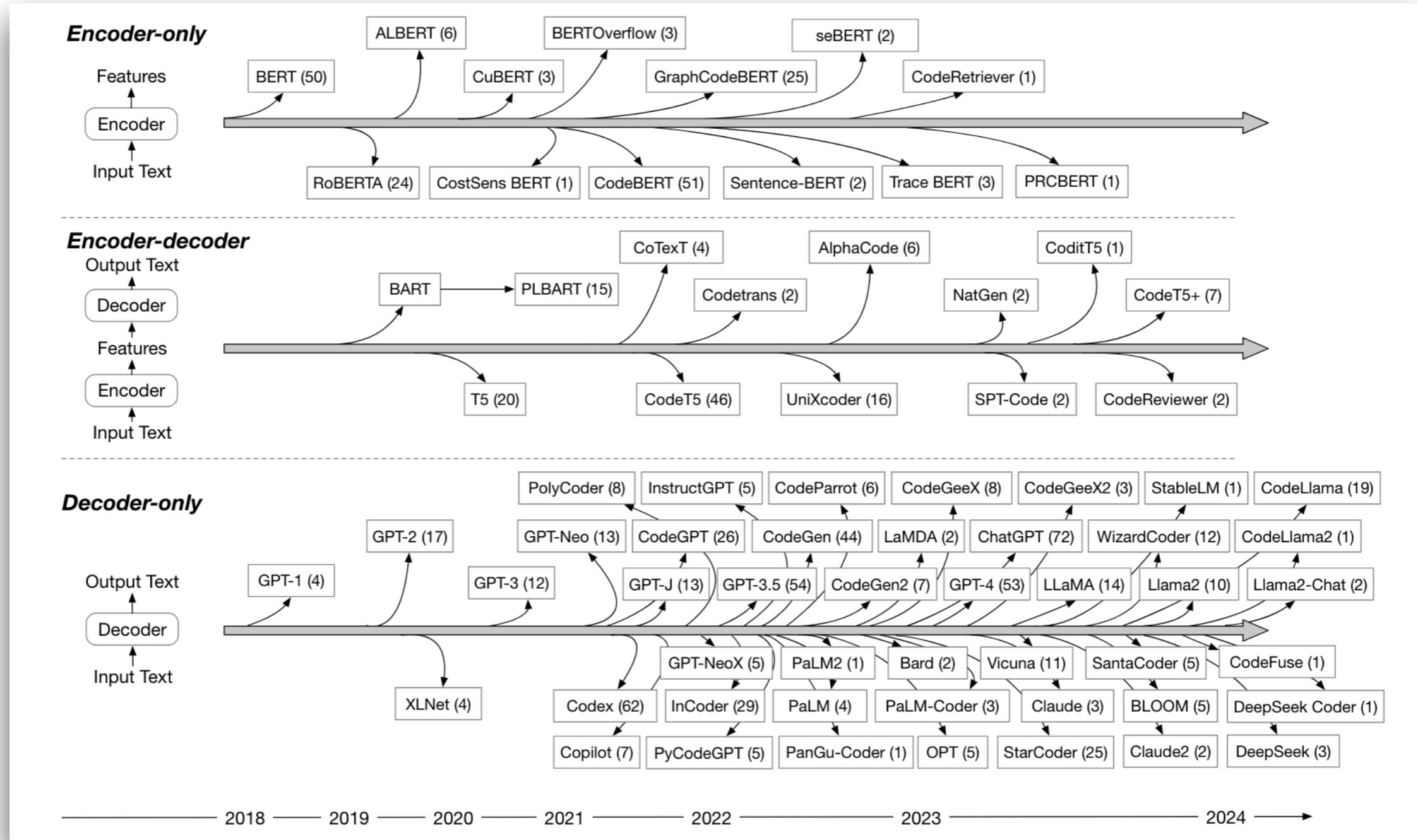
1.0 que são LLMs?



