



# LLMs and their applications

---

JADS

ERIK TROMP - UNDERSTANDING

# LLMs

---

What are LLMs (Large Language Models)?

# LLMs

---

You already know this 😊

# LLMs

Well, in essence we have seen them already

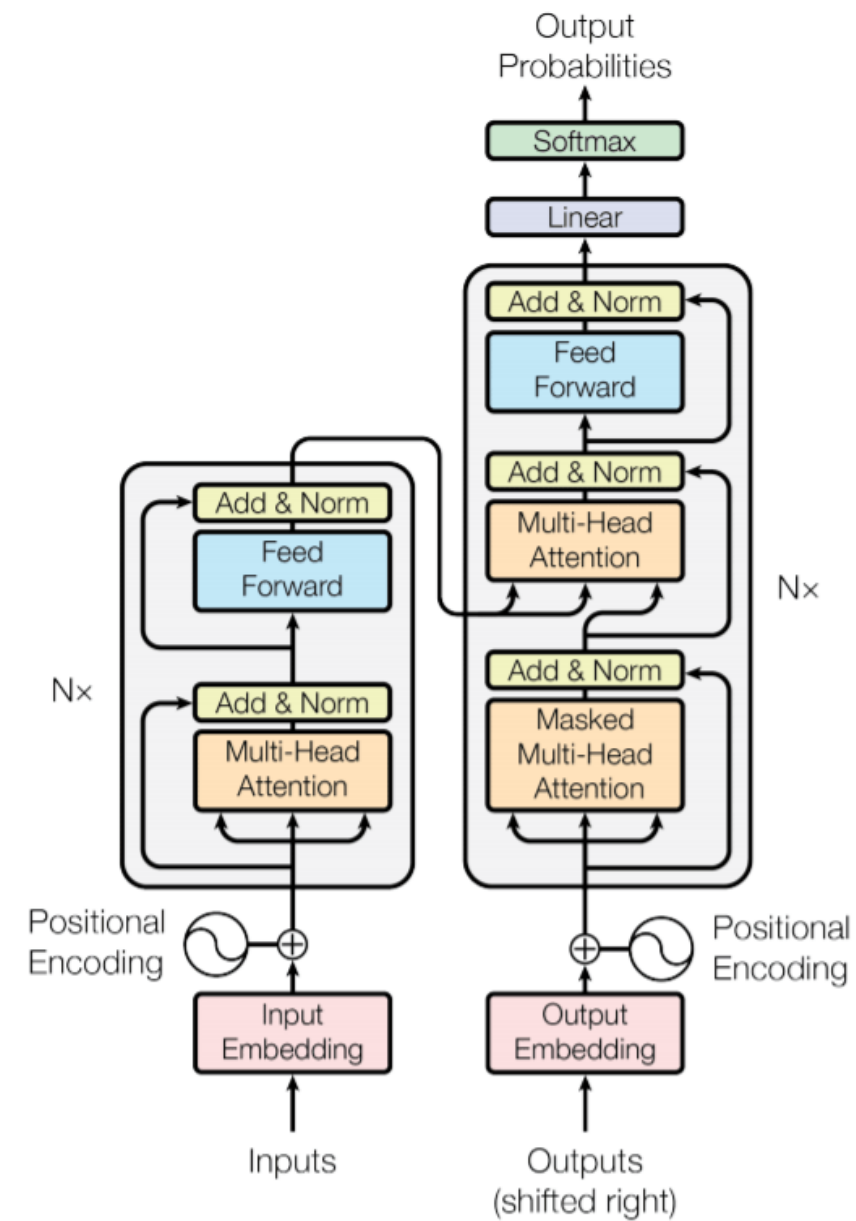


Figure 1: The Transformer - model architecture.



