BCS - Hertfordshire

AI State Of Play

Richard Simon
Ameer Ogirimah

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)...

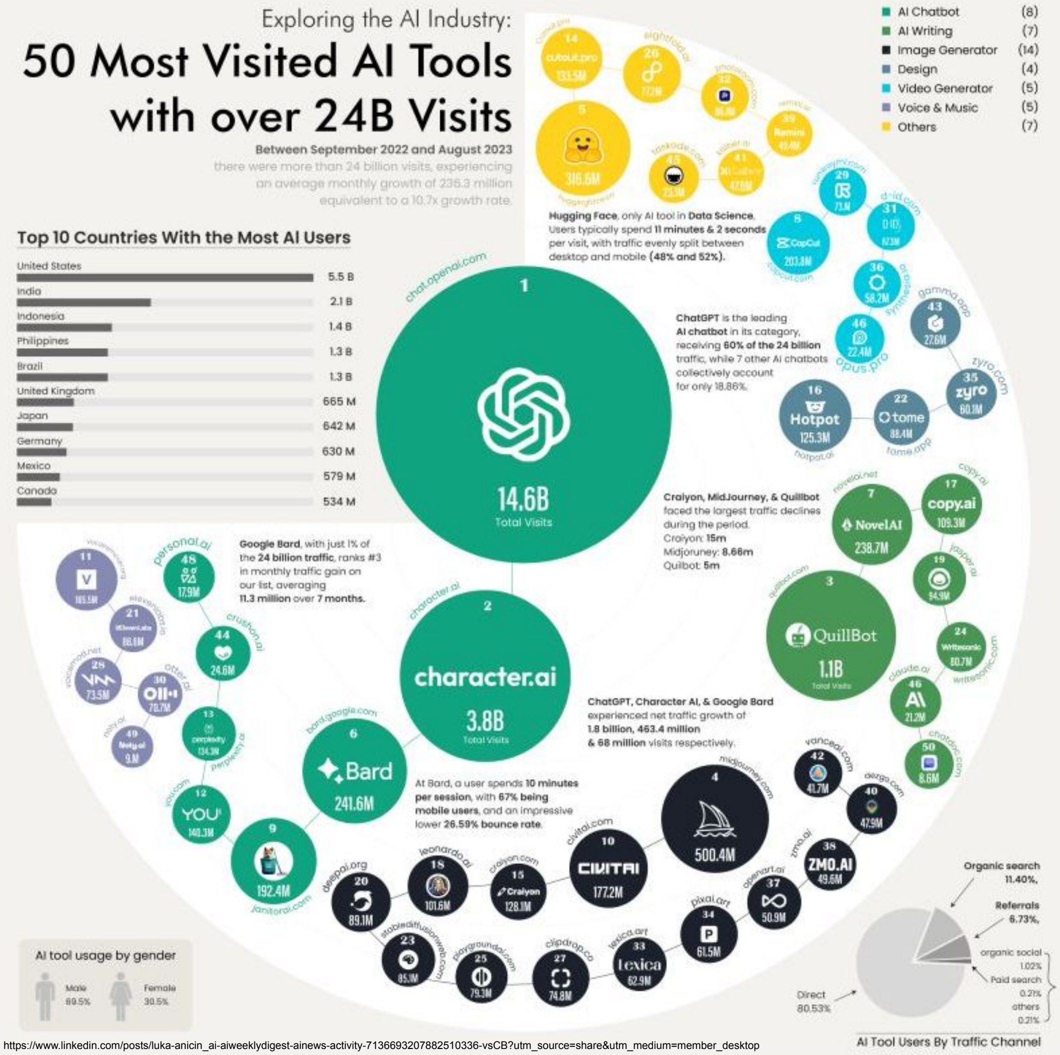
6. Limitations of LLMs

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?



Exploring the AI Industry:

50 Most Visited AI Tools with over 24B Visits

Between September 2022 and August 2023 there were more than 24 billion visits, experiencing an average monthly growth of 236.3 million equivalent to a 10.7x growth rate.

Hugging Face, only AI tool in **Data Science**. Users typically spend **11 minutes & 2 seconds** per visit, with traffic evenly split between desktop and mobile **(48% and 52%).**

ChatGPT is the leading **AI chatbot** in its category, receiving **60% of the 24 billion** traffic, while 7 other AI chatbots collectively account for only 18.86%.

Cralyon, MidJourney, & Quillbot faced the largest traffic declines during the period. Craiyon: **15m** Midjoruney: **8.66m** Quilbot: **5m**

Google Bard, with just 1% of the **24 billion traffic**, ranks #3 in monthly traffic gain on our list, averaging **11.3 million** over **7 months.**

ChatGPT, Character AI, & Google Bard experienced net traffic growth of **1.8 billion, 463.4 million & 68 million** visits respectively.

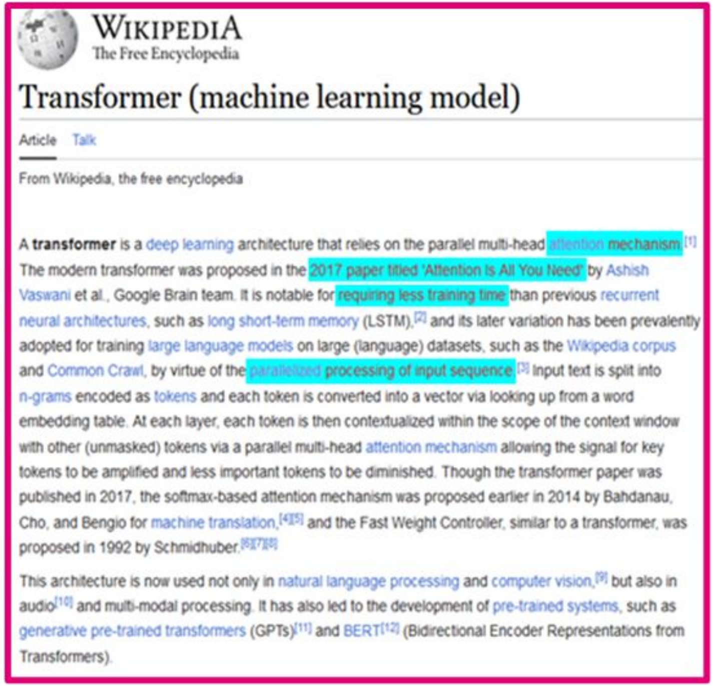
At Bard, a user spends **10 minutes per session**, with **67% being mobile users**, and an impressive lower **26.59% bounce rate**.

