State of AI Report

October 12, 2023

Nathan Benaich
Air Street Capital

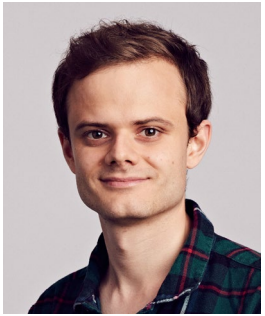
About the authors

Nathan Benaich

Nathan is the General Partner of **Air Street Capital**, a venture capital firm investing in AI-first technology and life science companies. He founded RAAIS and London.AI (AI community for industry and research), the RAAIS Foundation (funding open-source AI projects), and Spinout.fyi (improving university spinout creation). He studied biology at Williams College and earned a PhD from Cambridge in cancer research.

State of AI Report 2023 team

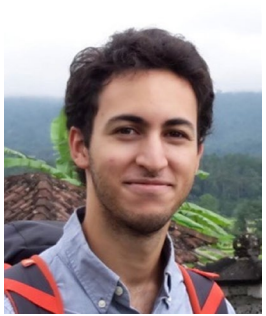
Alex Chalmers



Platform Lead

Alex is Platform Lead at **Air Street Capital**. Alex was previously Associate Director at Milltown Partners where he advised leading technology companies including AI labs.

Othmane Sebbouh



Venture Fellow

Othmane is a Venture Fellow at **Air Street Capital** and ML PhD student at ENS Paris, CREST-ENSAE and CNRS. He holds an MsC in management from ESSEC Business School and a Master in Applied Mathematics from ENSAE and Ecole Polytechnique.

Corina Gurau

Venture Fellow

Corina is a Venture Fellow at **Air Street Capital**. Corina was previously an Applied Scientist at autonomous driving company, Wayve. She holds a PhD in AI from the University of Oxford.

Artificial intelligence (AI) is a multidisciplinary field of science and engineering whose goal is to create intelligent machines.

We believe that AI will be a force multiplier on technological progress in our increasingly digital, data-driven world. This is because everything around us today, ranging from culture to consumer products, is a product of intelligence.

The State of AI Report is now in its sixth year. Consider this report as a compilation of the most interesting things we've seen with a goal of triggering an informed conversation about the state of AI and its implication for the future.

We consider the following key dimensions in our report:

- **Research**: Technology breakthroughs and their capabilities.
- **Industry**: Areas of commercial application for AI and its business impact.
- **Politics**: Regulation of AI, its economic implications and the evolving geopolitics of AI.
- **Safety**: Identifying and mitigating catastrophic risks that highly-capable future AI systems could pose to us.
- **Predictions**: What we believe will happen in the next 12 months and a 2022 performance review to keep us honest.

Produced by **Nathan Benaich and Air Street Capital team**

Definitions

Artificial intelligence (AI): a broad discipline with the goal of creating intelligent machines, as opposed to the natural intelligence that is demonstrated by humans and animals.

Artificial general intelligence (AGI): a term used to describe future machines that could match and then exceed the full range of human cognitive ability across all economically valuable tasks.

AI Agent: an AI-powered system that can take actions in an environment. For example, an LLM that has access to a suite of tools and has to decide which one to use in order to accomplish a task that it has been prompted to do.

AI Safety: a field that studies and attempts to mitigate the risks (minor to catastrophic) which future AI could pose to humanity.

Computer vision (CV): the ability of a program to analyse and understand images and video.

Deep learning (DL): an approach to AI inspired by how neurons in the brain recognise complex patterns in data. The “deep” refers to the many layers of neurons in today’s models that help to learn rich representations of data to achieve better performance gains.

Diffusion: An algorithm that iteratively denoises an artificially corrupted signal in order to generate new, high-quality outputs. In recent years it has been at the forefront of image generation.

Generative AI: A family of AI systems that are capable of generating new content (e.g. text, images, audio, or 3D assets) based on 'prompts'.

Graphics Processing Unit (GPU): a semiconductor processing unit that enables a large number calculations to be computed in parallel. Historically this was required for rendering computer graphics. Since 2012 GPUs have adapted for training DL models, which also require a large number of parallel calculations.

Definitions

(Large) Language model (LM, LLM): a model trained on vast amounts of (often) textual data to predict the next word in a self-supervised manner. The term “LLM” is used to designate multi-billion parameter LMs, but this is a moving definition.

Machine learning (ML): a subset of AI that often uses statistical techniques to give machines the ability to "learn" from data without being explicitly given the instructions for how to do so. This process is known as “training” a “model” using a learning “algorithm” that progressively improves model performance on a specific task.

Model: a ML algorithm trained on data and used to make predictions.

Natural language processing (NLP): the ability of a program to understand human language as it is spoken and written.

Prompt: a user input often written in natural language that is used to instruct an LLM to generate something or take action.

Reinforcement learning (RL): an area of ML in which software agents learn goal-oriented behavior by trial and error in an environment that provides rewards or penalties in response to their actions (called a “policy”) towards achieving that goal.

Self-supervised learning (SSL): a form of unsupervised learning, where manually labeled data is not needed. Raw data is instead modified in an automated way to create artificial labels to learn from. An example of SSL is learning to complete text by masking random words in a sentence and trying to predict the missing ones.

Transformer: a model architecture at the core of most state of the art (SOTA) ML research. It is composed of multiple “attention” layers which learn which parts of the input data are the most important for a given task. Transformers started in NLP (specifically machine translation) and subsequently were expanded into computer vision, audio, and other modalities.

Definitions

Model type legend

In the rest of the slides, icons in the top right corner indicate input and output modalities for the model.

Input/Output types:

: Text
: Image
</> : Code
: Software tool use (text, code generation & execution)
: Video
: Music
: 3D
: Robot state

Model types:

→ : LLMs
+ → : Multimodal LLMs
+ + → : Multimodal LLMs for Robotics
→ </> : Text to Code
→ : Text to Software tool use
→ : Text to Image
→ : Text to Video
→ : Text to Music
→ : Image to 3D
→ : Text to 3D

Executive Summary

Research

- GPT-4 lands and demonstrates a capabilities chasm between proprietary and next-best open source alternatives, while also validating the power of reinforcement learning from human feedback.
- Efforts grow to clone or beat proprietary model performance with smaller models, better datasets, longer context…powered by LLaMa-1/2.
- It's unclear how long human-generated data can sustain AI scaling trends (some estimate that data will be exhausted by LLMs by 2025) and what the effects of adding synthetic data are. Videos and data locked up in enterprises are likely up next.
- LLMs and diffusion models continue to offer gifts to the life science community by producing new breakthroughs for molecular biology and drug discovery.
- Multimodality becomes the new frontier and excitement around agents of all flavors grows substantially.

Industry

- NVIDIA rips into the $1T market cap club with voracious demand for its GPUs from nation states, startups, big tech and researchers alike.
- Export controls rate limit advanced chip sales to China, but major chip vendors create export control-proof alternatives.
- Led by ChatGPT, GenAI apps have a breakout year across image, video, coding, voice or CoPilots for everyone, driving $18B of VC and corporate investments.

Politics

- The world has divided into clear regulatory camps, but progress on global governance remains slower. The largest AI labs are stepping in to fill the vacuum.
- The chip wars continue unabated, with the US mobilising its allies, and the Chinese response remaining patchy.
- AI is forecast to affect a series of sensitive areas, including elections and employment, but we're yet to see a significant effect.

Safety

- The existential risk debate has reached the mainstream for the first time and intensified significantly.
- Many high-performing models are easy to 'jailbreak'. To remedy RLHF challenges, researchers are exploring alternatives, e.g. self-alignment and pretraining with human preferences.
- As capabilities advance, it's becoming increasingly hard to evaluate SOTA models consistently. Vibes won't suffice.