<https://www.youtube.com/watch?v=eMlx5fFNoYc>

@gustavopinto

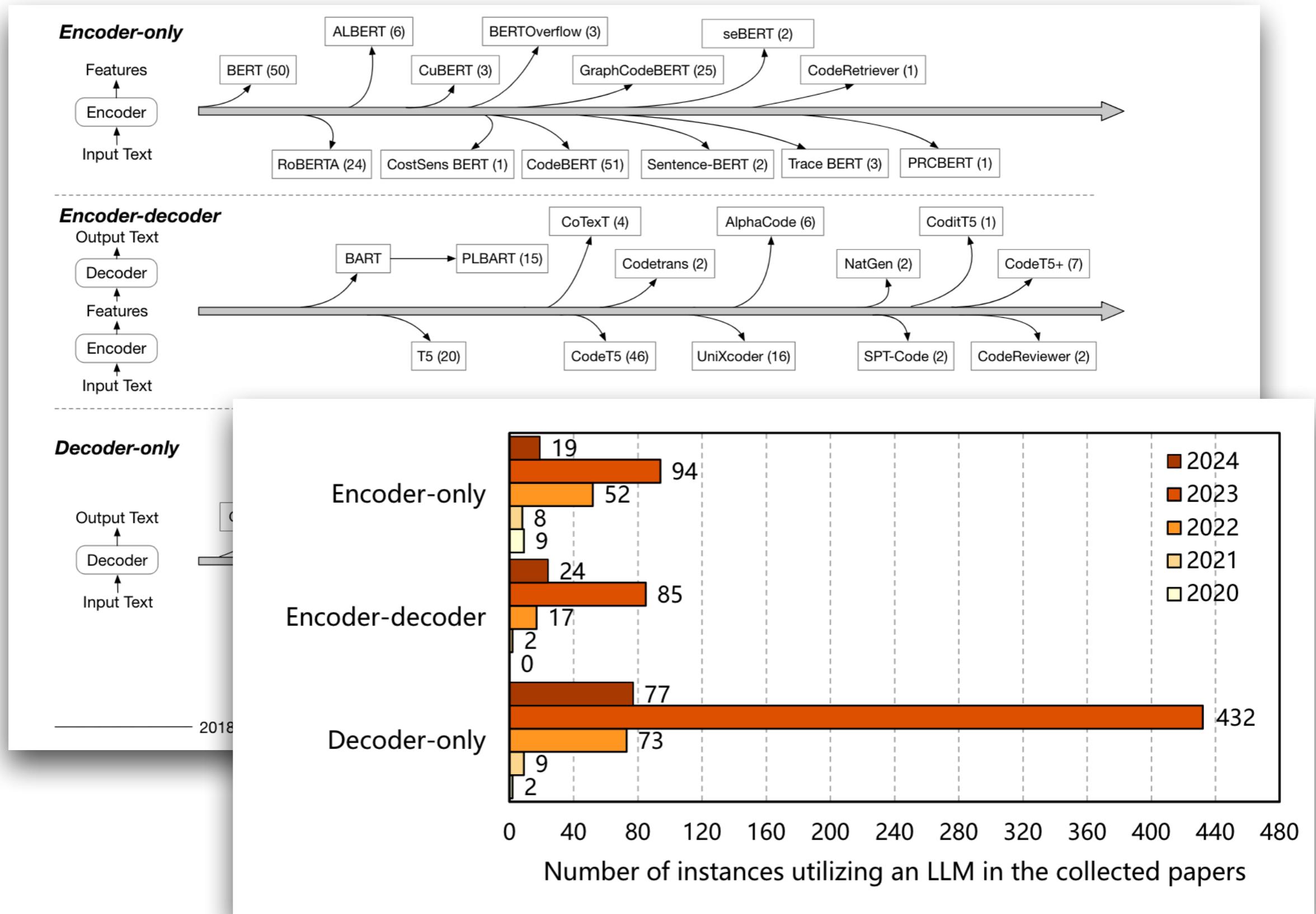
1.0 que São LLMs?



<https://arxiv.org/pdf/2308.10620>

@gustavopinto

1.0 que São LLMs?



