IEEE OEB-LMAG “AI Hands-on Workshop”

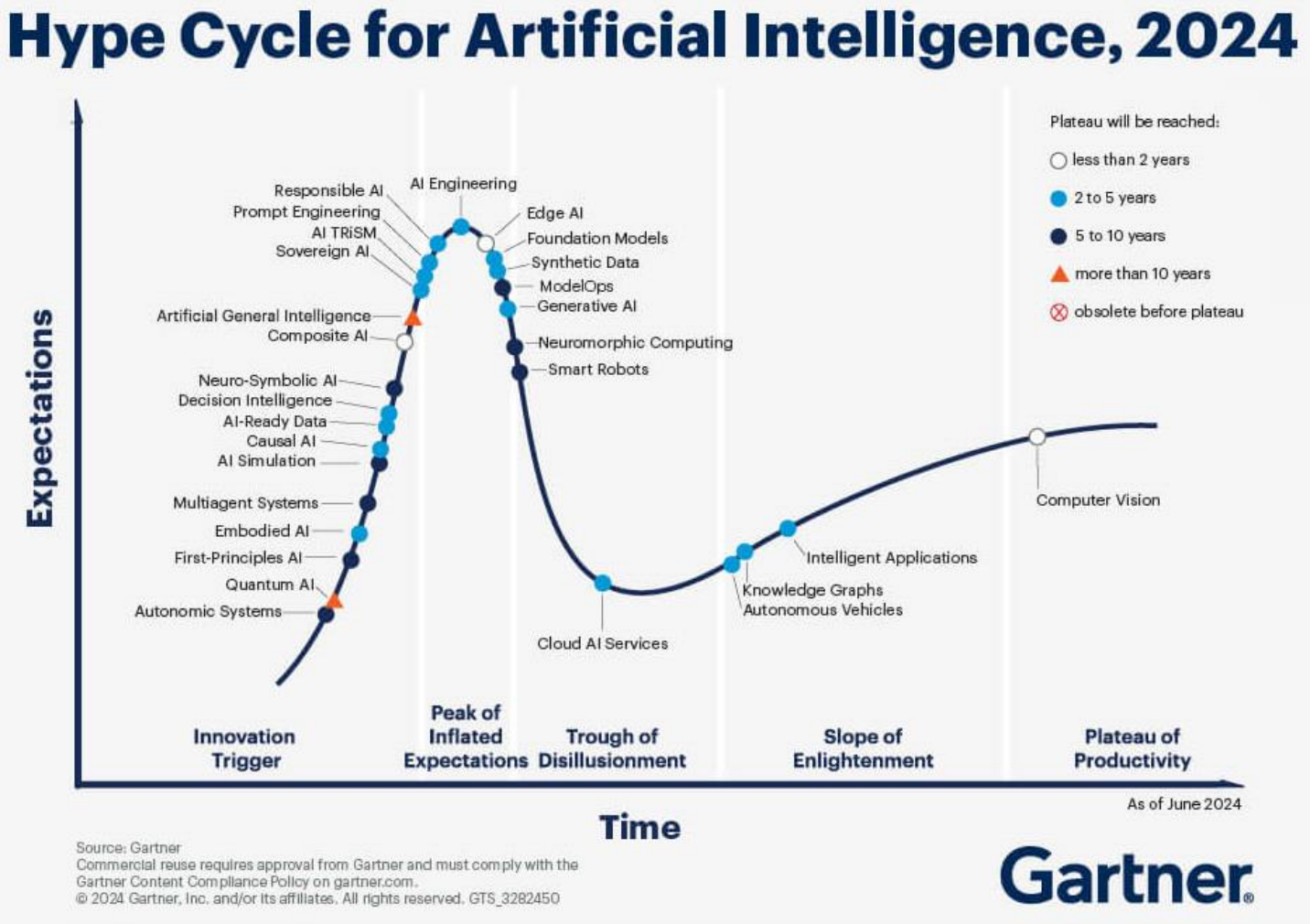
January 22, 2025 11:00AM - 2:00PM
Held at Beeb’s Sports Bar in Livermore, CA

Additional Resources:

[all files can be downloaded from: https://tinyurl.com/IEEE-OEBLM-AI]

The Gartner “Hype Cycle Chart” for AI (June 2024)

This shows Gartner’s assessment of the various AI Technologies, and how far away they are from reality!