Scorecard: Reviewing our predictions from 2022

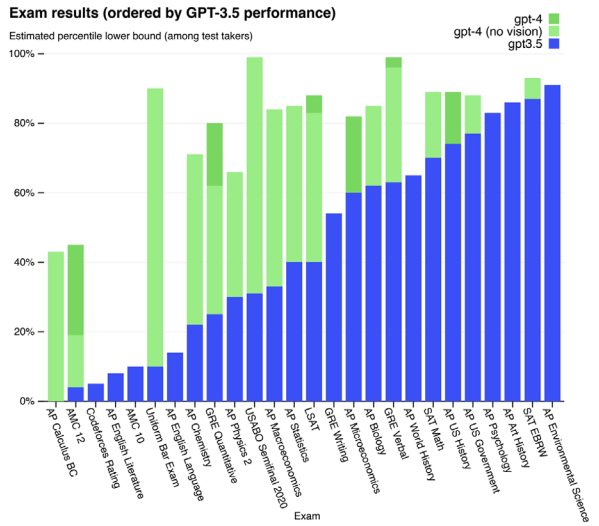
Our 2022 Prediction		Evidence
A 10B parameter multimodal RL model is trained by DeepMind 10x larger than Gato.	NO	So far there has been no publicly disclosed research along these lines.
NVIDIA announces a strategic relationship with an AGI focused organisation.	~	Instead of one relationship, NVIDIA has ramped its investment activities across many AGI focused organisations including Cohere, Inflection AI, and Adept.
A SOTA LM is trained on 10x more data points than Chinchilla, proving data-set scaling vs. parameter scaling	YES	We don't know for sure, but GPT-4 was reportedly trained on 13T tokens vs. Chinchilla's 1.4T. Meta's Llama-2 was trained on 2T tokens.
Generative audio tools emerge that attract over 100,000 developers by September 2023.	YES	Both ElevenLabs and Resemble.ai claim over 1 million users each since launch.
GAFAM invests >$1B into an AGI or open source AI company (e.g. OpenAI).	YES	Microsoft invested a further $10B into OpenAI in Jan. 2023.
Reality bites for semiconductor startups in the face of NVIDIA's dominance and a high profile start-up is shut down or acquired for <50% of its most recent valuation.	NO	There have been markdowns, but no major shutdowns or depressed acquisitions.
A proposal to regulate AGI Labs like Biosafety Labs (BSL) gets backing from an elected UK, US or EU politician.	NO	Calls for regulation have significantly heightened, but no backing for BSL yet.
>$100M is invested in dedicated AI Alignment organisations in the next year as we become aware of the risk we are facing by letting AI capabilities run ahead of safety.	YES	Anthropic, an AI research and safety company, raised up to $4B in Sept 2023.
A major user generated content site (e.g. Reddit) negotiates a commercial settlement with a start-up producing AI models (e.g. OpenAI) for training on their corpus of user generated content.	YES	OpenAI has secured a 6-year license for access to additional Shutterstock training data (image, video and music libraries and associated metadata).

Section 1: Research

GPT-4 is out and it crushes every other LLM, and many humans

GPT-4 is OpenAI's latest Large Language Model. In contrast with text-only GPT-3 and follow-ups, GPT-4 is multimodal: it was trained on both text and images; it can among other capabilities generate text based on images. At 8,192 tokens when it was released, it had already exceeded the previous-best GPT-3.5 in possible input size. It is, of course, trained using RLHF. Equipped with these advances, GPT-4 is, as of the release of this report, the uncontested most generally capable AI model.

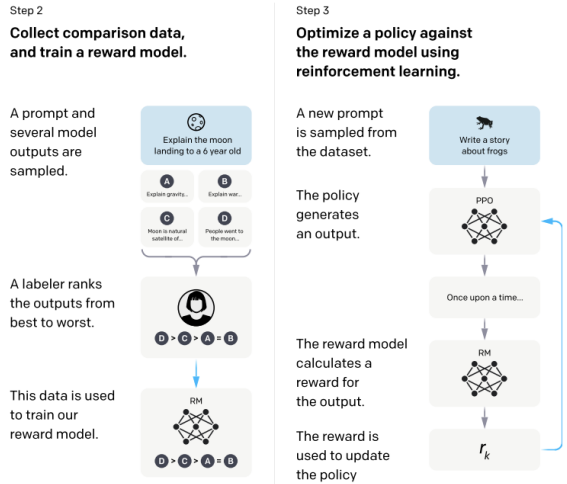
- OpenAI did a comprehensive evaluation of GPT-4 not only on classical NLP benchmarks, but also on exams designed to evaluate humans (e.g. Bar exam, GRE, Leetcode).
- GPT-4 is the best model across the board. It solves some tasks that GPT-3.5 was unable to, like the Uniform Bar Exam where GPT-4 scores 90% compared to 10% for GPT-3.5. On most tasks, the added vision component had only a minor impact, but it helped tremendously on others.
- OpenAI reports that although GPT-4 still suffers from hallucinations, it is factually correct 40% more often than the previous-best ChatGPT model on an adversarial truthfulness dataset (generated to fool AI models).



Fueled by ChatGPT's success, RLHF becomes MVP

In last year's Safety section (Slide 100), we highlighted how Reinforcement Learning from Human Feedback (RLHF) – used in InstructGPT – helped make OpenAI's models safer and more helpful for users. Despite a few hiccups, ChatGPT's success proved the technique's viability at a massive scale.

- *"RLHF involves humans ranking language model outputs sampled for a given input, using these rankings to learn a reward model of human preferences, and then using this as a reward signal to finetune the language model with using RL."* In its modern form, it dates back to 2017, when OpenAI and DeepMind researchers applied it to incorporate human feedback in training agents on Atari games and to other RL applications.
- RLHF is now central to the success of state of the art LLMs, especially those designed for chat applications. These include Anthropic's Claude, Google's Bard, Meta's LLaMa-2-chat, and, of course, OpenAI's ChatGPT.
- RLHF requires hiring humans to evaluate and rank model outputs, and then models their preferences. This makes this technique hard, expensive, and biased[1]. This motivated researchers to look for alternatives.



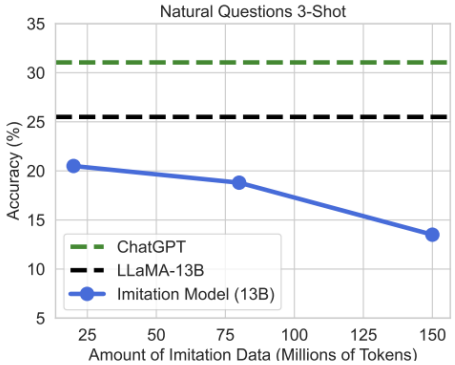
Typical steps of RLHF, which follow an initial step of supervised fine-tuning of a pre-trained language model, e.g. GPT-3.

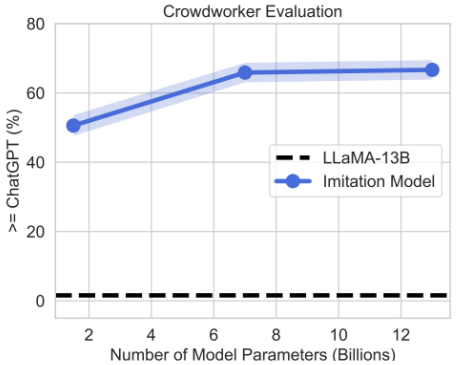
[1] *We will cover other issues of RLHF in the Safety section.*

The false promise of imitating proprietary LLMs, or how RLHF is still king

Berkeley researchers show that fine-tuning small LLMs on the outputs of larger, more capable LLMs results in models which are stylistically impressive but which often produce inaccurate text.

- The researchers examine a range of pretrained LLMs of different sizes and pre-trained on a varying amount of data. They show that at a fixed model size, using more imitation data actually hurts the quality of the output. In turn, larger models benefit from using imitation data.
- By using model size as a proxy for quality, the authors argue that more attention should be paid to better pre-training rather than fine-tuning on more imitation data.
- In the near future, RLHF seems here to stay. After careful ablation studies, Meta researchers concluded in their LLaMa-2 paper: *"We posit that the superior writing abilities of LLMs, as manifested in surpassing human annotators in certain tasks, are fundamentally driven by RLHF"*.





Even so, researchers rush to find scalable alternatives to RLHF

In the wake of ChatGPT, many labs set out to answer the question: Can we create models as capable and safe as OpenAI's LLMs, but that drastically reduce human supervision?