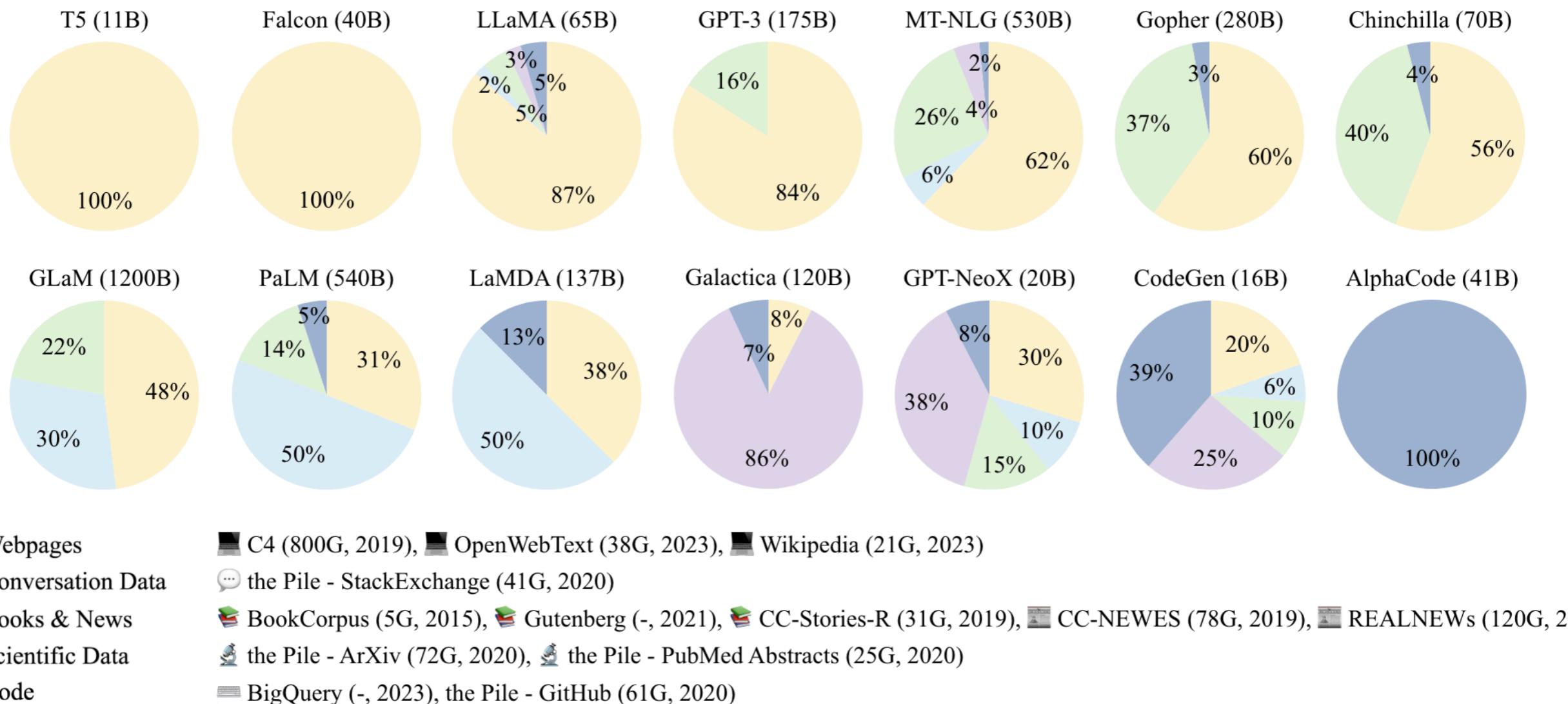


Fig. 6: Ratios of various data sources in the pre-training data for existing LLMs.