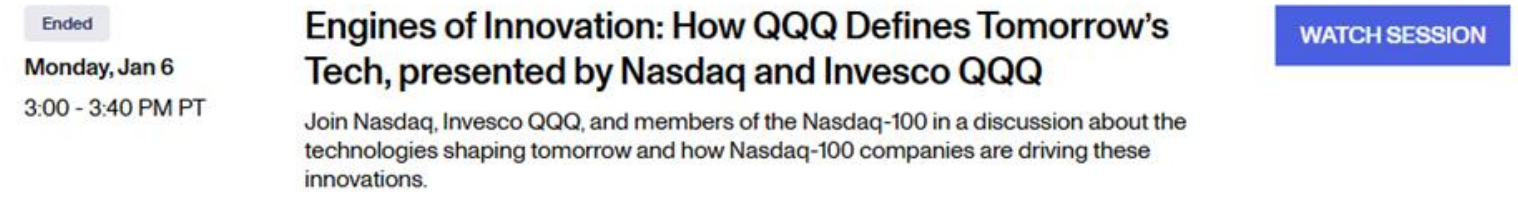


Presentations from 2025's Consumer Electronics Show

(CES, the largest tradeshow in the USA, held in early January each year)
https://live.ces.tech/ *(this site does work, but you have to explore it to find the sessions for each day – there are a lot there!)*

Here are two examples of these presentations:

The Engines of Innovation Session is interesting, as you get some insight into why AI is driving innovation at various companies!
https://www.ces.tech/videos/2025/january/engines-of-innovation-how-qqq-defines-tomorrow-s-tech-presented-by-nasdaq-and-invesco-qqq



This NVIDIA Keynote is a mind blowing 1.5 hour tour by NVIDIA's CEO of their latest AI chips and capabilities!
https://www.ces.tech/videos/2025/january/nvidia-keynote/

2025 Predictions: Enterprises, Researchers and Startups Home In on Humanoids, AI Agents as Generative AI Crosses the Chasm

NVIDIA experts across accelerated computing, data science and research predict multimodal models will speed industry innovation and efficiency.
https://blogs.nvidia.com/blog/generative-ai-predictions-2025-humanoids-agents/
Cliff Edwards | December 5, 2024

NVIDIA Releases NIM Microservices to Safeguard Applications for Agentic AI | NVIDIA Blog

The best AI productivity tools by category

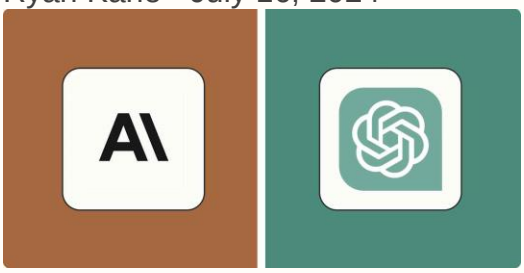
https://zapier.com/blog/best-ai-productivity-tools/

- Chatbots (ChatGPT, Claude, Meta AI)
- Search engines (Perplexity, Google AI Overviews, Arc Search)
- Content creation (Jasper, Anyword, Writer)
- Grammar checkers and rewording tools (Grammarly, Wordtune, ProWritingAid)
- Video creation and editing (Runway, Descript, Wondershare Filmora)
- Image generation (DALL·E 3, Midjourney, Ideogram)
- Social media management (FeedHive, Vista Social, Buffer)
- Voice and music generation (ElevenLabs, Suno, AIVA)
- Knowledge management and AI grounding (Mem, Notion AI Q&A, Personal AI)
- Task and project management (Asana, Any.do, BeeDone)
- Transcription and meeting assistants (Fireflies, Avoma, tl;dv)
- Scheduling (Reclaim, Clockwise, Motion)
- Email (Shortwave, Microsoft Copilot Pro for Outlook, Gemini for Gmail)
- Slide decks and presentations (Tome, Beautiful.ai, Slidesgo)
- Resume builders (Teal, Enhancv, Kickresume)
- Automation (Zapier)
- Other AI productivity tools

Claude vs. ChatGPT: What's the difference? [2024]

https://zapier.com/blog/claude-vs-chatgpt/

Ryan Kane · July 16, 2024



When OpenAI released the first iteration of ChatGPT in late 2022, it quickly became the fastest-growing app ever, amassing over one hundred million users in its first two months. Of all the

competing large language models (LLMs) ChatGPT has inspired—and there are many—its closest rival in terms of performance is Claude, which launched in 2023.