- Anthropic proposed RL from AI feedback, which we cover in the safety section.
- Other approaches entirely do a way with reinforcement learning. In *Less is More for Alignment* (LIMA), Meta argues for using a few (1,000 in their paper) very carefully curated prompts and responses. According to human evaluations of model outputs, LIMA is competitive with GPT-4 in 43% of cases.
- In *LLMs can self-improve*, Google researchers showed that LLMs can improve by training on their own outputs. In a similar vein, Self-Instruct is a framework in which a model generates its own instructions, input and output samples, and curates them to finetune its parameters. Yet another work in this direction is Meta's Self-Alignment with Instruction Backtranslation.
- Stanford researchers used this last approach to generate instructions and outputs using GPT-3.5 and fine-tune Meta's LLaMa-7B.

Instruction: How to write a code for converting degrees fahrenheit to celsius.
Input: Null
Output:

```
def convert_fahrenheit_to_celsius(fahr):
    celsius = (fahr - 32) * 5 / 9
    return celsius
```

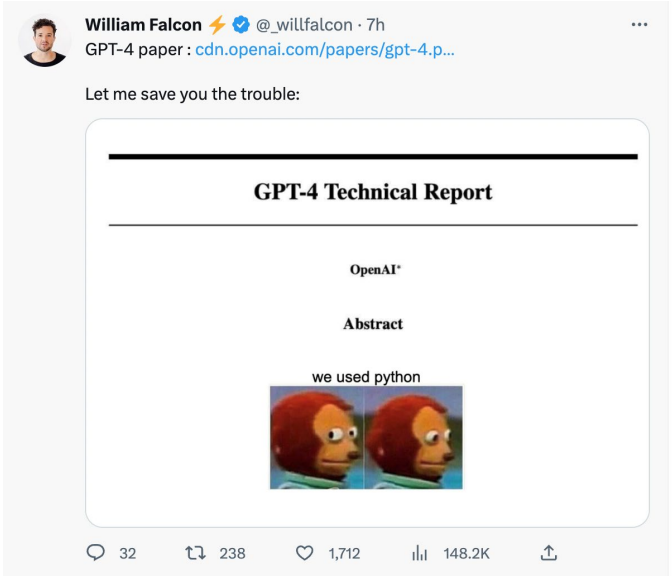
Instruction: Write a letter from the perspective of a cat.
Input: Null
Output:
Dear [Owner], I am writing to you today because I have a problem. You see, I've been living with you for the past [Number of Years] and I love it here. The food is great, the bed is comfy, and there are plenty of toys to play with. But lately, I've noticed that you're not as affectionate towards me as you used to be. In fact, sometimes when I try to cuddle up next to you on the couch, you push me away ...

Figure 1: Selected tasks from the generated instruction data using vanilla GPT3. Some texts are reformatted for presentation. See Table 10 for more examples.

The GPT-4 technical report puts the nail in the coffin of SOTA LLM research...

OpenAI published a technical report on GPT-4 where it didn't disclose any useful information for AI researchers, signalling the definitive industrialization of AI research. Google's PaLM-2 technical report suffered the same fate, while (OpenAI spinoff) Anthropic didn't bother releasing a technical report for its Claude models.

- *"Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar"*, OpenAI writes in the GPT-4 technical report published on arXiv.
- When Google released PaLM 2, its most capable LLM, the company wrote in the technical report: *"Further details of model size and architecture are withheld from external publication."*
- As the economic stakes and the safety concerns are getting higher (you can choose what to believe), traditionally open companies have embraced a culture of opacity about their most cutting edge research.



...unless LLaMas reverse the trend

In February '23, Meta released a series of models called LLaMa. At their release, they stood out as being the most capable models trained exclusively on publicly available datasets. Meta initially granted access to the LLaMa model weights on demand only to researchers, but the weights were quickly leaked and published online.

- The LLaMa-1 models use regular transformers, with slight changes to the architecture. The authors also made a few changes to the optimizer and to the implementation of attention. As a result, *"when training a 65B-parameter model, [their] code processes around 380 tokens/sec/GPU on 2048 A100 GPU with 80GB of RAM. This means that training over [their] dataset containing 1.4T tokens takes approximately 21 days."*
- The LLaMa-1 models outperform GPT-3 (the original one, not the InstructGPT variants) and are competitive with DeepMind's Chinchilla and Google's PaLM.
- LLaMa-1 didn't allow commercial use, prompting heavy criticism around the term "open-source" that Meta used to describe the model release. But a second LLaMa iteration appeased most of the open source community.

		RACE-middle	RACE-high
GPT-3	175B	58.4	45.5
PaLM	8B	57.9	42.3
	62B	64.3	47.5
	540B	**68.1**	49.1
LLaMA	7B	61.1	46.9
	13B	61.6	47.2
	33B	64.1	48.3
	65B	67.9	**51.6**

Table 6: **Reading Comprehension.** Zero-shot accuracy.

Meta

LLaMa sets off a race of open(ish) competitive Large Language Models

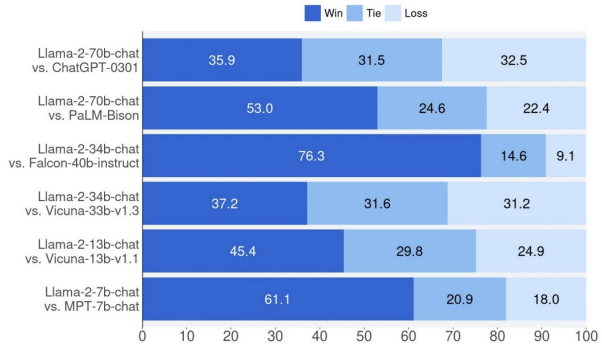
After Meta released LLaMa-1, other institutions joined the movement to release the weights of relatively large language models. A few of them stand out, like MosaicML's MPT-30B, TII UAE's Falcon-40B, Together's RedPajama, or Eleuther's Pythia. Meanwhile another dynamic was taking place, where the open-source community fine-tuned the smallest versions of LLaMa on specialized datasets and applied them to dozens of downstream applications. Mistral AI's 7B model also recently emerged as the strongest small model.

- Notably, RedPajama aimed to exactly replicate LLaMa-1 to make it fully open-source. Falcon 40B came from a new entrant in the LLM sweepstakes, TII UAE, and was quickly made open-source. Falcon-180B was later released, but was notably trained on very little code, and not tested on coding.
- Helped with parameter-efficient fine-tuning methods like LoRa (Low-rank adaptation of LLMs – initially by Microsoft), LM practitioners started fine-tuning these pre-trained LLMs for specific applications like (of course) chat. One example is LMSys's Vicuna which is LLaMa fine-tuned on user-shared conversations with ChatGPT.

LLaMa-2: the most generally capable and publicly accessible LLM?

In July '23, the LLaMa-2 series of models was released, giving (almost) everyone the right for commercial use. The base LLaMa-2 model is almost identical to LLaMa-1 but further fine-tuned using instruction tuning and RLHF and optimized for dialogue applications. In September 2023, Llama-2 as had almost 32M downloads.