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

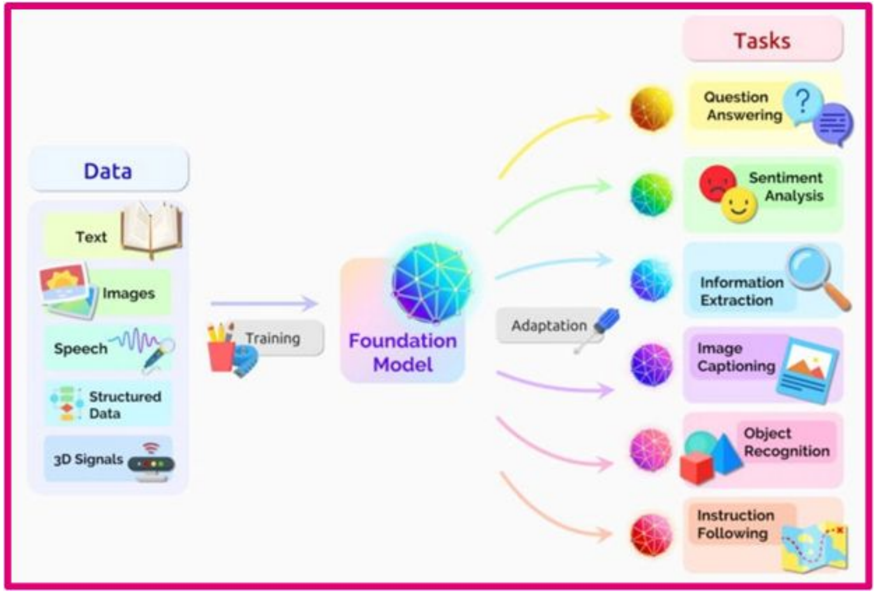
Transformer (machine learning model)

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]

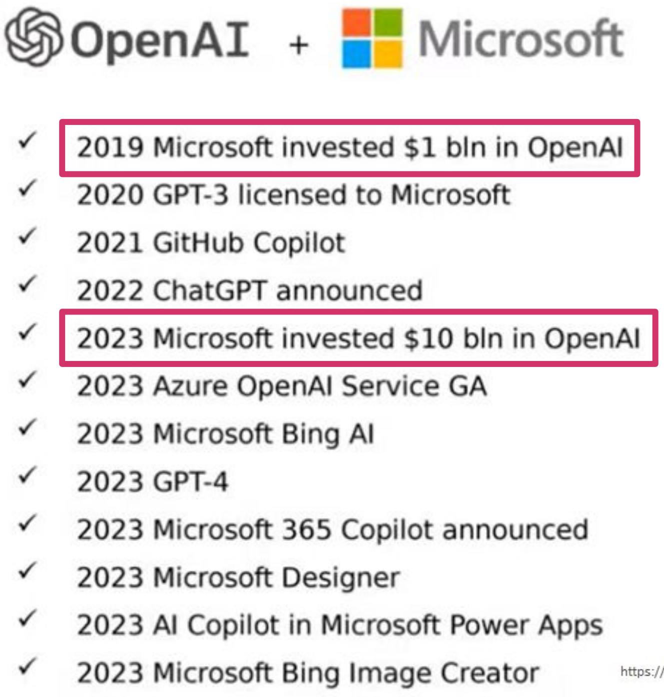
This architecture is now used not only in natural language processing and computer vision,[9] but also in audio[10] and multi-modal processing. It has also led to the development of pre-trained systems, such as generative pre-trained transformers (GPTs)[11] and BERT[12] (Bidirectional Encoder Representations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



Pattern > Inference > Response.

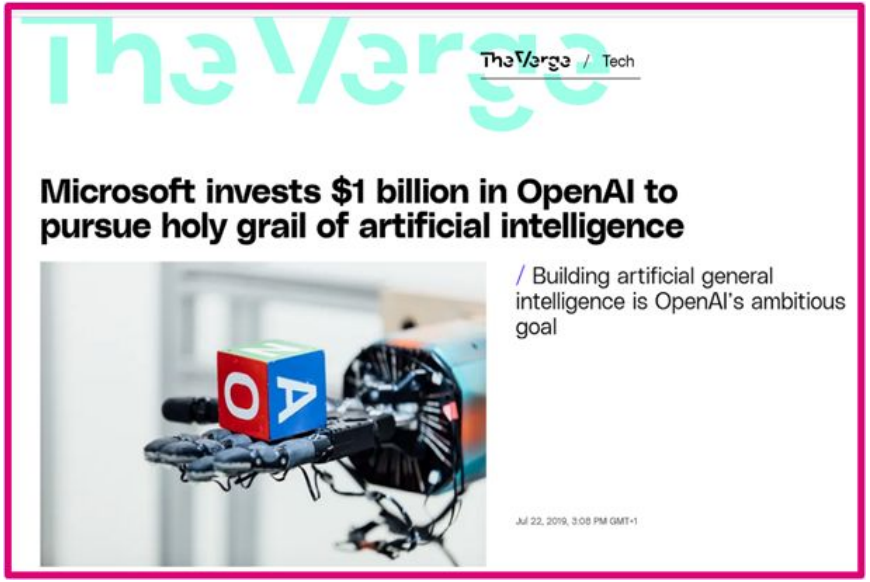
https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/

