BCS - Hertfordshire

AI State Of Play

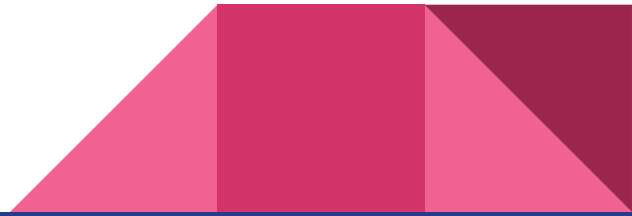
Richard Simon
Ameer Ogirimah

University of Hertfordshire

Talk overview + notes

- Pre-ChatGPT
- ChatGPT and the rise of its competitors
- LLM Current State of Play
- Sustainably Building LLM-Based Applications using the following method:
  - Fine-tuning open-source models with your domain data
  - Using available LLM APIs to build
  - Integrating LLM Software solutions into your Application

- The Future Prospects and Possible Challenges of the LLM Community



Agenda (draft)

1. The big players in the LLM community

Microsoft/OpenAI, Google, Facebook.

1 slide each - state of play, what they're up to, what's happening and what they're bringing out in 2024

2. Available models: Open-source and Closed-source

Summary list the models, number of params, release dates, model types (GPTs = language, text, image, etc.)

3. Relevant Libraries and Frameworks

Summary list of PyTorch (FB), Langchain, TensorFlow (Google), etc.

4. Requirements for building a model

Computing power, domain data (raw data, 'enriched data', etc), Vector Database

5. Different ways models are being used (LLMs)

Sentiment Analysis, Text Generation/Summary, Code Gen, Specific Domains (health, law, etc)...

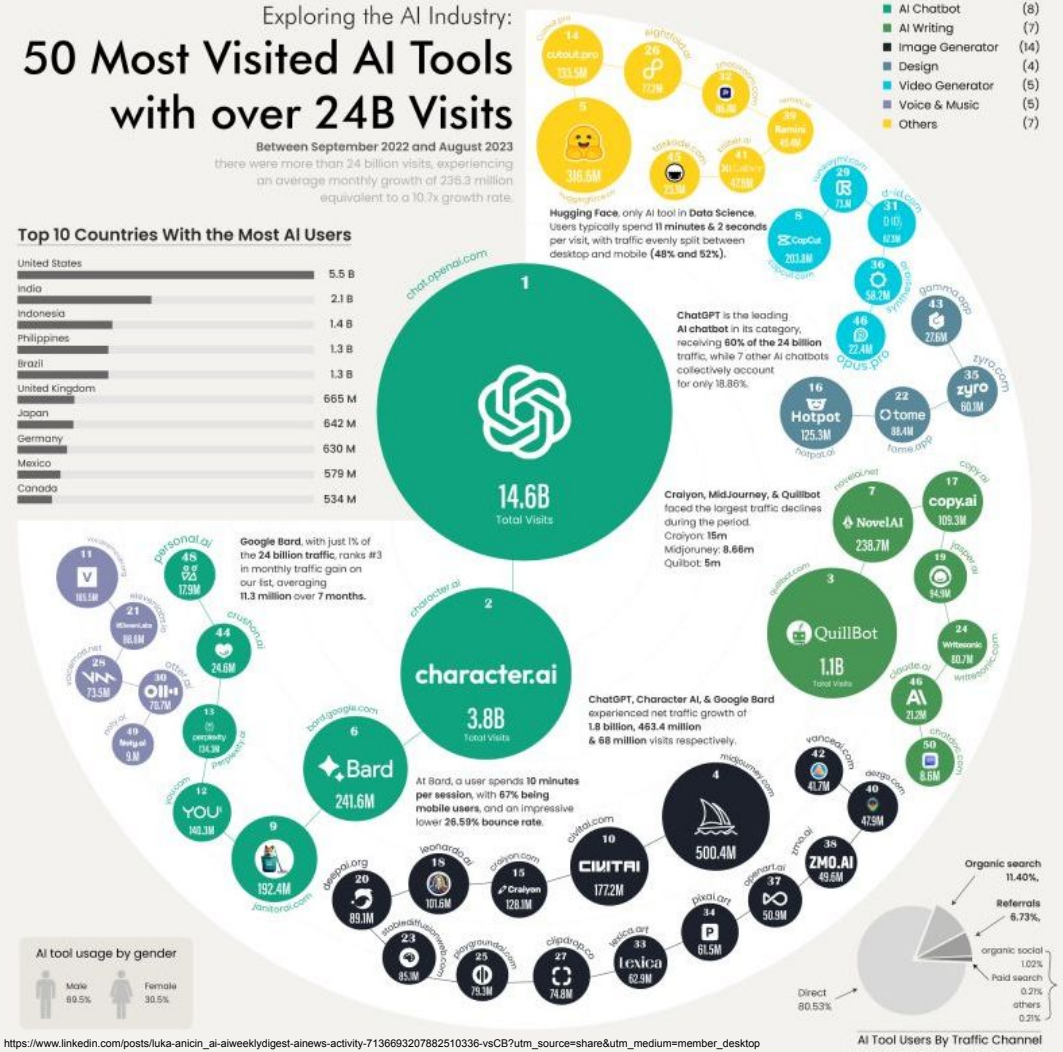
6. Limitations of LLMs

Bias, Hallucinations, Hardware resources(GPU, TPU, and associated energy)…

7. And Opportunities available due to the growth of LLMs

Commercial opportunities, Agents (OpenAI GPTs)...