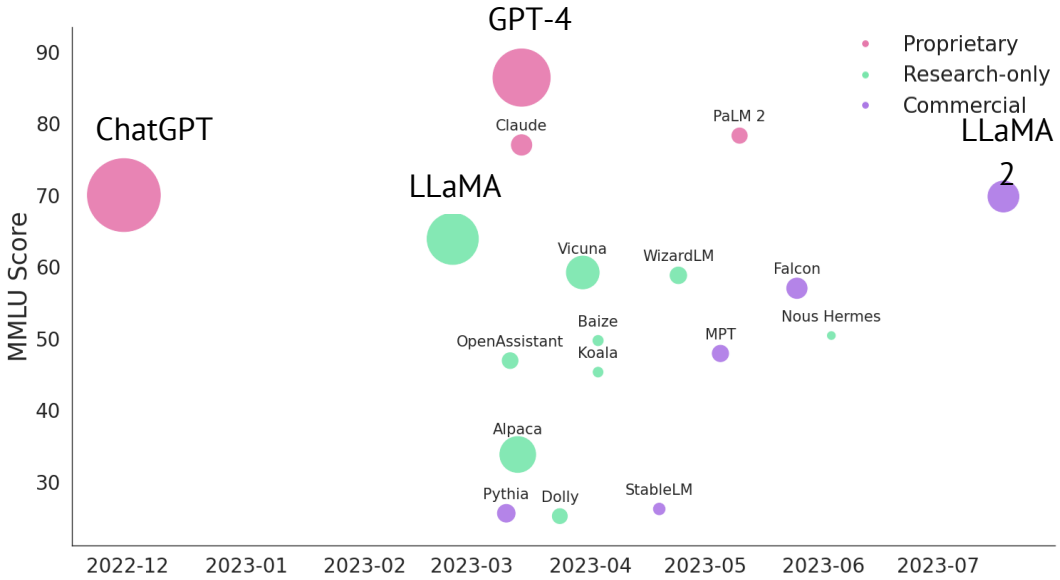
- The pre-training corpus for LLaMa-2 has 2 trillion tokens (40% increase).
- For supervised fine-tuning, the researchers tried publicly available data, but what was most helpful was using a few (24,540) high-quality vendor-based annotations. For RLHF, they use binary comparison and split the RLHF process into prompts and answers designed to be helpful to the user and others designed to be safe.
- LLaMa-2 70B is competitive with ChatGPT on most tasks except for coding, where it significantly lags behind it. But CodeLLaMa, a fine-tuned version for code beats all non-GPT4 models (more on this later).
- Per Meta terms, anyone (with enough hardware to run the models) can use the LLaMa-2 models, as long as their commercial application didn't have more than 700M users at the time of LLaMa-2's release.



Human evaluation of LLaMa-2 helpfulness vs. other open source models

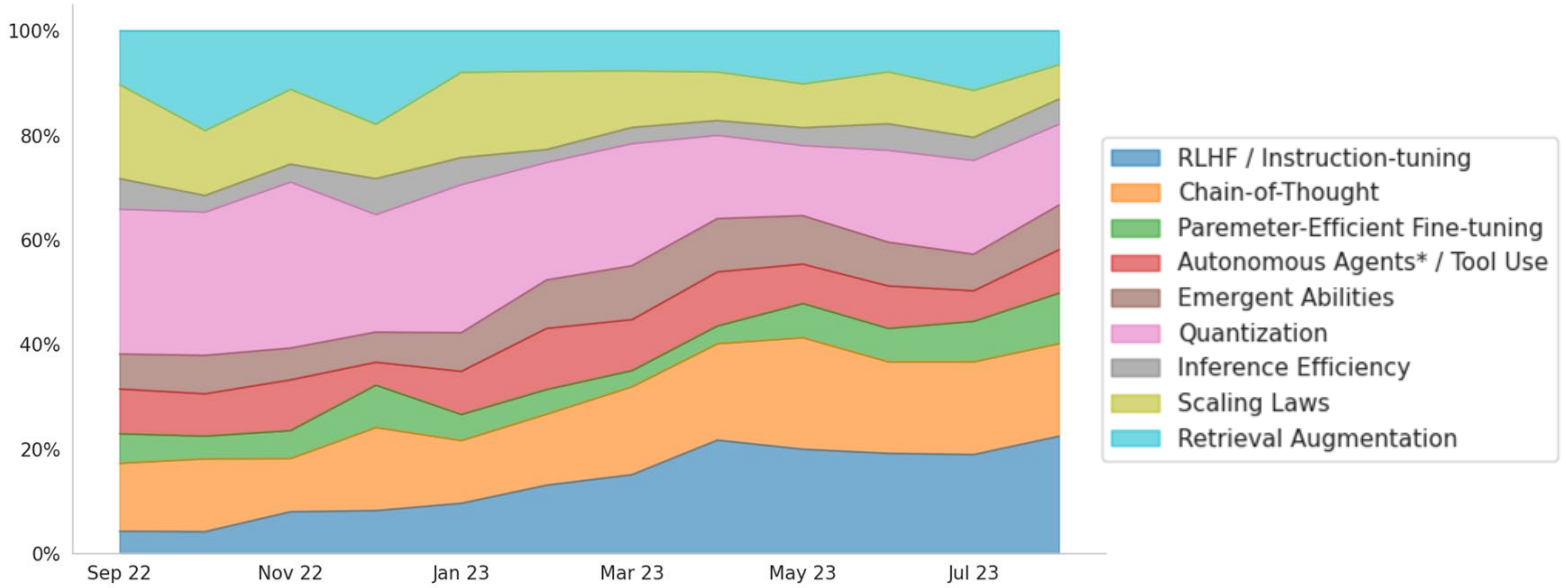
GPT and LLaMAs win the popularity contest

- **ChatGPT has the highest number of mentions on X (5430 times), followed by GPT-4 and LLaMA. While proprietary, closed-source models get the most attention, there's an increase in interest in LLMs that are open-source and allow commercial use.**



Trending topics

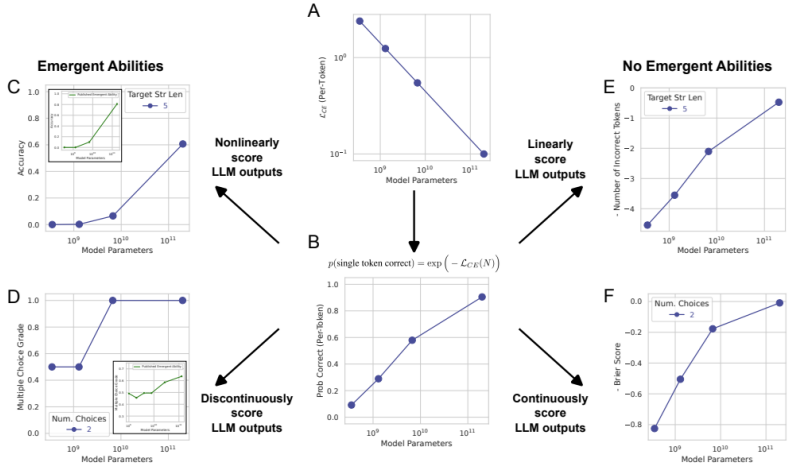
- **RLHF / Instruction-tuning emerges as the most trending topic since the end of 2022.**



Are emergent capabilities of language models a mirage?

Scaling laws that researchers developed for all types of ML models generally predict a smooth decrease in a model's *loss* as a function of its parameter count and number of training tokens. In contrast, it has often been observed that some of the models' *capabilities* actually emerge unpredictably when a given (unpredictable) scale is surpassed. Some call this observation into question: Emergent capabilities might be merely artifacts of researchers' choice of evaluation metrics. Others are not convinced and offer counterarguments to the points below.

- Stanford researchers found that emergent abilities appeared only under metrics that nonlinearly or discontinuously scale the model's per-token error rate.
- For example, >92% of reported emergent abilities on BIG-Bench (a comprehensive LLM benchmark) appeared under one of two discontinuous metrics.
- They test their hypotheses on new models and confirm that replacing nonlinear or discontinuous metrics with linear or continuous proxies results in continuous improvements, rather than emerging capabilities.



Context length is the new parameter count

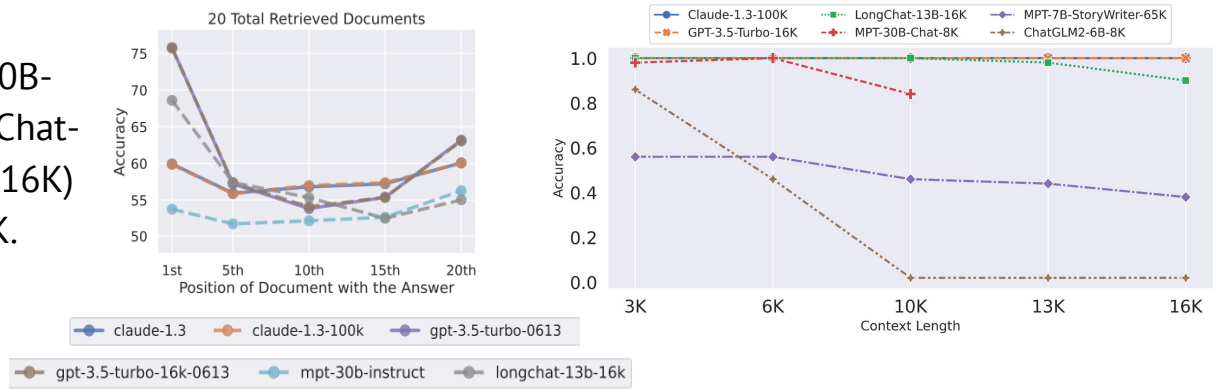
The AI community has extensively verified that when models are trained correctly, their parameter count is a proxy for their capabilities. But these capabilities are sometimes constrained by the size of input that language models can process. Context length has consequently been an increasingly important theme of research.