# LLMs

Well, in essence we have seen them already

But then with billions of parameters

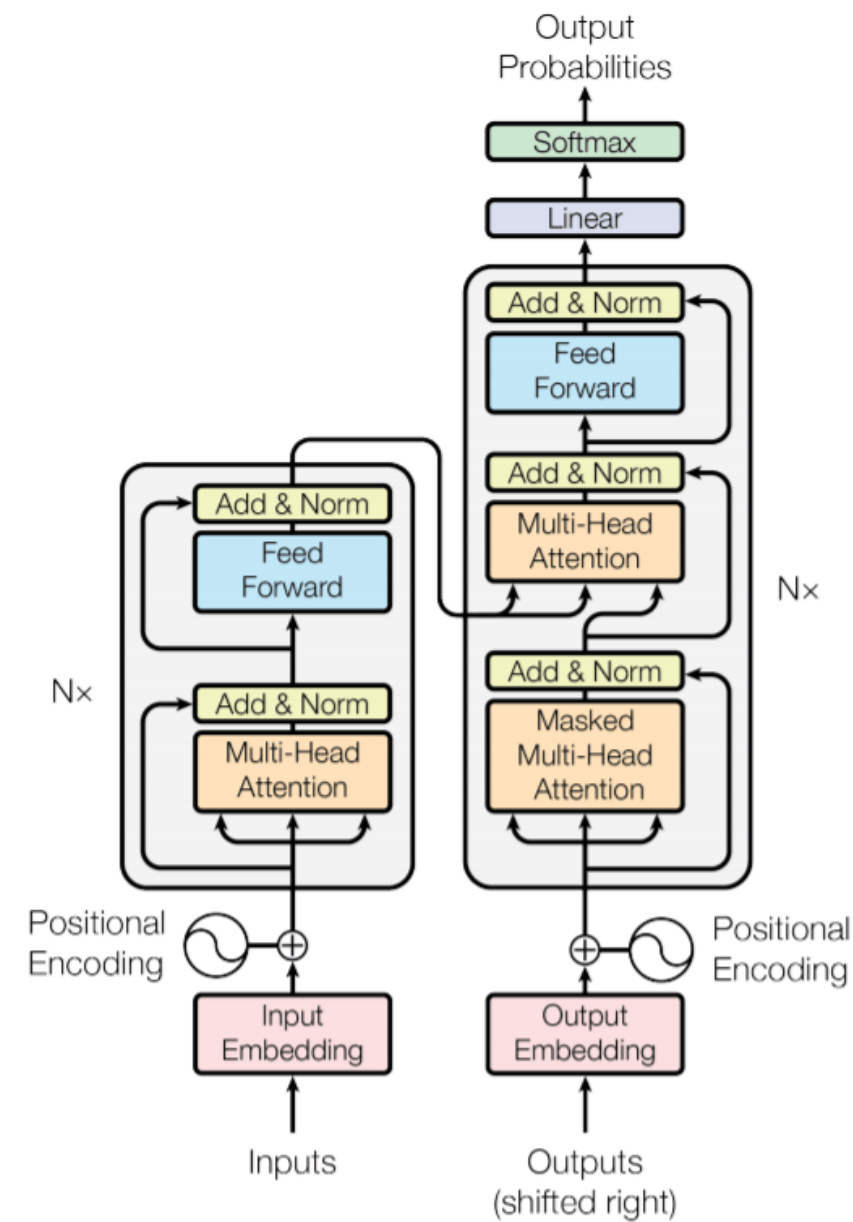


Figure 1: The Transformer - model architecture.

