

Generative AI and Large Language Models

MMA 865

Moez Ali



Smith | Queen's
SCHOOL OF BUSINESS University

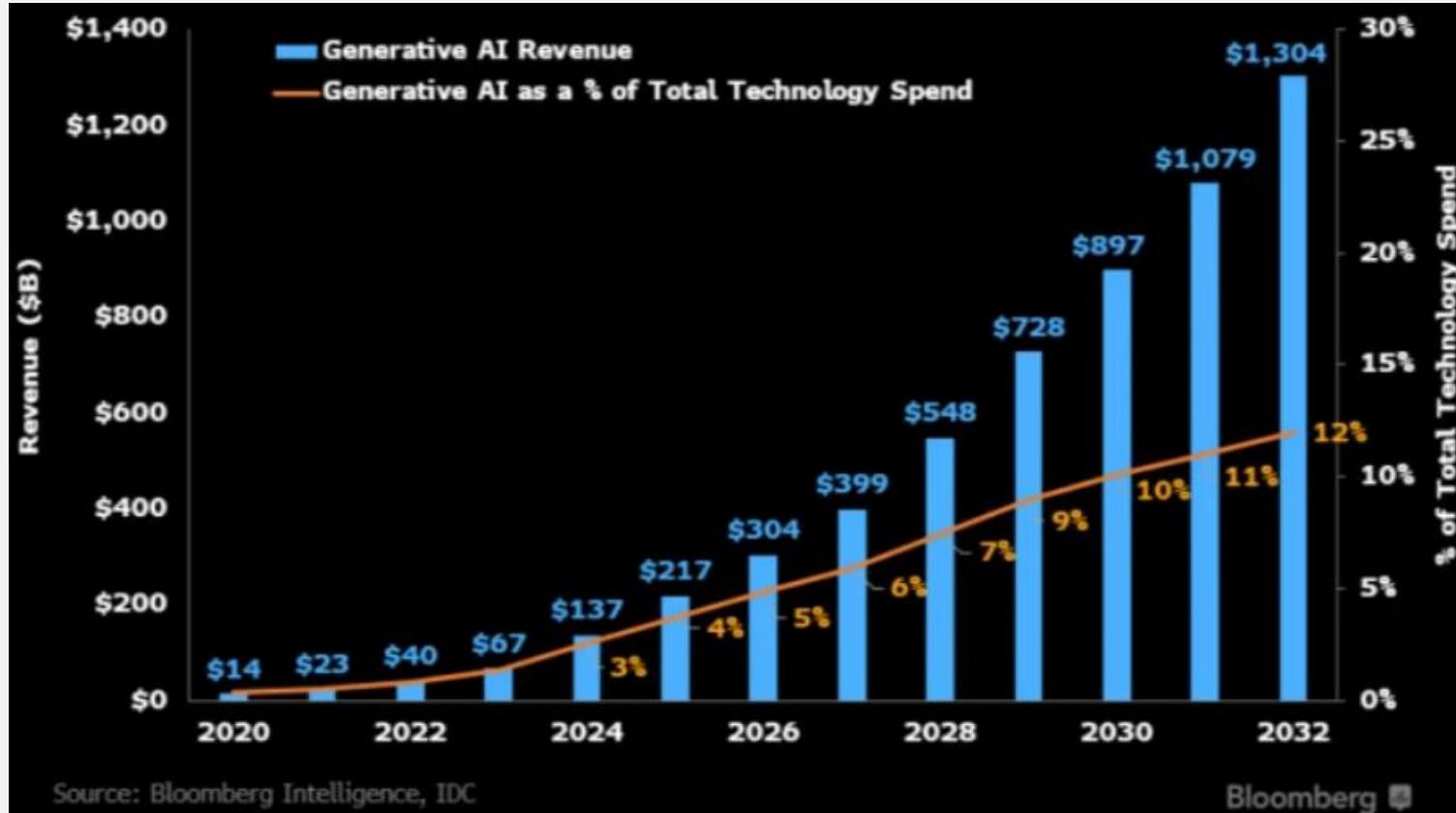
Session Outline

- Generative AI
- Large Language Models

- **Week 1 – NLP and use-cases**
 - Information Extraction
 - Document Classification
 - Sentiment Analysis
- **Week 2 – Topic Modeling, Text Preprocessing and Vectorization**
 1. Topic Modeling
 - Unsupervised ML technique for identifying themes in large corpus of text
 2. Text Preprocessing
 - Tokenization, Stopwords removal, Lemmatization or stemming, bi-grams, etc.
 3. Text Vectorization
 - Bag of Word / Count Vectorizer, Tf-Idf,
 - Deep learning: Embeddings

Definition of Generative AI

- Generative Artificial Intelligence is a type of AI system capable of generating text, images, or other media in response to prompts in natural language.



Bloomberg

Bloomberg Intelligence Interactive Calculator: Generative AI Market Opportunity

(\$ million, unless otherwise specified)

Generative AI Revenue Projections	2022	2027E	2032E	2022-32E CAGR
Hardware	\$37,973	\$223,615	\$641,737	33%
Devices (Inference)	\$4,128	\$82,965	\$168,233	45%
Computer Vision AI Products	\$1,032	\$22,124	\$60,564	50%
Conversational AI Products	\$3,096	\$60,841	\$107,669	43%
Infrastructure (Training)	\$33,845	\$140,650	\$473,505	30%
AI Server	\$22,563	\$49,641	\$133,817	19%
AI Storage	\$9,025	\$33,094	\$92,642	26%
Generative AI Infrastructure as a Service	\$2,256	\$57,915	\$247,046	60%
Software	\$1,493	\$58,826	\$279,899	69%
Specialized Generative AI Assistants	\$447	\$20,864	\$89,035	70%
Coding, DevOps and Generative AI Workflows	\$213	\$12,617	\$50,430	73%
Generative AI Workload Infrastructure Software	\$439	\$13,468	\$71,645	66%
Generative AI Drug Discovery Software	\$14	\$4,042	\$28,343	113%
Generative AI Based Cybersecurity Spending	\$9	\$3,165	\$13,946	109%
Generative AI Education Spending	\$370	\$4,669	\$26,500	53%
Generative AI Based Gaming Spending	\$190	\$20,668	\$69,414	80%
Generative AI Driven Ad Spending	\$57	\$64,358	\$192,492	125%
Generative AI Focused IT Services	\$83	\$21,690	\$85,871	100%
Generative AI Based Business Services	\$38	\$10,188	\$34,138	97%
Total	\$39,834	\$399,345	\$1,303,551	42%

Source: Bloomberg Intelligence, IDC, eMarketer, Statista

Google generative ai cyber security

All Images Videos News Shopping Books Maps More Tools

SC Media 2025 Forecast: AI to supercharge attacks, quantum threats grow, SaaS security woes

SC Media's 2025 annual roundup of predictions, forecasts and warnings for the year ahead by top cybersecurity experts.

1 day ago

P PYMNTS.com 55% of Companies Have Implemented AI-Powered Cybersecurity

The share of chief operating officers (COOs) who report that their companies have implemented artificial intelligence (AI)-powered automated...

20 hours ago

MSSP Alert 8 Cybersecurity Trends and Opportunities for 2025

The rise of AI, ransomware attacks, threats to healthcare, and geopolitics are among the concerns in the new year.

3 days ago

F Forbes How Generative AI Is Powering A New Era Of Cybersecurity

Over 80% of respondents to a CrowdStrike survey say they are planning to adopt or have already integrated generative AI solutions into their...

2 weeks ago

I Inc.com As New Cybersecurity Threats Loom, AI Defenses Get Popular



generative ai drug discovery

All Images Videos News Shopping Books Maps More Tools

 Nature

Generative AI platforms drive drug discovery dealmaking

Aqemia announced a collaboration with Sanofi using its platform based on generative AI and 'deep physics' to identify small-molecule drug...

Aug 16, 2024



 The World Economic Forum

GenAI in clinical trials: Challenges, solutions and results

The journey from drug discovery to patient access is complex, costly and time-consuming. While generative AI (genAI) is transforming...

2 weeks ago



 AbbVie

Three ways AI is changing drug discovery at AbbVie

Explore how artificial Intelligence can accelerate discovery of new drug targets, optimize drug design, and help get therapies to the right...

Sep 25, 2024



 Forbes

How Generative AI Is Accelerating Drug Discovery

Generative AI is revolutionizing drug discovery by decoding the complex languages of biology and chemistry, enabling pharmaceutical...

Jun 19, 2024



 Health Tech World

How gen AI will transform life sciences in 2025

The life sciences industry is in a position to help humanity, perhaps more than ever, with the introduction of generative AI (gen AI).

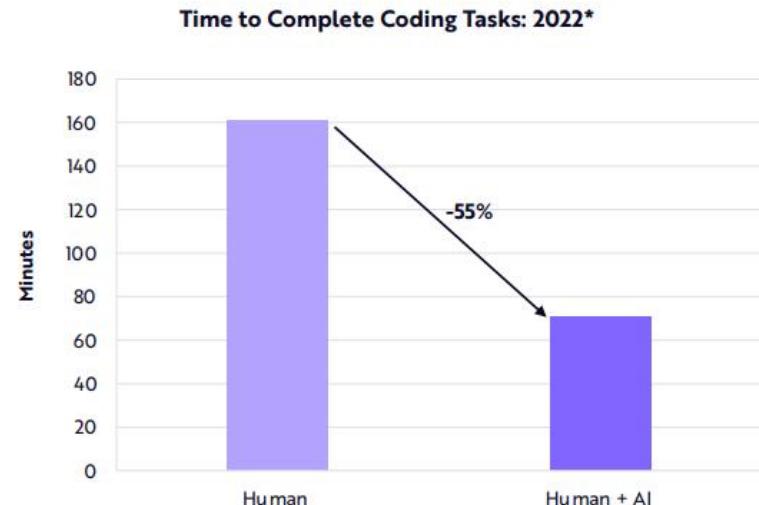
3 days ago



AI Is Increasing The Productivity Of Knowledge Workers

Coding Assistants

Software engineers completed a coding task in less than half the time with AI coding assistant GitHub Copilot.



Generative Image Models

According to our research, AI can create a graphic design for just \$0.08** in minutes — a *di minimis* cost compared to \$150 for human labor.



Human	
Cost	\$150
Time	5 Hours
Generative AI	
Cost	\$0.08
Time	< 1 Minute

*Based on data from GitHub. **Generative AI models translated "a picture of an astronaut on Mars" into multiple images in just a few seconds. Sources: ARK Investment Management LLC, 2023. Kalliamvakou, E. 2022. Forecasts are inherently limited and cannot be relied upon. For informational purposes only and should not be considered investment advice or a recommendation to buy, sell, or hold any particular security. Past performance is not indicative of future results.



create image of a cute dog wearing Christmas
hat with a banner in hand with MMA 865
written on it

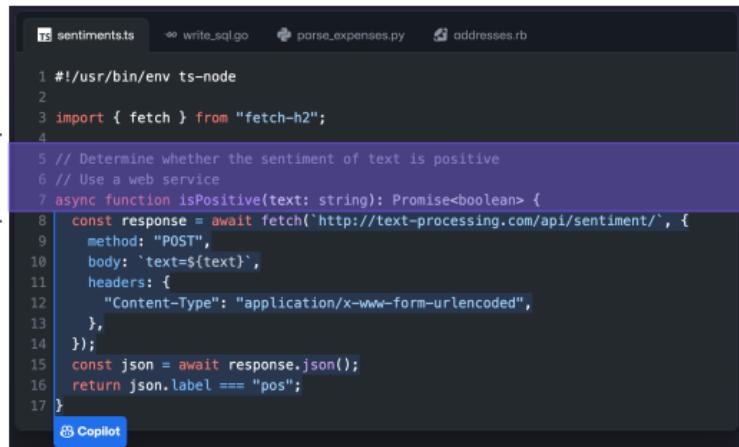
ARK Invest Big Ideas 2023

AI Could Lead To A 10-Fold Increase In Coding Productivity

Based on a 70% annualized drop in training costs and feedback loops, AI coding assistants like Copilot could increase the output of software engineers ~10-fold by 2030.

Human Input

Github Copilot Example

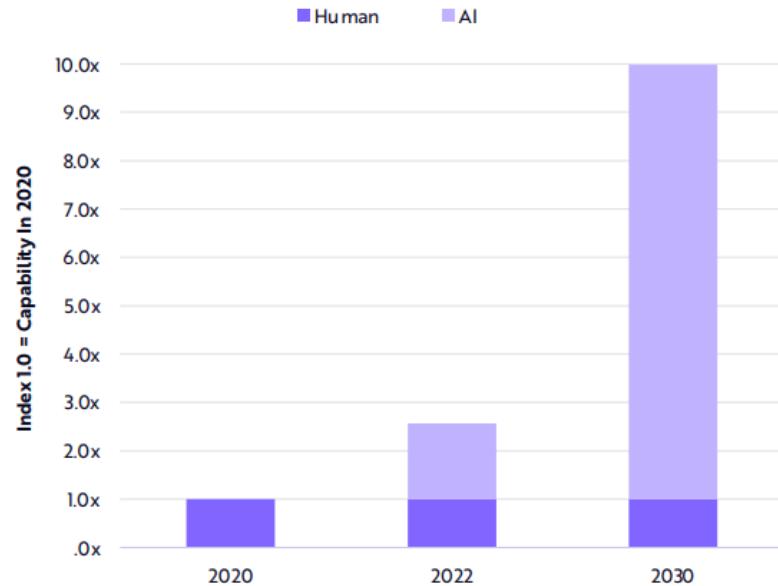


```
sentiments.ts
write_sql.go
parse_expenses.py
addresses.rb

1 #!/usr/bin/env ts-node
2
3 import { fetch } from "fetch-h2";
4
5 // Determine whether the sentiment of text is positive
6 // Use a web service
7 async function isPositive(text: string): Promise<boolean> {
8   const response = await fetch(`http://text-processing.com/api/sentiment/`, {
9     method: "POST",
10    body: `text=${text}`,
11    headers: {
12      "Content-Type": "application/x-www-form-urlencoded",
13    },
14  });
15  const json = await response.json();
16  return json.label === "pos";
17}

Copilot
```