<https://arxiv.org/pdf/2303.18223>

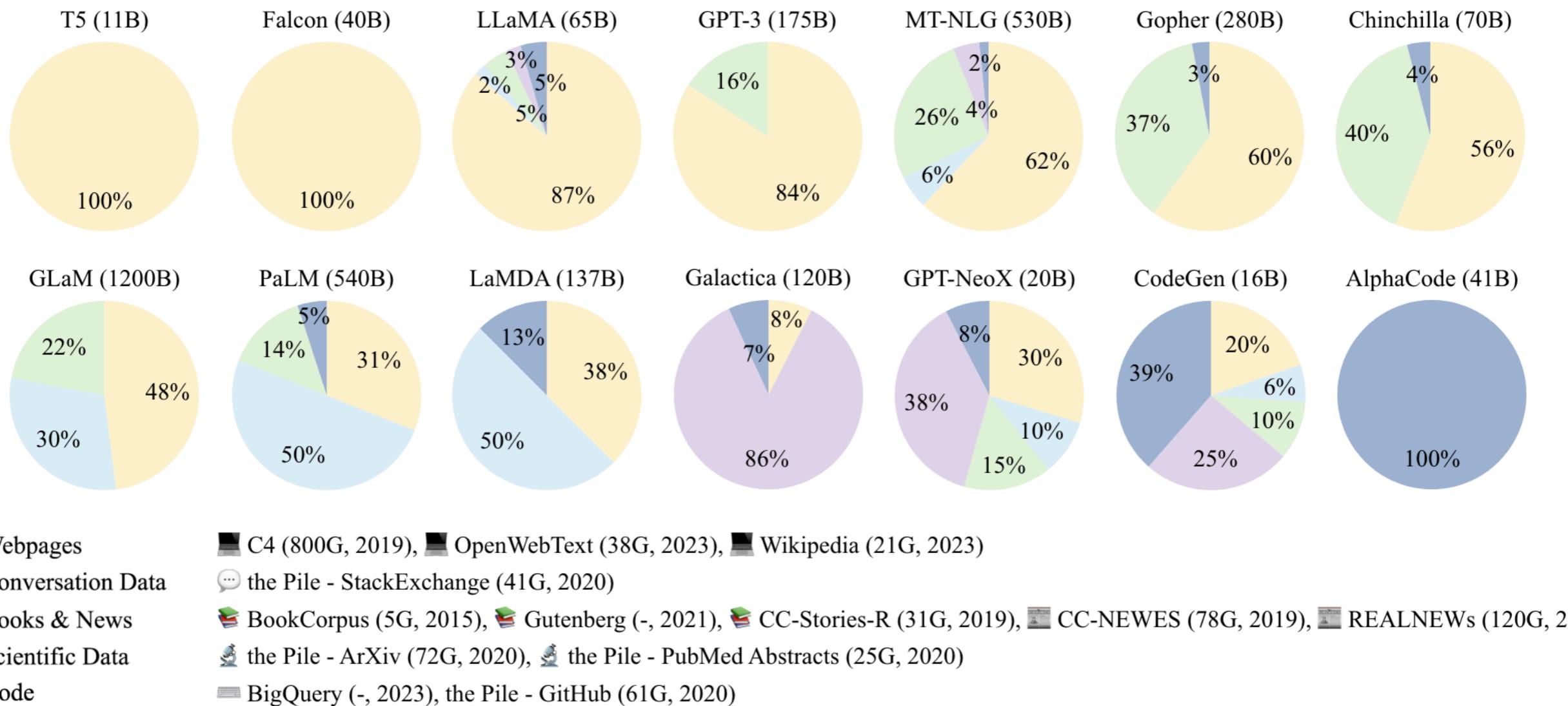


Fig. 6: Ratios of various data sources in the pre-training data for existing LLMs.

<https://arxiv.org/pdf/2303.18223>

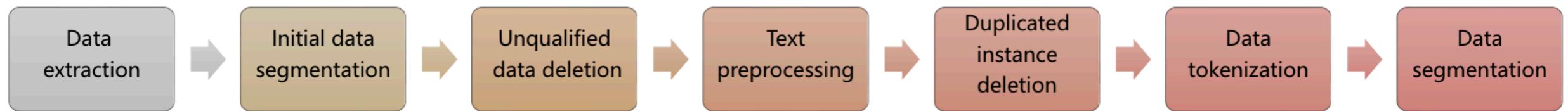


Fig. 7. The data preprocessing procedure for text-based datasets.

GenAI para Devs

The screenshot shows a Substack article page. At the top, it says "ML4SE". The main title is "Um panorama sobre Inteligência Artificial Generativas para devs". Below the title is a subtitle "Entendo os desafios e as opções de ferramentas". There is a profile picture of Gustavo Pinto, the author, with the name "GUSTAVO PINTO" and the date "JUL 14, 2023". Below the author info are social sharing icons for heart, comment, and share, along with a "Share" button and an ellipsis. A link to "Me siga no X" is also present. The article content starts with a paragraph about Generative Artificial Intelligence (GenAI) and its training process using GANs and Transformers. It then discusses the rapid growth of AI tools like Midjourney and ChatGPT compared to traditional web browsers. Following this, there is a section on how GenAI can aid software development teams in tasks like test creation and code generation. Finally, the article addresses the risks associated with GenAI, particularly regarding the lack of determinism and explainability of model outputs.