# LLMs

Well, in essence we have seen them already

But then with billions of parameters

Trained on hundreds of gigabytes to petabytes  
of data

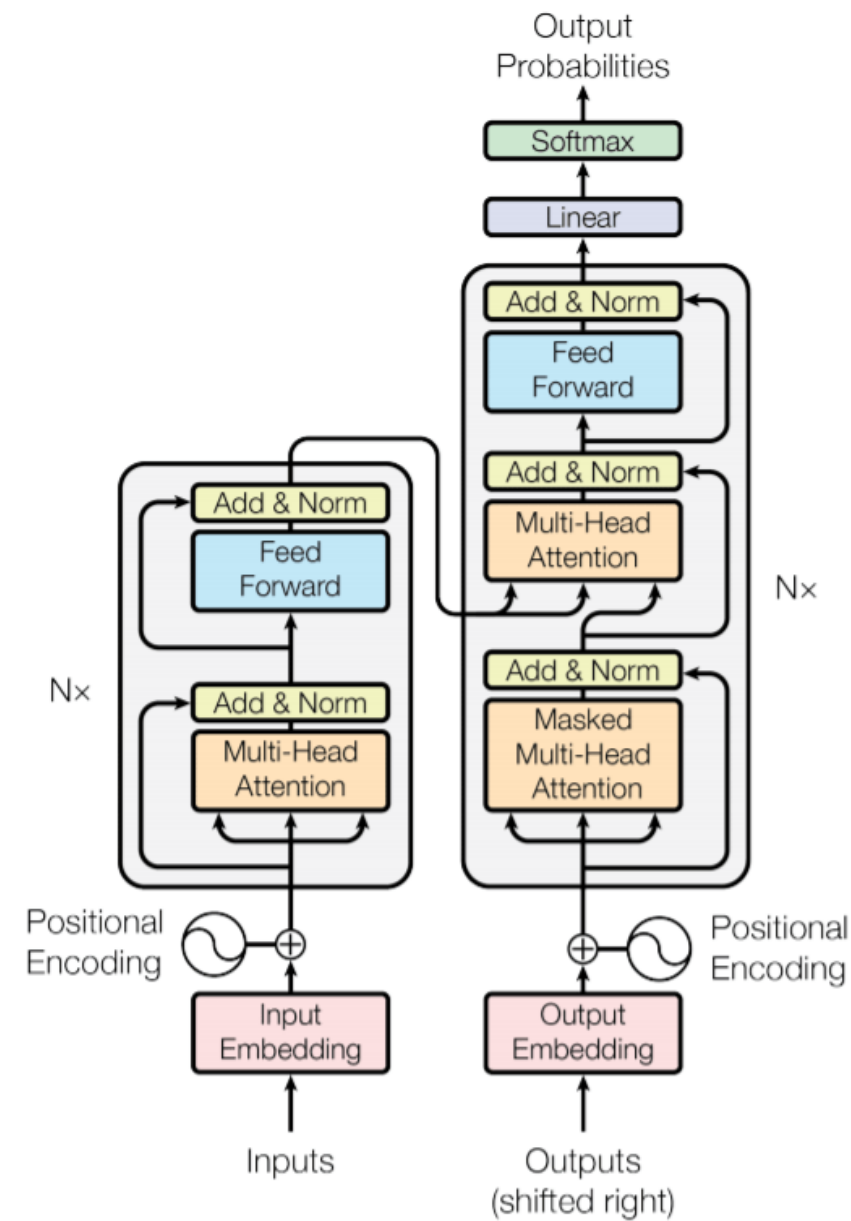


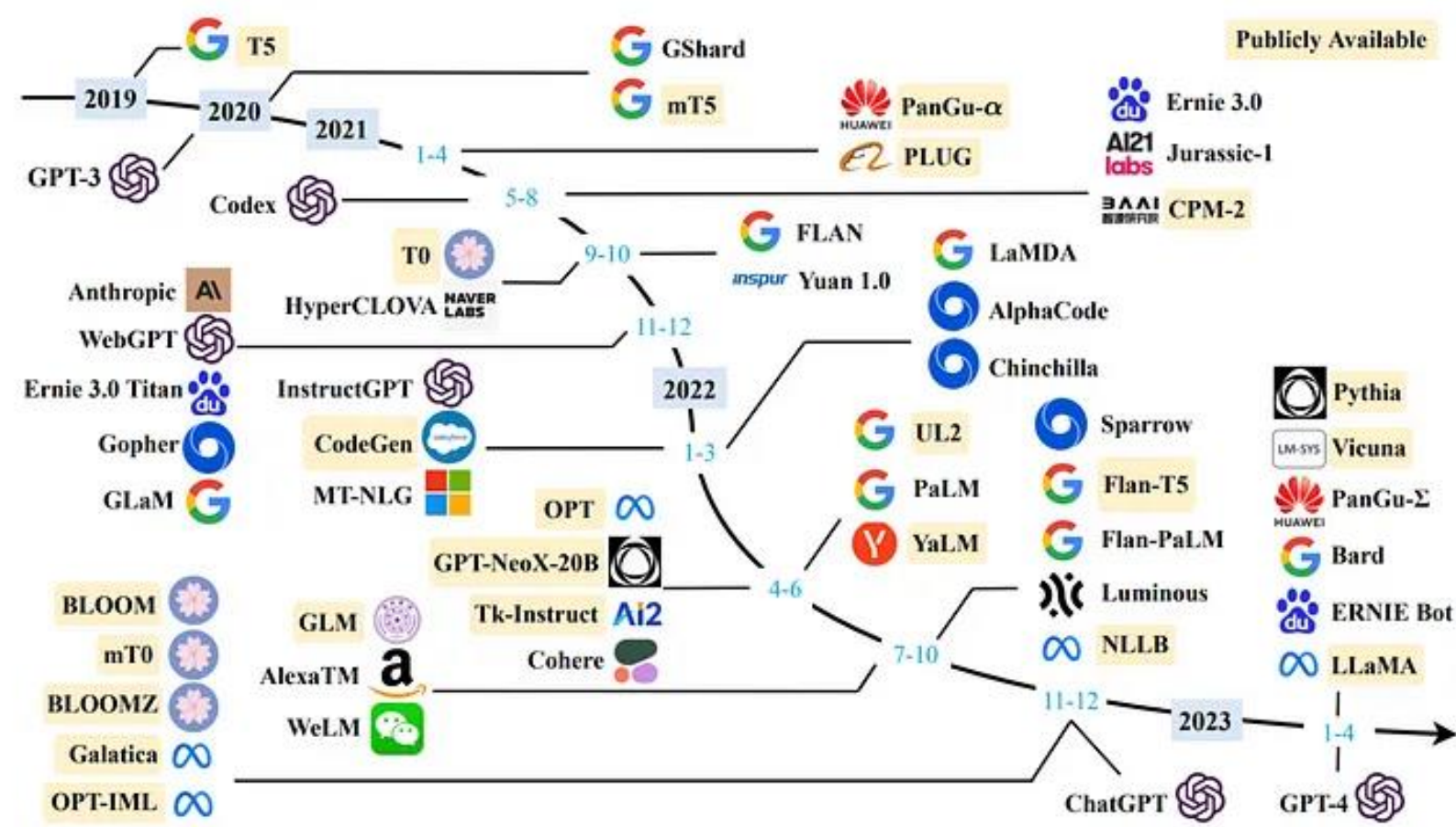
Figure 1: The Transformer - model architecture.