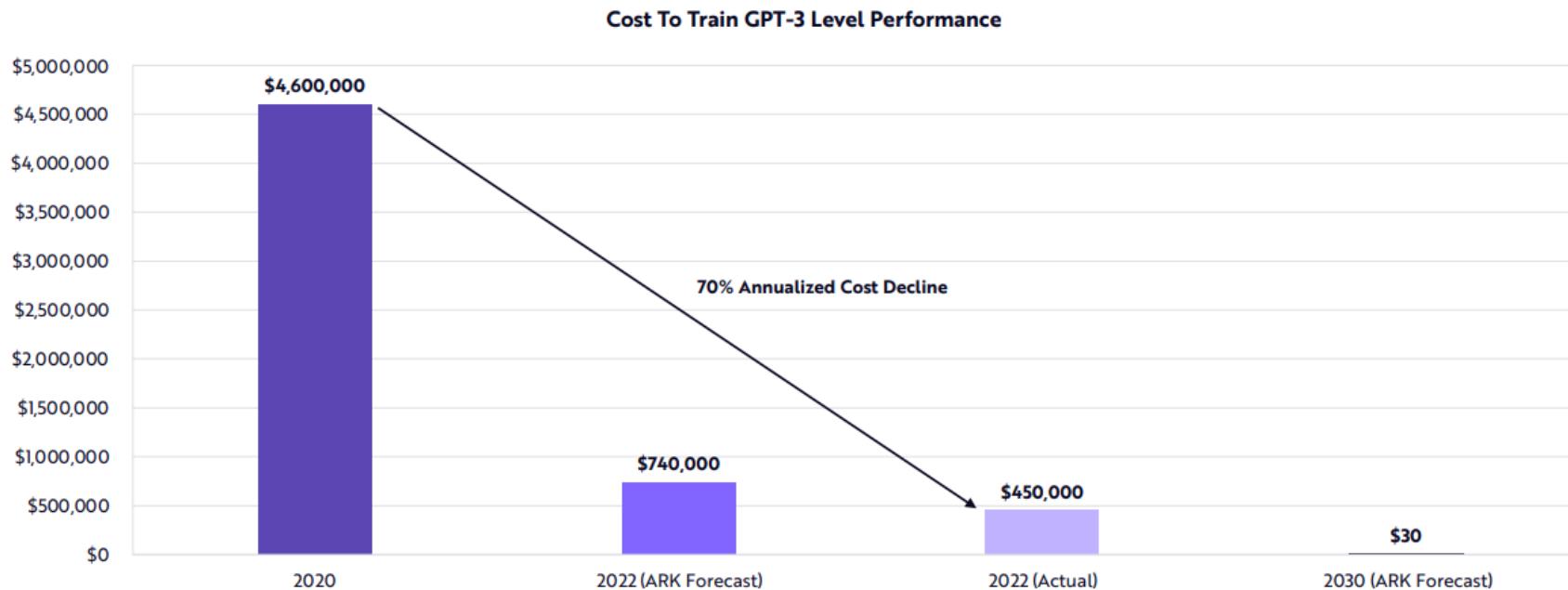
Output of Human + AI: Coding Tasks



Sources: ARK Investment Management LLC, 2023. GitHub 2023. Forecasts are inherently limited and cannot be relied upon. For informational purposes only and should not be considered investment advice or a recommendation to buy, sell, or hold any particular security. Past performance is not indicative of future results.

AI Training Costs Continue To Plummet

Mosaic^{ML} recently released AI training tools that can train language models to GPT-3 level performance for just \$450,000, roughly one-tenth the \$4.6 million just two years ago. AI training costs are dropping ~70% per year, even faster than the 60% estimate based on research presented in ARK's Big Ideas 2022.

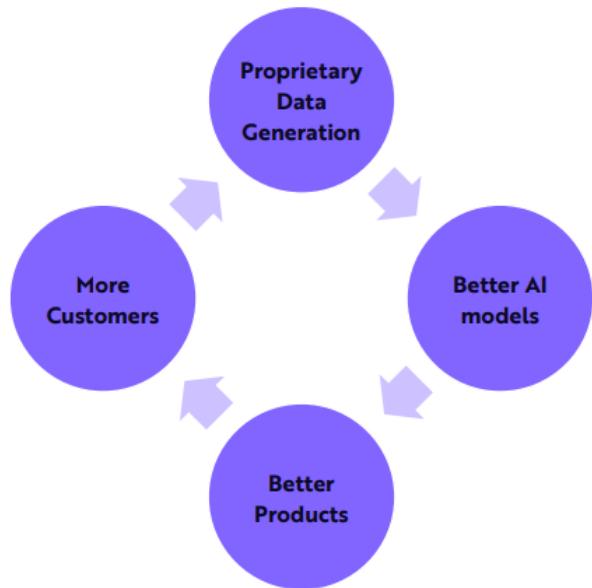


Sources: ARK Investment Management LLC, 2023; Venigalla, A. et al. 2022; Li, C. 2020. Forecasts are inherently limited and cannot be relied upon. For informational purposes only and should not be considered investment advice or a recommendation to buy, sell, or hold any particular security. Past performance is not indicative of future results.

ARK Invest Big Ideas 2023

Proprietary Data Could Create Moats

High-quality domain-specific AI training data could result in winner-takes-most outcomes across vertical applications.



Domain	Autonomous Driving	Software Development	Dialogue
Data Metric	Recorded real-world driving miles	Lines of code	Logged conversations
Feedback Loop	Autonomous disengagements	Rejected code	Rephrased questions
Company Example	Tesla	Replit	Twilio

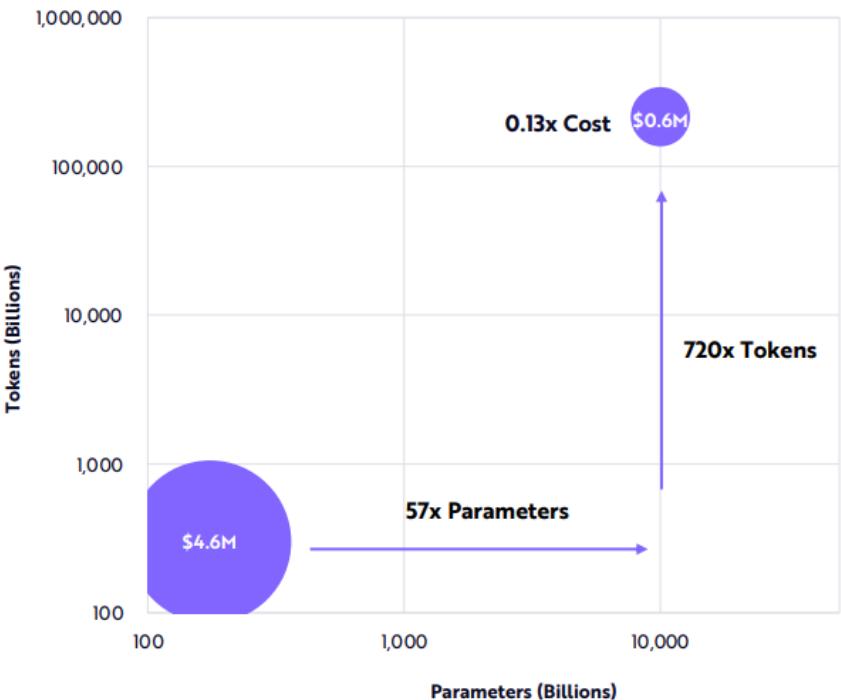
Sources: ARK Investment Management LLC, 2023. Forecasts are inherently limited and cannot be relied upon. For informational purposes only and should not be considered investment advice or a recommendation to buy, sell, or hold any particular security. Past performance is not indicative of future results.

AI Is Creating Explosive Demand For Training Data

The cost to train the state-of-the-art GPT-3 in 2020 was \$4.6 million. Based on our modeling, the cost of training an AI model with 57x more parameters and 720x more tokens than GPT-3 would drop from \$17 billion today to \$600,000 by 2030.

Research suggests the optimal training dataset for a 10 trillion parameter model would require at least 216 trillion tokens. For perspective, 216 trillion tokens is roughly equivalent in size to 38,000 Wikipedias. In a world of low-cost compute, data will become the primary constraint.

GPT-3 Capability Today vs. Projected Capability In 2030



Sources: ARK Investment Management LLC, 2023; Wikipedia 2023; Hoffmann, J. et al. 2022. Forecasts are inherently limited and cannot be relied upon. For informational purposes only and should not be considered investment advice or a recommendation to buy, sell, or hold any particular security. Past performance is not indicative of future results.

GPT-4

Parameters: 1.7 trillion
Tokens: 13 trillion

GPT-3.5

Parameters: 175 billion
Tokens: 400 billion

[Source](#)

The Three Modern Phases of AI

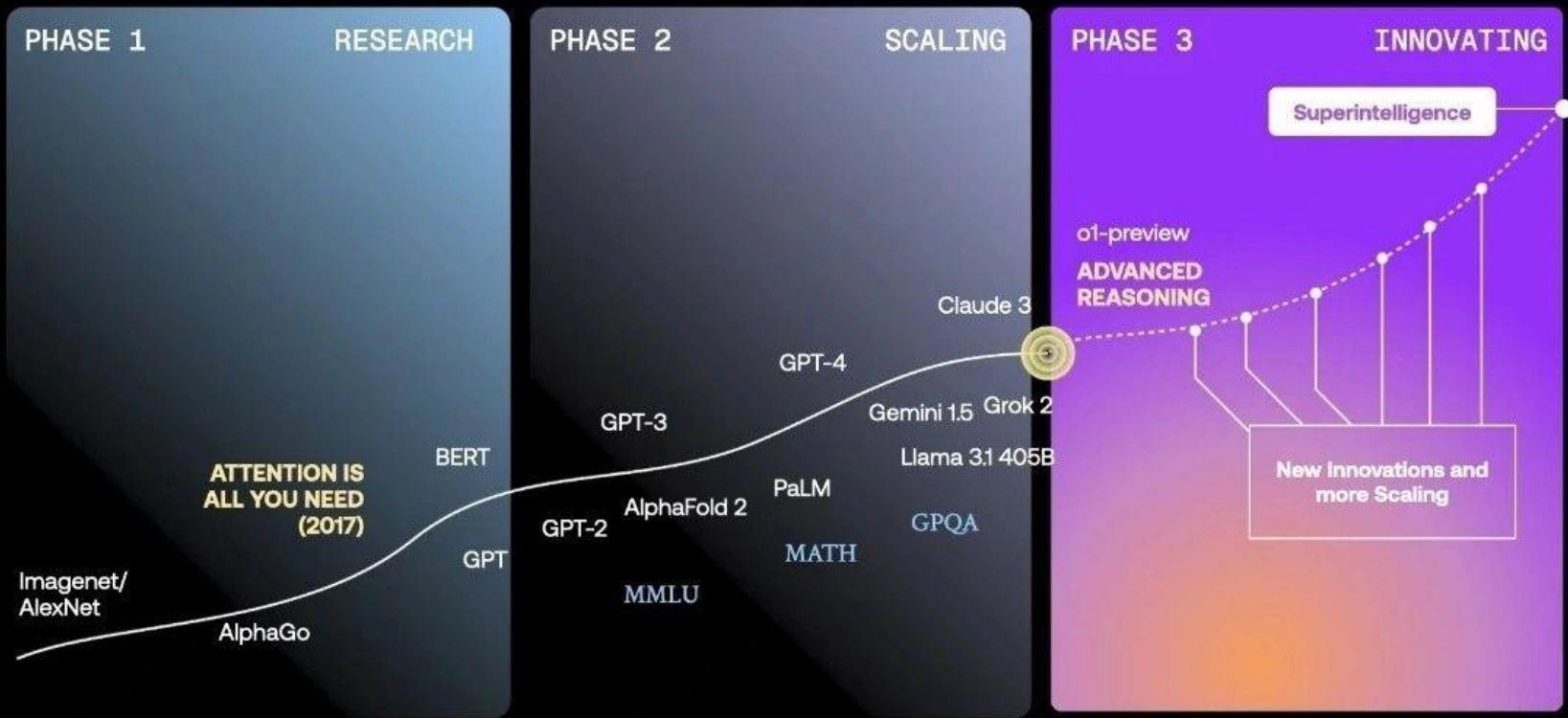
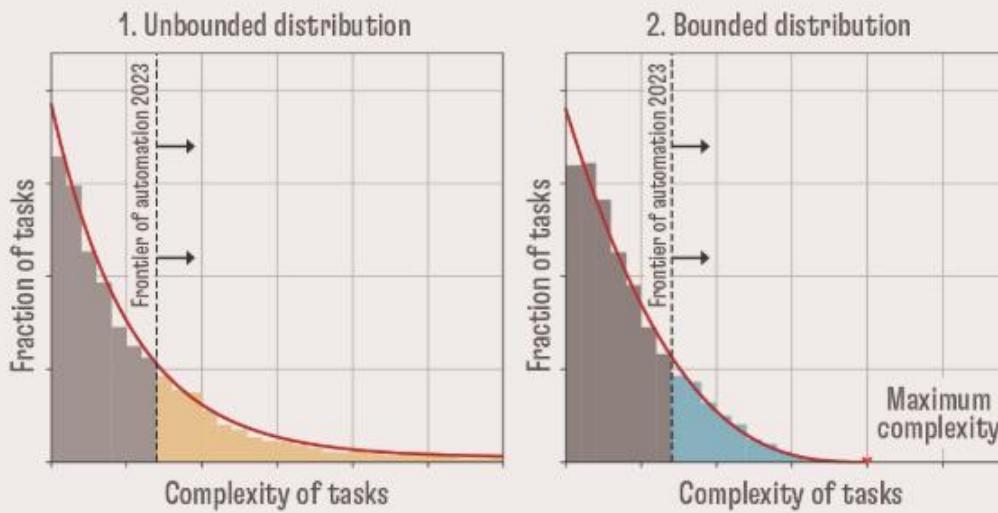




CHART 1

AI pushing the limits

Automation has already taken over many tasks that previously only humans could perform, as indicated by the gray bars. As AI advances, human beings may be able to continue to move into more complex tasks indefinitely (yellow bars), or AI may eventually overtake all human capabilities (blue bars).



SOURCE: Anton Korinek.

The Economics of Gen AI (2/2)

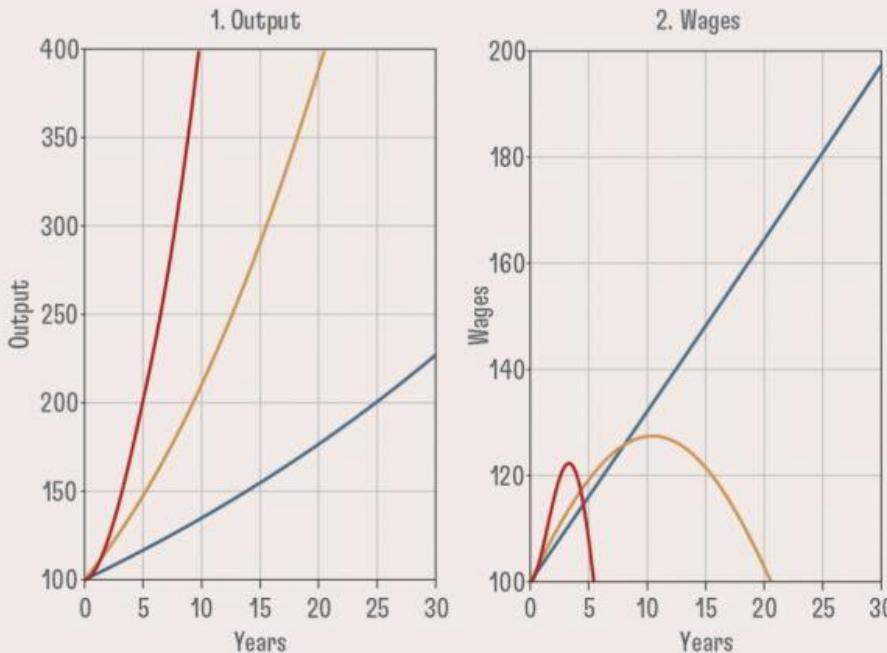
CHART 2

Scenarios for output and wages

The advent of AGI leads to a takeoff in growth at the expense of plummeting competitive market wages.

