OpenAI &
AI Advancements in 2023

Arlo – SMART AI

Ugo Jardonnet – 03/30/2023

INTRODUCTION

- ChatGPT is amazing
- OpenAI is at the top of the AI game but there is a lot of competition
- New Arlo Products are coming

What is ChatGPT ?

- A Large Language Model (LLM)
  - Text Completion
  - Translation
  - …
- ChatGPT = Generative Pretrained Transformer, optimized for Chat
- March 2023
  - GPT3 = LLM from 2019 (finetunable, not opt. for chat)
  - ChatGPT (free) = GPT-3.5-turbo
  - ChatGPT Plus = GPT4

- Training Database: 5Tb … 15Tb?
- Model weights: 175B … 1T?
- Runs on Nvidia HGX with 8 A100 (1 A100=25k$) -> H100

But also

- System prompt / prompt engineering
- Filters and plugins
- RLHF a specific training for dialog and safety

Risks:

Cost

Bias

Disinformation

Misuse

Economic shock

Environmental Impact

Requires:

Research

Regulation

Traceability

(Arlo Image Watermarking)

OpenAI Products

- GPT 3, 3.5, 4 – Chatbot and Text completion (LLM)
- Codex – Code completion (LLM)
- Whisper – Speech Recognition (Transformer)
- CLIP – Text Image Similarity (VLM based on ViT)
- Dalle2 – Image Generation (CLIP guided latent diffusion)

TRANSFORMER MODELS

OpenAI models are Transformer models (or based on Transformer models)

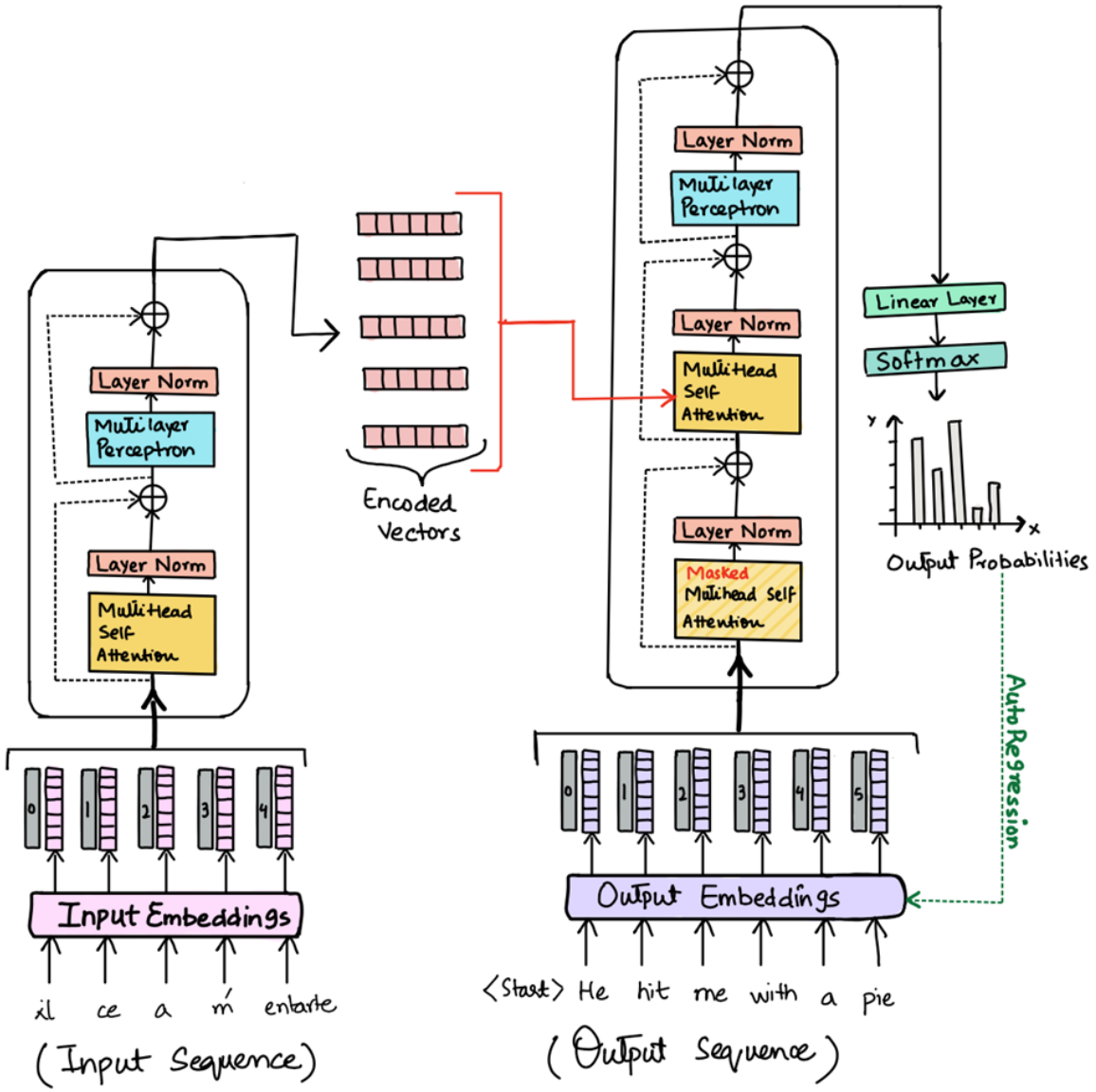
Transformer is the state-of-the-art model architecture for most AI tasks in 2023:

Text, image, audio, video…

Attention is all you need, Vaswani et al. 2017, Google, Toronto University, link

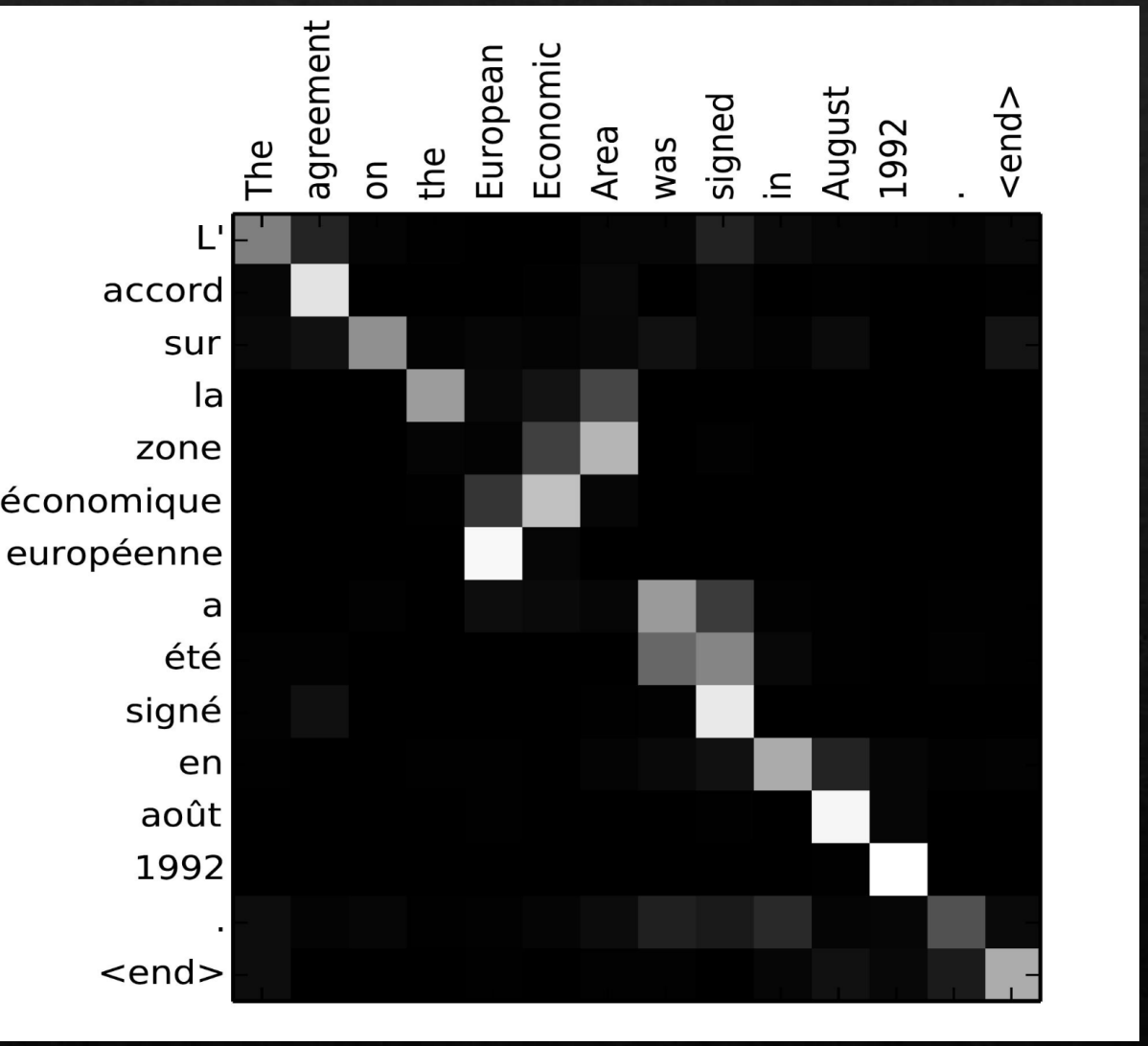
The Transformers Architecture[1]