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

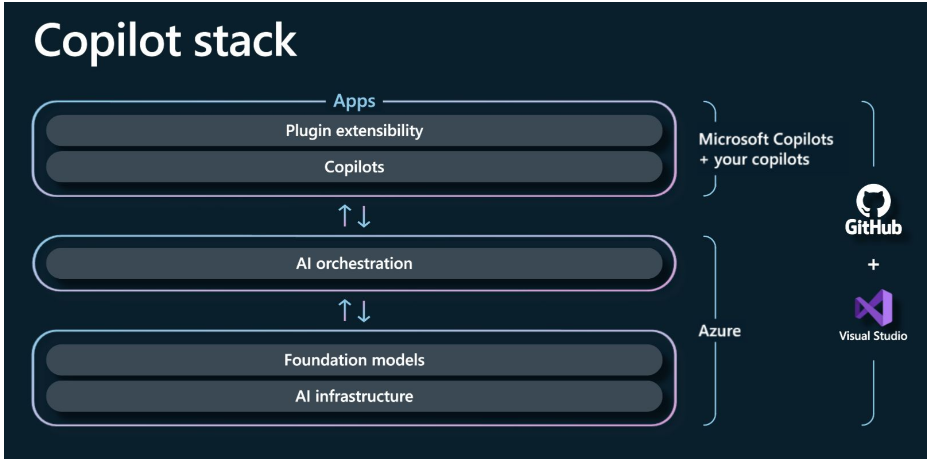
Fully committed to AI in all their platforms

Copilot - base platform for all products

DALL-E - ‘borrowed’ from OpenAI

ChatGPT3.5 + 4

ChatGPT4 Turbo - Multimodal model, available now



Copilot stack

Apps

Plugin extensibility

Copilots

AI orchestration

Foundation models

AI infrastructure

Microsoft Copilots + your copilots

Azure

GitHub + Visual Studio

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

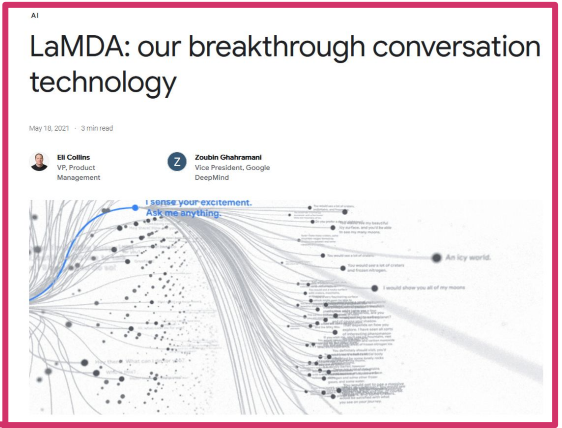
Google Bard

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



AI

LaMDA: our breakthrough conversation technology

May 18, 2021 · 3 min read

Eli Collins
VP, Product Management

Zoubin Ghahramani
Vice President, Google DeepMind

Medium

Google's Bard Will Kill ChatGPT — It is Microsoft Teams vs. Slack All Over Again.

History favors the winners, and you know how big Microsoft Teams is.

AL Anany · Follow
5 min read · Jan 20

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

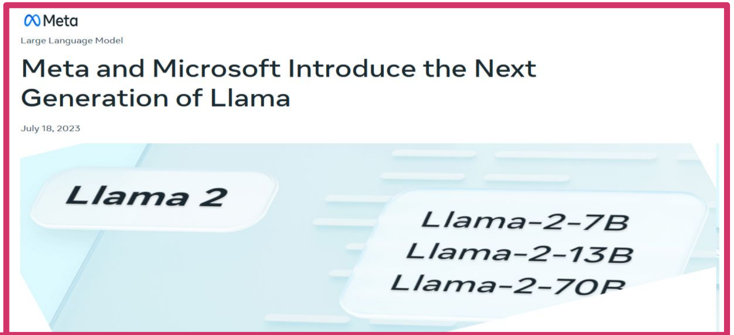
Llama/Llama2 - Open Source models

Future Releases:

Continue with Open Source strategy

Open Source attracts better talent

Put AI into every Meta service in 2024



Meta

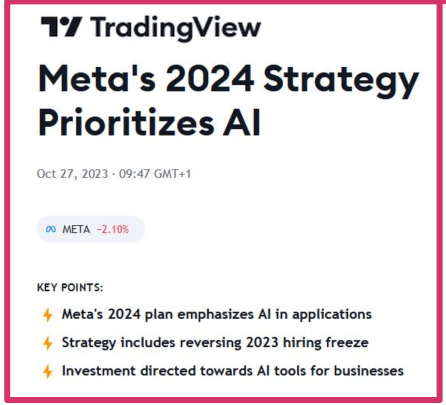
Large Language Model

Meta and Microsoft Introduce the Next Generation of Llama

July 18, 2023

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

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



TradingView

Meta's 2024 Strategy Prioritizes AI

Oct 27, 2023 · 09:47 GMT+1

META −2.10%

KEY POINTS:

- Meta's 2024 plan emphasizes AI in applications
- Strategy includes reversing 2023 hiring freeze
- Investment directed towards AI tools for businesses

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

Future Releases:

Trainium2 - AI Accelerator chip

Amazon Q - ???