Why now?



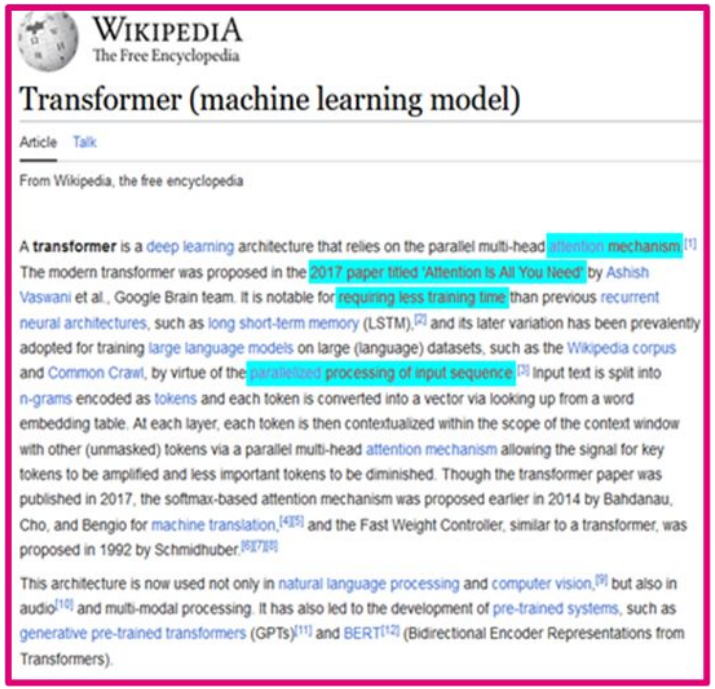
https://www.linkedin.com/posts/luka-anicin_ai-aiweeklydigest-ainews-activity-7136693207882510336-vsCB?utm_source=share&utm_medium=member_desktop

"It's an understatement to say that 2023 has been the year [of the return] of AI."

- Paul Bevan (Director of Infra Research @ Bloor Research)

bloor.com

What Has Changed?



WIKIPEDIA
The Free Encyclopedia

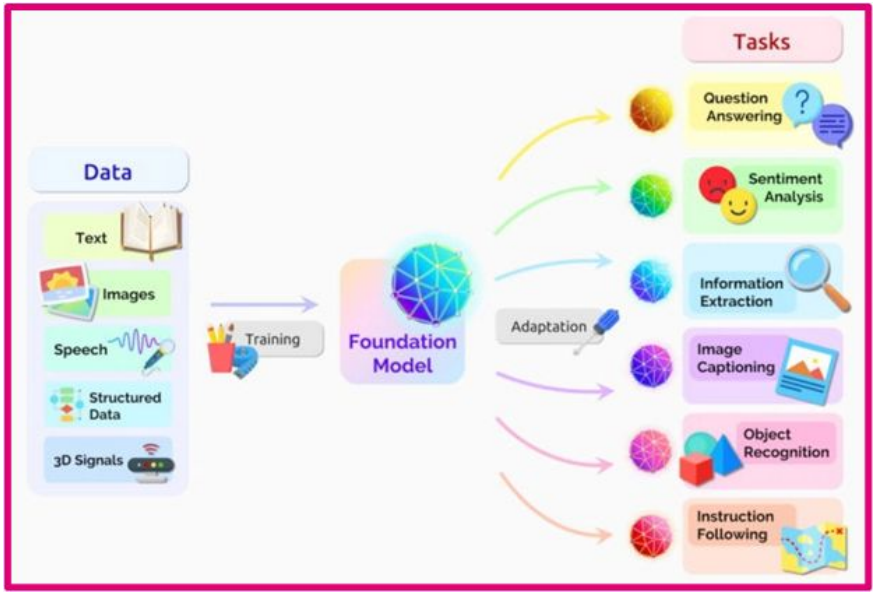
Transformer (machine learning model)

Article Talk

From Wikipedia, the free encyclopedia

A **transformer** is a deep learning architecture that relies on the parallel multi-head attention mechanism.[1] The modern transformer was proposed in the 2017 paper titled 'Attention is All You Need' by Ashish Vaswani et al., Google Brain team. It is notable for requiring less training time than previous recurrent neural architectures, such as long short-term memory (LSTM),[2] and its later variation has been prevalently adopted for training large language models on large (language) datasets, such as the Wikipedia corpus and Common Crawl, by virtue of the parallelized processing of input sequence.[3] Input text is split into n-grams encoded as tokens and each token is converted into a vector via looking up from a word embedding table. At each layer, each token is then contextualized within the scope of the context window with other (unmasked) tokens via a parallel multi-head attention mechanism allowing the signal for key tokens to be amplified and less important tokens to be diminished. Though the transformer paper was published in 2017, the softmax-based attention mechanism was proposed earlier in 2014 by Bahdanau, Cho, and Bengio for machine translation,[4][5] and the Fast Weight Controller, similar to a transformer, was proposed in 1992 by Schmidhuber.[6][7][8]

This architecture is now used not only in natural language processing and computer vision,[9] but also in audio[10] and multi-modal processing. It has also led to the development of pre-trained systems, such as generative pre-trained transformers (GPTs)[11] and BERT[12] (Bidirectional Encoder Representations from Transformers).