When I first compared them head-to-head in April 2024, Claude's Opus model held a slight edge over GPT-4. But in May 2024, ChatGPT closed the gap again by launching GPT-4o, a multimodal AI model; Claude quickly followed with the release of Claude 3.5 in June 2024.

Meet your new AI teammates

I've used ChatGPT and Claude regularly since each was released. And to compare these two AI juggernauts, I ran over a dozen tests to gauge their performance on different tasks, paying close attention to areas where GPT-4o and Claude 3.5 showed better—or worse—performance than their predecessors.

Here, I'll explain the strengths and limitations of Claude and ChatGPT, so you can decide which is best for you.

Note: OpenAI recently released GPT-4o mini—a smaller model that's faster and cheaper than 4o—along with o1—its newest series of models that's better at working through complex tasks. Because these models are still so new, this article focuses on comparing GPT-4o and Claude 3.5.

Claude vs. ChatGPT at a glance

Claude and ChatGPT are powered by similarly powerful LLMs and LMMs. They differ in some important ways, though: ChatGPT is more versatile, with features like image generation and internet access, while Claude offers cheaper API access and a larger context window (meaning it can process more data at once).

Here's a quick rundown of the differences between these two AI models:

	Claude	ChatGPT
Company	Anthropic	OpenAI
AI model	Claude 3.5 Sonnet Claude 3 Opus Claude 3 Haiku	GPT-4 GPT-4o GPT-4o mini
Context window	200,000 tokens (and up to 1,000,000 tokens for certain use cases)	128,000 tokens (GPT-4o)
Internet access	No	Yes
Image generation	No	Yes (DALL·E)
Supported languages	Officially, English, Japanese, Spanish, and French, but in my testing, Claude supported every language I tried (even less common ones like Azerbaijani)	95+ languages
Paid tier	$20/month for Claude Pro	$20/month for ChatGPT Plus

	Claude	ChatGPT
Team plans	$30/ user/month; includes Projects feature for collaboration	$30/user/month; includes workspace management features and shared custom GPTs
API pricing (for input)	- $15 per 1M input tokens and $75 per 1M output tokens (Claude 3 Opus) - $3 per 1M input tokens and $15 per 1M output tokens (Claude 3.5 Sonnet) - $0.25 per 1M input tokens and $1.25 per 1M output tokens (Claude 3 Haiku)	- $5 per 1M input tokens and $15 per 1M output tokens (GPT-4o) - $0.50 per 1M input tokens and $1.50 per 1M output tokens (GPT-3.5 Turbo) - $30 per 1M input tokens and $60 per 1M output tokens (GPT-4)

To compare the performance of one LLM to another, AI firms use benchmarks like standardized tests. OpenAI's benchmarking of GPT-4o shows impressive performances on LLM-specific tests like the MMLU, which measures undergraduate-level knowledge, and HumanEval, which measures coding ability. Meanwhile, Anthropic has published a head-to-head comparison of Claude, ChatGPT, Llama, and Gemini that shows its Claude 3.5 Sonnet model edging out GPT-4o on most tests.