# LLMs

## Landscape of LLM Companies

Category	Company	Found	Funding (06/2023)	LLMs	License	Serving
Leading LLM Companies	OpenAI	2015	\$11.3B	GPT-3.5, GPT-4 (ChatGPT)	Licensed	API
	Anthropic	2021	\$1.5B	Claude	Licensed	API
	Cohere	2019	\$435M	Command	Licensed	API
	Cerebras	2015	\$715M	Cerebras-GPT	Open Source	Local
	Aleph Alpha	2019	\$31M	Luminous	Licensed	API
	AI21 Labs	2017	\$118.5M	Jurassic-2	Licensed	API
Big Commercial Players	Google	1998	-	PALM-2 (Bard)	Licensed	API
	Meta	2004	-	LLaMA	Open Source (nonCommercial)	Local
	Databricks	2013	\$3.5B	Dolly	Open Source	Local
Others	HuggingFace	2017	\$160M	HuggingChat, Bloom, StarCoder	Open Source	Local
	StabilityAI	2019	\$114M	StableLM	Open Source	Local
	EleutherAI	2020	-	GPT-J, GPT-NeoX	Open Source	Local
	MosaicML	2021	\$64M	MPT-7B	Open Source	Local
	H2O.ai	2011	\$251M	-	-	-
	LMSYS	2011	-	Vicuna	Open Source	Local
	Hippocratic	2023	\$50M	-	Licensed	-