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

Amazon SageMaker

Amazon Bedrock

PartyRock

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

2 Available Models

Available Models

ChatGPT 3/4 Turbo

DALL-E

TTS

Whisper

Moderation

Copilot Platform

GitHub Copilot

Azure OpenAI

Bing Chat

365 Copilot

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

IBM Granite

Nvidia:

StyleGAN3

EG3D

Megatron 530B LLM

Computer Vision:

Detectron2

DensePose

Language:

Seamless; Llama

3 Libraries + Frameworks

Torch

Pytorch

MxNet

Tensorflow

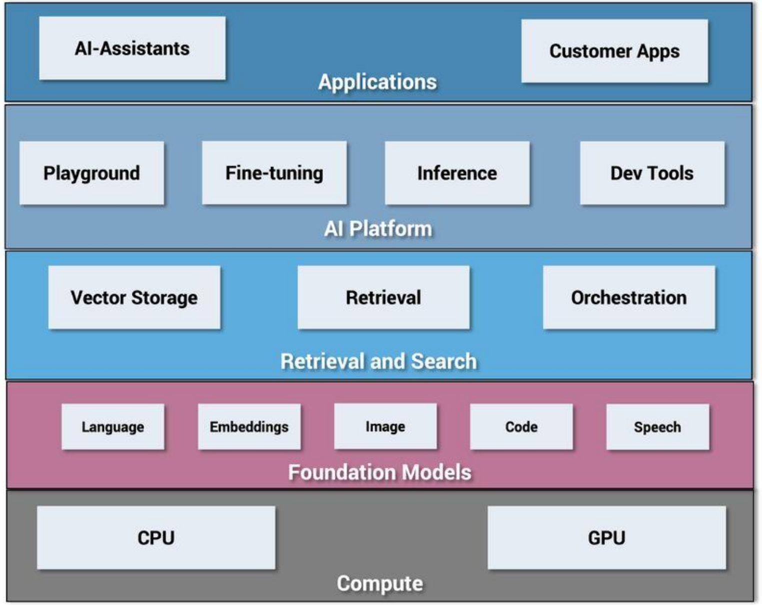
Keras

Langchain

Nemo (Nvidia Cloud Native Framework)

4 Requirements for building a model

Typical Architecture



AI-Assistants

Customer Apps

Applications

Playground

Fine-tuning

Inference

Dev Tools

AI Platform

Vector Storage

Retrieval

Orchestration

Retrieval and Search

Language

Embeddings

Image

Code

Speech

Foundation Models

CPU

GPU

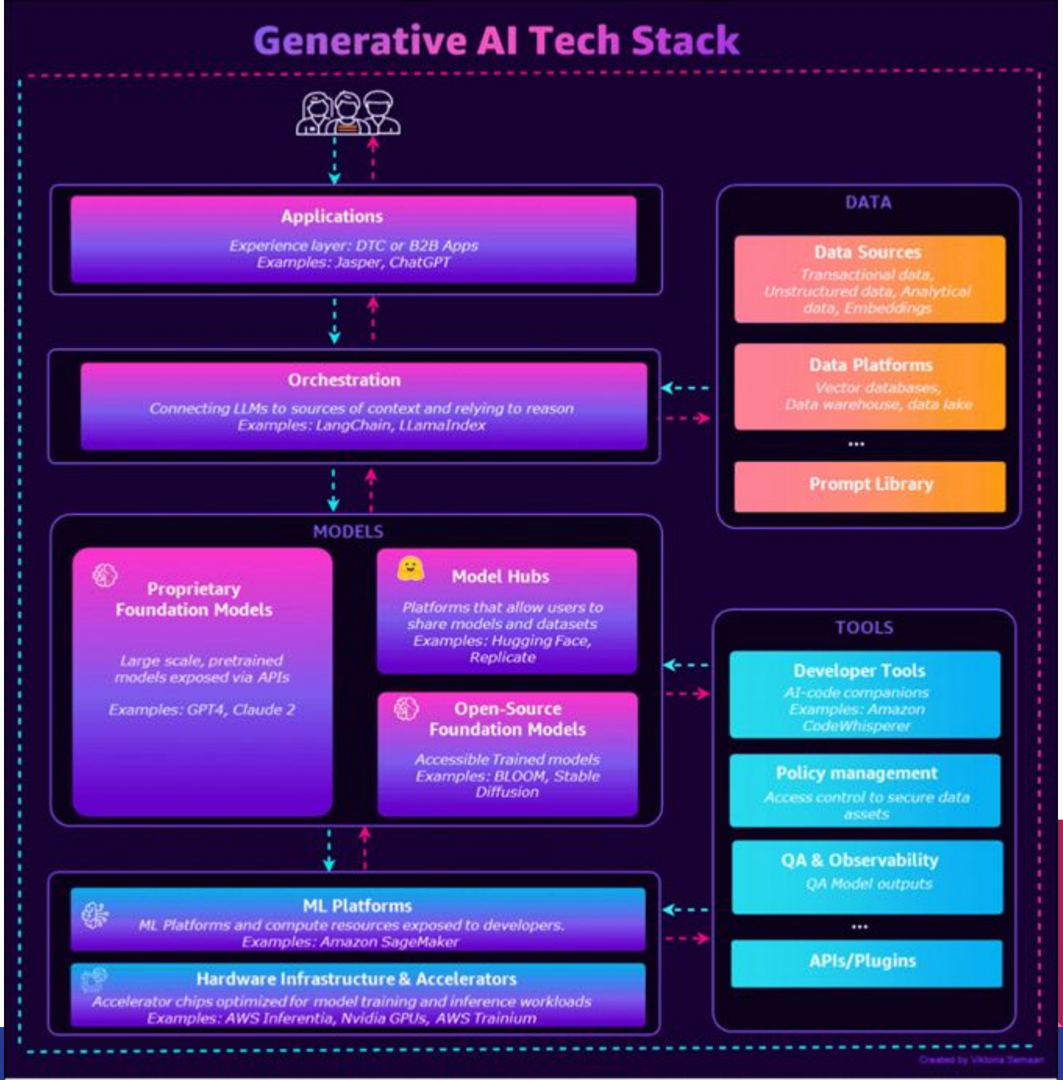
Compute

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



Generative AI Tech Stack

Applications
Experience layer: DTC or B2B Apps
Examples: Jasper, ChatGPT

Orchestration
Connecting LLMs to sources of context and relying to reason
Examples: LangChain, LLamaIndex

MODELS

Proprietary Foundation Models
Large scale, pretrained models exposed via APIs
Examples: GPT4, Claude 2

Model Hubs
Platforms that allow users to share models and datasets
Examples: Hugging Face, Replicate

Open-Source Foundation Models
Accessible Trained models
Examples: BLOOM, Stable Diffusion

ML Platforms
ML Platforms and compute resources exposed to developers.
Examples: Amazon SageMaker

Hardware Infrastructure & Accelerators
Accelerator chips optimized for model training and inference workloads
Examples: AWS Inferentia, Nvidia GPUs, AWS Trainium

DATA

Data Sources
Transactional data, Unstructured data, Analytical data, Embeddings

Data Platforms
Vector databases, Data warehouse, data lake

...