Pattern > Inference > Response.

https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/

Why ChatGPT Now?

OpenAI + Microsoft

- ✓ 2019 Microsoft invested $1 bln in OpenAI
- ✓ 2020 GPT-3 licensed to Microsoft
- ✓ 2021 GitHub Copilot
- ✓ 2022 ChatGPT announced
- ✓ 2023 Microsoft invested $10 bln in OpenAI
- ✓ 2023 Azure OpenAI Service GA
- ✓ 2023 Microsoft Bing AI
- ✓ 2023 GPT-4
- ✓ 2023 Microsoft 365 Copilot announced
- ✓ 2023 Microsoft Designer
- ✓ 2023 AI Copilot in Microsoft Power Apps
- ✓ 2023 Microsoft Bing Image Creator

The Verge / Tech

Microsoft invests $1 billion in OpenAI to pursue holy grail of artificial intelligence

/ Building artificial general intelligence is OpenAI's ambitious goal

Jul 22, 2019, 3:08 PM GMT+1

https://www.youtube.com/watch?v=LwLnhg0fna8

https://www.theverge.com/2019/7/22/20703578/microsoft-openai-investment-partnership-1-billion-azure-artificial-general-intelligence-agi

1 LLM Big Players [state of play]

Big Players: Microsoft

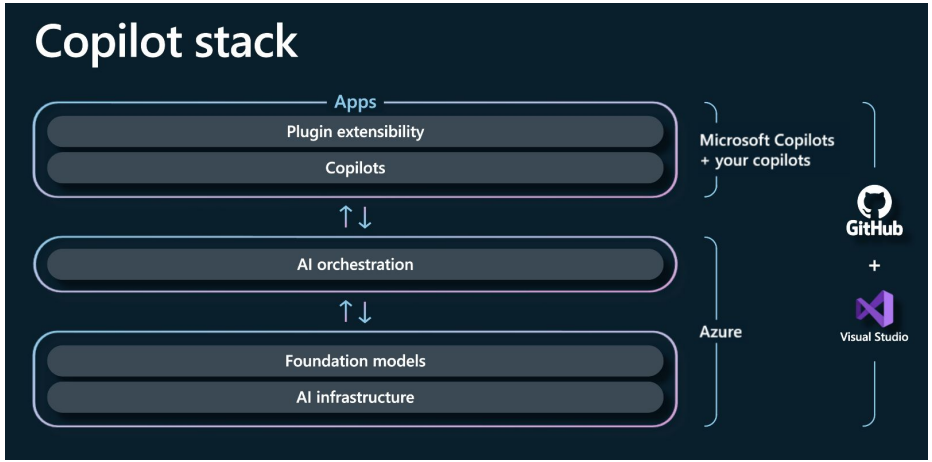
Fully committed to AI in all their platforms

Copilot - base platform for all products

DALL-E - ‘borrowed’ from OpenAI

ChatGPT3.5 + 4

ChatGPT4 Turbo - Multimodal model, available now



Future Releases:

Windows 10 + 11 Copilot (Beta available now)

ChatGPT5

Azure Maia - purpose-built hyperscale data centre for AI accelerator silicon + Maia 100 - AI Accelerator chip

Cobalt - Cloud CPU (ARM-based) - general purpose

OpenAI Project Q* (Artificial General Intelligence?)

Big Players: Google

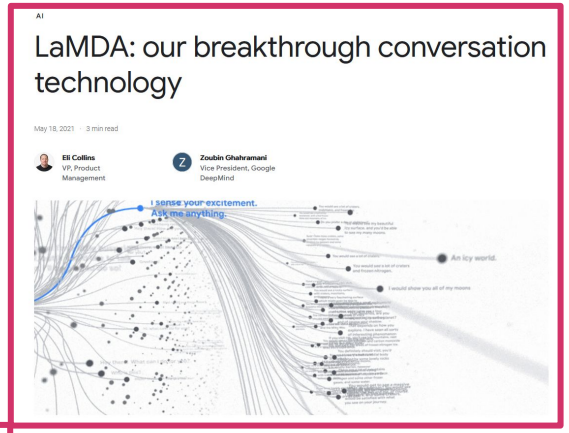
Google Brain research lab proposed **Transformer** architecture in 2017

Bard - based on:

LaMDA (Language Model for Dialogue Applications)

Transformer Neural Network Architecture

Google still playing catch-up with OpenAI/Microsoft



https://blog.google/technology/ai/lamda/

https://blog.research.google/2017/08/transformer-novel-neural-network.html

https://entreprenal.com/googles-sparrow-will-kill-chatgpt-it-is-microsoft-teams-vs-slack-all-over-again-da8c5a69c58f

Future Releases:

Gemini

Makersuite (Available in US only, to be released soon)

Big Players: Meta (FB)

Meta AI - incorporating AI into all Meta services

AI-driven FB + Insta feeds/recommendations

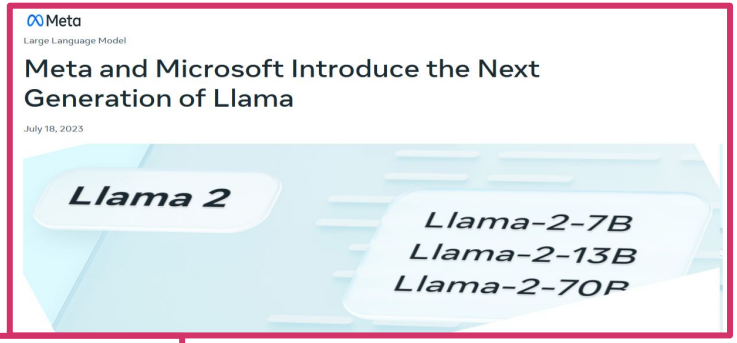
Llama/Llama2 - Open Source models

Future Releases:

Continue with Open Source strategy

Open Source attracts better talent

Put AI into every Meta service in 2024



https://ai.meta.com/blog/llama-2/

https://ai.meta.com/llama/

https://www.tradingview.com/news/tradingview:ca3da3396094b:0-meta-s-2024-strategy-prioritizes-ai/

https://www.fool.com/investing/2023/11/01/mark-zuckerberg-reveals-advantage-meta-ai-strategy/

Big Players: Amazon Web Services

SageMaker - ML model training for devs = IaaS

BedRock - Foundation Models for GenAI = ‘AIPaaS’

Augment with Fine Tuning, RAG and Agents

PartyRock - App Builder for BedRock = Low Code/No Code

https://aws.amazon.com/generative-ai/

Future Releases:

Trainium2 - AI Accelerator chip

Amazon Q - ???

Graviton4 - Cloud CPU (ARM-based, 30% faster than predecessor)

2 Available Models

Available Models

ChatGPT 3/4 Turbo

DALL-E

TTS

Whisper

Moderation

Copilot Platform

GitHub Copilot

Azure OpenAI

Bing Chat

365 Copilot

Microsoft OpenAI

Imagen (Text-to-Speech Diffusion model)

Chirpy (Speech model)

Codey (Code completion and generation)

Muse (Text-to-Speech Transformer model)

Vertex AI Model training and deployment platform

Google

IBM Granite

Nvidia:

StyleGAN3

EG3D

Megatron 530B LLM

IBM NVIDIA

Computer Vision:

Detectron2

DensePose

Language:

Seamless; Llama

Meta

3 Libraries + Frameworks

Torch

Pytorch

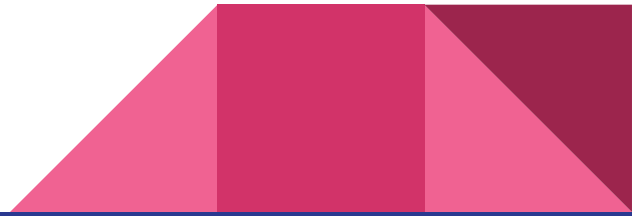
MxNet

Tensorflow

Keras

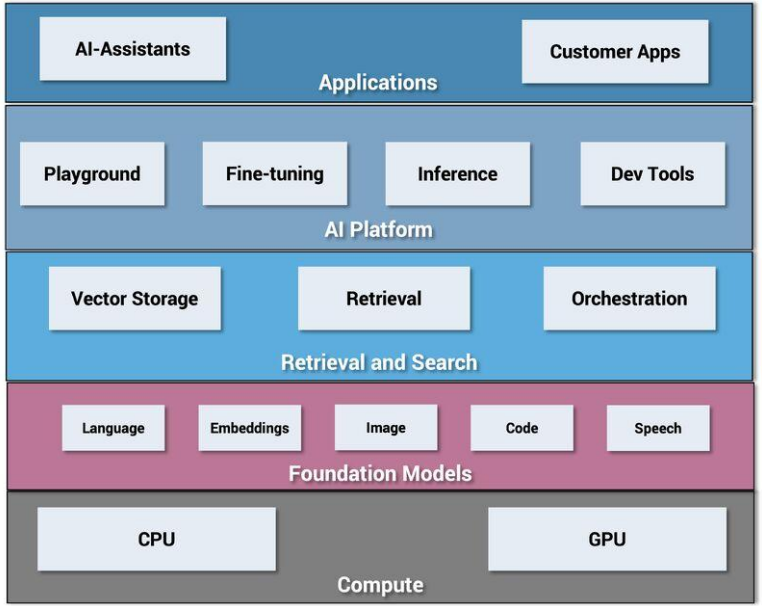
Langchain

Nemo (Nvidia Cloud Native Framework)