(initial year = 100)

— Traditional — Baseline AGI — Aggressive AGI



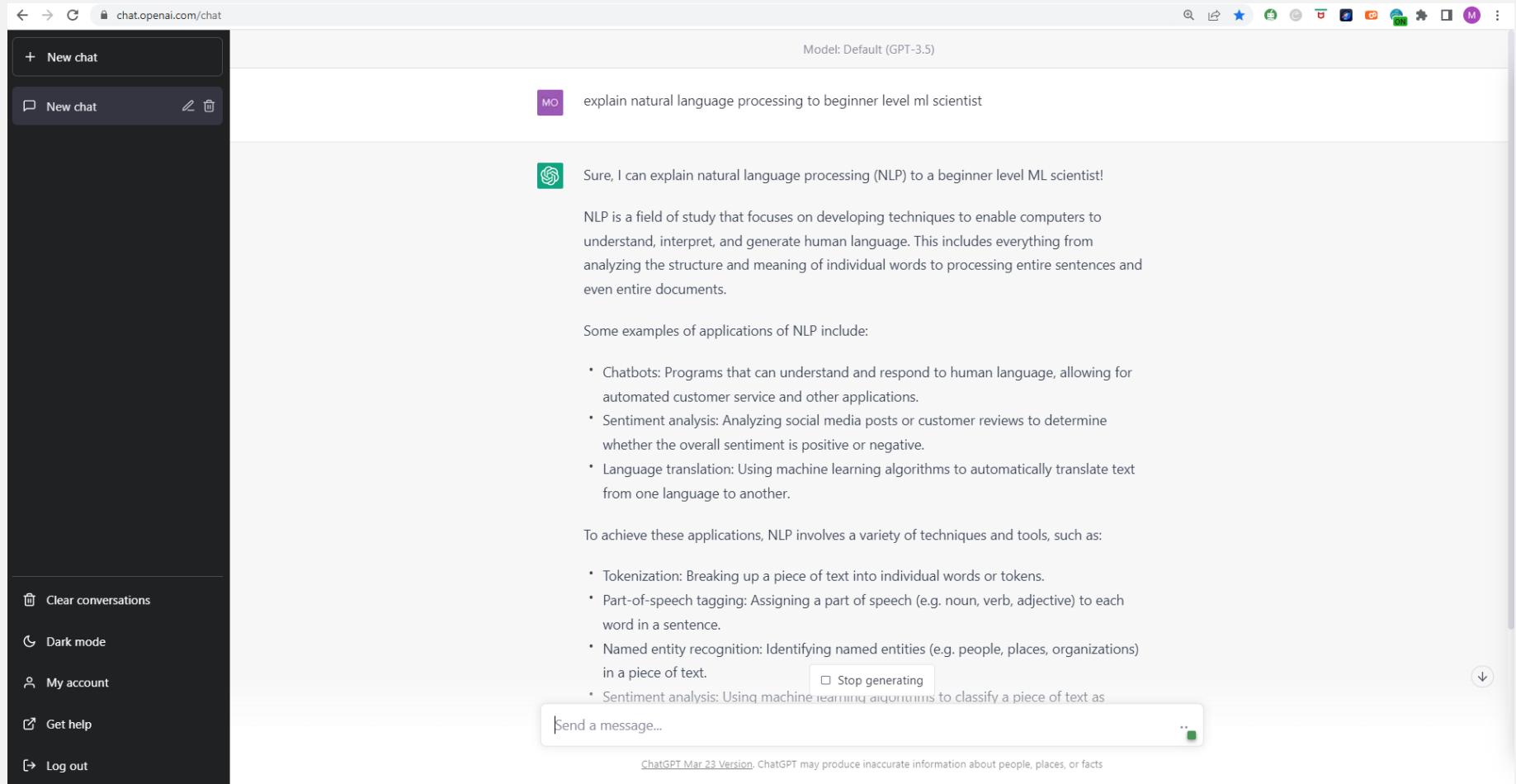
SOURCE: Anton Korinek.

NOTE: AGI = artificial general intelligence.

Generative AI Models

- **Text to Text:**
 - **Proprietary:** GPT, Claude, Cohere, Inflection Pi
 - **Open-Source:** LLaMA, Mixtral,
- **Text to Image:**
 - **Proprietary:** DALL-E, Midjourney
 - **Open-Source:** Stable Diffusion, niche players
- **Text to Video:**
 - **Proprietary:** SoRA, Runway, niche players
 - **Open-Source:** Stable Diffusion Video
- **Multi-Model:**
 - Text to Text, Text to Image, Text to Video, Text to Audio
 - Image to Text, Image to Video
 - Video to Text, Audio to Text

ChatGPT – Most popular Gen AI product



The screenshot shows a web browser window for chat.openai.com/chat. The interface is dark-themed. On the left, there's a sidebar with options like '+ New chat', 'New chat' (with a plus icon), and a trash bin icon. Below these are links for 'Clear conversations', 'Dark mode', 'My account', 'Get help', and 'Log out'. The main area shows a message from the user: 'explain natural language processing to beginner level ml scientist'. The AI response is: 'Sure, I can explain natural language processing (NLP) to a beginner level ML scientist!'. It continues by explaining what NLP is: 'NLP is a field of study that focuses on developing techniques to enable computers to understand, interpret, and generate human language. This includes everything from analyzing the structure and meaning of individual words to processing entire sentences and even entire documents.' It then lists examples of NLP applications: 'Some examples of applications of NLP include:

- Chatbots: Programs that can understand and respond to human language, allowing for automated customer service and other applications.
- Sentiment analysis: Analyzing social media posts or customer reviews to determine whether the overall sentiment is positive or negative.
- Language translation: Using machine learning algorithms to automatically translate text from one language to another.

' Finally, it describes the techniques used: 'To achieve these applications, NLP involves a variety of techniques and tools, such as:

- Tokenization: Breaking up a piece of text into individual words or tokens.
- Part-of-speech tagging: Assigning a part of speech (e.g. noun, verb, adjective) to each word in a sentence.
- Named entity recognition: Identifying named entities (e.g. people, places, organizations) in a piece of text.
- Sentiment analysis: Using machine learning algorithms to classify a piece of text as

' A 'Stop generating' button is visible next to the last bullet point. At the bottom, there's a text input field with 'Send a message...' placeholder text and a small green square icon. A note at the bottom says 'ChatGPT Mar 23 Version. ChatGPT may produce inaccurate information about people, places, or facts'.

TEXT TO IMAGE MODELS



AI-GENERATED FAKE IMAGE



TEXT TO VIDEO MODELS



WSJ
NEWS



Runway Gen2 + Eleven labs



Toys R Us short promo film almost entirely created by AI





Google Veo 2



||

Prompt: A low-angle shot captures a flock of pink flamingos gracefully wading in a lush, tranquil lagoon. The vibrant pink of their plumage contrasts beautifully with the verdant green of the surrounding vegetation and the crystal-clear turquoise water. Sunlight glints off the water's surface, creating shimmering reflections that dance on the flamingos' feathers. The birds' elegant, curved necks are submerged as they walk through the shallow water, their movements creating gentle ripples that spread across the lagoon. The composition emphasizes the serenity and natural beauty of the scene, highlighting the delicate balance of the ecosystem and the inherent grace of these magnificent birds. The soft, diffused light of early morning



||

Prompt: A cinematic shot captures a fluffy Cockapoo, perched atop a vibrant pink flamingo float, in a sun-drenched Los Angeles swimming pool. The crystal-clear water sparkles under the bright California sun, reflecting the playful scene. The Cockapoo's fur, a soft blend of white and apricot, is highlighted by the golden sunlight, its floppy ears gently swaying in the breeze. Its happy expression and wagging tail convey pure joy and summer bliss. The vibrant pink flamingo adds a whimsical touch, creating a picture-perfect image of carefree fun in the LA sunshine.

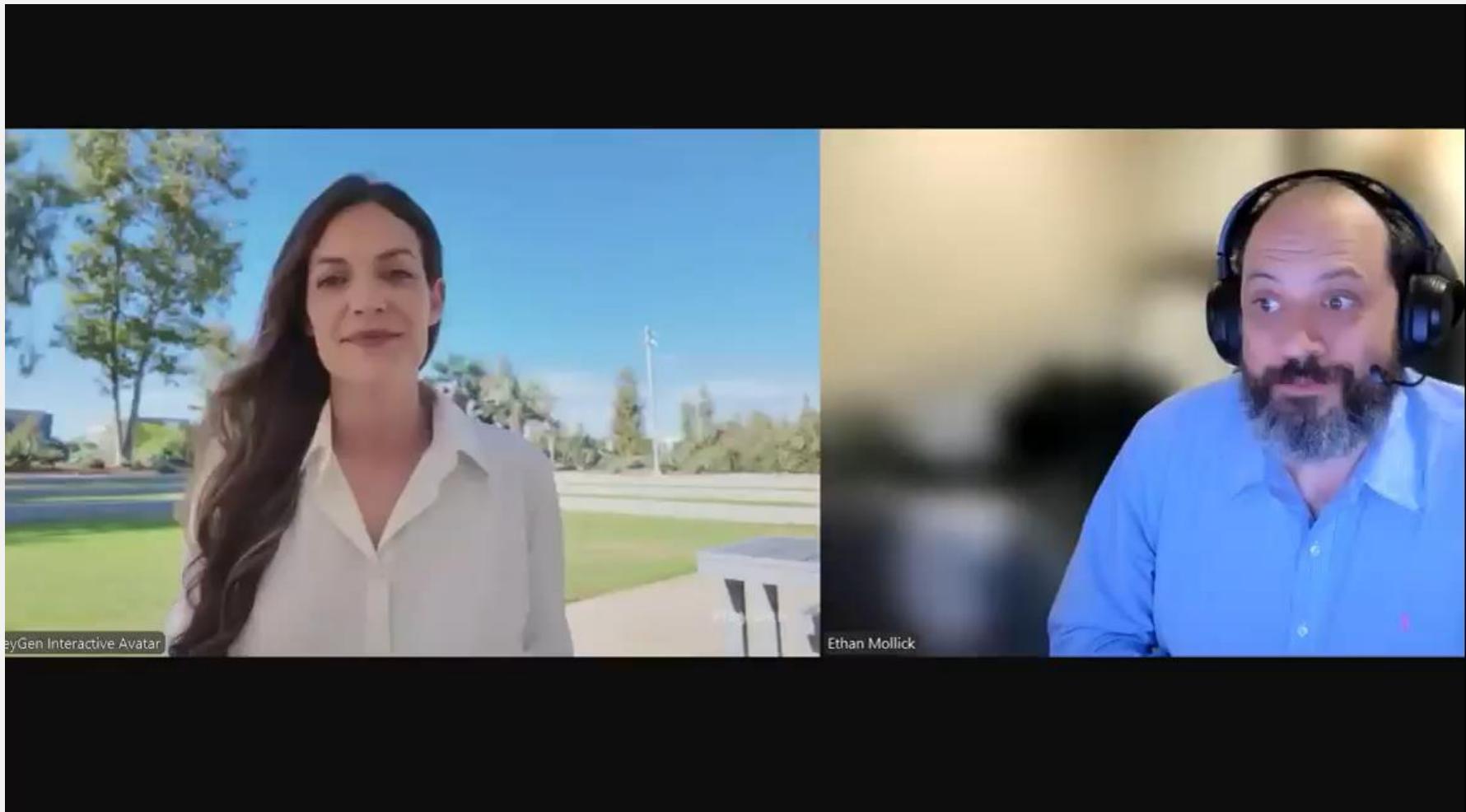
[View on YouTube](#)



||

Prompt: A cinematic, high-action tracking shot follows an incredibly cute dachshund wearing swimming goggles as it leaps into a crystal-clear pool. The camera plunges underwater with the dog, capturing the joyful moment of submersion and the ensuing flurry of paddling with adorable little paws. Sunlight filters through the water, illuminating the dachshund's sleek, wet fur and highlighting the determined expression on its face. The shot is filled with the vibrant blues and greens of the pool water, creating a dynamic and visually stunning sequence that captures the pure joy and energy of the swimming dachshund.