Prompt Library

TOOLS

Developer Tools
AI-code companions
Examples: Amazon CodeWhisperer

Policy management
Access control to secure data assets

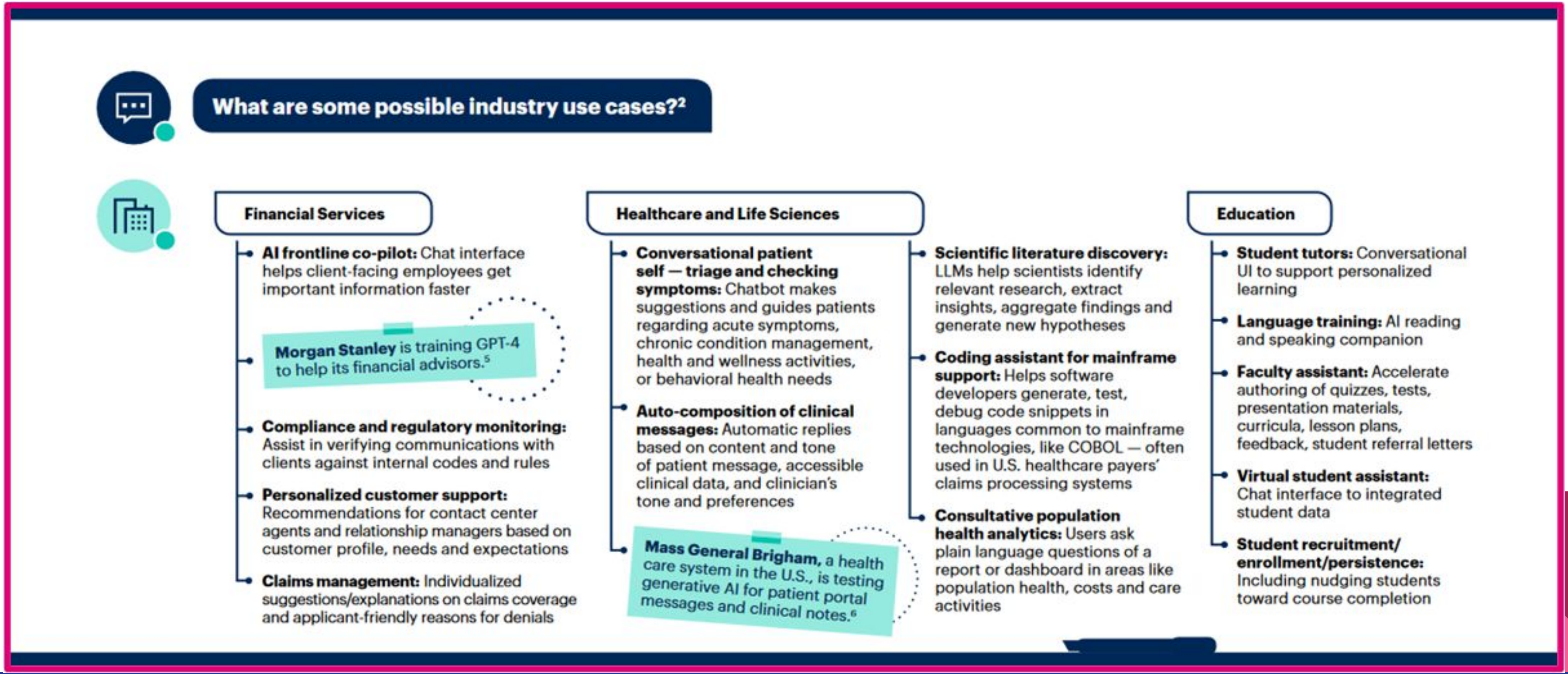
QA & Observability
QA Model outputs

...

APIs/Plugins

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

5 - Usecases



What are some possible industry use cases?[2]

Financial Services

- **AI frontline co-pilot:** Chat interface helps client-facing employees get important information faster
- **Morgan Stanley** is training GPT-4 to help its financial advisors.[5]
- **Compliance and regulatory monitoring:** Assist in verifying communications with clients against internal codes and rules
- **Personalized customer support:** Recommendations for contact center agents and relationship managers based on customer profile, needs and expectations
- **Claims management:** Individualized suggestions/explanations on claims coverage and applicant-friendly reasons for denials