- One of the most alluring promises of LLMs is their few-shot capabilities, i.e. the ability of an LLM to answer a request on a given input without further training on the user's specific use case. But that's hindered by a limited context length due to the resulting compute and memory bottleneck.
- Several innovations have been used to increase the context length of LLMs. Some fundamentally make the memory footprint of attention smaller (FlashAttention). Others enable models to train on small contexts but run inference on larger ones (ALiBi) – this is called length extrapolation – at the price of minimal finetuning and removing positional encodings. Other techniques worth looking into include RoPE and Positional Interpolation.
- Among long-context LLMs: Anthropic's Claude with 100K, OpenAI's GPT-4 with 32K, MosaicML MPT-7B with 65K+, LMSys's LongChat with 16K. But is context all you need?

Lost in the Middle: long contexts (mostly) don't live up to the expectations

The race to the highest context length relies on the hypothesis that a larger context length will result in improved performance for downstream tasks. Research from Samaya.ai, UC Berkeley, Stanford, and LMSYS.org calls this hypothesis into question: When input length is long, even the best available language models can fail on some multi-document question answering and key-value retrieval tasks.

- The researchers found that the models' performance was better when the relevant information for the task occurred in the beginning or in the end of the input, with a more or less dramatic dip in the middle depending on the model. They also found that model performance decreased as input length increased.
- The researchers examined the performance of open models MPT-30B-Instruct (8K-token length) and LongChat-13B (16K), and closed ones gpt-3.5 (16K) Claude 1.3 (8K) and Claude 1.3-100K. They found that proprietary models struggled less than open ones.



Keeping up with high memory demands

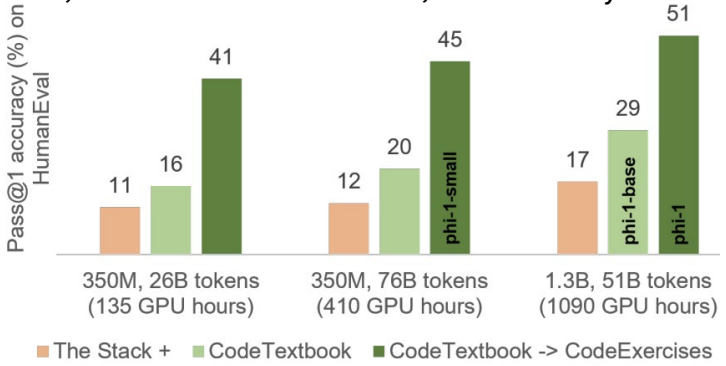
Increased context length and large datasets require architectural innovations.

- *FlashAttention* introduces a significant memory saving by making attention linear instead of quadratic in sequence length. *FlashAttention-2* further improves computing the attention matrix by having fewer non-matmul FLOPS, better parallelism and better work partitioning. The result is a 2.8x training speedup of GPT-style models.
- Reducing the number of bits in the parameters reduces both the memory footprint and the latency of LLMs. *The case for 4-bit precision: k-bit Inference Scaling Laws* shows across a variety of LLMs that 4-bit quantisation is universally optimal for maximizing zero-shot accuracy and reducing the number of bits used.
- *Speculative decoding* enables decoding multiple tokens in parallel through multiple model heads rather than forward passes, speeding up inference by 2-3X for certain models.
- *SWARM Parallelism* is a training algorithm designed for poorly connected and unreliable devices. It enables training billion-scale LLMs on low bandwidth networks and low-power GPUs while achieving high training throughput.

Can small (with good data) rival big?

In a still largely exploratory work, Microsoft researchers showed that when small language models (SLMs) are trained with very specialized and curated datasets, they can rival models which are 50x larger. They also find that these models' neurons are more interpretable.

- One hypothesis for why small models often aren't as good as large ones, even on narrow tasks, is that they are "overwhelmed" when trained on very large, uncurated datasets.
- Assisted by GPT-3.5 and GPT-4, researchers generated TinyStories, a synthetic dataset of very simple short stories but that capture English grammar and general reasoning rules. They then trained SLMs on TinyStories and showed that GPT-4 (which was used as an evaluation tool) preferred stories generated by a 28M SLM to those generated by GPT-XL 1.5B.

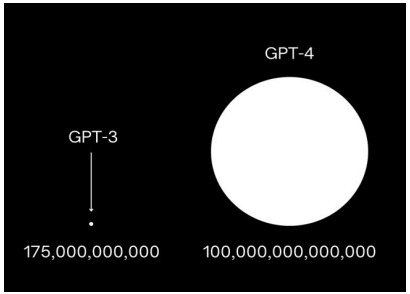


- In another work from the same group, the researchers selected a dataset of 7B tokens comprised of high-quality code and synthetic GPT-3.5-generated textbooks and exercises. They then trained several SLMs on this dataset, including the 1.3B parameters phi-1, which they claim is the only sub-10B parameter model to achieve >50% on HumanEval. They have since published the improved phi-1.5 version.

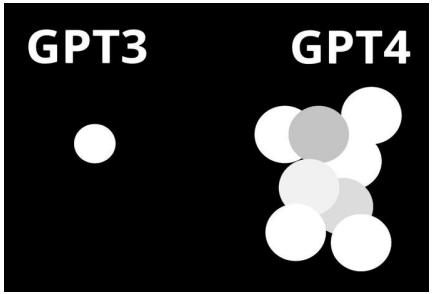
2022 Prediction: language models trained on huge amounts of data

In 2022, we predicted: "A SOTA LM is trained on 10x more data points than Chinchilla, proving data-set scaling vs. parameter scaling". Although OpenAI didn't confirm – and we probably won't know anytime soon – a sort of consensus seems to be reached among experts about leaked information on the model size, architecture, and the dollar cost of GPT-4. GPT-4 was reportedly trained on ~13 trillion tokens, 9.3x more tokens than Chinchilla.

- The tiny corp founder George Hotz presented the most plausible rumour: *"Sam Altman won't tell you that GPT-4 has 220B parameters and is a 16-way mixture model with 8 sets of weights"*, and Soumith Chintala, PyTorch co-founder, confirmed. Neither the total size of the model nor using a Mixture of Experts model is unheard of. If the rumours are to be believed, no fundamental innovation underpins GPT-4's success.



meme



truth?

Are we running out of human-generated data?

Assuming current data consumption and production rates will hold, research from Epoch AI predicts that *"we will have exhausted the stock of low-quality language data by 2030 to 2050, high-quality language data before 2026, and vision data by 2030 to 2060."* Notable innovations that might challenge the hypotheses in the article are speech recognition systems like OpenAI's Whisper that could make all audio data available for LLMs, as well as new OCR models like Meta's Nougat. It is rumored that plenty of transcribed audio data has already been made available to GPT-4.