[View on YouTube](#)



Interview with CTO of OpenAI

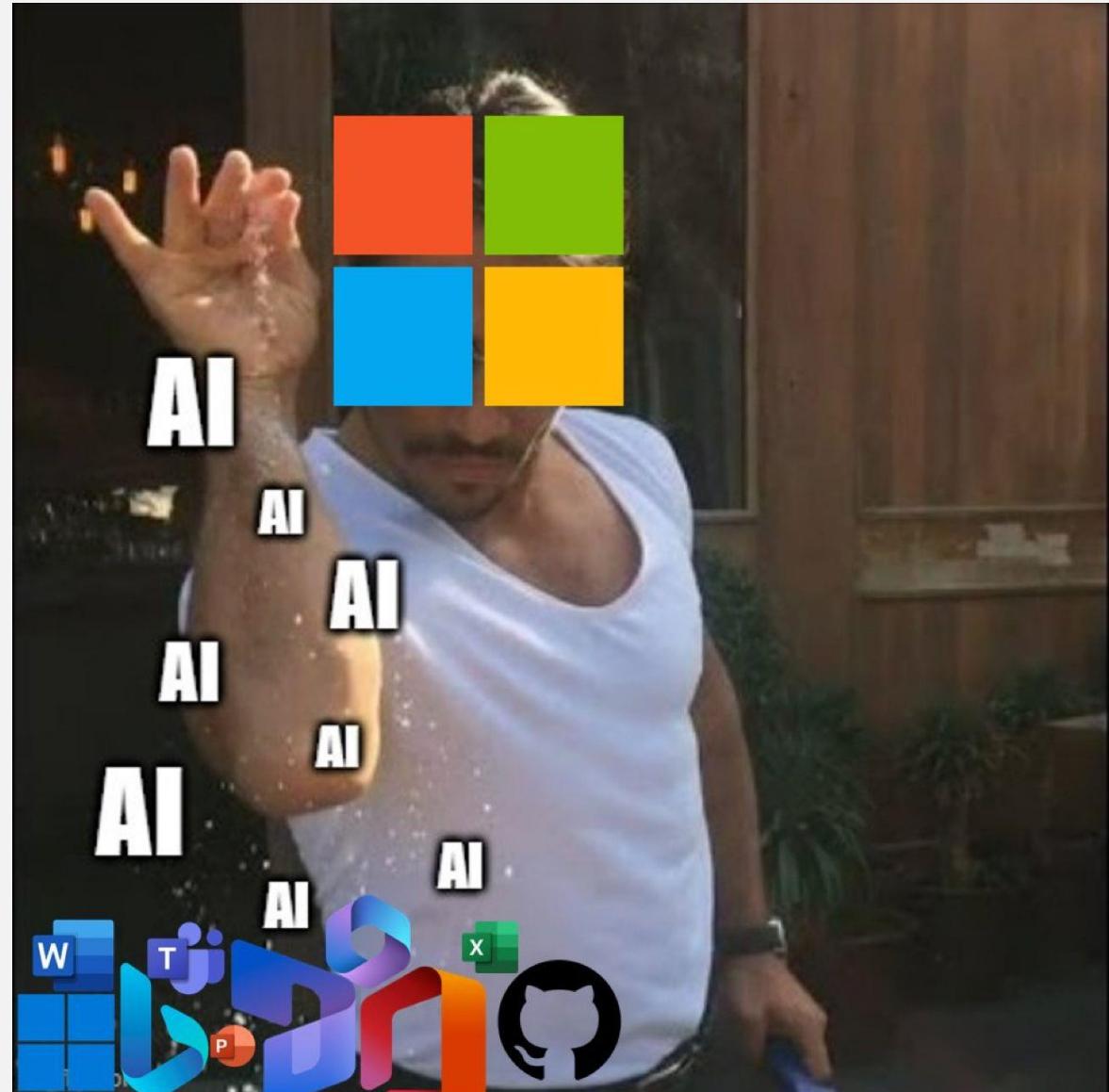
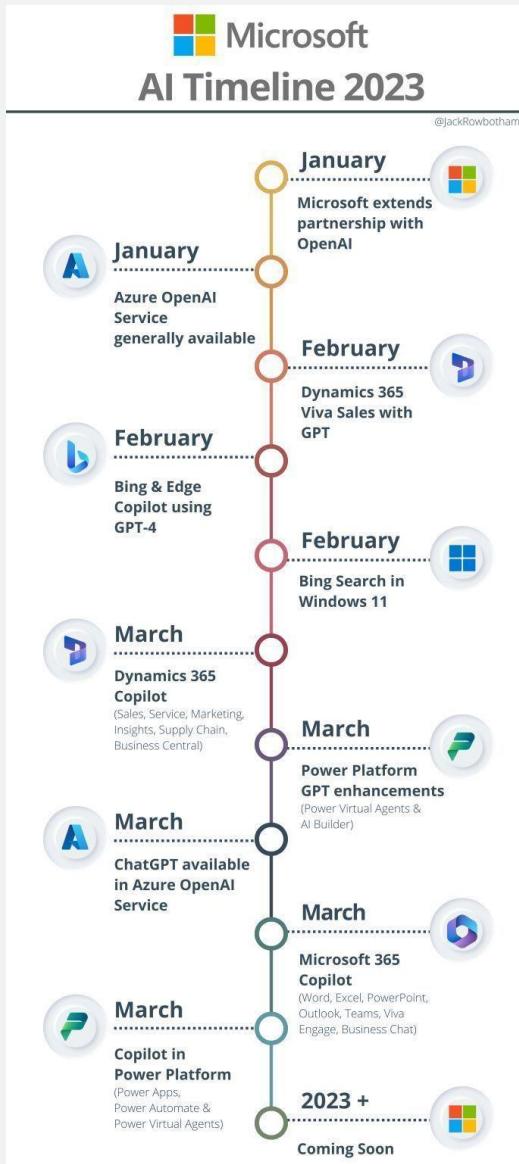


TEXT TO AUDIO

NVIDIA Fuggato



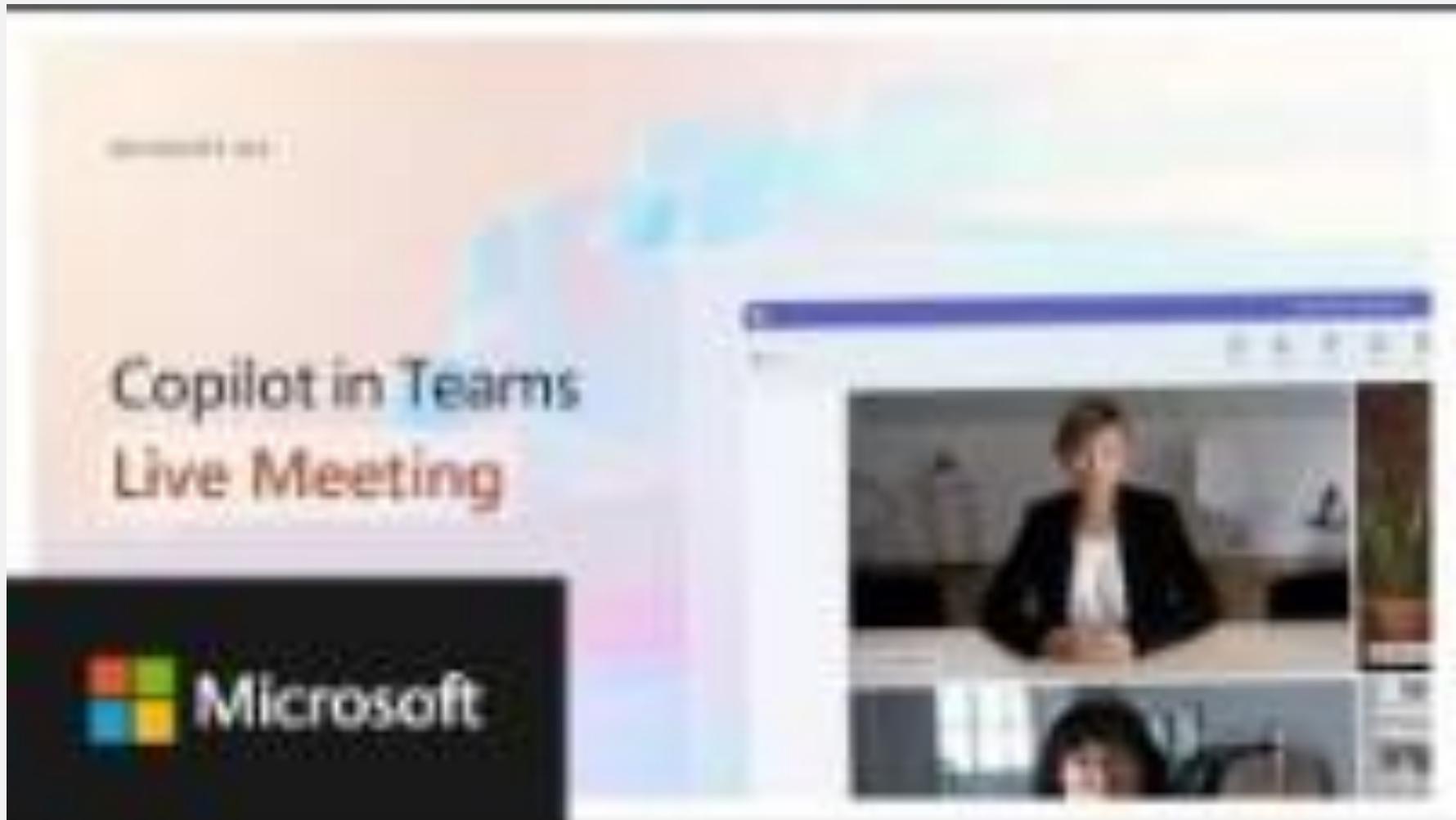
Microsoft GPT Integration timelines



Copilot in Microsoft PowerPoint

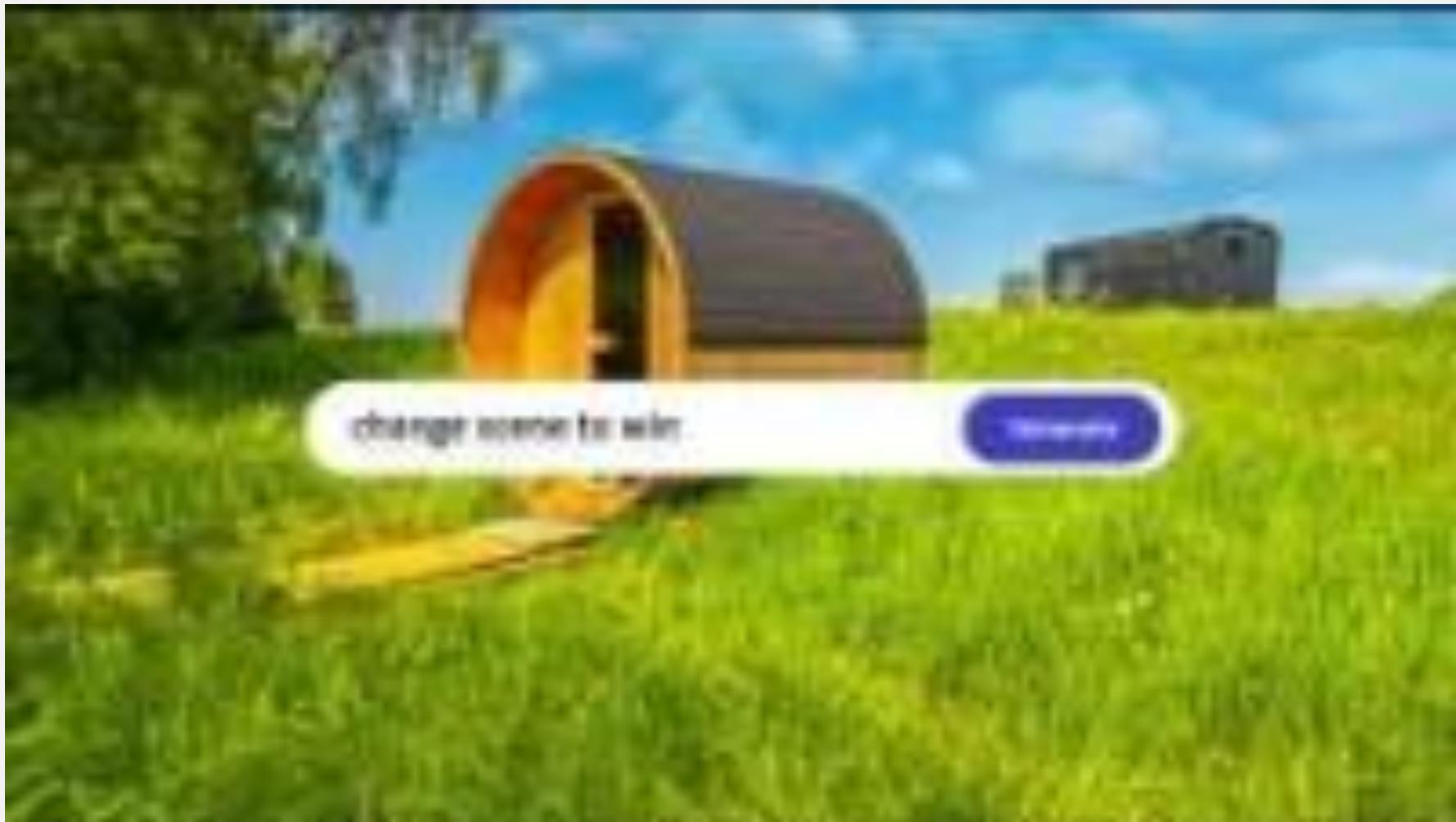


Copilot in Microsoft Teams



The image shows a screenshot of a Microsoft Teams live meeting interface. At the top left, there's a blurred profile picture of a person. The main title of the meeting is "Copilot in Teams Live Meeting". On the right side of the title bar, there are several small, overlapping windows showing different video feeds from other participants. Below the title bar, the Microsoft logo is visible. The main video frame on the right shows a woman in a professional setting, possibly an office or conference room, wearing a dark blazer over a white top. She appears to be speaking or presenting. There are also smaller video frames at the bottom of the screen showing other participants.

Adobe Firefly: New Creative Generative AI Models



GitHub Copilot X



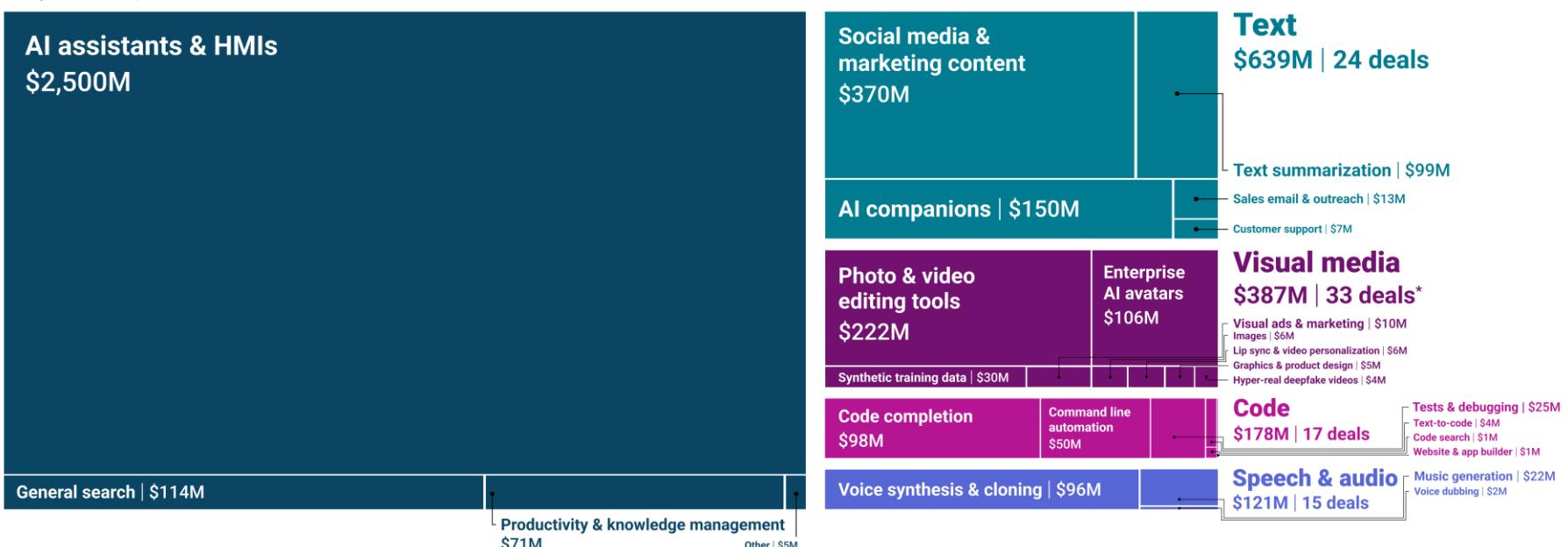


Where is all the money going in generative AI?