# LLMs





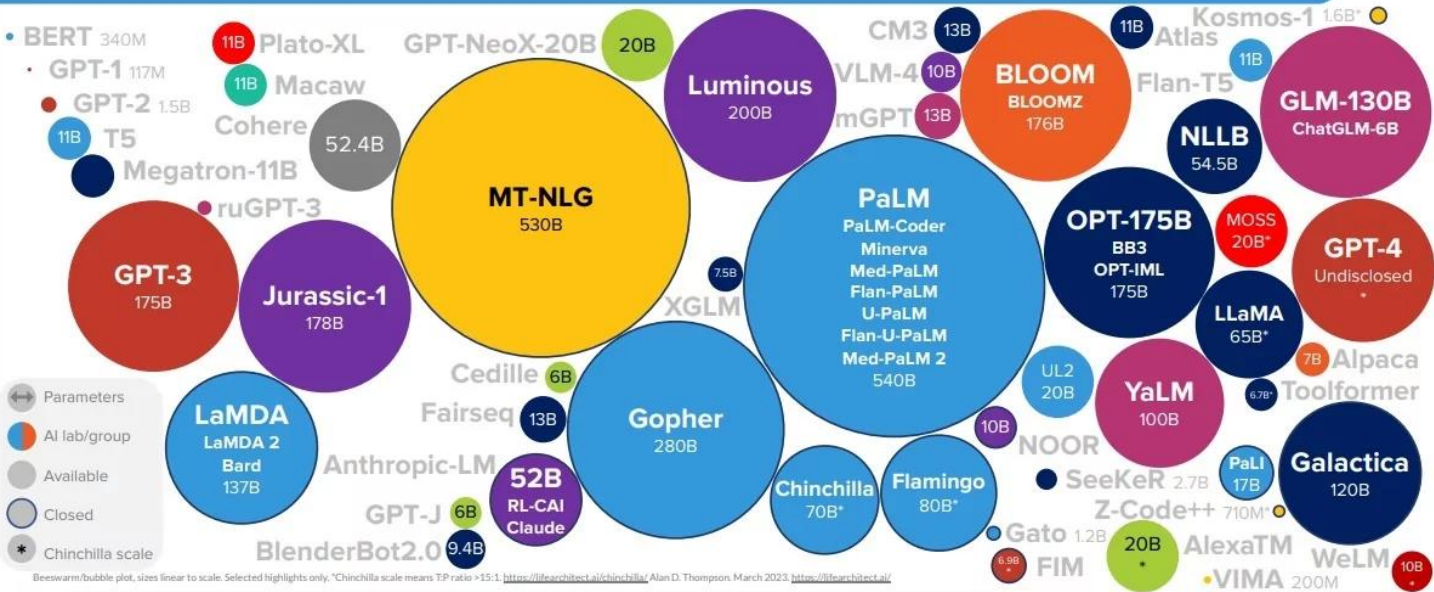
# LLMs

---

So how big are these models?

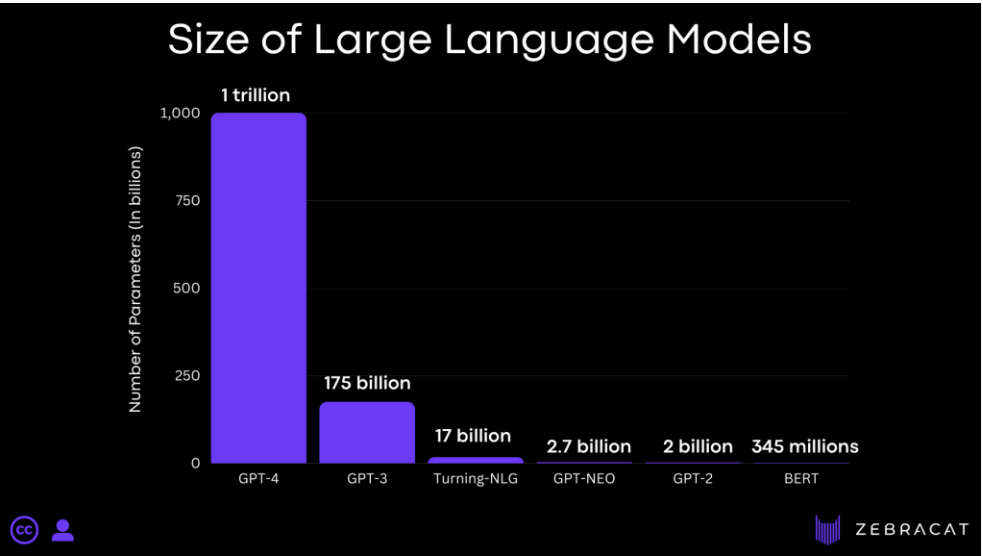
# LLMs

## LANGUAGE MODEL SIZES TO MAR/2023



[LifeArchitect.ai/models](https://life architect.ai/models)

## Size of Large Language Models



# LLMs

---

What are they trained on?

# LLMs

---

📄 Wikipedia

📄 Common crawl

📄 The Pile

# LLMs

---

## Wikipedia

Interesting, well-structured but rather small (20GB for English)

## Common crawl

## The Pile



# LLMs

---

## ❑ Wikipedia

Interesting, well-structured but rather small (20GB for English)

## ❑ Common crawl

Websites crawled monthly, petabytes in total, tens of terabytes per month

## ❑ The Pile

# LLMs

---

## ❑ Wikipedia

Interesting, well-structured but rather small (20GB for English)

## ❑ Common crawl

Websites crawled monthly, petabytes in total, tens of terabytes per month

## ❑ The Pile

A combined (22) dataset of 825GB, often used as standard

# LLMs

---

What more data do we have?

# LLMs

---

But be warned!

# LLMs

---

## *Franzen, Grisham and Other Prominent Authors Sue OpenAI*

The suit, filed with the Authors Guild, accuses the A.I. company of infringing on authors' copyrights, claiming it used their books to train its ChatGPT chatbot.

## **Authors' lawsuit against OpenAI could 'fundamentally reshape' artificial intelligence, according to experts**

Plaintiffs include famous authors like George R.R. Martin and Jodi Piccoult.