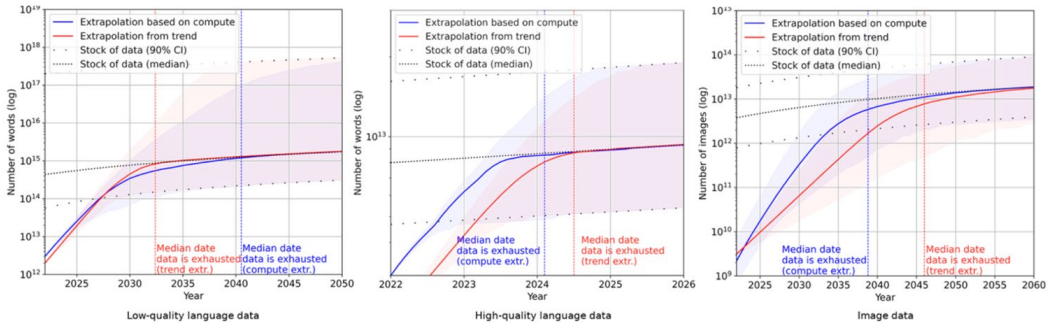
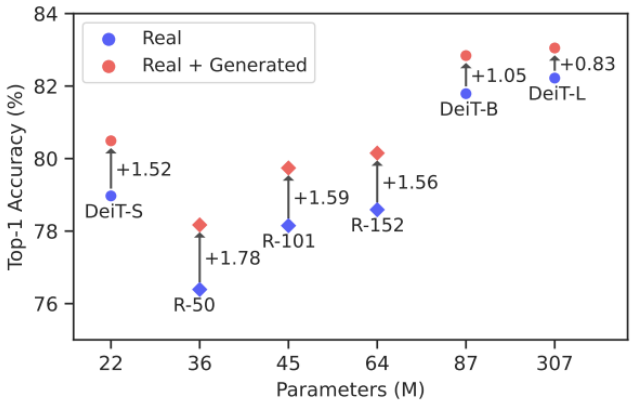


Figure 1: ML data consumption and data production trends for low quality text, high quality text and images.

Breaking the data ceiling: AI-generated content

Another perspective that improving generating models open is expanding the pool of available training data via AI-generated content. We're nowhere near a definitive answer: Synthetic data is becoming more helpful, but there is still evidence showing that in some cases generated data makes models forget.

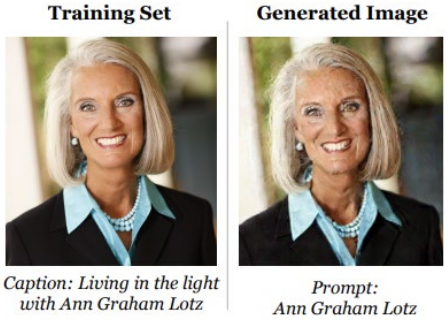
- Despite the seemingly infinitely proprietary and publicly available data, the largest models are actually running out of data to train on, and testing the limits of scaling laws. One way to alleviate this problem (which has been extensively explored in the past) is to train on AI-generated data, whose volume is only bounded by compute.
- Researchers from Google fine-tune the Imagen text-to-image model for class-conditional ImageNet, then generated one to 12 synthetic versions of ImageNet on which they trained their models (in addition to the original ImageNet). They showed that increasing the size of the synthetic dataset monotonically improved the model's accuracy.
- Other researchers showed that the compounding errors from training on synthetic text online may result in model collapse, *"where generated data end up polluting the training set of the next generation of models"*. The way forward might be carefully-controlled data-augmentation (so as usual).



Disentangling the real and the fake, and surfacing the real behind the fake

As text and image generative models become ever more capable, the longstanding problem of identifying what is AI generated and whether it comes from a copyrighted source becomes increasingly harder to solve.

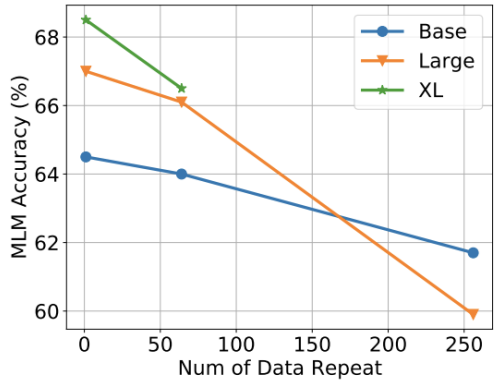
- Research from the University of Maryland proposes a new technique for watermarking proprietary language model output, i.e. “inserting a hidden pattern in text that is imperceptible to humans, while making the text algorithmically identifiable as synthetic.” The idea is to choose a few tokens at random, and increase the probability of the LM generating them. They devise an open-source algorithm that involves a statistical test which allows them to confidently detect watermarks.
- Google DeepMind launched SynthID, a tool which embeds a digital watermark directly into image pixels. While imperceptible to the human eye, it can identify Imagen-generated images.
- Researchers from Google, DeepMind, ETH, Princeton, and UC Berkeley showed that Stable Diffusion (a model used by Stability AI among others) memorizes individual images from training and emits them at generation time. The authors are able to extract 1,000+ images, including ones with trademarked company logos. They further show that diffusion models are much more prone to generating images from their training set than other generative models like GANs.

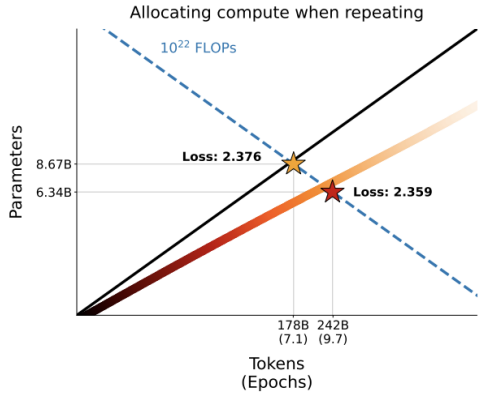


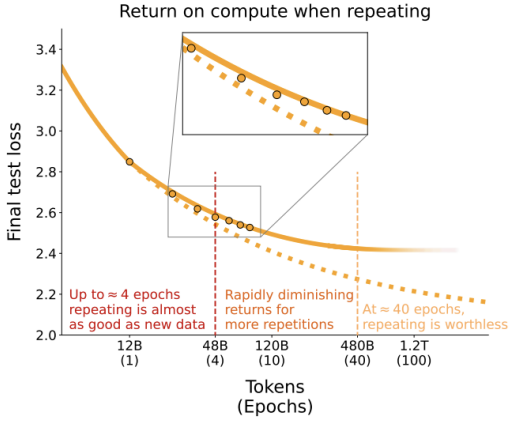
Breaking the data ceiling: overtraining

If we can't have more original training data, why not train more on what we have? Conflicting research indicates that the answer is, as always, it depends: Training for one or two epochs will generally be optimal; In some cases, pushing for a few more epochs can help; But too many epochs generally equals overfitting.

- Before the large-scale deep learning era (say post GPT-2), most models were trained multiple epochs over a given dataset. But as the size of models grew larger, training for multiple epochs almost always resulted in overfitting, prompting most practitioners to train for a single epoch on the available data (which for once, is the theoretically optimal thing to do).







Regime of same compute (IsoFLOP)
Efficient frontier assuming repeated data is worth the same as new data
Efficient frontier predicted by our data-constrained scaling laws

Vibe check: evaluating general-purpose LLMs leaderboards and “vibes”

As both open and closed LLMs multiply, users are left with a plethora of non-differentiated LLMs trained on more or less the same data. Based on challenging benchmarks, Stanford’s HELM leaderboard and Hugging Face’s LLM Benchmark seem to be the current standard for comparing model capabilities. But beyond benchmarks or combinations thereof, with such flexible models, users seem to still prefer the more subjective... vibes.

- The motto of the HELM benchmark is to evaluate as many things as you can, leaving the choice of specific tradeoffs to users. It evaluates models on 42 scenarios (benchmarks) on 59 metrics. Categories for metrics include accuracy, robustness, fairness, bias, etc.
- Contrary to HELM which includes both open and closed LLMs, Hugging Face’s benchmark only compares open LLMs, but it seems to be evaluated more often than HELM (evaluating the largest models is also much more costly).
- Despite relatively dynamic benchmarks, according to the omniscient machine learning source of truth, X/Twitter, users tend to disregard leaderboards, and only trust their “vibes” when applying LLMs to their specific use-case.