Healthcare and Life Sciences

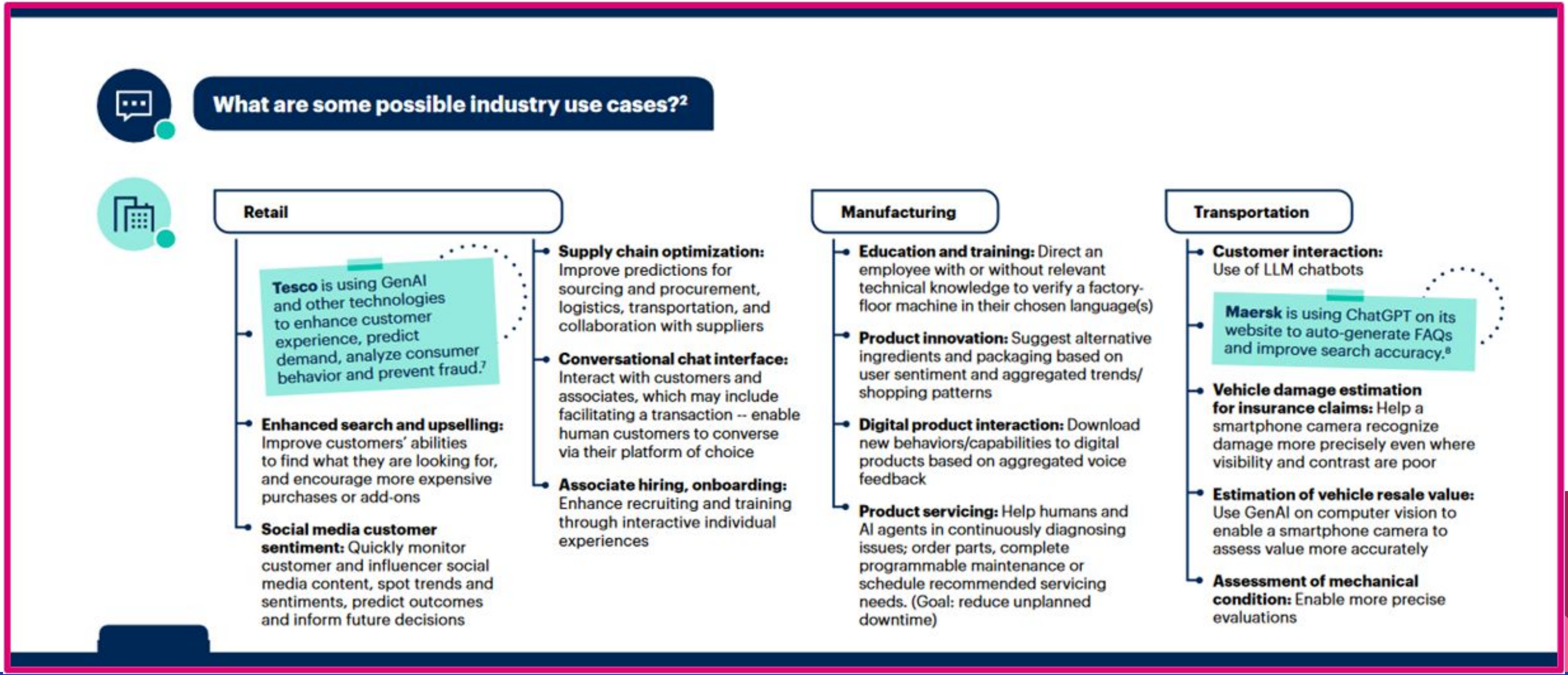
- **Conversational patient self — triage and checking symptoms:** Chatbot makes suggestions and guides patients regarding acute symptoms, chronic condition management, health and wellness activities, or behavioral health needs
- **Auto-composition of clinical messages:** Automatic replies based on content and tone of patient message, accessible clinical data, and clinician's tone and preferences
- **Mass General Brigham,** a health care system in the U.S., is testing generative AI for patient portal messages and clinical notes.[6]
- **Scientific literature discovery:** LLMs help scientists identify relevant research, extract insights, aggregate findings and generate new hypotheses
- **Coding assistant for mainframe support:** Helps software developers generate, test, debug code snippets in languages common to mainframe technologies, like COBOL — often used in U.S. healthcare payers' claims processing systems
- **Consultative population health analytics:** Users ask plain language questions of a report or dashboard in areas like population health, costs and care activities

Education

- **Student tutors:** Conversational UI to support personalized learning
- **Language training:** AI reading and speaking companion
- **Faculty assistant:** Accelerate authoring of quizzes, tests, presentation materials, curricula, lesson plans, feedback, student referral letters
- **Virtual student assistant:** Chat interface to integrated student data
- **Student recruitment/ enrollment/persistence:** Including nudging students toward course completion

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

5 - Usecases



What are some possible industry use cases?[2]

Retail

Tesco is using GenAI and other technologies to enhance customer experience, predict demand, analyze consumer behavior and prevent fraud.[7]

- **Enhanced search and upselling:** Improve customers' abilities to find what they are looking for, and encourage more expensive purchases or add-ons
- **Social media customer sentiment:** Quickly monitor customer and influencer social media content, spot trends and sentiments, predict outcomes and inform future decisions
- **Supply chain optimization:** Improve predictions for sourcing and procurement, logistics, transportation, and collaboration with suppliers
- **Conversational chat interface:** Interact with customers and associates, which may include facilitating a transaction -- enable human customers to converse via their platform of choice
- **Associate hiring, onboarding:** Enhance recruiting and training through interactive individual experiences

Manufacturing

- **Education and training:** Direct an employee with or without relevant technical knowledge to verify a factory-floor machine in their chosen language(s)
- **Product innovation:** Suggest alternative ingredients and packaging based on user sentiment and aggregated trends/ shopping patterns
- **Digital product interaction:** Download new behaviors/capabilities to digital products based on aggregated voice feedback
- **Product servicing:** Help humans and AI agents in continuously diagnosing issues; order parts, complete programmable maintenance or schedule recommended servicing needs. (Goal: reduce unplanned downtime)

Transportation

- **Customer interaction:** Use of LLM chatbots

Maersk is using ChatGPT on its website to auto-generate FAQs and improve search accuracy.[8]