## **'New York Times' considers legal action against OpenAI as copyright tensions swirl**

AUGUST 16, 2023 · 5:53 PM ET



# LLMs

---

Training a single LLM from scratch will cost you around 1M-10Ms

We call this a foundation model

# LLMs

---

Then you run your “InstructGPT”

# LLMs

---

Then you run your “InstructGPT”

50k-100k examples

# LLMs

---

Then you run your “InstructGPT”

50k-100k examples

Gets you a conversational model

# LLMs

---

Then you finetune using RLHF



# LLMs

---

Then you finetune using RLHF

Hundreds to thousands of examples

# LLMs

---

Then you finetune using RLHF

Hundreds to thousands of examples

Gives you conversational towards your use-case

# LLMs

---

(Or nowadays, use GPTs)

# LLMs

---

## **Creators are leaking their data by using Custom GPTs**

Protect your prompts and data from leakage

One of many examples

<https://medium.com/@mcraddock/creators-are-leaking-their-data-by-using-custom-gpts-fbaa530c89ee>

# LLMs

---

LLaMA (Meta)

2048 x A100 GPUs

23 Days

+/- \$30M

GPT 3

1024 x A100 GPUs

34 Days

+/- \$4.6M

Falcon 40B

384 x A100 GPUs

60 Days

+/- \$2M



# LLMs

---





You want to run an LLM yourself for inference?



# LLMs

---

## GPU RAM required

 LLaMA 65B	<b>260GB</b>
 Falcon 40B	<b>160GB</b>
 LLaMA 33B	<b>132GB</b>
 Falcon 7B	<b>28GB</b>

# LLMs

---

## Challenges in LLMs

- ❑ Self-hosting near impossible
- ❑ Data security/privacy
- ❑ Model licensing
- ❑ Context windows (we are going further and further though)
- ❑ Hallucination
- ❑ Bias and fairness privacy

# LLMs

---

So far, this was Large Language Models...

# LLMs

---

So far, this was Large Language Models...

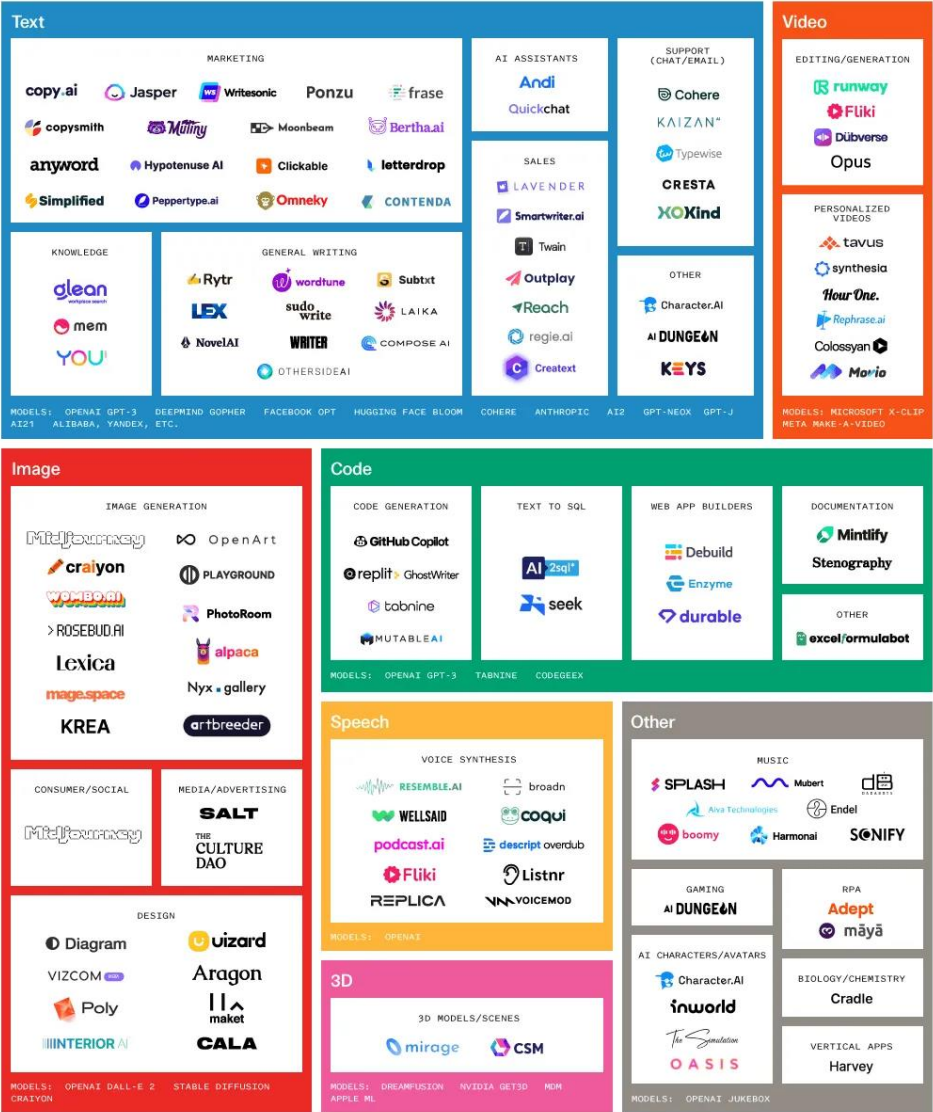
But GenAI is more than language!