4 Requirements for building a model

Typical Architecture

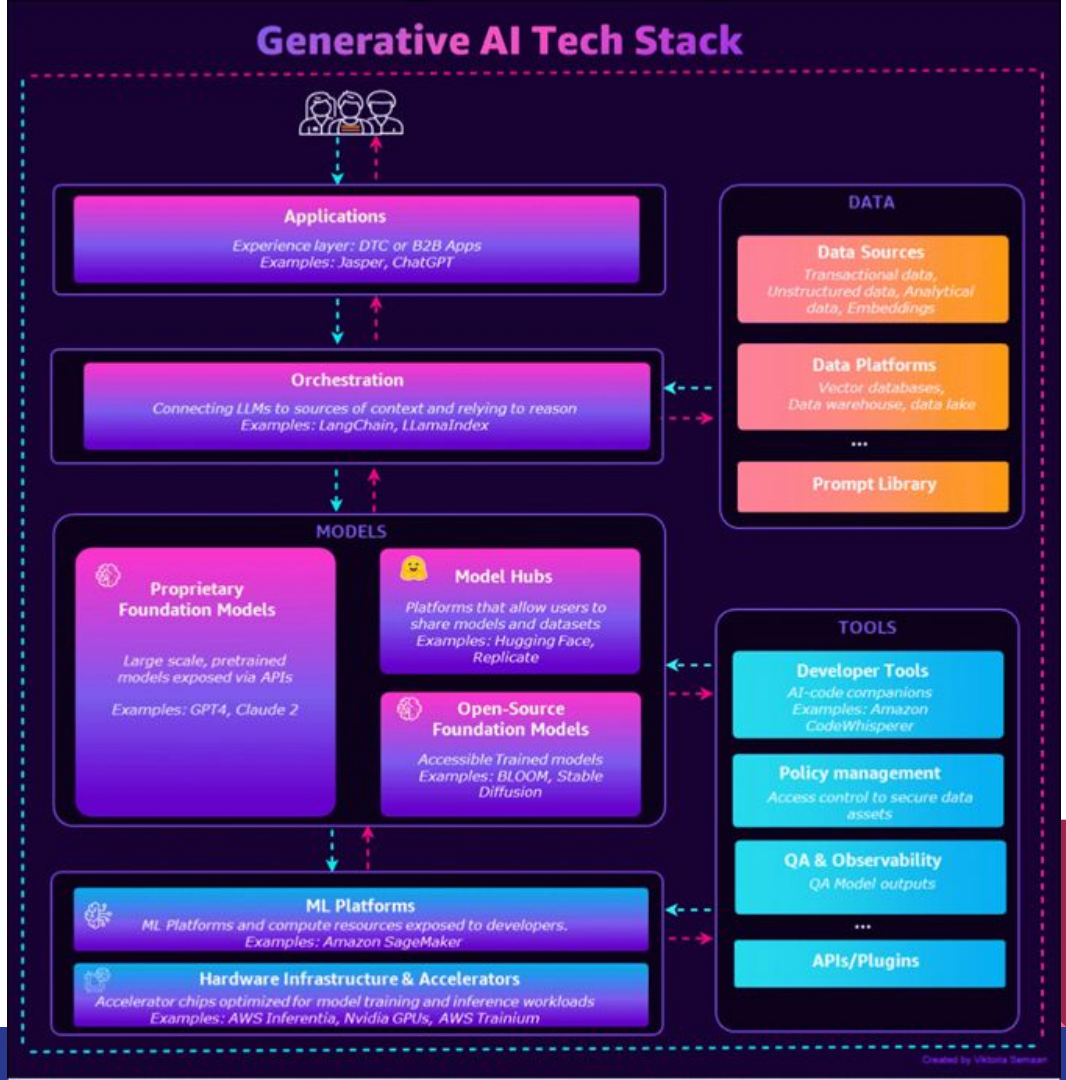


Courtesy of Janakiram MSV

https://www.linkedin.com/posts/janakiramm_here-is-the-big-picture-of-the-modern-genai-activity-7137698886344724480-ZutC?utm_source=share&utm_medium=member_desktop