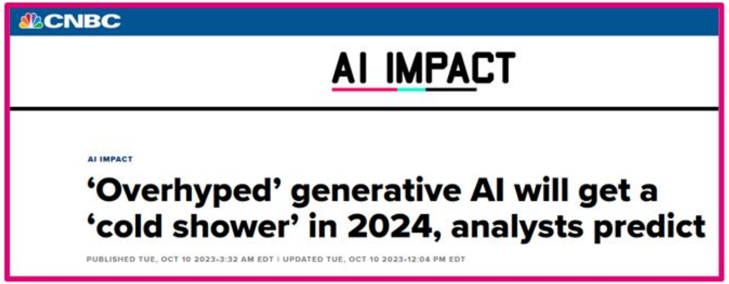
- **Vehicle damage estimation for insurance claims:** Help a smartphone camera recognize damage more precisely even where visibility and contrast are poor
- **Estimation of vehicle resale value:** Use GenAI on computer vision to enable a smartphone camera to assess value more accurately
- **Assessment of mechanical condition:** Enable more precise evaluations

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



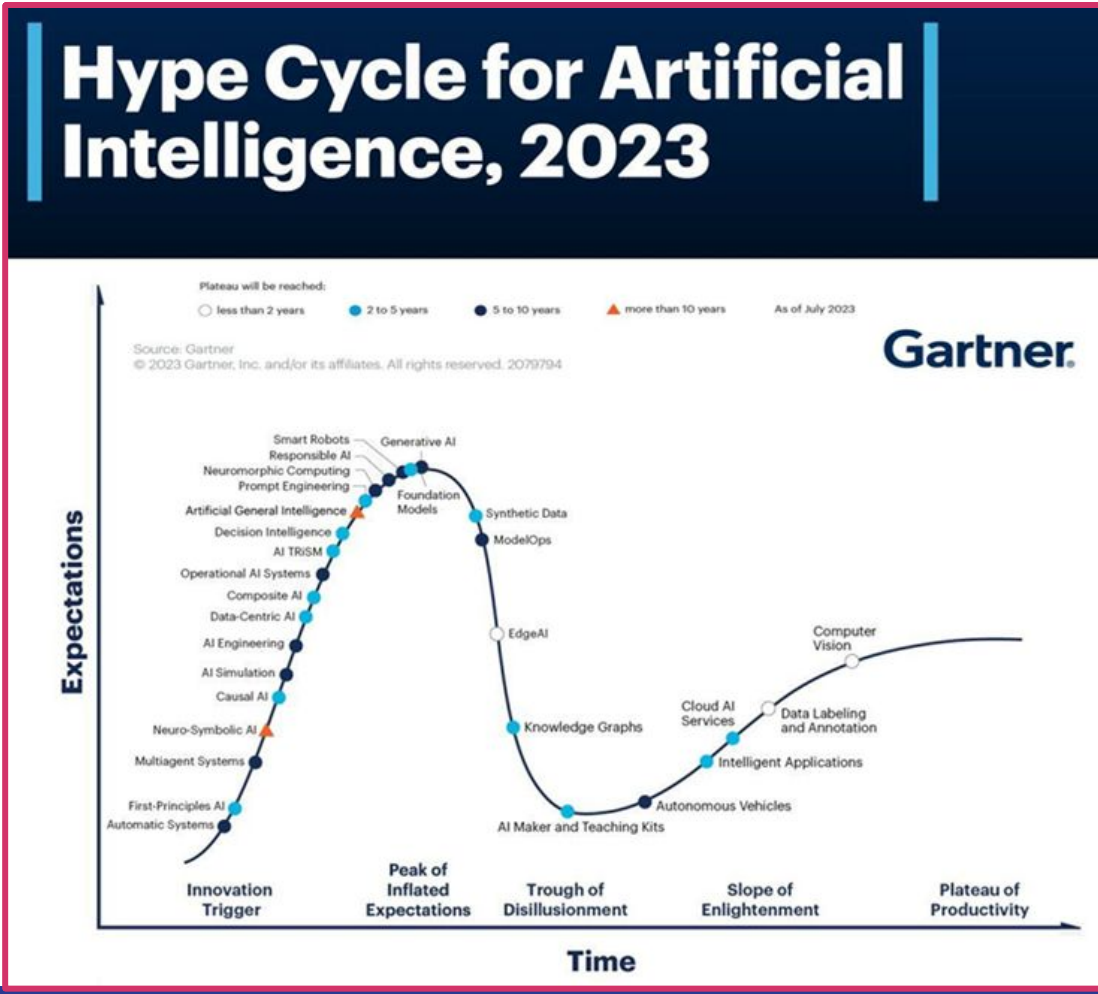
CNBC

AI IMPACT

AI IMPACT

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

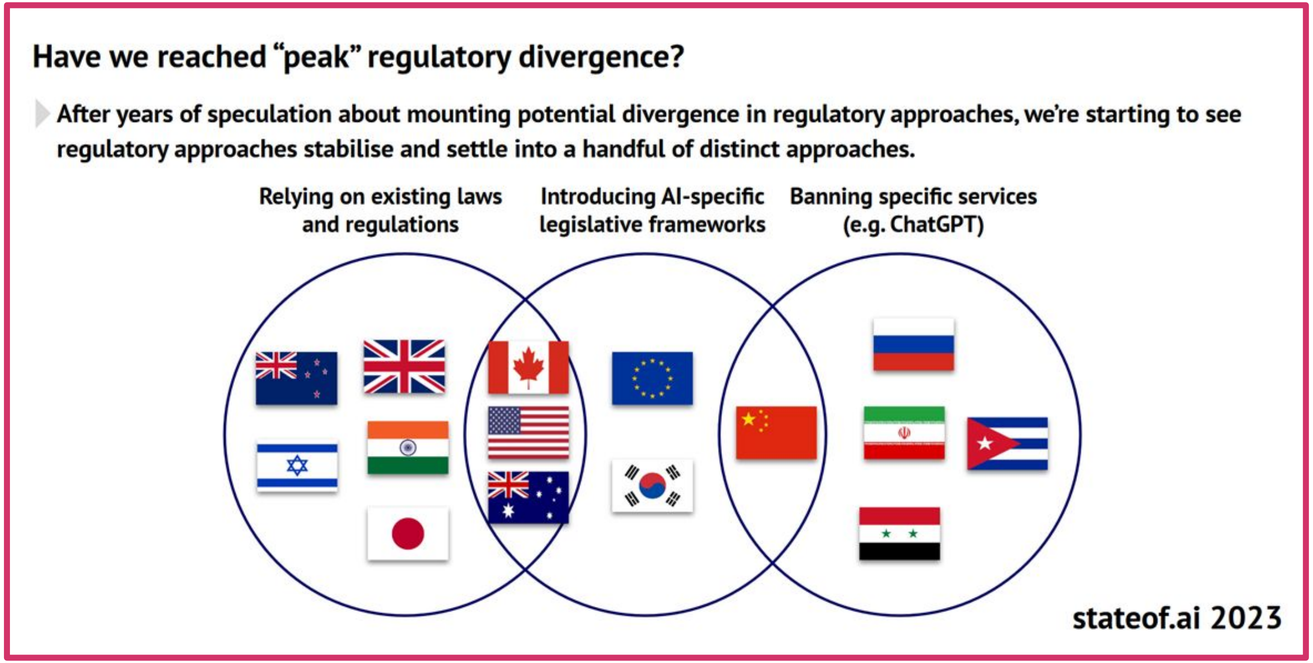
PUBLISHED TUE, OCT 10 2023-3:32 AM EDT | UPDATED TUE, OCT 10 2023-12:04 PM EDT



Hype Cycle for Artificial Intelligence, 2023

Source: Gartner
© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. 2079794

6 - Limitations of LLMs



Have we reached "peak" regulatory divergence?

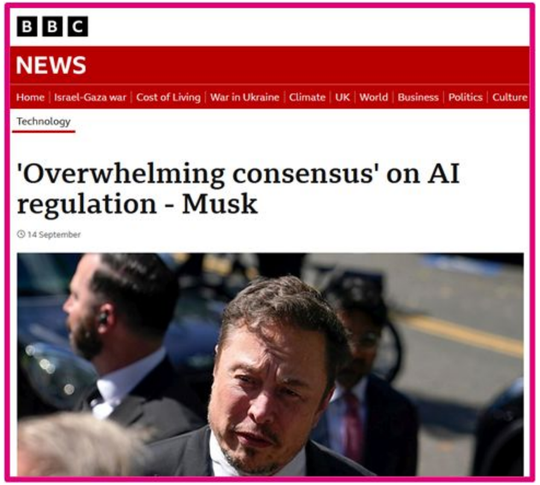
- After years of speculation about mounting potential divergence in regulatory approaches, we're starting to see regulatory approaches stabilise and settle into a handful of distinct approaches.

Relying on existing laws and regulations

Introducing AI-specific legislative frameworks

Banning specific services (e.g. ChatGPT)