	Claude 3.5 Sonnet	Claude 3 Opus	GPT-4o	Gemini 1.5 Pro	Llama-400b (early snapshot)
Graduate level reasoning *GPQA, Diamond*	59.4%* 0-shot CoT	50.4% 0-shot CoT	53.6% 0-shot CoT	—	—
Undergraduate level knowledge *MMLU*	88.7%** 5-shot 88.3% 0-shot CoT	86.8% 5-shot 85.7% 0-shot CoT	— 88.7% 0-shot CoT	85.9% 5-shot —	86.1% 5-shot —
Code *HumanEval*	92.0% 0-shot	84.9% 0-shot	90.2% 0-shot	84.1% 0-shot	84.1% 0-shot
Multilingual math *MGSM*	91.6% 0-shot CoT	90.7% 0-shot CoT	90.5% 0-shot CoT	87.5% 8-shot	—
Reasoning over text *DROP, F1 score*	87.1 3-shot	83.1 3-shot	83.4 3-shot	74.9 Variable shots	83.5 3-shot Pre-trained model
Mixed evaluations *BIG-Bench-Hard*	93.1% 3-shot CoT	86.8% 3-shot CoT	—	89.2% 3-shot CoT	85.3% 3-shot CoT Pre-trained model
Math problem-solving *MATH*	71.1% 0-shot CoT	60.1% 0-shot CoT	76.6% 0-shot CoT	67.7% 4-shot	57.8% 4-shot CoT
Grade school math *GSM8K*	96.4% 0-shot CoT	95.0% 0-shot CoT	—	90.8% 11-shot	94.1% 8-shot CoT

Image source: Anthropic

While these benchmarks are undoubtedly useful, some machine learning experts speculate that this kind of testing overstates the progress of LLMs. As new models are released, they may (perhaps accidentally) be trained on their own evaluation data. As a result, they get better and better at standardized tests—but when asked to figure out new variations of those same questions, they sometimes struggle.

To get a sense for how each model performs on common daily-use tasks, I devised my own comparisons. Here's a high-level overview of what I found.

Task	Winner	Observations
Creativity	Claude	Claude's default writing style is more human-sounding and less generic.
Proofreading and fact-checking	Claude	Both do a good job spotting errors, but Claude is a better editing partner because it presents mistakes and corrections more clearly.
Image processing	Tie	Neither Claude nor ChatGPT is 100% accurate at identifying objects in images, and both have issues with counting. As long as you don't need absolute precision, both models provide remarkable insights into uploaded images.
Logic and reasoning	ChatGPT	From math to physics to riddles, both LLMs perform capably. But GPT-4o is a more trustworthy partner than Claude 3.5 for complex equations.
Emotion and ethics	Tie	Earlier iterations of Claude felt more "human" and empathetic, but Claude 3.5 and GPT-4o take an equally robotic approach.
Analysis and summaries	ChatGPT	While Claude 3.5 officially has a larger context window, in my tests, GPT-4o went far beyond its stated limits and was able to process much larger documents than Claude. GPT-4o also provided more accurate analysis.
Coding	Claude	Claude 3.5 is a more capable coding assistant, and its Artifacts feature provides a handy (and interactive) user interface that lets you immediately see the results of your code.
Integrations	ChatGPT	From its native DALL·E image generation tool to its internet access and third-party GPTs, ChatGPT's capabilities go beyond Claude's standard offering.

Read on to learn more about how Claude and ChatGPT performed on each task.