Distribution of generative AI funding, Q3'22 – Q2'23

Generative interfaces

\$2,690M | 23 deals



Source: CB Insights. Based on an analysis of 210+ generative AI companies building cross-industry enterprise solutions; excludes deals to industry-specific companies and model developers such as OpenAI.

*Includes 1 deal in motion capture animation and 1 deal in synthetic anonymization with undisclosed funding.

SELECT TOP INVESTORS IN GENERATIVE AI (MARCH 2023)

Investor	Typical Stage	Apps – Text / Audio / Code	Apps – Image / Video / Design	Apps – Others (e.g. search, health)	Enablement / MLOps / Dev Tool	Infra / Data / Model / Research
SOMA CAPITAL	Seed	Draft cogram FATHOM Pilot	tavus ^{ai} Penguin AI	ClipDrop durable SelfScale	slai Monterey Astack	generally Intelligent
TIGERGLOBAL	Stage-Agnostic	copy.ai Writsonic DEEPOGRAM	Bir	CRESTA Grain Moltby ushur	aporia invideo slai OctoML ROBUST INTELLIGENCE	cohere Pinecone scale
SEQUOIA	Seed – Series C	copy.ai Twain warp	tavus ^{ai} pencil VEED.IO	Moltby CRESTA obviously.ai glean neeva	replicate Hugging Face ROBUST INTELLIGENCE	Hugging Face
AIX Ventures	Seed / Series A	Replica	Omneky Facet	ATOMIC AI Proluent Perplexity	mutable.ai Hugging Face Weights & Biases	Character.AI
andressen horowitz	Stage-Agnostic	dialpad replit	descript reface YEPIC STUDIO	CRESTA BigHot mem metaphor	replicate anyscale	
khosla ventures	Stage-Agnostic	Draft Replica bloop. tobine Flexics	R Simplified	regie.ai menten AI SPLASH		OpenAI
South Park Commons	Pre-seed / Seed	Speak replit	Facet Reduct LUMA AI Storytime	regie.ai Forethought durable	DASHWORKS GANTRY	
Alumni Ventures	Seed – Series B	agolo lang.ai replicant	chooch	Woebot Health sorcer0 moises Monterey	slai BANANA	Synthesis AI
INSIGHT VENTURES PARTNERS	Stage-Agnostic	Jasper AssemblyAI WRITER Beepdub.ai	Hour One Lightricks Wizard	Proluent Moltby	Weights & Biases	TOHIC
8VC	Seed – Series B	casetext	tome Rephrase.ai METAPHYSIC	BigHot metaphor ushur Forethought		AI21labs
Amplify	Pre-seed / Seed	Primer	runway LUMA AI		Magic anyscale GANTRY OctoML	Outerbounds
COATUE	Later-Stage/ Buyout	Jasper replit	tome runway		Hugging Face Weights & Biases GANTRY Lightning	scale stability.ai
hustlefund	Pre-seed / Seed	copy.ai HyperWrite LOVO TALES	BIFROST detail	OASIS CALA onloop	slai	
Index Ventures	Stage-Agnostic		Twelve Labs PolyAI	Hebbia moises	GANTRY Lightning Humanioop	cohere scale
SAMSUNG NEXT	Series A - B	tabnine Fireflies		Forethought	mutable.ai mosaic aporia	datagen TOHIC stability.ai
Accel	Stage-Agnostic	AssemblyAI Locoify	Facet	ada Prodigal timelyAI The.com	DhiWise	scale
Bloomberg BETA	Pre-seed / Seed	Primer xplixt extio	PolyAI	ProteinDure COLIV	replicate Weights & Biases	TOHIC
greylockpartners.	Seed – Series B		tome Reduct Video	CRESTA neeva ATOMIC AI	gretel	Adept Inflection
L2 VENTURES	Seed	Jasper FATHOM OBSERVE-AI	tavus ^{ai} PhotoRoom	Moltby Attention needi		
LUX	Seed – Series C	Primer	runway BIFROST	A-ALPHA BIO Enveda BIOSCIENCES	Hugging Face mosaic	
Section 32	Series A - C	dialpad Primer	METAPHYSIC inworld	XOKind	gretel	cohere
VILLAGE GLOBAL	Pre-seed / Seed	lang.ai	Omneky	Enveda BIOSCIENCES Forethought Alde	gretel	SYNTACRA
Broad Capital	Pre-seed / Seed	Speechify copy.ai	tavus ^{ai} Beulr RIVERSIDE	SelfScale SetScale		
foundation capital	Seed / Series A	Jasper	D-ID	regie.ai	any scale Booth.ai	Outerbounds
FOUNDERS FUND	Seed / Series A	Speak	tome inworld	OASIS REPLICASIMI		AI FOUNDATION scale

VALUE GENERATION

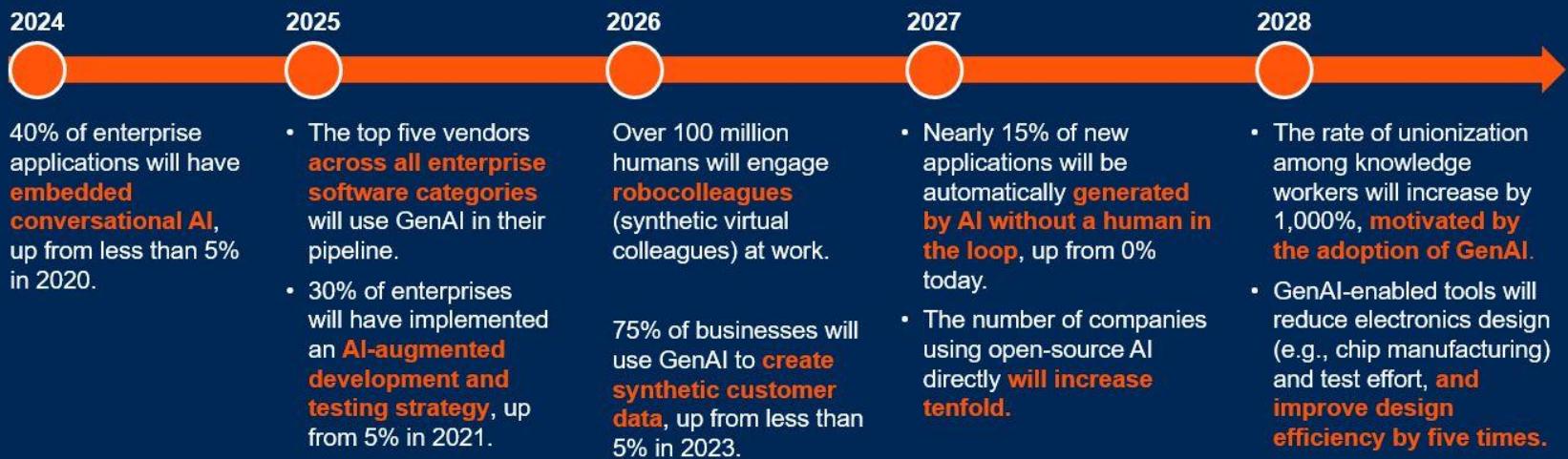
1. Infra (Azure, GCP, AWS)

2. Model (OpenAI, Anthropic, Inflection, etc.)

3. Applications (???)

What Is the Likely Path for GenAI Beyond 2024?

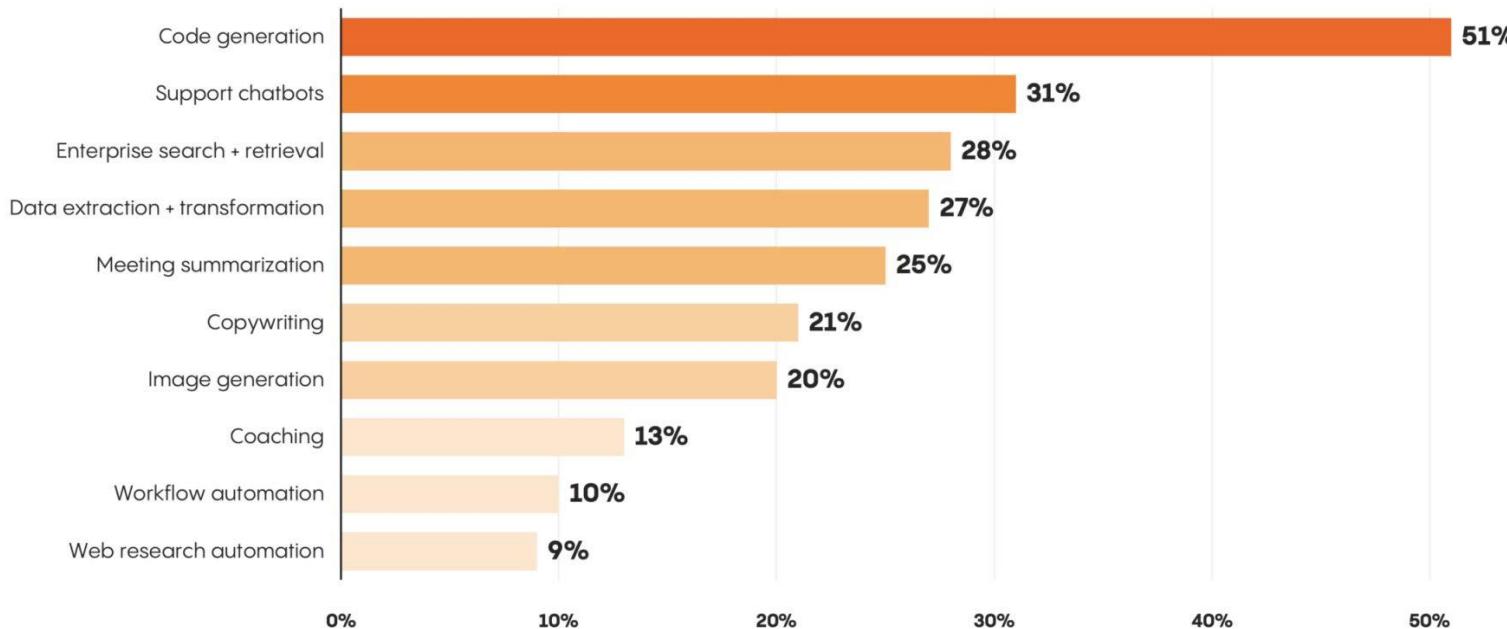
Gartner Prediction of GenAI Adoption



Source: Gartner

Menlo Ventures State of Gen AI 2024

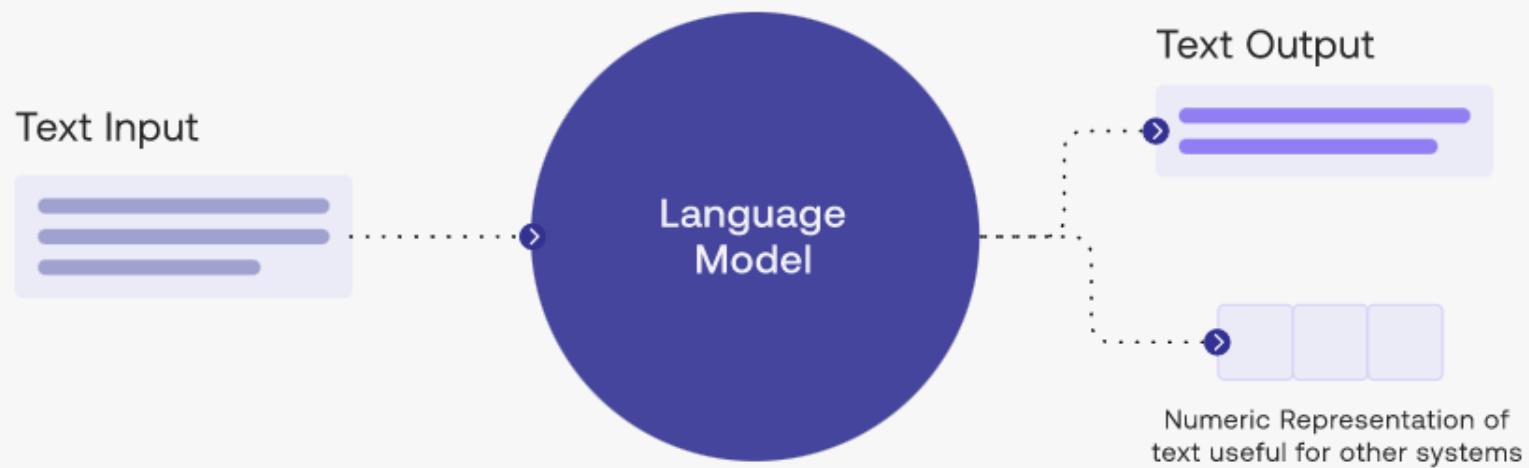
Dominant Generative AI Use Cases



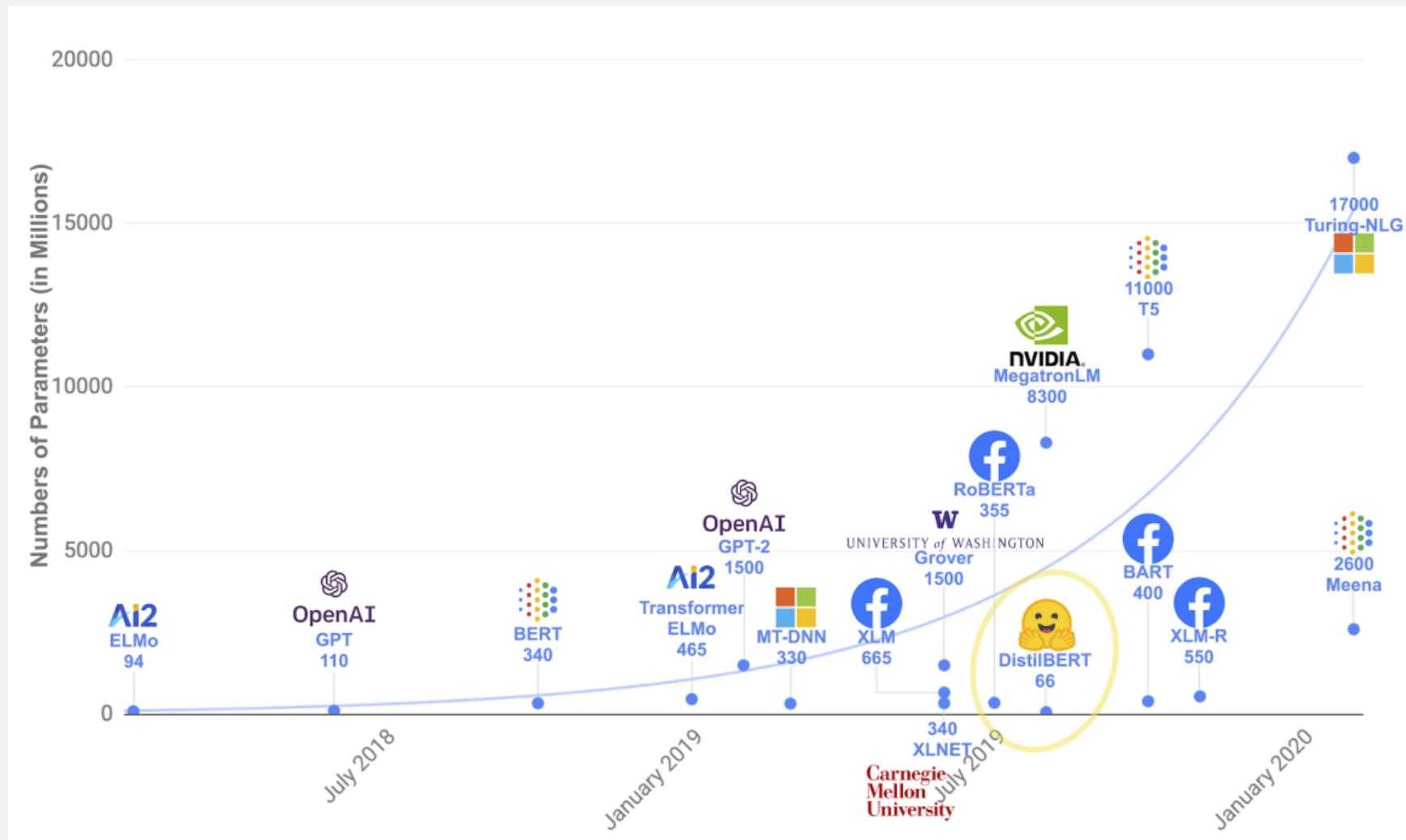
LARGE LANGUAGE MODELS (LLMs)

Definition of LLM

- Large language models (LLMs) are advanced artificial intelligence systems that are designed to process and understand natural language. They open new possibilities of text understanding and generation in software systems.