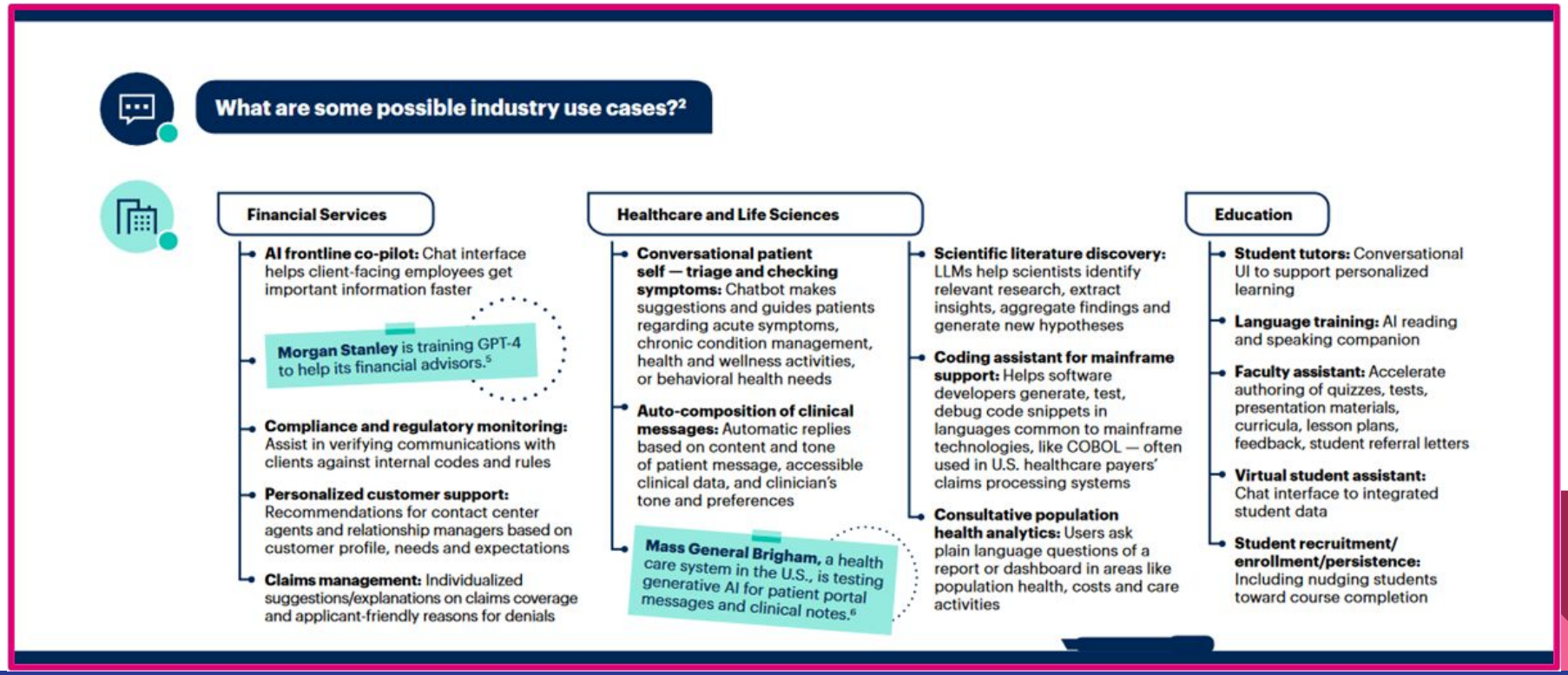
Courtesy of Victoria Semaan

https://www.linkedin.com/posts/semaan_genai-generativeai-ai-activity-7115322938991538177-mk4i?utm_source=share&utm_medium=member_desktop



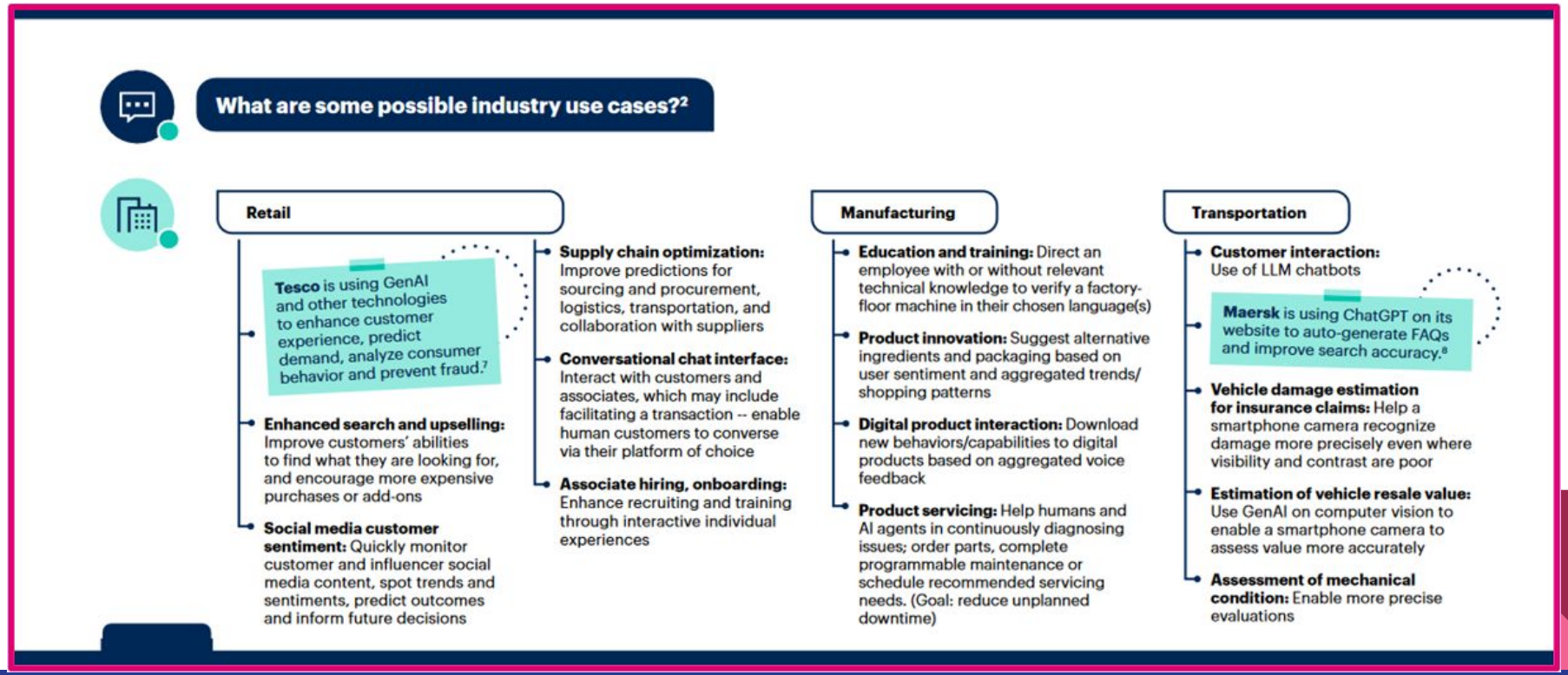
5 Different ways models are being used [Gartner usecases report]

5 - Usecases



https://www.linkedin.com/posts/gartner_generativeai-leadership-ai-activity-7118330499311251456--26m?utm_source=share&utm_medium=member_desktop

5 - Usecases



https://www.linkedin.com/posts/gartner_generativeai-leadership-ai-activity-7118330499311251456--26m?utm_source=share&utm_medium=member_desktop

6 Limitations of LLMs

[1 - infra: compute resources, power, sec/privacy;
2 - models: hallucinations, bias, limited knowledge base,