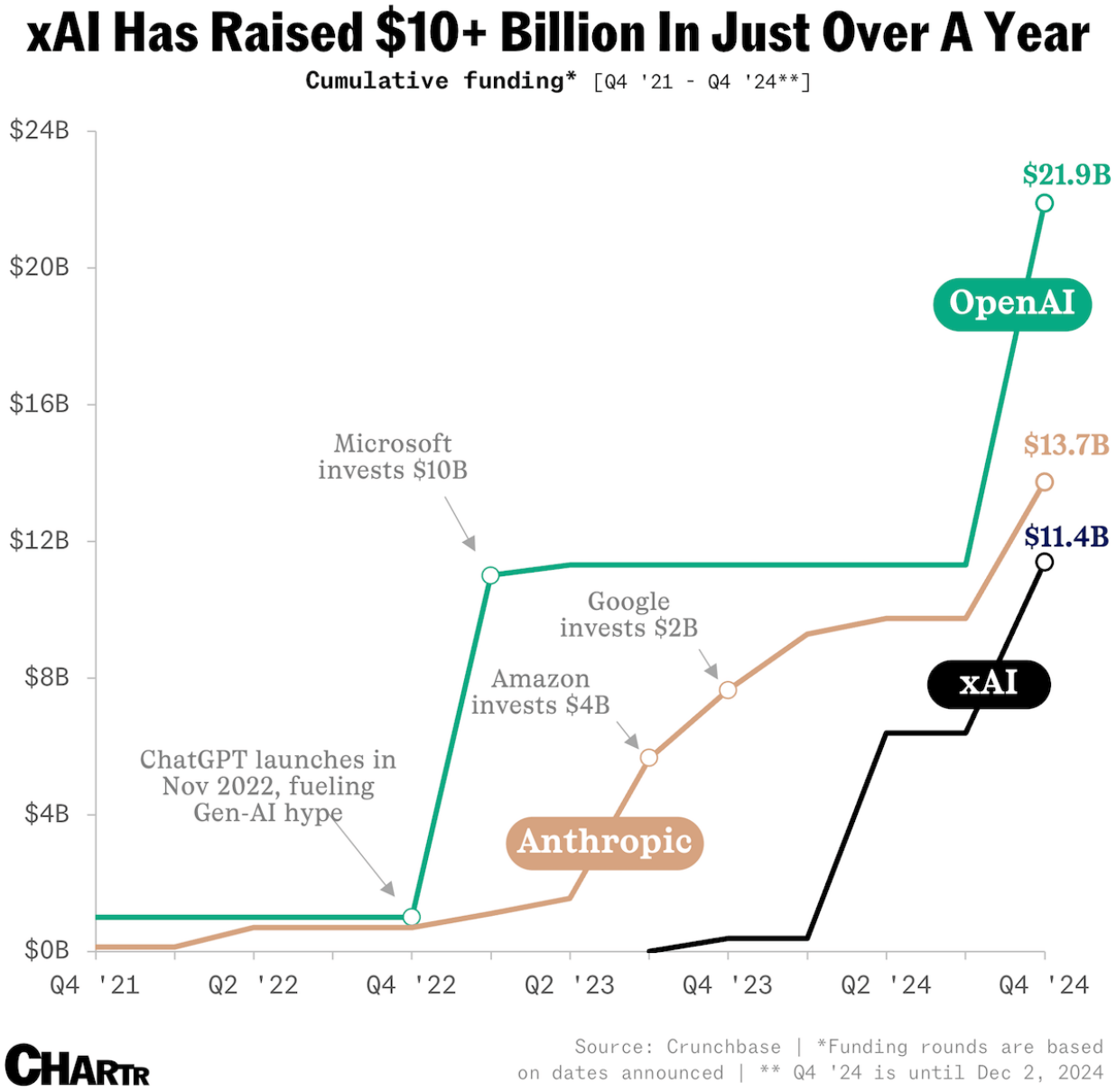
Elon Musk’s xAI has raised more than $11 billion in record time

Elon Musk's feud with OpenAI — the nonprofit he co-founded — has escalated, again

https://sherwood.news/tech/elon-musks-xai-raised-usd11-billion-in-record-time/

Last Friday, Musk filed an injunction to halt OpenAI’s for-profit transition, accusing it of orchestrating a “**group boycott**” that blocked funding for his own AI venture, xAI. In October, the Financial Times **reported** that OpenAI had discouraged investors from backing rival AI startups during its latest funding round.

But, even if Sam Altman and co. have been forcing investors to commit to monogamy and invest only in OpenAI, you wouldn’t exactly say xAI has struggled to find backers.



In just 16 months since its July 2023 launch, xAI has raised **~$11 billion** — a milestone that took OpenAI around eight years and the Amazon-backed Anthropic nearly four years to achieve. Indeed, xAI’s latest funding

round catapulted its valuation to **$50 billion**, **according** to the Wall Street Journal. That surpasses Anthropic's $19 billion **valuation** and the valuations of public heavyweights like **Ford** ($43 billion), **Kroger** ($43 billion), and **Lululemon** ($42 billion). It's also more than the $44 billion Musk **paid** for X just two years ago.
Why the scramble for cash?
In 2024, if you want to compete in AI, you need to be willing to pour billions into physical *stuff* — AI chips and data centers: xAI's latest $5 billion will partially **fund** the purchase of 100,000 Nvidia chips for its recently completed data center in Memphis. Meanwhile, Anthropic is on track to **build** one of the world's largest AI supercomputers, and OpenAI is **expanding** its footprint across the US Midwest and Southwest.
According to McKinsey & Company's October **research**, demand for AI-specific data centers is projected to grow 33% annually through 2030, and could eventually account for 70% of global data center demand. TL;DR: AI is an expensive game, and xAI is leaning hard on Musk's name to compete.