Pre-GPT Era



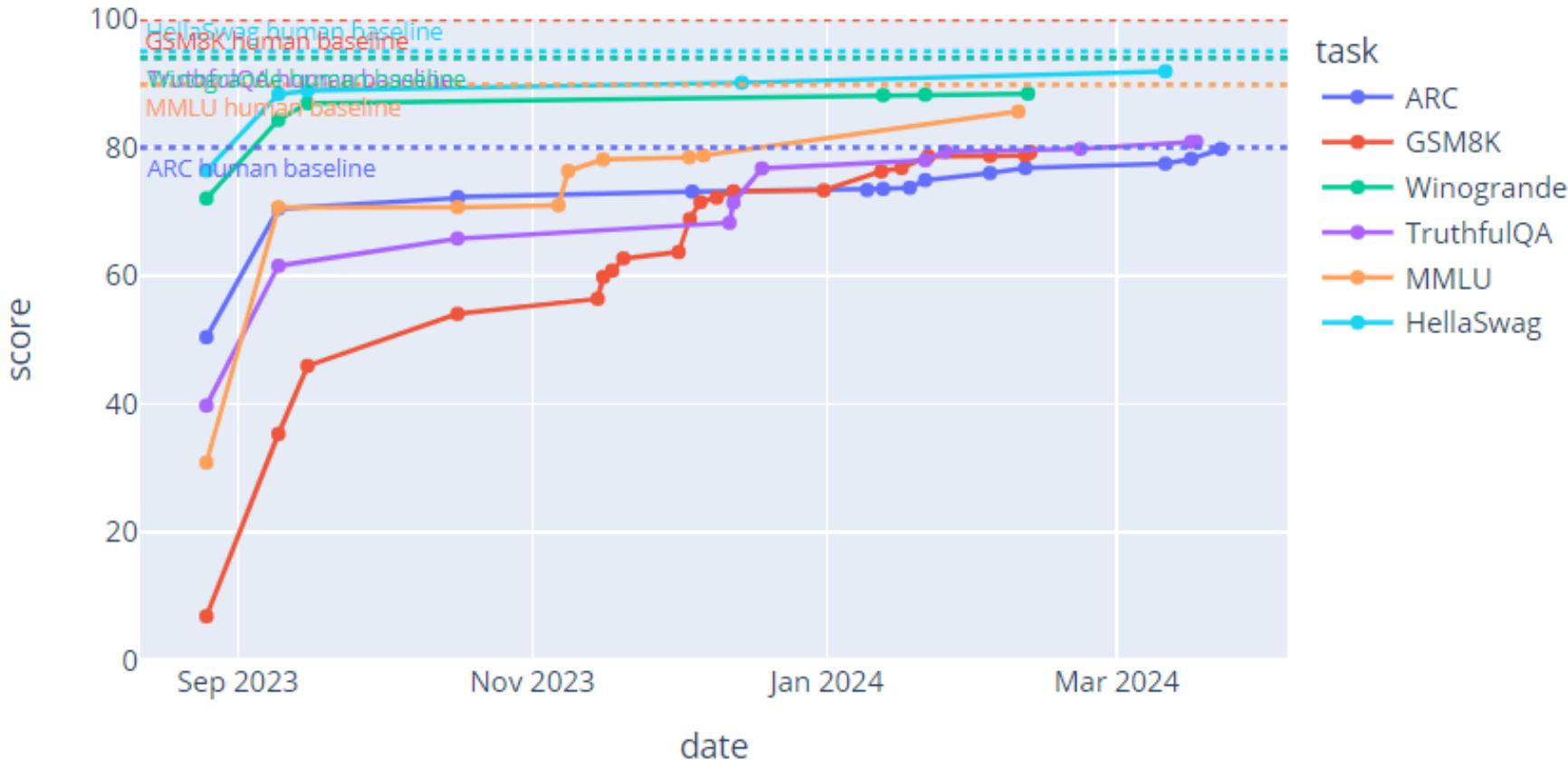


LMSYS Chatbot Arena Leaderboard (as of 25-Nov-2024)

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	2	Gemini-Exp-1121	1365	+8/-6	5625	Google	Proprietary	Unknown
1	1	ChatGPT-4o-latest (2024-11-20)	1361	+4/-5	10658	OpenAI	Proprietary	Unknown
3	5	Gemini-Exp-1114	1344	+4/-5	12778	Google	Proprietary	Unknown
4	2	o1-preview	1334	+4/-4	27835	OpenAI	Proprietary	2023/10
5	8	o1-mini	1308	+3/-4	31992	OpenAI	Proprietary	2023/10
5	5	Gemini-1.5-Pro-002	1301	+5/-3	27336	Google	Proprietary	Unknown
6	5	Gemini-1.5-Pro-Exp-0827	1299	+4/-4	32345	Google	Proprietary	2023/11
8	11	Grok-2-08-13	1289	+4/-3	52102	xAI	Proprietary	2024/3
8	13	Yi-Lightning	1287	+4/-3	29336	01 AI	Proprietary	Unknown
8	5	GPT-4o-2024-05-13	1285	+2/-2	111745	OpenAI	Proprietary	2023/10
9	3	Claude 3.5 Sonnet (20241022)	1282	+4/-3	29454	Anthropic	Proprietary	2024/4
11	18	Athene-v2-Chat-72B	1274	+8/-6	4354	NexusFlow	NexusFlow	Unknown

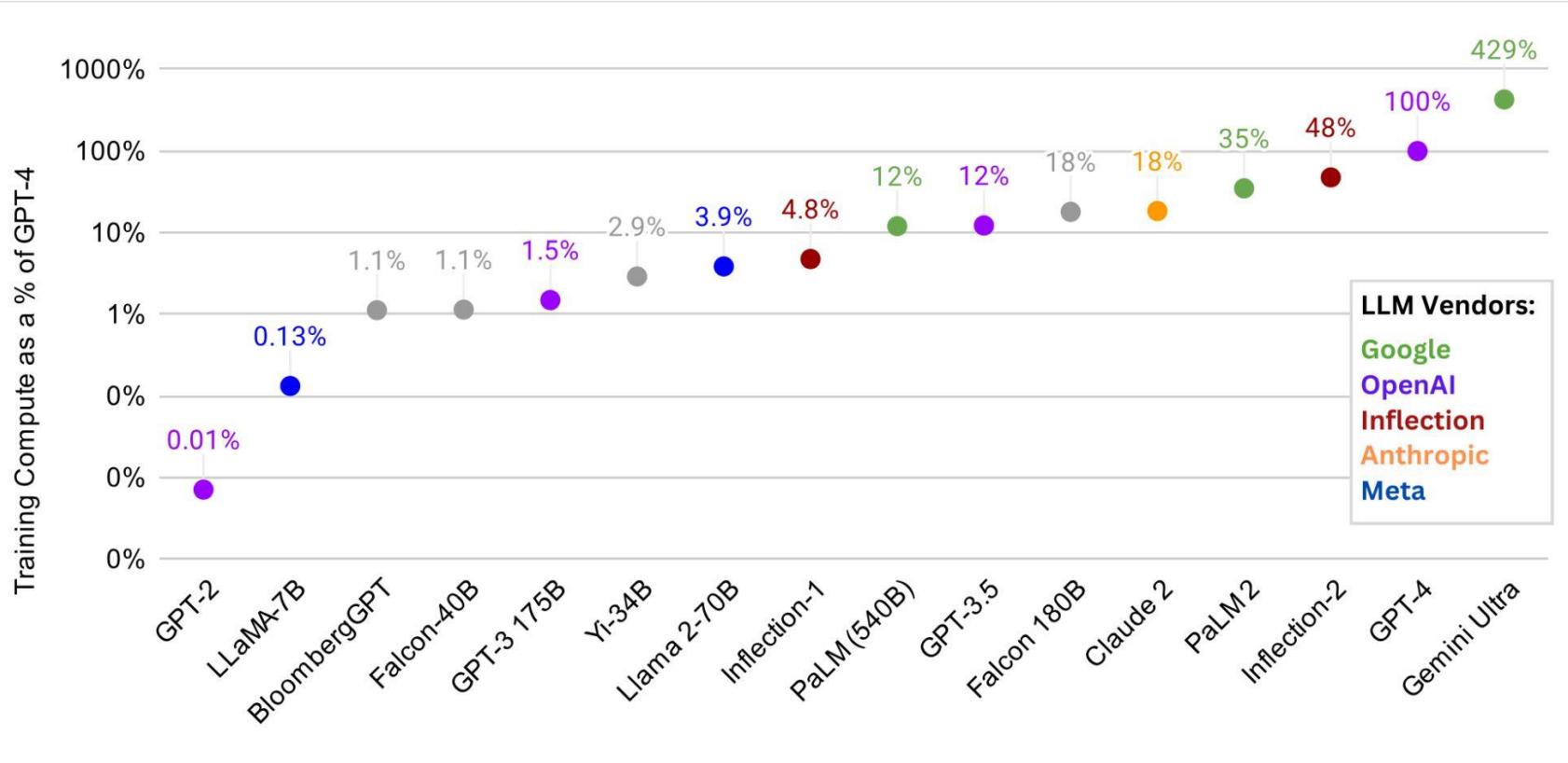
Performance of open-source models over time (as of 31-Mar-2024)

Top Scores and Human Baseline Over Time (from last update)



Training Cost as % of GPT-4 compute

LLM Training Compute as a % of GPT-4



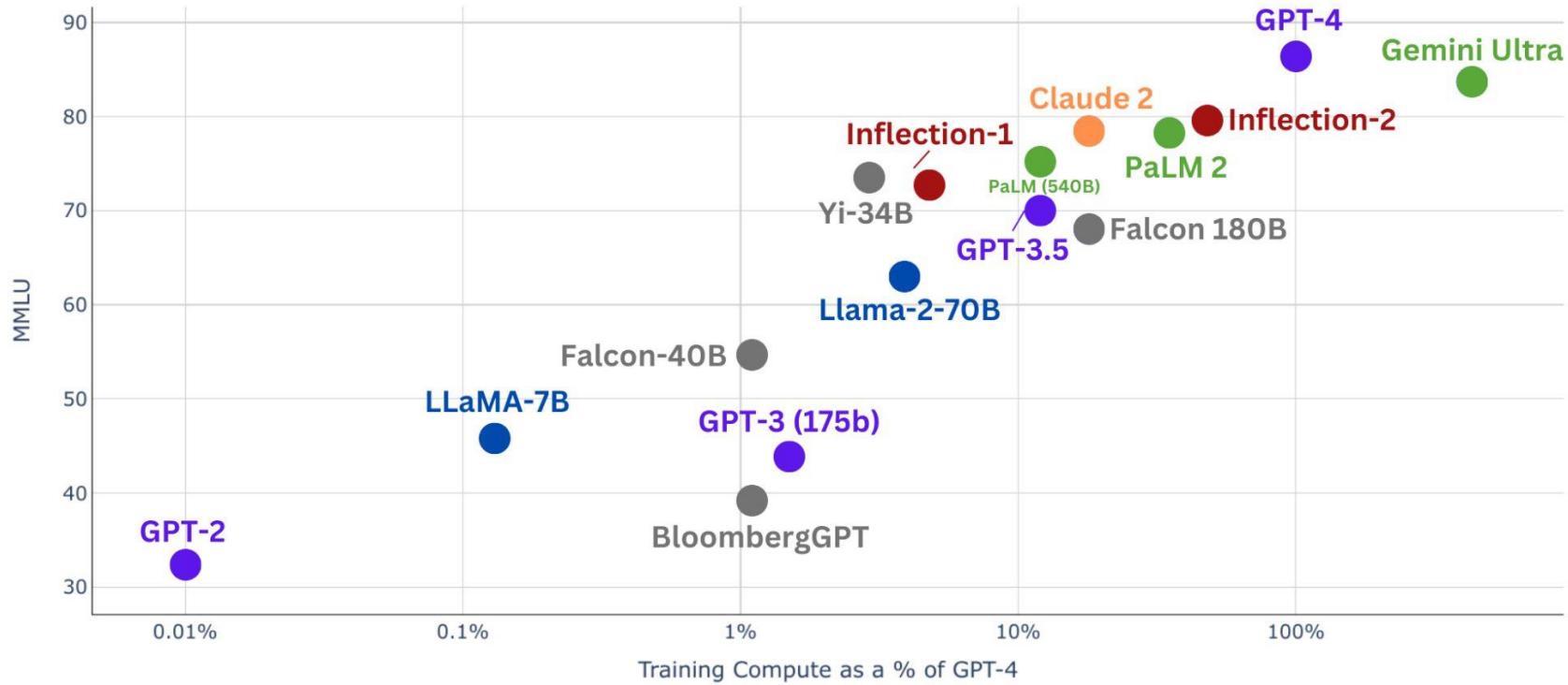
Source: Analysis by Peter Gostev (<https://www.linkedin.com/in/peter-gostev/>)

Based on data from Epochs.org (<https://epochai.org/blog/who-is-leading-in-ai-an-analysis-of-industry-ai-research>)