# LLMs

# The Generative AI Application Landscape v2

A work in progress



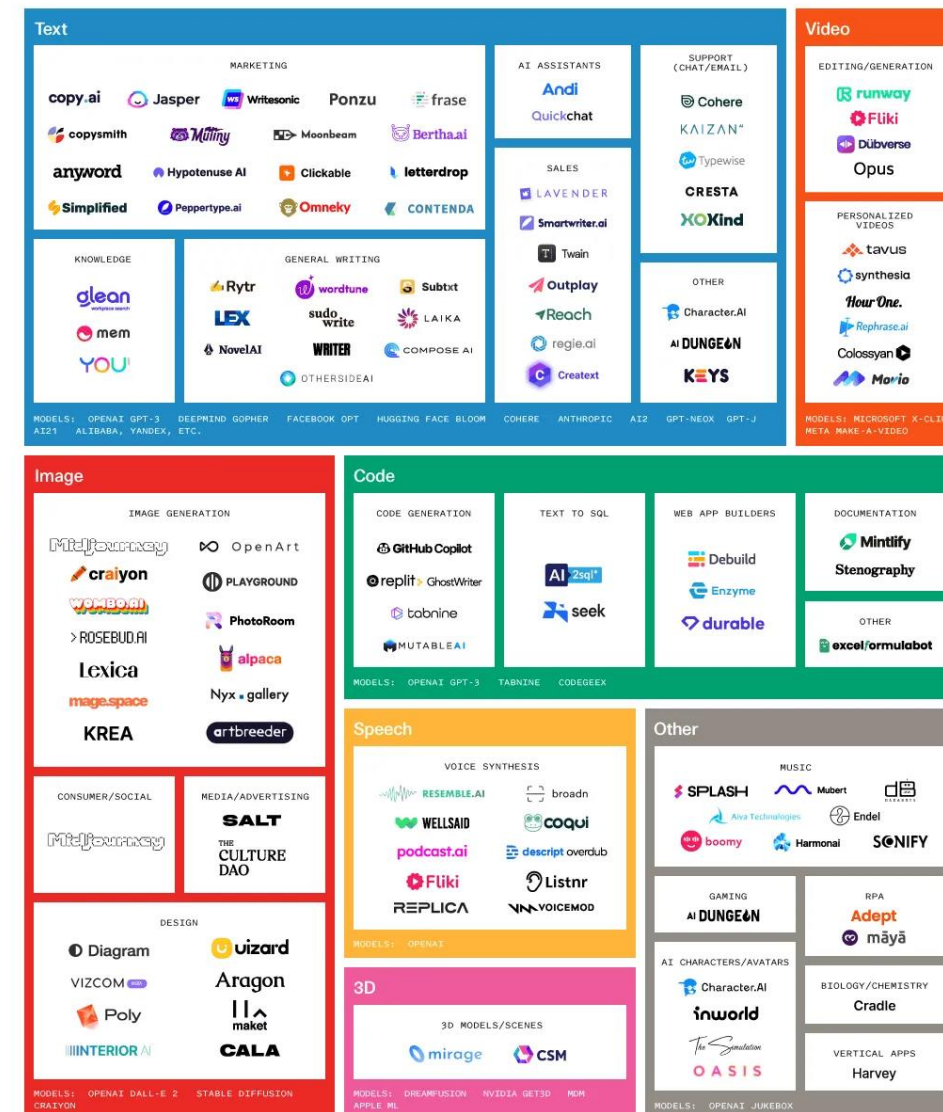
# LLMs

What about?

- ❑ Image generation?
- ❑ 3D generation?
- ❑ (Text-to-)Speech generation?
- ❑ Music generation?
- ❑ Video generation?
- ❑ “Talking head”/lipsync generation?