6 - Limitations of LLMs

CNBC

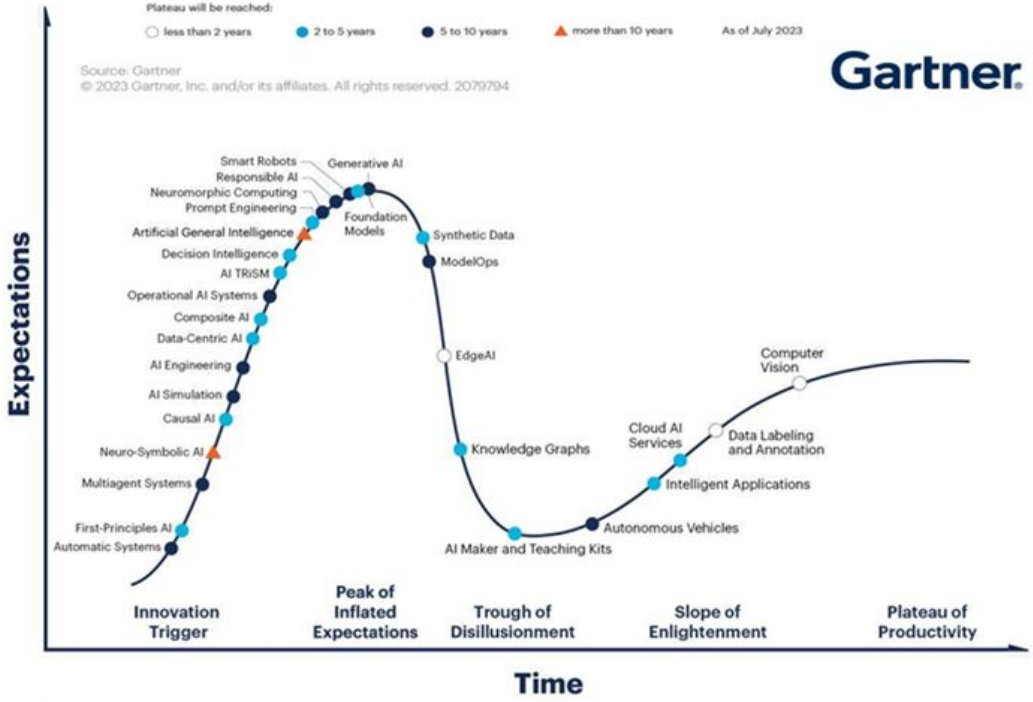
AI IMPACT

AI IMPACT

‘Overhyped’ generative AI will get a ‘cold shower’ in 2024, analysts predict

PUBLISHED TUE, OCT 10 2023•3:32 AM EDT | UPDATED TUE, OCT 10 2023•12:04 PM EDT

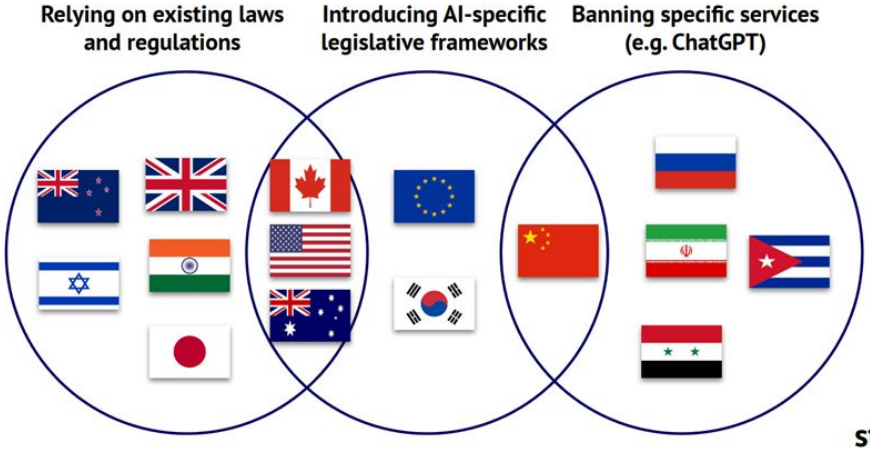
Hype Cycle for Artificial Intelligence, 2023



6 - Limitations of LLMs

Have we reached "peak" regulatory divergence?

- After years of speculation about mounting potential divergence in regulatory approaches, we're starting to see regulatory approaches stabilise and settle into a handful of distinct approaches.



stateof.ai 2023

6 - Limitations of LLMs

https://www.theregister.com/2023/09/29/nuclear_powered_datacenters/

https://www.theregister.com/2023/10/18/hyperscale_datacenter_capacity/

https://www.theregister.com/2023/09/25/microsoft_nuclear_energy_manager_job/

6 - Limitations of LLMs

Hallucinations

- Situations were a model makes up totally wrong inferences

Bias

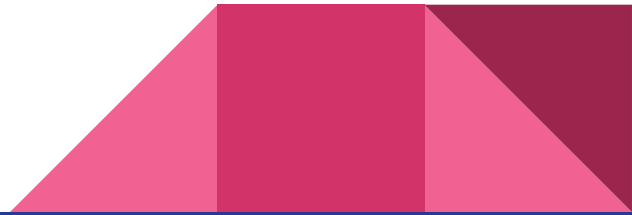
- Models are dependent on the data they attained with, and struggle with unseen data

Limited Knowledge Base

- Models are mostly trained using data available on the internet, and their knowledge is limited to that