ML4SE

Um panorama sobre Inteligência Artificial Generativas para devs

Entendo os desafios e as opções de ferramentas

GUSTAVO PINTO JUL 14, 2023

1 1 Share ...

[Me siga no X](#) | [Me siga no LinkedIn](#) | [Apoie a Newsletter](#) | [Solicite uma consultoria](#)

A Inteligência Artificial Generativa (do Inglês, Generative Artificial Intelligence, ou simplesmente “GenAI”) é um sub-ramo da Inteligência Artificial que utiliza de técnica como [GANs](#) ou [Transformers](#) para treinar modelos de aprendizado de máquina em grandes conjuntos de dados. O objetivo é aprender padrões nos dados de treinamento, a fim de serem capazes de gerar dados altamente realísticos.

Diversos produtos que utilizam de técnicas de GenAI —como o Midjourney, ChatGPT e CoPilot— ganharam rapidamente milhões de usuários. O ChatGPT, por exemplo, levou somente dois meses para alcançar 100 milhões de usuários. Como comparação, a internet demorou cerca de 80 meses para ter seus primeiros 100 milhões de usuários. Parte deste destaque está relacionado a capacidade em gerar imagens, textos e código de alta qualidade.

Dados seus desempenhos em diferentes atividades, pesquisadores e profissionais das indústrias em explorado como estas ferramentas podem acelerar a produtividade de times de desenvolvimento de software. Por exemplo, por experiência, percebemos que ferramentas de GenAIs podem melhorar a qualidade do processo de desenvolvimento, 1) ajudando a criar testes a partir de requisitos, 2) detalhando a descrição de requisitos, ou como são mais comumente conhecidos 3) gerando código para uma funcionalidade.

Por outro lado, GenAI pode criar novos riscos uma vez que seus resultados não são nem determinístico (ou seja, pequenas variações nos dados de entrada ou nas condições do modelo podem levar a saídas diferentes) nem explicáveis (ou seja, é difícil compreender como exatamente o modelo chegou a uma determinada saída ou decisão). Estudos recentes indicam que uso de ferramentas como [Co-Pilot](#) pode invariavelmente gerar código com bugs, ou testes de baixíssima cobertura.

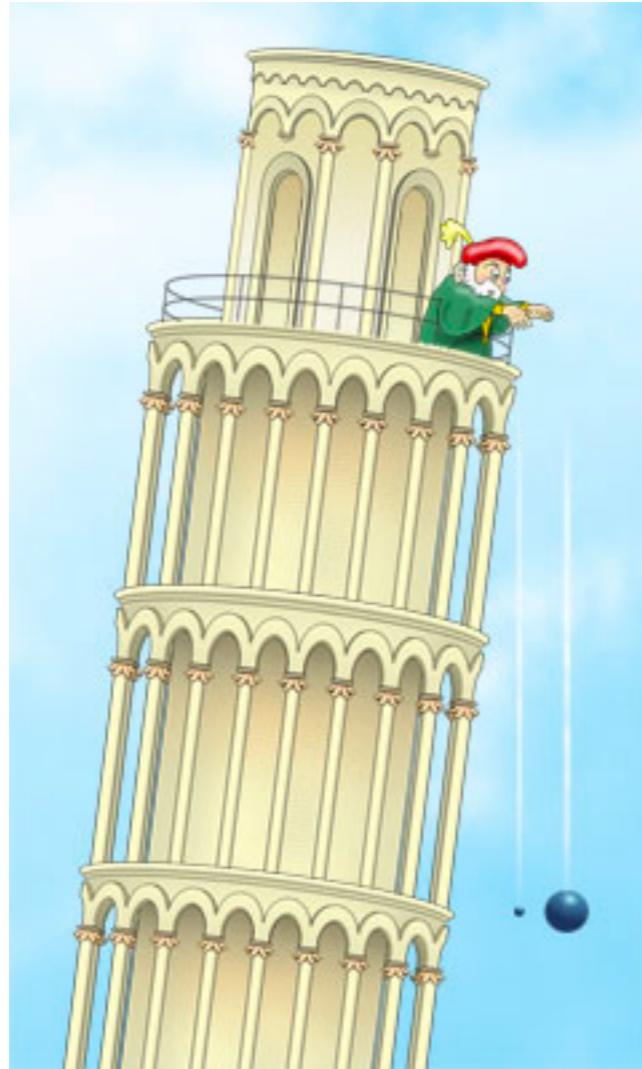
- Benefícios da GenAI
- Ferramentas de GenAI
- Riscos de GenAI

1.0 que é um modelo?