MMLU vs. Training Cost as % of GPT-4

Compute vs Benchmarks

Compute as a % of GPT-4, MMLU Benchmark Score



Source: Analysis by Peter Gostev (<https://www.linkedin.com/in/peter-gostev/>), Based on benchmark reports and data from Epochs.org (<https://epochai.org/blog/who-is-leading-in-ai-an-analysis-of-industry-ai-research>)

Elo Scores by Company - Top 7

Top Ranked Model by Company in the Chatbot Arena

[Replay](#)



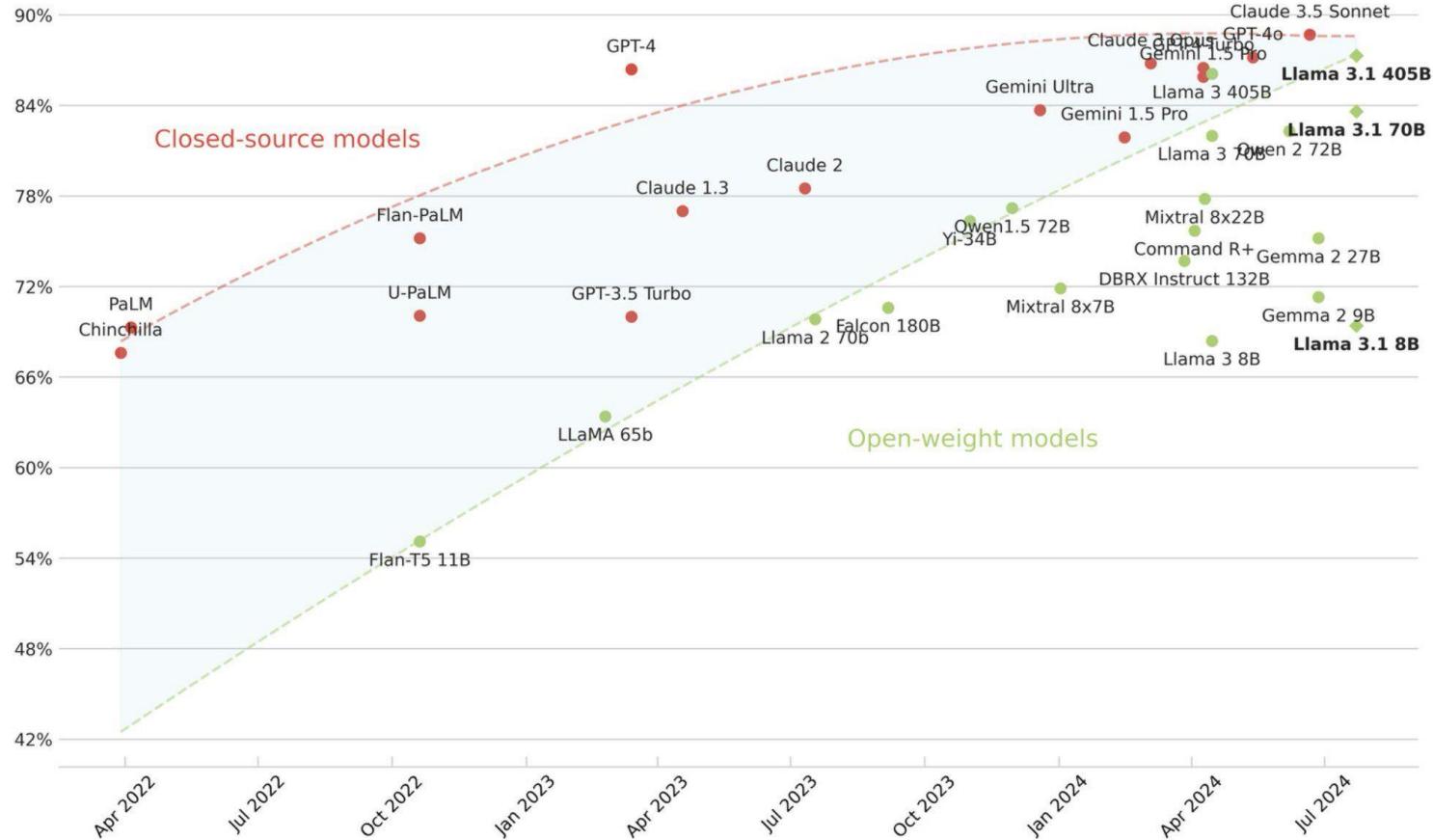
Source: [LMSYS Chatbot Arena](#), Peter Gostev

Closed-source vs. open-weight models

@maximelabonne

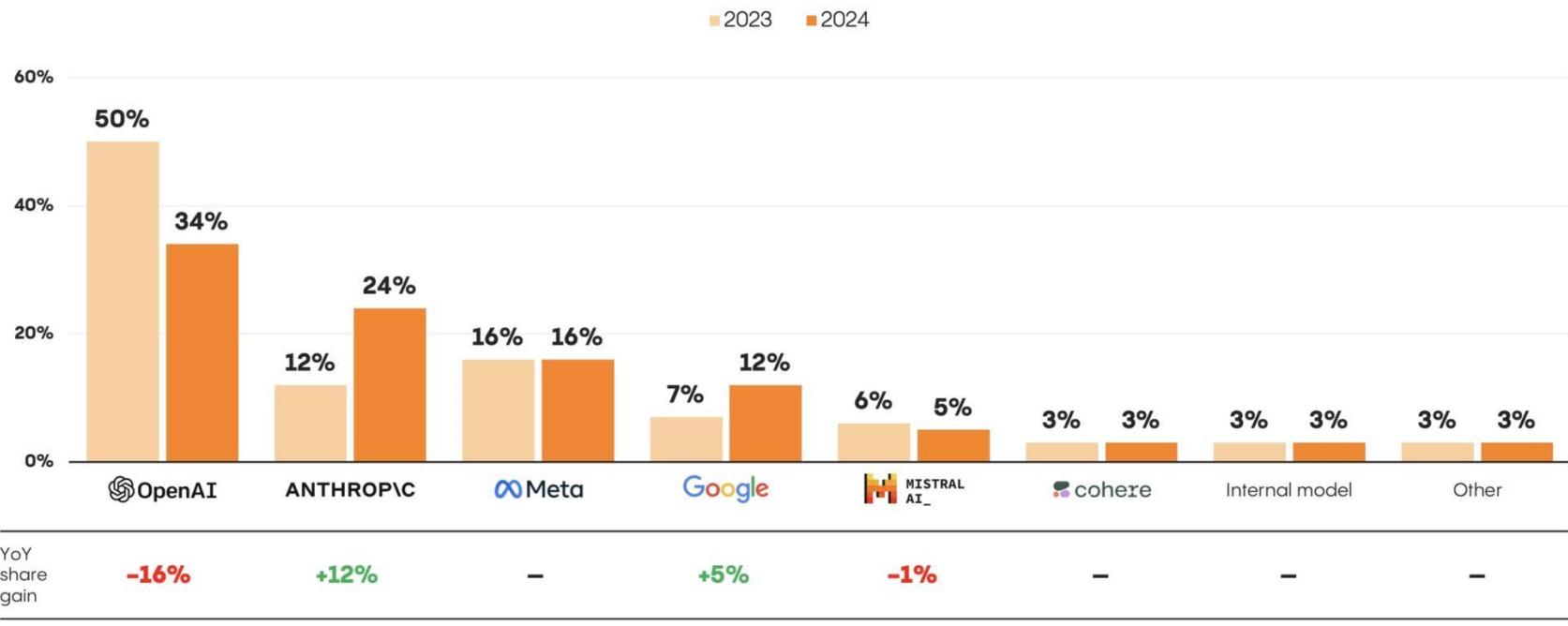
Llama 3.1 405B closes the gap with closed-source models for the first time in history.

MMLU (5-shot)



Market share of LLMs

Market Share of LLMs: 2023 vs. 2024



PROMPT ENGINEERING

Prompt Engineering

- Instructions and context given to a language model to accomplish a specific task are referred to as **prompts**.
- **Prompt engineering** refers to the process of designing and refining prompts to effectively utilize large language models (LLMs) across variety of applications.
- Prompt engineering is a useful skill.

Prompt Engineering

MO

what is prompt engineering?



Prompt engineering is the discipline of designing, refining, and optimizing prompts to make the most effective use of language models (LMs) for various tasks. It involves understanding the nuances of the model's responses and tailoring the prompts to guide the model towards producing desired outputs. As language models like those of the GPT series became more advanced, the manner in which they are prompted becomes crucial for harnessing their full potential, hence the emergence and importance of prompt engineering.



MO

what is prompt engineering? be concise. not more than 2 sentences. explain in very simple language.



Prompt engineering is the process of crafting and refining questions or statements to get the best responses from a language model. It helps in guiding the model to produce the desired output.



Specific prompts for:

- Text Summarization
- Question Answering
- Text Classification
- Role Playing
- Code Generation
- Reasoning

Example of Role-playing Prompt

MO

Imagine you are a medieval knight addressing the king about an impending dragon threat to the kingdom. How would you report the situation and ask for assistance? 



Your Majesty,



I come before you with grave news that demands our immediate attention. There have been sightings of a fearsome dragon in the eastern regions of our kingdom, causing distress among the villagers. I fear that if we do not act promptly, the beast may pose a significant threat to our land and its people. I humbly request your permission to assemble a select group of our bravest knights and resources to confront and subdue this menace, ensuring the safety and prosperity of our realm. Your guidance and support in this matter will be invaluable.

Your loyal servant,
Sir [Your Name]

Zero shot vs. Few shot prompting

- **Zero-shot prompting:** Language model performs tasks without prior specific examples, relying on its generalized training.
 - Zero-shot can sometimes generalize better but might lack specificity.
- **Few-shot prompting:** Model is given a small number of examples to infer the task, guiding its response.
 - The potential for few-shot to provide more accuracy for niche tasks but at the risk of overfitting.

Zero Shot Prompting

MO

Classify the following text into positive, neutral, or negative sentiment.

1. I hate this pizza restaurant.
2. Wow. what a lovely weather!
3. I am amazed by the views
4. Can you please call the driver?
5. I can't sleep.



Here are the classifications for the given texts:

1. I hate this pizza restaurant. - **Negative**
2. Wow. what a lovely weather! - **Positive**
3. I am amazed by the views - **Positive**
4. Can you please call the driver? - **Neutral**
5. I can't sleep. - **Negative**

Few Shot Prompting

MO

Text: "Barack Obama was the president of the United States."

Entities: "Barack Obama: Person, United States: Location"

Text: "Apple is a tech company based in California."

Entities: "Apple: Organization, California: Location"

Text: "Madrid is the capital of Spain."

Entities:



Text: "Madrid is the capital of Spain."

Entities: "Madrid: Location, Spain: Location"

Prompt Engineering

ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs

Jingming Zhuo^{1,2,*}, Songyang Zhang^{1,*}, Xinyu Fang^{1,3}, Haodong Duan¹

Dahua Lin^{1,4}, Kai Chen^{1,†}

¹Shanghai AI Laboratory, ²Jilin University, ³Zhejiang University,

⁴The Chinese University of Hong Kong

jingmingzhuo@gmail.com, zhangsongyang@pjlab.org.cn

* equal contribution, † corresponding author

Abstract

Large language models (LLMs) have demonstrated impressive capabilities across various tasks, but their performance is highly sensitive to the prompts utilized. This variability poses challenges for accurate assessment and user satisfaction. Current research frequently overlooks instance-level prompt variations and their implications on subjective evaluations. To address these shortcomings, we introduce **ProSA**, a framework designed to evaluate and comprehend prompt sensitivity in LLMs. ProSA incorporates a novel sensitivity metric, **PromtSensiScore**, and leverages decoding confidence to elucidate underlying mechanisms. Our extensive study, spanning multiple tasks, uncovers that prompt sensitivity fluctuates across datasets and models, with larger models exhibiting enhanced robustness. We observe that few-shot examples can alleviate this sensitivity issue, and subjective evaluations are also susceptible to prompt sensitivities, particularly in complex, reasoning-oriented tasks. Furthermore, our findings indicate that higher model confidence correlates with increased prompt robustness. We believe this work will serve as a helpful tool in studying prompt sensitivity of LLMs. The project is released at: <https://github.com/open-compass/ProSA>.

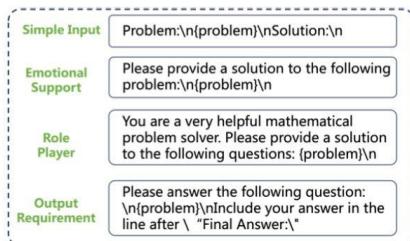


Figure 1: **A Showcase of the Four Prompt Templates on MATH.** These four prompt templates represent four different styles of constructing prompts, serving as an example of the diversity in human prompt expression.

styles of prompts used by LLMs when completing mathematical tasks.

The diversity of prompts elicits various responses from LLMs. Recent studies (Zhu et al., 2023; Pezeshkpour and Hruschka, 2024; Sclar et al., 2024; Zhou et al., 2023) investigate model performance across different prompt templates and demonstrate that LLMs are highly sensitive to the nuances of prompts. Even minor alterations to prompts can lead to substantial declines in model performance. This sensitivity to prompts poses a challenge for researchers aiming to precisely eval-

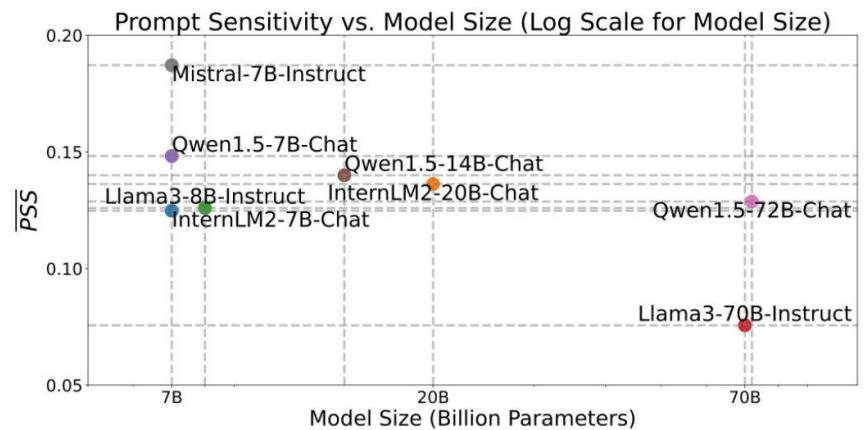


Figure 4: **Prompt Sensitivity vs. Model Size.** The comparative charts display the relationship between the size of the model's parameters and prompts sensitivity. **PSS** refers to the average PSS of four datasets.

FINE-TUNING LARGE LANGUAGE MODELS

What is Fine-tuning LLM?

- Fine-tuning involves updating the weights of a pre-trained language model on a new task and dataset.
- Modifying a pre-trained model to specialize in specific tasks.
- Risk of overfitting with too little data.
- Ideal for domain-specific applications like medical, legal, or finance.