Large Reasoning Models by AI Maker Space

https://www.canva.com/design/DAGcR9XBWgk/9SW-O0ckDoQ67yggqWu9ag/view

By The End of the Session:

- Understand **reasoning** in LLMs
- Understand **Chain-of-Thought** and how the idea applies to reasoning **models like o1**
- Learn about the *research that led up to o1* and *what's been going on to take reasoning to the next level*; where does this lead?

Ndea's AI Breakthrough: Learning Like Humans, Smarter Than Ever

François Chollet, the mastermind behind the Keras AI framework, has co-founded **Ndea**, a pioneering AI lab, after leaving Google. Partnering with Mike Knoop of Zapier, Chollet aims to combine deep learning with program synthesis to create AI that learns as efficiently as humans. The lab's unique approach seeks to eliminate bottlenecks in AI development, pushing the boundaries of artificial intelligence.

Top AI Shops Fail Transparency Test

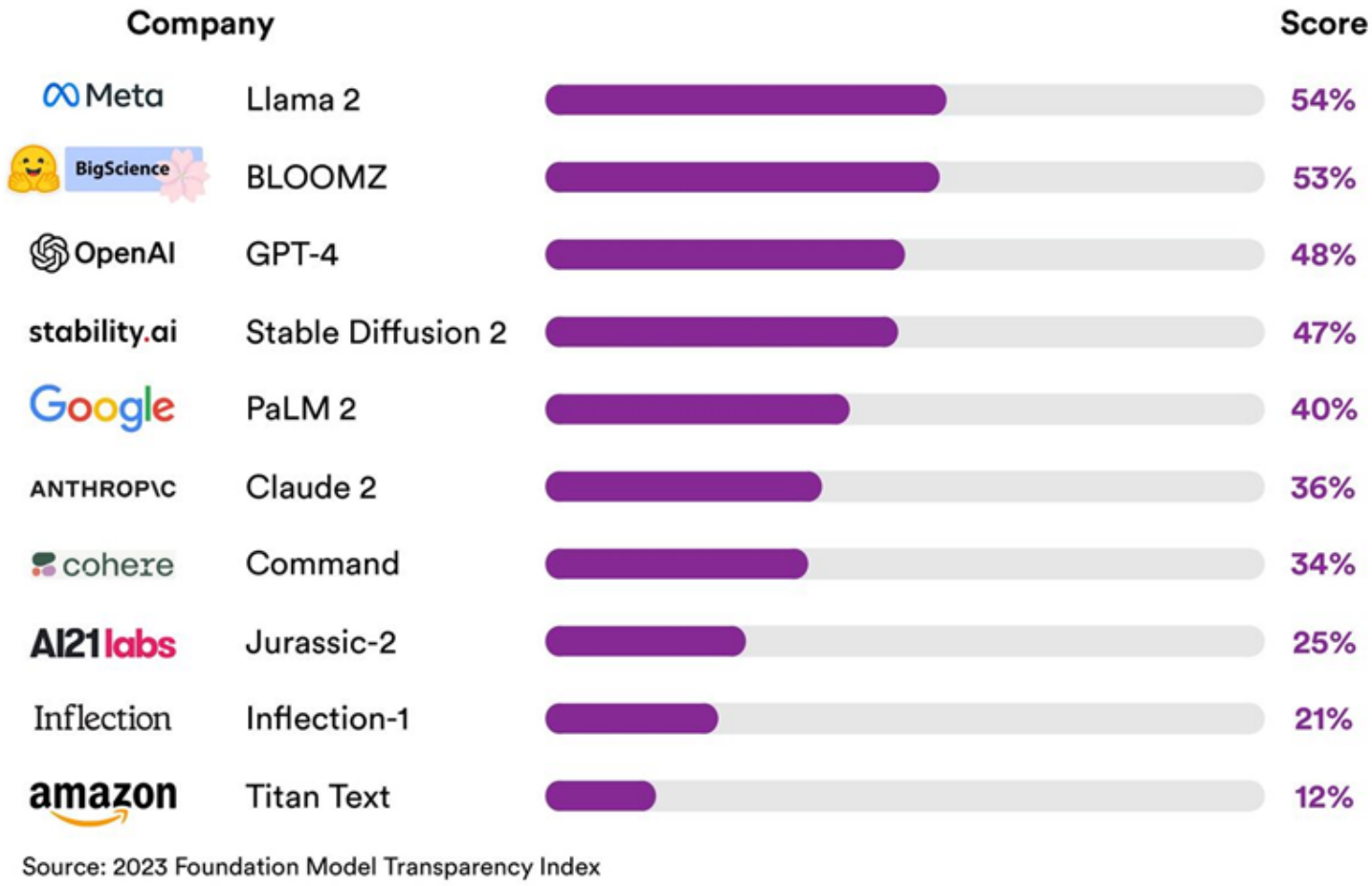
Stanford transparency index rates Meta, OpenAI, and others on 100 indicators