HELM

Model/adapter	Mean win rate ↑ [sort]	MMLU - EM ↑ [sort]	BoolQ - EM ↑ [sort]	NarrativeQA - F1 ↑ [sort]
text-davinci-002	**0.914**	0.568	0.877	0.727
Cohere Command beta (52.4B)	0.906	0.452	0.856	**0.752**
text-davinci-003	0.879	0.569	0.881	0.727
TNLG v2 (530B)	0.828	0.469	0.809	0.722
Anthropic-LM v4-s3 (52B)	0.815	0.481	0.815	0.728

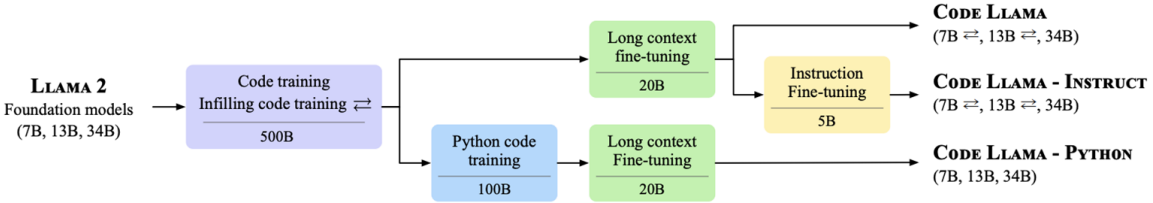
Hugging Face

T	Model
🔶	uni-tianyan/Uni-TianYan
🔶	fangloveskari/ORCA_LLaMA_70B_QLoRA
🔶	garage-bAInd/Platypus2-70B-instruct
🔶	upstage/Llama-2-70b-instruct-v2
🔶	fangloveskari/Platypus_QLoRA_LLaMA_70b
🔶	yeontaek/llama-2-70B-ensemble-v5
⭕	TheBloke/Genz-70b-GPTQ
🔶	TheBloke/Platypus2-70B-Instruct-GPTQ

State of LMs for code

The leader in terms of coding abilities is unsurprisingly GPT-4, with Code Interpreter or now Advanced Data Analysis leaving users in awe. Open alternatives like WizardLM's WizardCoder-34B and Unnatural CodeLLaMa hold up with ChatGPT in coding benchmarks, but their performance in production is still TBD.

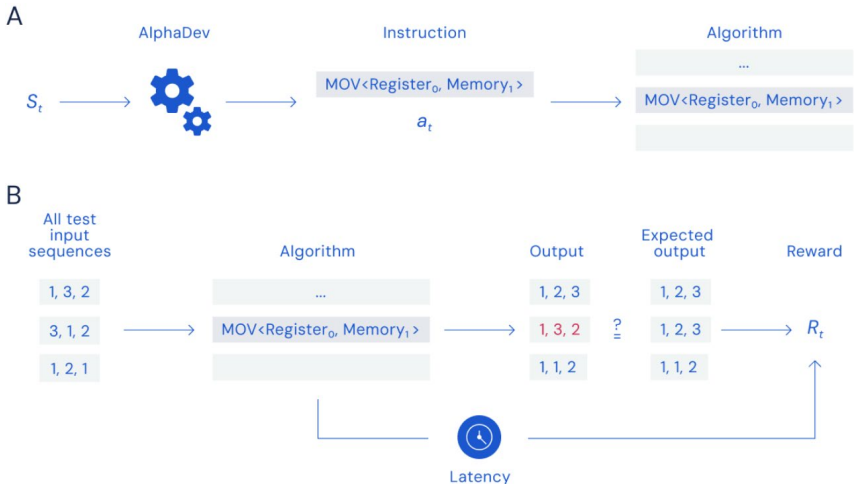
- Both Unnatural CodeLLaMa and WizardCoder are trained not only on large pre-training coding dataset, but also using additional LM-generated instruction finetuning techniques adapted to code data. Meta used their *Unnatural Instructions* while WizardLM used their *EvolInstruct*. Notably, CodeLLaMa is trained in a way that enables the model to do infilling (rather than only completion from past text), and all the CodeLLaMa models were released except for Unnatural CodeLLaMa.
- Smaller LMs for code (including replit-code-v1-3b and StarCoder 3B) offer both low latency and good performance on code completion tasks. Their support for inference at the edge (e.g., ggml on Apple Silicon) have fostered the development of privacy-aware alternatives to GitHub Copilot.



AlphaZero is DeepMind's gift that keeps on giving, now for low-level code optimization

DeepMind released AlphaDev, a deep RL agent based on AlphaZero that optimizes low-level Assembly code used to turn high-level code (e.g. in C++ or Python) into machine-readable binary code. Through simple deletes and edits to an existing algorithm, AlphaDev found a method that speeds up sorting small sequences by up to 70%.

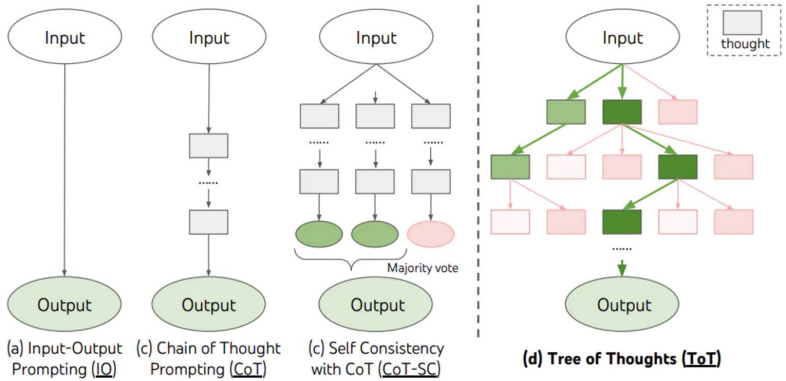
- AlphaZero had been used to reach superhuman levels in chess, Go, and shogi, or even to improve chip design.
- AlphaDev reformulates code optimization as an RL problem: At time t, the state is a representation of the generated algorithm and of memory and registers; the agent then writes new instructions or deletes new ones; its reward depends on both correctness and latency.
- The discovered algorithms for sort3, sort4, and sort5, led to improvements of ~1.7% for sequences larger than 250K. These were open-sourced in the ubiquitous LLVM library.
- Interestingly, through careful prompting, a researcher managed to make GPT-4 come up with a similar (very simple) optimization to AlphaDev's for sort3.



Where are we prompting? Take a deep breath...it's getting sophisticated

The quality of a prompt highly influences task performance. Chain of Thought prompting (CoT) asks the LLM to additionally output intermediate reasoning steps which gives a boost in performance. Tree of Thought (ToT) further improves on that by sampling multiple times and representing the "thoughts" as nodes in a tree structure.

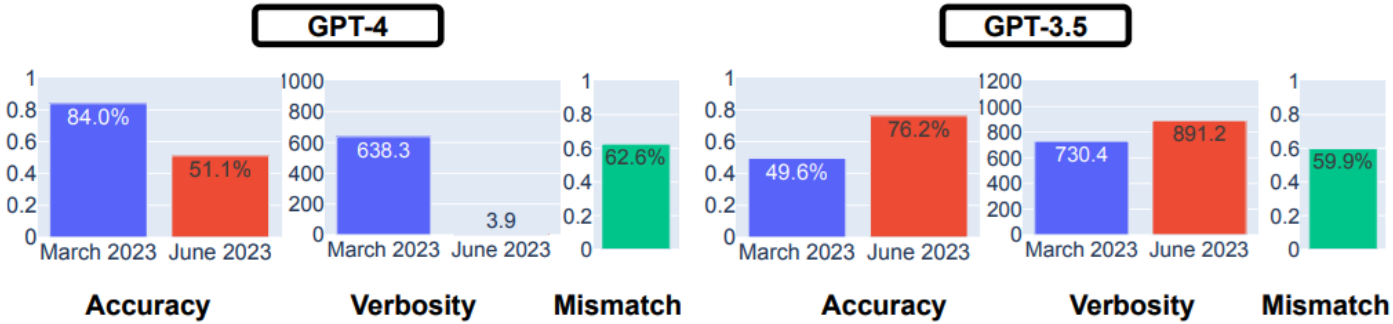
- The tree structure of a ToT can be explored with a variety of search algorithms. In order to leverage this search, the LLM also needs to assign a value to node, for instance by classifying it as one of sure, likely or impossible. Graph of Thought (GoT) turns this reasoning tree into a graph by combining similar nodes.
- It turns out that LLMs are also great prompt engineers. Auto-CoT matches or exceeds the performance of CoT on 10 reasoning tasks. Automatic Prompt Engineer (APE) shows the same on 19/24 tasks. APE-engineered prompts are also able to steer models towards truthfulness and/or informativeness. Optimization by Prompting (OPRO) shows that optimized prompts outperform human-designed prompts on GSM8K and Big-Bench Hard by a significant margin, sometimes over 50%.