stateof.ai 2023



BBC

NEWS

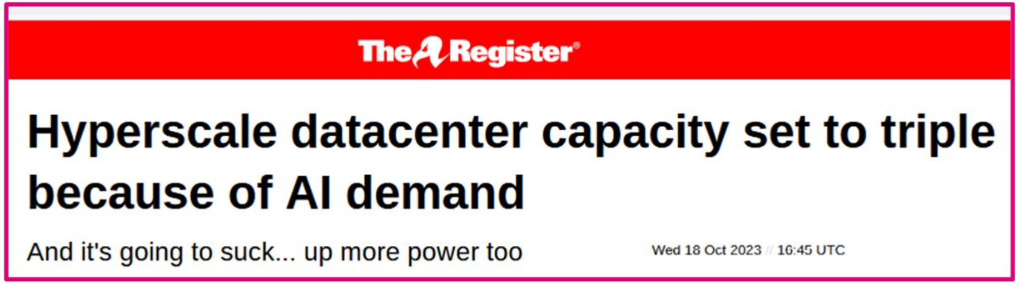
Home | Israel-Gaza war | Cost of Living | War in Ukraine | Climate | UK | World | Business | Politics | Culture

Technology

'Overwhelming consensus' on AI regulation - Musk

14 September

6 - Limitations of LLMs

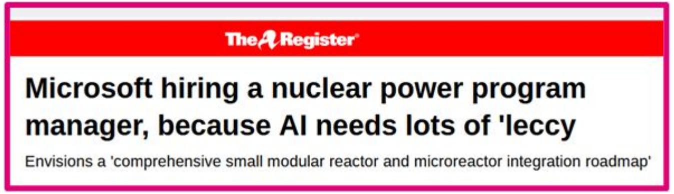


The Register

Hyperscale datacenter capacity set to triple because of AI demand

And it's going to suck... up more power too

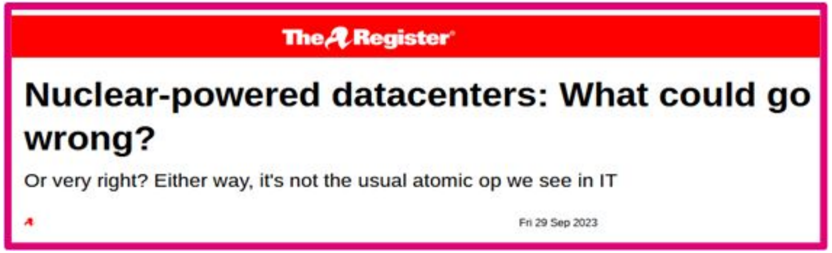
Wed 18 Oct 2023 // 16:45 UTC



The Register

Microsoft hiring a nuclear power program manager, because AI needs lots of 'leccy

Envisions a 'comprehensive small modular reactor and microreactor integration roadmap'



The Register

Nuclear-powered datacenters: What could go wrong?

Or very right? Either way, it's not the usual atomic op we see in IT

Fri 29 Sep 2023

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

Opportunities due to growth of LLMs

- OpenAI
  - GPTs (Agents)
  - OpenAI Marketplace

Summary

GenerativeAI models are here to stay

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 and privacy still to come

Enterprises require an AI Strategy to navigate safely

Thank you!

Q&A

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