https://spectrum.ieee.org/ai-ethics

Eliza Strickland | 22 Oct 2023 4 min read

Eliza Strickland is a senior editor at IEEE Spectrum covering AI and biomedical engineering

foundation models large language models artificial intelligence openai meta ai ethics transparency



These 10 large "foundation models" graded by a new AI transparency index all had failing scores. Stanford Center for Research on Foundation Models

In July and September, 15 of the biggest AI companies signed on to the White House's voluntary commitments to manage the risks posed by AI. Among those commitments was a promise to be more transparent: to share information "across the industry and with governments, civil society, and academia," and to publicly report their AI systems' capabilities and limitations. Which all sounds great in theory, but what does it mean in practice? What exactly is transparency when it comes to these AI companies' massive and powerful models? Thanks to a report spearheaded by Stanford's Center for Research on Foundation Models (CRFM), we now have answers to those questions. The foundation models they're interested in are general-purpose creations like OpenAI's GPT-4 and Google's PaLM 2, which are trained on a huge amount of data and can be adapted for many different applications. The Foundation Model Transparency Index graded 10 of the biggest such models on 100 different metrics of transparency.

The highest total score goes to Meta's Llama 2, with 54 out of 100.

They didn't do so well. The highest total score goes to Meta's Llama 2, with 54 out of 100. In school, that'd be considered a failing grade. "No major foundation model developer is close to providing adequate transparency," the researchers wrote in a blog post, "revealing a fundamental lack of transparency in the AI industry."

Rishi Bommasani, a PhD candidate at Stanford's CRFM and one of the project leads, says the index is an effort to combat a troubling trend of the past few years. "As the impact goes up, the transparency of these models and companies goes down," he says. Most notably, when OpenAI versioned-up from GPT-3 to GPT-4, the company wrote that it had made the decision to withhold all information about "architecture (including model size), hardware, training compute, dataset construction, [and] training method."

The 100 metrics of transparency (listed in full in the blog post) include upstream factors relating to training, information about the model's properties and function, and downstream factors regarding the model's distribution and use. "It is not sufficient, as many governments have asked, for an organization to be transparent when it releases the model," says Kevin Klyman, a research assistant at Stanford's CRFM and a coauthor of the report. "It also has to be transparent about the resources that go into that model, and the evaluations of the capabilities of that model, and what happens after the release."

To grade the models on the 100 indicators, the researchers searched the publicly available data, giving the models a 1 or 0 on each indicator according to predetermined thresholds. Then they followed up with the 10 companies to see if they wanted to contest any of the scores. "In a few cases, there was some info we had missed," says Bommasani.

Spectrum contacted representatives from a range of companies featured in this index; none of them had replied to requests for comment as of our deadline.