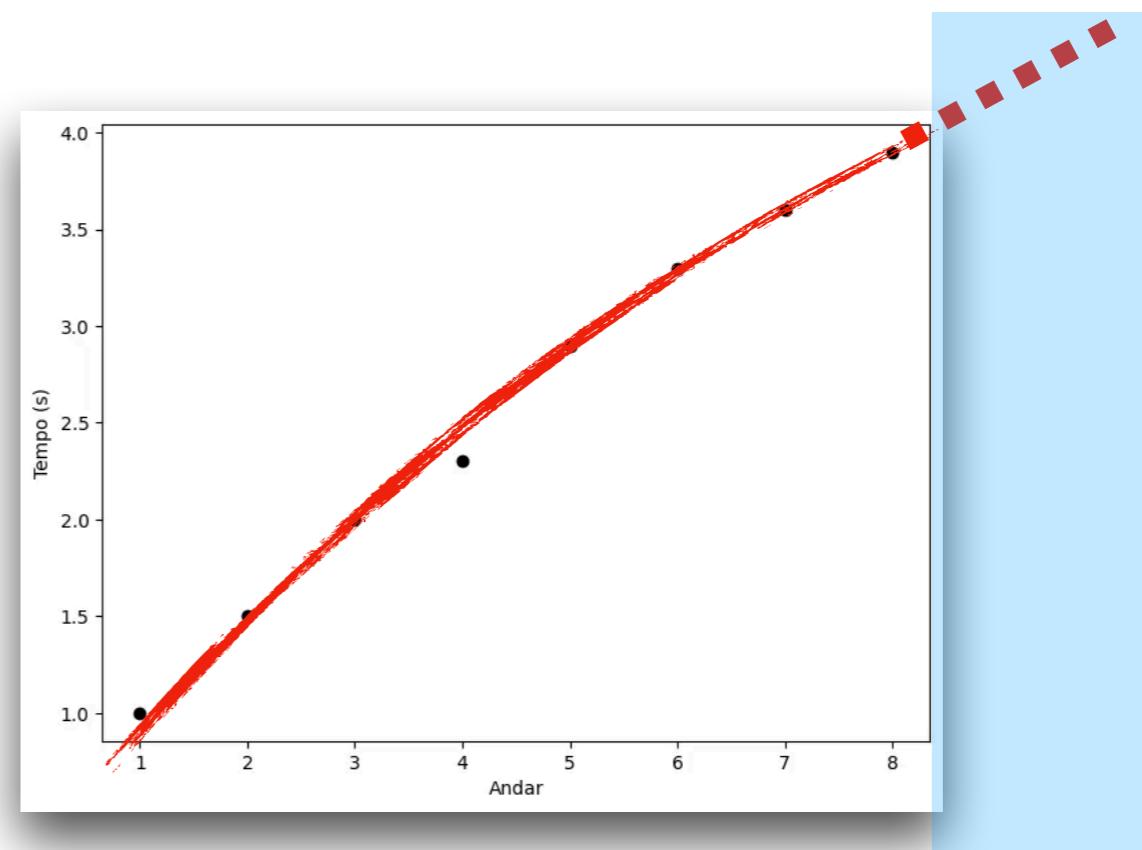


@gustavopinto

1. O que é um modelo?



Galileu, que observou que o tempo de descida das bolas é independente da suas massa!



?



@gustavopinto

1. O que é um modelo?

Um **modelo** de aprendizado de máquina é uma representação matemática que foi **treinada** em um **conjunto de dados** fornecidos como entrada para fazer certos tipos de **previsões**.

$$Y = M(A, D)$$

Regressão Linear

Vetor numérico

1. O que é um modelo?

Mas o que seria um:
Grande Modelo de Linguagem (ou LLM)?

$$Y = M(A, D)$$

Bilhões (ou trilhões)

Redes Neurais

The diagram illustrates the components of a large language model. It features a mathematical equation $Y = M(A, D)$. Above the equation, the word "Bilhões" (Billions) is written in red, with a red arrow pointing from it to the function M . Below the equation, the words "Redes Neurais" (Neural Networks) are written in red, with another red arrow pointing from them to the same function M .