- Auto regressive = Predicts next word at each call
- Attention based



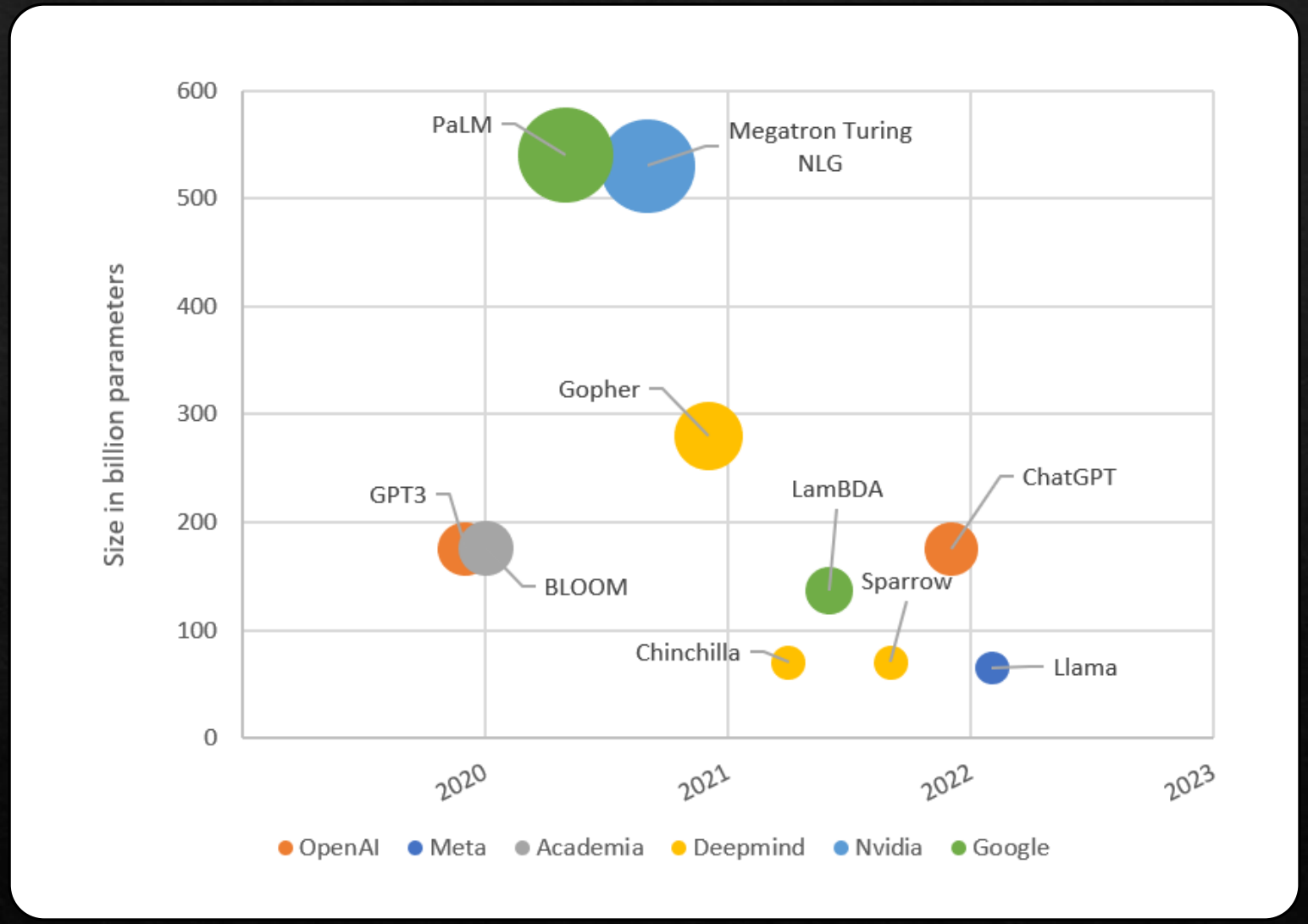
ATTENTION LAYER

- Attention layer has quadratic complexity
- Can be efficiently processed on GPU but still requires a lot of processing power



Competition

- New LLMs are released al the time
- Deepmind Sparrow, papers but no demo
- A few models are already above 1000B parameters: Google Switch Transformer, GPT4?



Much smaller, high performance, models are coming…

		BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA
GPT-3	175B	60.5	81.0	-	78.9	70.2	68.8	51.4	57.6
Gopher	280B	79.3	81.8	50.6	79.2	70.1	-	-	-
Chinchilla	70B	83.7	81.8	51.3	80.8	74.9	-	-	-
PaLM	62B	84.8	80.5	-	79.7	77.0	75.2	52.5	50.4
PaLM-cont	62B	83.9	81.4	-	80.6	77.0	-	-	-
PaLM	540B	**88.0**	82.3	-	83.4	**81.1**	76.6	53.0	53.4
LLaMA	7B	76.5	79.8	48.9	76.1	70.1	72.8	47.6	57.2
	13B	78.1	80.1	50.4	79.2	73.0	74.8	52.7	56.4
	33B	83.1	82.3	50.4	82.8	76.0	**80.0**	**57.8**	58.6
	65B	85.3	**82.8**	**52.3**	**84.2**	77.0	78.9	56.0	**60.2**

Table 3: **Zero-shot performance on Common Sense Reasoning tasks.**

Meta LLaMA: Open and Efficient Foundation Language Models, Touvron et al. , Feb 2023, link

ARLO AI in Development

Bibliography

[1] ‘Attention is all you need, Vaswani et al. 2017, Google, Toronto University, link

[2] ‘Learning Transferable Visual Models From Natural Language Supervision’, Radford et al., 2021, link

[3] ‘High-Resolution Image Synthesis with Latent Diffusion Models’, Rombach et al. , 2022, link

Sparks of Artificial General Intelligence: Early experiments with GPT-4, Bubeck et al. , March 2023, link

https://futureoflife.org/open-letter/pause-giant-ai-experiments/