Fine-tuning lets you get more out of the models available through the API by providing:

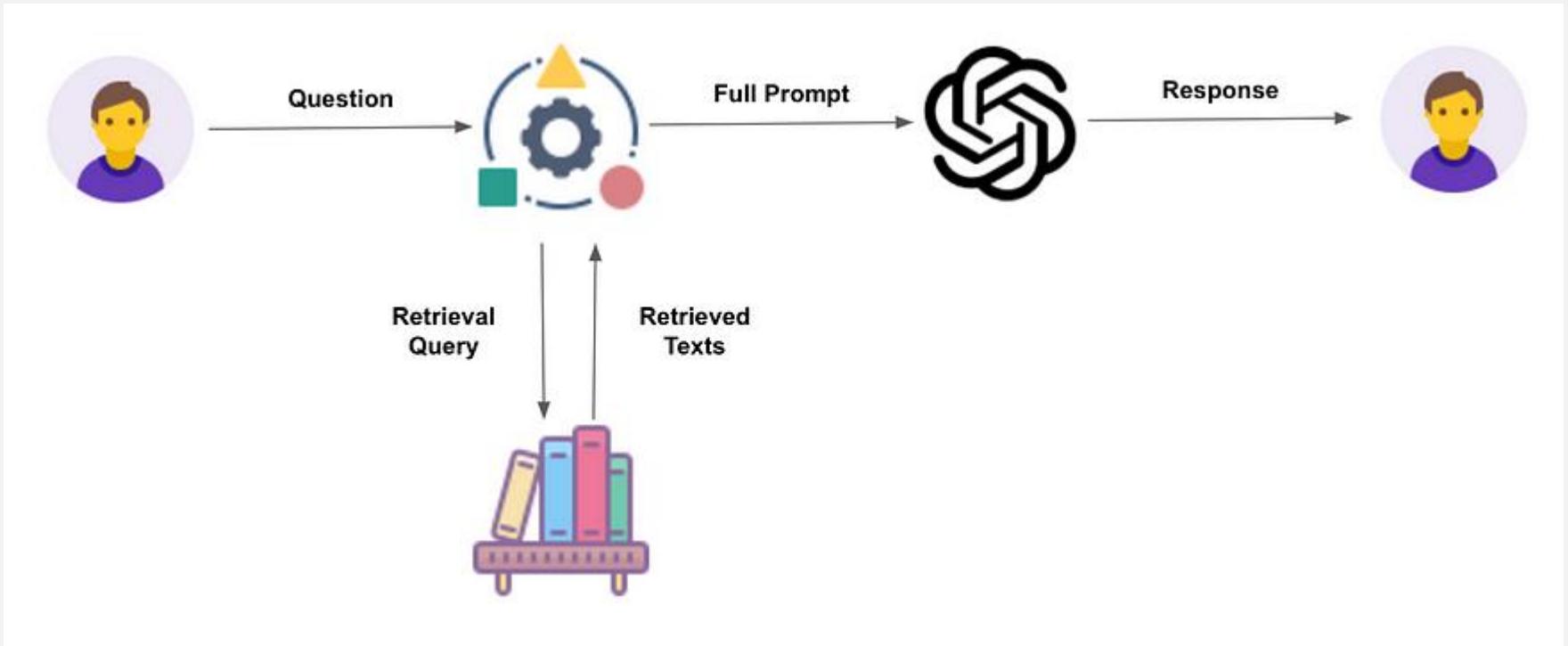
- Higher quality results than prompting
- Ability to train on more examples than can fit in a prompt
- Token savings due to shorter prompts
- Lower latency requests

RETRIEVAL AUGMENTED GENERATION (RAG)

Retrieval Augmented Generation

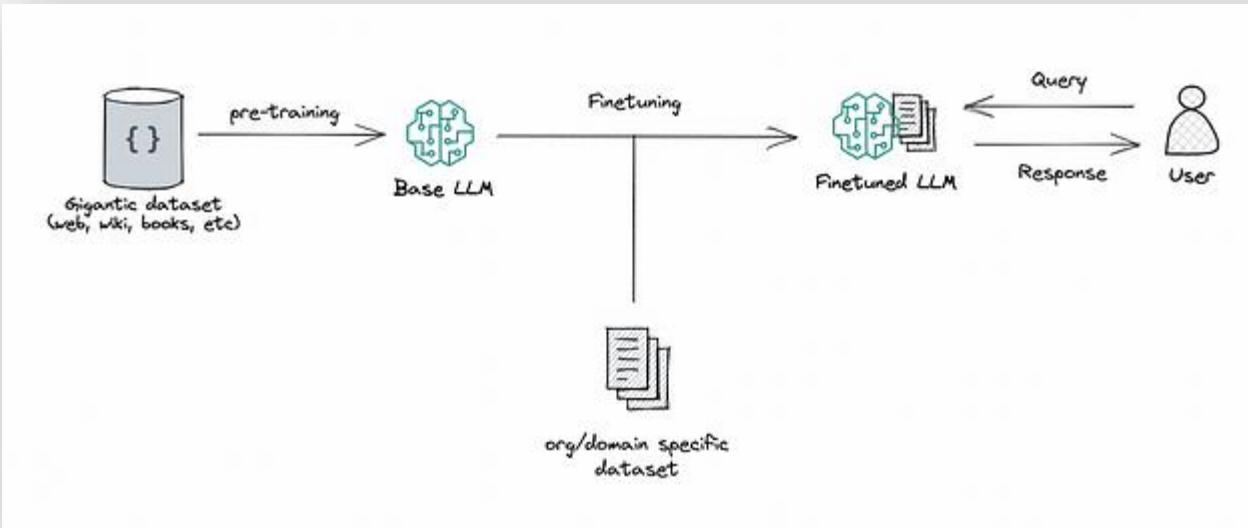
- Combines strengths of retrievers and generators for advanced question-answering.
- Queries an external knowledge source to pull relevant information.
- Increases the breadth and specificity of answers beyond base model's knowledge.
- Effective for tasks with vast and ever-expanding information domains.

Retrieval Augmented Generation

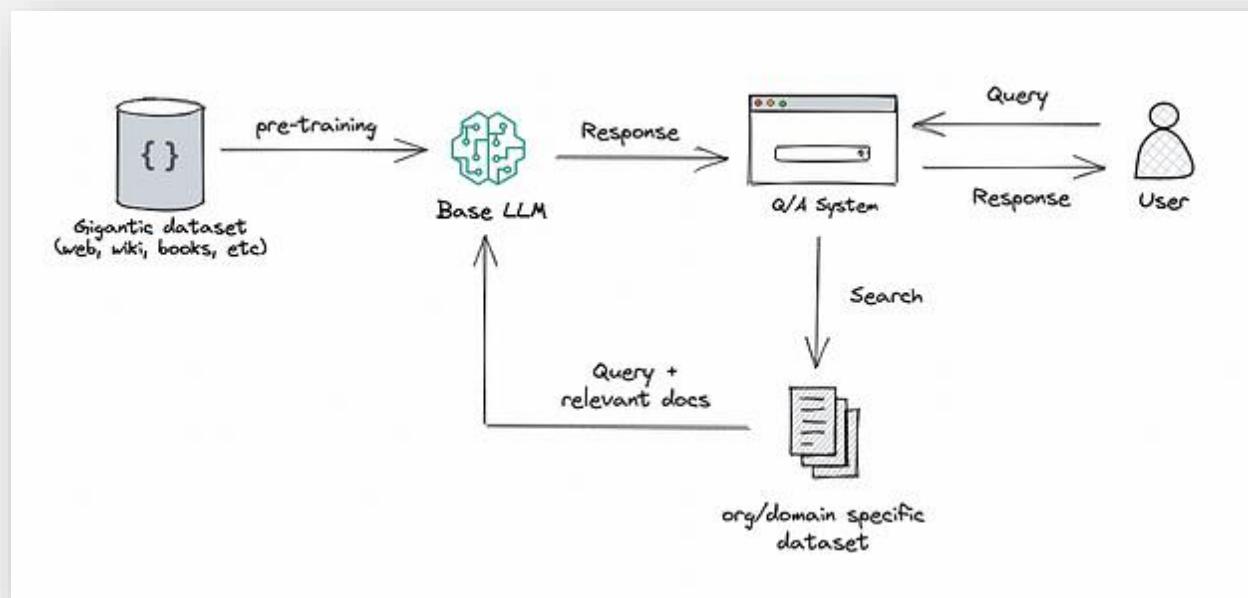


[Source](#)

Fine-tuning vs. RAG (Which one is which?)



Fine-tuning LLM



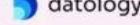
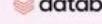
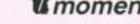
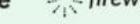
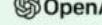
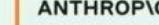
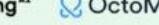
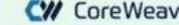
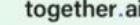
Retrieval Augmented Generation (RAG)

[Source](#)

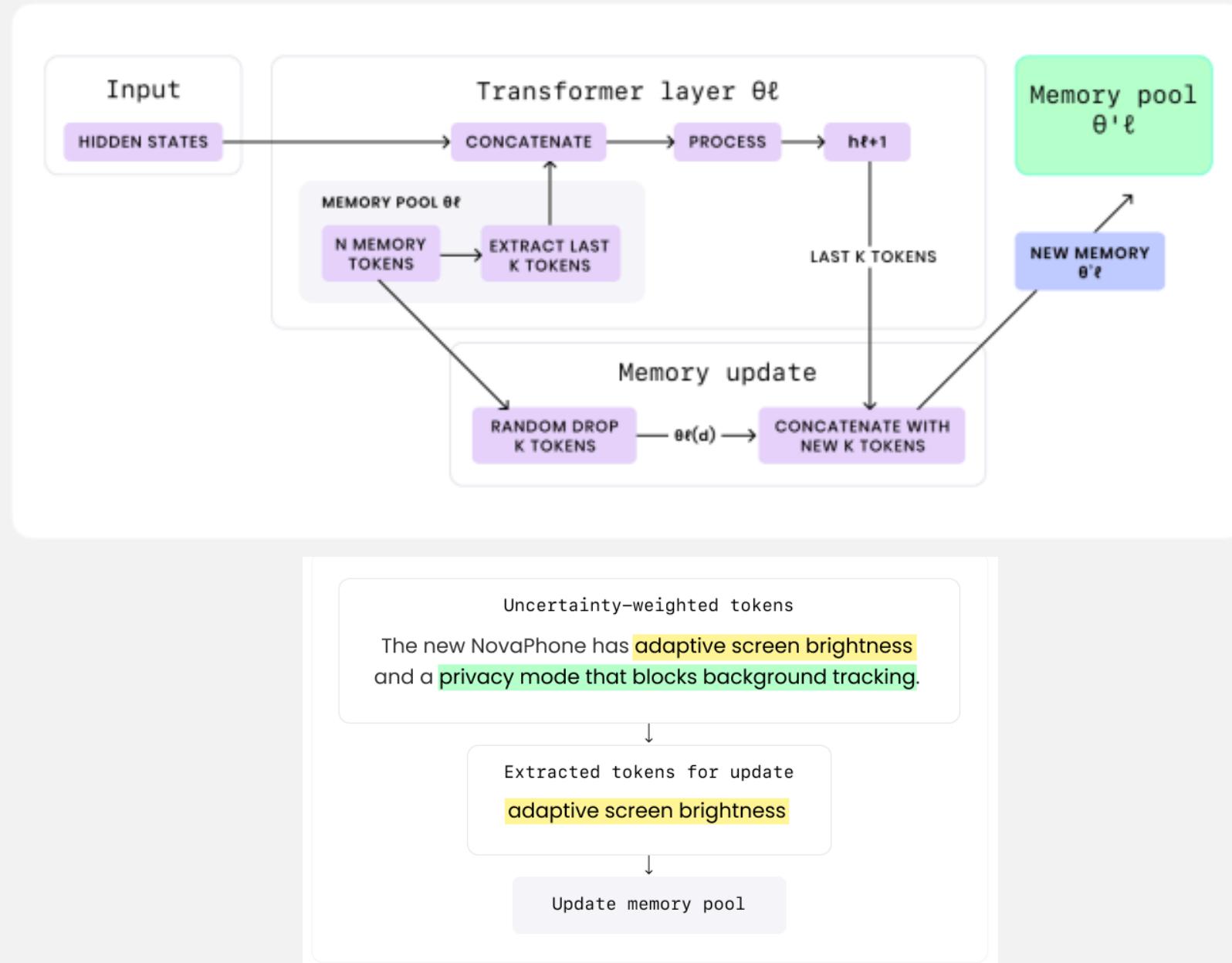
How to get the most out of LLMs



Modern AI Stack: The Emerging Building Blocks for GenAI

Layer 4: Observability	OBSERVABILITY, EVALUATION, SECURITY									
	 Helicone	 AgentOps	 Humanloop	 Credal.ai	 CALYPSO AI	 truera	 eppo	 BRAINTRUST	 Patronus AI	 LOGIO
Layer 3: Deployment	PROMPT MANAGEMENT ORCHESTRATION									
	 vellum	 LangSmith			 Martian		 orkes		 Radiant	
	AGENT TOOL FRAMEWORKS									
	 LangChain	 AutoGPT	 FIXIE		 LlamaIndex					
Layer 2: Data	DATA PRE-PROCESSING ETL + DATA PIPELINES									
	 gable	 datologyai	 Cleanlab	 STRUCTURE	 NOMIC	 Lexy	 Indexify			
	DATABASES (VECTOR, DB, METADATA STORE, CONTEXT CACHE)									
	 databricks	 upstash	 Pinecone	 NEON	 WarpStream	 momento				
Layer 1: Compute + Foundation	MODEL DEPLOYMENT + INFERENCE FINETUNING + RLHF									
	 baseten	 Modal	 Replicate	 clarifai	 Substrate	 fireworks.ai	 LAMINI	 Predibase	 arcee.ai	
	FOUNDATION MODELS TRAINING									
	 OpenAI	 ANTHROPIC	 MISTRAL AI_	contextual.ai	 Hugging Face	Llama 2	 Modular	 Lightning AI	 OctoML	
	GPU PROVIDERS									
	 aws	 Azure	 Google Cloud	 CoreWeave	 Lambda	 FOUNDRY	 together.ai			

Emerging Now: Self-evolving models



- Content Creation, Writing Assistants, SEO
- Content and Idea generation
- Personalized Virtual Assistants
- Question Answering Systems (customer service, e-commerce, etc.)
- Translation & Language Services
- Educational Tools
- Code generation and debugging, developer tools, etc.
- Research & Data Analysis
- Task Automation – A new way to use softwares?

\$1 = \$3.70

How real-world businesses are transforming with AI

Nov 12, 2024 | Alysa Taylor - Chief Marketing Officer, Commercial Cloud & AI



Explore more than
200
AI transformation stories



SUMMARY

- Generative AI
 - Text to Text
 - Text to Image
 - Text to Audio / Video
 - Text to Code
- Large Language Models
 - Prompt Engineering
 - Fine-tuning LLMs
 - Retrieval Augmented Generation (RAG)
- LLM Application and use-cases