Real-life Data Deficiency

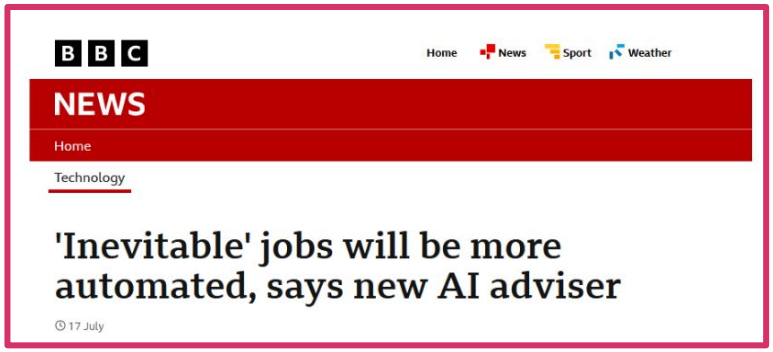
- We could run out of data to feed into LLMs - as early as 2025
- Maybe forced to switch to ‘Synthetic Data’
- Stateof.ai Report (Air Street Capital)



7 Growth Opportunities due to growth of LLMs [jobs - AI/ML Engineers, Prompt Engineers, Data Scientists, Task Automation + Efficiency; Commercial appval


MASTER PLAN:
DESTROY
CHATGPT
Google
28:21
MASTER PLAN:
DESTROY
GOOGLE
Microsoft
27:50
Daily Mail
A.I. 'COULD WIPE OUT HUMANITY'
Threat 'as bad as nuclear war' MUST be tackled, say tech bosses
Daily Mail STANDING UP FOR FREE SPEECH
Academic at Oxford Union with 3 bodyguards ... price of free speech in Britain
THE TIMES
What the Brits behind Succession did next
Boos for trans rights protest at Oxford Union
AI pioneers fear extinction
i
AI creators fear the extinction of humanity

Growth Opportunities - Jobs



"The whole world will have to rethink the way in which people work…There will be winners or [sic] losers on a global basis in terms of where the jobs are as a result of AI."

- Ian Hogarth

(Head of UK Gov AI Task Force)

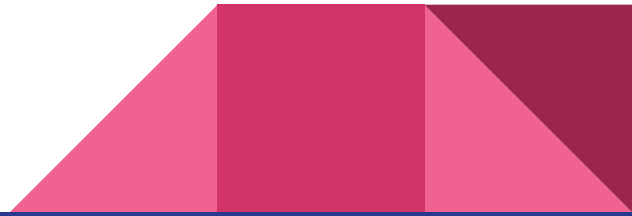
https://www.bbc.co.uk/news/technology-66128106

'- Goldman Sachs

https://www.bbc.co.uk/news/technology-65102150

Opportunities due to growth of LLMs

- OpenAI
  - GPTs (Agents)
  - OpenAI Marketplace



Conclusions

Conclusions

GenerativeAI models are here to stay - Pandora's box is open!

Vendors will continue to innovate and offer services + solutions in AI

2024 could be the year when the dust settles on the AI hype - or perhaps not!

A vast number of Open Source models available for experimentation

All verticals can benefit from AI - gravitate to best-fit solutions

Regulation on safety, sovereignty, privacy and jobs still to come

Enterprises require an AI Strategy to navigate safely

Thank you!

Q&A

Link to Hands-on GitHub Repo

https://github.com/Ogirimah/Generative-AI-Workshop

Back-up

Microsoft/OpenAI

GPT - GPT-4 Turbo

DALL-E

TTS

Whisper

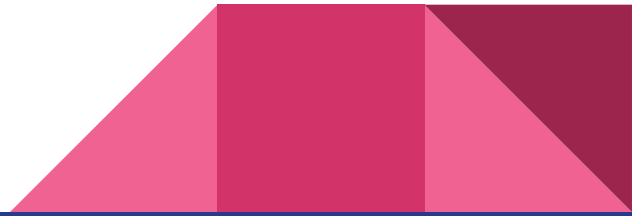
Moderation

Copilot

Github Copilot

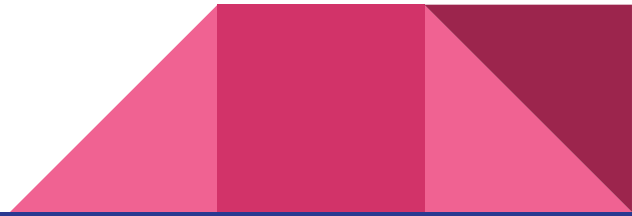
Multiple models available (ChatGPT 3.5, ChatGPT 4, DALLe, Azure AI)

Bing Chat, GitHub Copilot, 365 Copilot, …etc



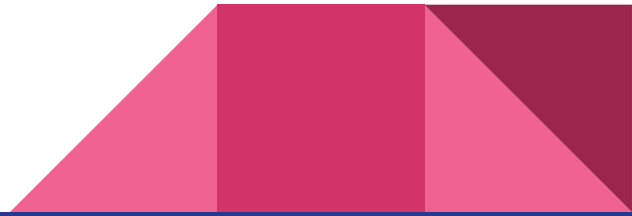
Google

- Imagen
- Muse
- Chirp
- Codey



Meta

- Computer Vision
  - Detectron 2
  - DensePose
- Language
  - Seamless
  - Llama



IBM and Nvidia

- IBM
  - Granite
- Nvidia
  - StyleGAN3
  - EG3D
  - Megatron 530B LLM