I. O que é um modelo?

The screenshot shows the Hugging Face homepage. At the top, there's a search bar and navigation links for Models, Datasets, Spaces, Docs, Solutions, Pricing, Log In, and Sign Up. The main area features a large banner with a smiling emoji and the text "The AI community building the future." Below the banner, it says "The platform where the machine learning community collaborates on models, datasets, and applications." To the right of the banner is a sidebar with categories like NLP, Computer Vision, and Audio. The right side of the page displays a list of trending models, such as "meta-llama/Llama-2-7b" and "stabilityai/stable-diffusion-xl-base-1.0". Below this is a section titled "Trending on 😊 this week" with cards for "tiiuae/falcon-180B", "AI Comic Factory", "fka/awesome-chatgpt-prompts", and others. At the bottom, there are links to "Browse 300k+ models", "Browse 100k+ applications", and "Browse 50k+ datasets".

<https://huggingface.co/>

- GitHub para LLMs
- Todos os principais modelos estão disponíveis



@gustavopinto

I. O que é um modelo?

Hello world

ML4SE

Criando o modelo de aprendizado de máquina mais simples do mundo

GUSTAVO PINTO MAY 12, 2023

Share ...

Me siga no X | Me siga no LinkedIn | Apoie a Newsletter | Solicite uma consultoria

O termo “modelo” no contexto de aprendizado de máquina é recorrentemente utilizado para indicar a utilização dos algoritmos em um conjunto de dados, empregados em tarefas de classificação ou previsões, por exemplo.

ML4SE é uma newsletter focada em ajudar devs a mastigar técnicas de aprendizado de máquina. Para receber novos posts, se inscreva na newsletter.

✓ Subscribed

Mas o que seria esse tal “modelo”?

Suponha por um instante que você quer similar o estudo de Galileu, em que ele investigou quanto tempo demoraria a trajetória de duas bolas (com o mesmo volume mas diferente massas) que foram lançadas de diferentes andares da torre de Pisa, até alcançar o chão.

Se você for um experimentalista, talvez sua primeira tentativa seria subir na torre de Pisa e jogar as duas bolas, várias vezes, de cada um dos 8 andares da torre. Ao final desse dia de trabalho, você potencialmente chegaria a mesma conclusão que Galileu, que observou que o tempo de descida das bolas é independente da suas massas. Se fizéssemos um gráfico das anotações dos tempos de descida da bola por andar, talvez chegássemos em algo parecido com a figura abaixo.

<https://ml4se.substack.com/p/criando-um-modelo-de-ml>

ML4SE

Sumarizando READMEs de projetos de software livre

GUSTAVO PINTO JUN 29, 2023

Share ...

Me siga no X | Me siga no LinkedIn | Apoie a Newsletter | Solicite uma consultoria

Sumarização é o processo de condensar um texto ou documento para uma forma mais concisa, mantendo as informações essenciais e principais ideias.

A tarefa de sumarização envolve a geração de um resumo que captura o significado e o contexto do texto original, permitindo aos leitores obter uma visão geral do conteúdo sem a necessidade de ler todo o documento.

A sumarização pode ser realizada de duas formas principais: sumarização extrativa (produz um resumo extraído frases que representam em conjunto as informações mais relevantes do conteúdo original) e abstrativa (produz um resumo gerando novas frases do documento que resumem a ideia principal).

A sumarização é uma técnica que pode ser aplicada em diversos contextos, como resumos de notícias, resumos de artigos científicos ou livros. No processo de desenvolvimento de software, podemos também utilizar técnicas de sumarização em áreas como:

- Extração de requisitos: A sumarização pode ser usada para extrair as informações relevantes dos documentos de requisitos. Isso pode ajudar os times de desenvolvimento a identificar rapidamente os principais objetivos ou restrições do sistema.
- Análise de logs: Em sistemas de registro de eventos (logs), A sumarização pode ser aplicada para extrair as informações de registros de eventos (logs) relevantes, facilitando a compreensão de padrões de uso da aplicação ou até detectando eventuais eventos críticos no software.

<https://ml4se.substack.com/p/sumarizando-readmes-de-projetos-de>