> *In July and September, 15 of the biggest AI companies signed on to the White House's voluntary commitments to manage the risks posed by AI. Among those commitments was a promise to be more transparent: to share information "across the industry and with governments, civil society, and academia," and to publicly report their AI systems' capabilities and limitations. Which all sounds great in theory, but what does it mean in practice? What exactly is transparency when it comes to these AI companies' massive and powerful models?*
>
> *Thanks to a report spearheaded by Stanford's Center for Research on Foundation Models (CRFM), we now have answers to those questions. The foundation models they're interested in are general-purpose creations like OpenAI's GPT-4 and Google's PaLM 2, which are trained on a huge amount of data and can be adapted for many different applications. The Foundation Model Transparency Index graded 10 of the biggest such models on 100 different metrics of transparency.*
>
> *The highest total score goes to Meta's Llama 2, with 54 out of 100.*
>
> *They didn't do so well. The highest total score goes to Meta's Llama 2, with 54 out of 100. In school, that'd be considered a failing grade. "No major foundation model developer is close to providing adequate transparency," the researchers wrote in a blog post, "revealing a fundamental lack of transparency in the AI industry."*
>
> *Rishi Bommasani, a PhD candidate at Stanford's CRFM and one of the project leads, says the index is an effort to combat a troubling trend of the past few years. "As the impact goes up, the transparency of these models and companies goes down," he says. Most notably, when OpenAI versioned-up from GPT-3 to GPT-4, the company wrote that it had made the decision to withhold all information about "architecture (including model size), hardware, training compute, dataset construction, [and] training method."*
>
> *The 100 metrics of transparency (listed in full in the blog post) include upstream factors relating to training, information about the model's properties and function, and downstream factors regarding the model's distribution and use. "It is not sufficient, as many governments have asked, for an organization to be transparent when it releases the model," says Kevin Klyman, a research assistant at Stanford's*

CRFM and a coauthor of the report. "It also has to be transparent about the resources that go into that model, and the evaluations of the capabilities of that model, and what happens after the release."

To grade the models on the 100 indicators, the researchers searched the publicly available data, giving the models a 1 or 0 on each indicator according to predetermined thresholds. Then they followed up with the 10 companies to see if they wanted to contest any of the scores. "In a few cases, there was some info we had missed," says Bommasani.

Spectrum contacted representatives from a range of companies featured in this index; none of them had replied to requests for comment as of our deadline.

"Labor in AI is a habitually opaque topic. And here it's very opaque, even beyond the norms we've seen in other areas."
—Rishi Bommasani, Stanford

The provenance of training data for foundation models has become a hot topic, with several lawsuits alleging that AI companies illegally included authors' copyrighted material in their training data sets. And perhaps unsurprisingly, the transparency index showed that most companies have not been forthcoming about their data. The model Bloomz from the developer Hugging Face got the highest score in this particular category, with 60 percent; none of the other models scored above 40 percent, and several got a zero.

Dimensions of Transparency	Meta LLaMA 2	BigScience BLOOMZ	OpenAI GPT-4	stability.ai Stable Diffusion 2	ANTHROPIC Claude 2	Google PaLM 2	cohere Command	AI21labs Jurassic 2	Inflection Inflection-1	amazon Titan
Data	40%	60%	20%	40%	0%	20%	20%	0%	0%	0%
Labor	29%	71%	14%	14%	29%	0%	0%	0%	0%	0%
Compute	57%	14%	14%	57%	0%	0%	14%	0%	0%	0%
Methods	75%	100%	50%	100%	75%	75%	0%	0%	0%	0%
Model Basics	100%	100%	50%	83%	67%	67%	50%	50%	50%	33%
Model Access	67%	100%	67%	100%	33%	33%	33%	33%	0%	33%
Capabilities	60%	80%	100%	40%	80%	80%	60%	60%	60%	20%
Risks	57%	0%	57%	14%	29%	29%	29%	29%	0%	0%
Mitigations	33%	0%	33%	0%	33%	33%	0%	0%	0%	0%
Distribution	71%	71%	57%	71%	57%	57%	57%	43%	43%	29%
Usage Policy	40%	20%	80%	40%	80%	60%	40%	20%	80%	20%
Feedback	33%	33%	33%	33%	33%	33%	33%	33%	33%	0%
Impact	14%	14%	14%	14%	0%	0%	14%	14%	14%	0%

This heatmap chart shows how the 10 models were scored on 13 categories of indicators. The heatmap shows how the 10 models did on categories ranging from data to impact. Stanford Center for Research on Foundation Models