# The Generative AI Application Landscape v2

A work in progress



# LLMs

---

So we have seen GenAI for language, how do we do it for these fields?



# LLMs

---

How do the techniques differ from language?

# LLMs

They don't 😊

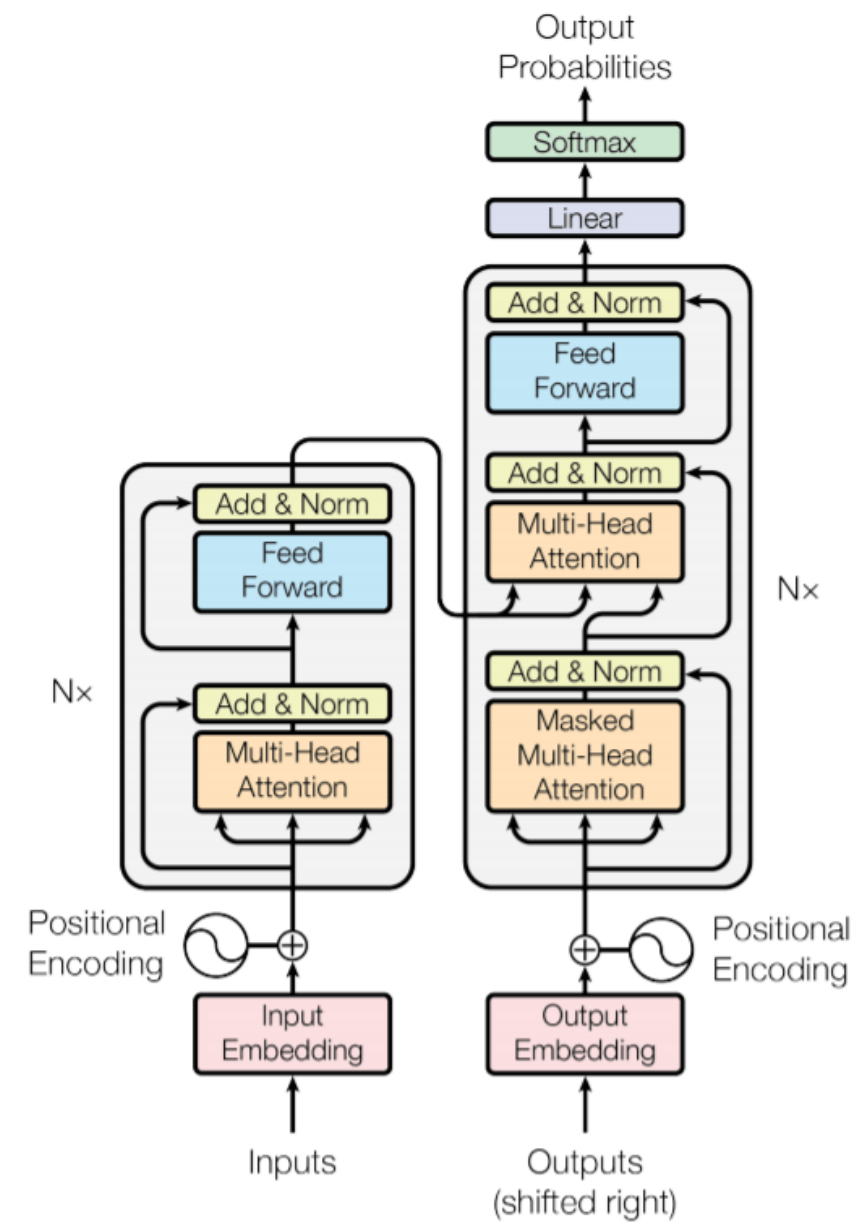


Figure 1: The Transformer - model architecture.

# LLMs

---

## Examples

- ❑ ChatGPT / Claude / Bard – text (code: Co-Pilot)
- ❑ Dall-E / Midjourney / StableDiffusion – image
- ❑ Whisper / Speedbrain – speech
- ❑ Gen-2 by runway / D.ID – video / talking head
- ❑ Suites: HeyGen / Canva / Adobe CC

# Questions?

---

Thanks for this DL ride

[erik@understandling.com](mailto:erik@understandling.com)

<https://www.understandling.com/>

<https://www.linkedin.com/in/eriktromp/>

<https://www.udemy.com/course/the-definitive-intro-to-big-data-science/?referralCode=47E37006CFC5B5599874>

<https://www.udemy.com/course/every-data-architecture-is-the-same/?referralCode=CD1F8BF2562089D63617>