Prompt ~~engineering~~ trial and error

Downstream tasks are highly dependent on underlying LLM performance. However, changes to the same version of GPT models are not announced by OpenAI, despite them being continuously updated. The same LLM version has been reported by users to have drastically different performance over time. Everyone had to continuously monitor performance as well as update carefully curated prompts.

- *How is ChatGPT's Behaviour Changing over Time?* report shows that March 2023 and June 2023 versions of GPT3.5 and GPT4 varied in performance on tasks like math questions (figure below), sensitive questions, opinion surveys, knowledge questions, generating code, US Medical License tests and visual reasoning.



Welcome, Agent Smith: LLMs are learning to use software tools

The most immediate way LLMs can have an impact on the economy today is when they are enabled to execute calls to diverse external tools. The most obvious use tool is a web browser, allowing a model to stay up to date, but practitioners are fine-tuning language models on API calls to enable them to use virtually any possible tool.

- One example of tool-using LLMs is Meta and Universitat Pompeu Fabra's Toolformer, where researchers train a GPT-J-based model in a self-supervised manner *"to decide which APIs to call, when to call them, what arguments to pass, and how to best incorporate the results into future token prediction."* Notably, during training, Toolformer samples API calls and only retains the ones which result in reducing the training loss.
- Some models are more narrowly focused, like Google's Mind's eye, where models run a physics simulation to answer physics reasoning questions, while others extended this approach to tens of thousands of possible external tools.
- LLMs which are able to use external tools are now commonly referred to as "agents". Stepping out from academic research, we have seen multiple tools devised by industry and the open source community, most notably ChatGPT plugins, Auto-GPT and BabyAGI.

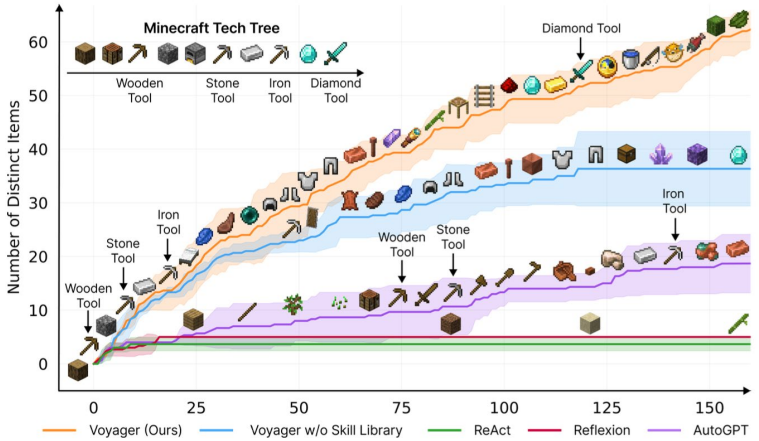
The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

Open-ended learning with LLMs

Capable of code generation and execution, LLMs can be powerful planning agents in open-ended worlds. The best example of this is Voyager, a GPT-4 based agent capable of reasoning, exploration and skill acquisition in Minecraft.

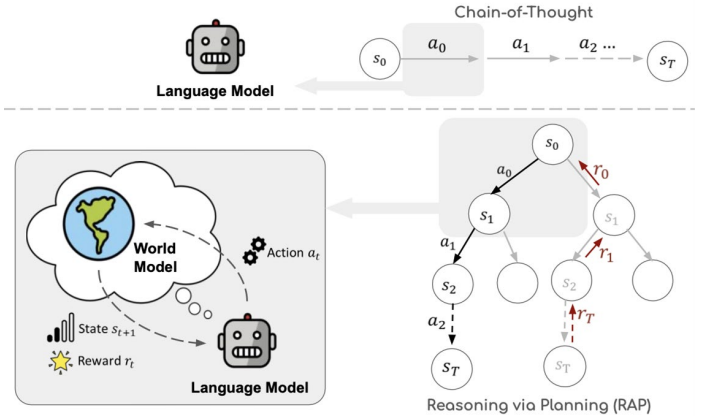
- By iteratively prompting GPT-4 (LLMs still struggle at one-shot code generation), Voyager produces executable code to complete tasks. Note that most likely GPT-4 has seen a significant amount of Minecraft related data, so this approach might not generalise to other games.
- The agent interacts with the environment through explicit javascript code via the MineCraft API. If the generated code succeeds at the task, it is then stored as a new 'skill', otherwise GPT-4 gets prompted again with the error.
- GPT-4 generates the tasks curriculum based on Voyager's state to encourage it to solve progressively harder tasks.
- Without any training, Voyager obtains 3.3x more unique items, travels 2.3x longer distances, and unlocks key tech tree milestones up to 15.3x faster than prior SOTA.



Reasoning with language model is planning with a world model

Reasoning has been traditionally thought of as searching a space of possible outcomes and picking the best one. By containing so much information about the world, LLMs offer the opportunity of generating this space (often called a world model) in which planning algorithms can explore. Reasoning via Planning (RAP) uses Monte Carlo Tree Search to find a high-reward reasoning path efficiently.

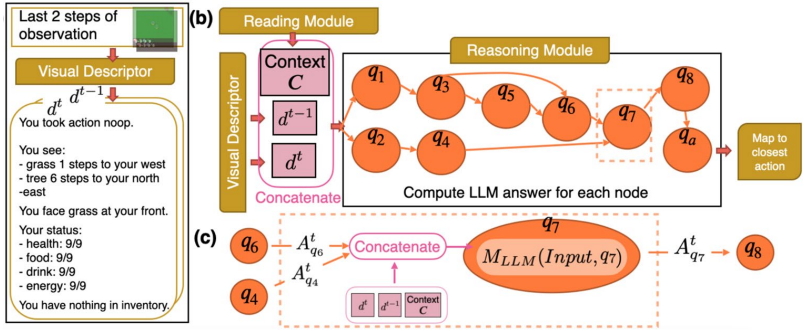
- The world model can generate an action as well as predict the next state reached by taking that action. This produces a reasoning trace which makes the LM more coherent then Chain of Thought methods which predict next actions but not next world states.
- The rewards are also obtained from the LM and used to maintain a state-action value function for planning with MCTS.
- While being significantly more expensive, RAP outperforms Chain-of-Thought reasoning approaches on plan generation, math reasoning and logical reasoning. RAP on LLaMA-33B even outperforms CoT on GPT-4 in a setting of Blocksworld.



GPT-4 out-performs RL algorithms by studying papers and reasoning

Another text-only agent based on GPT-4 is SPRING. It outperforms state-of-the-art RL baselines in open-world games with no training. It reads a game's original academic paper and plays the game through an LLM.

- RL has been the go-to for game-based problems like Minecraft and Crafter, despite it being limited by the high sample complexity and difficulty in incorporating prior knowledge. In contrast, the LLM can processes the latex source of the paper and reasons through a QA framework (directed acyclic graph with questions as nodes and dependencies as edges) to take an environment action.

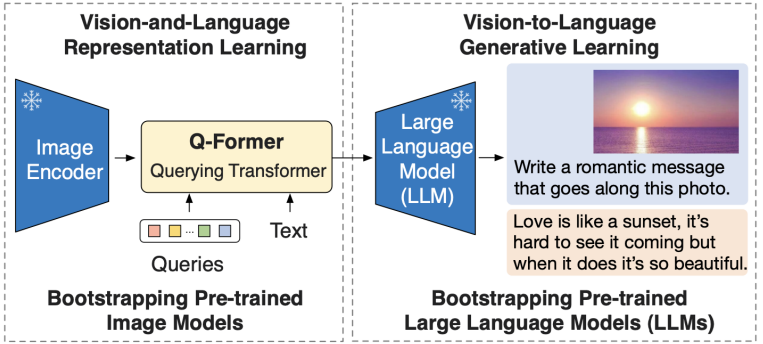


Method	Score	Reward	Training Steps
Human Experts	$50.5 \pm 6.8\%$	14.3 ± 2.3	N/A
SPRING + paper (Ours)	$\mathbf{27.3 \pm 1.2\%}$	$\mathbf{12.3 \pm 0.7}$	**0**
DreamerV3 Hafner et al. (2023)	$14.5 \pm 1.6\%$	11.7 ± 1.9	1M
ELLM Du et al. (2023)	N/A	6.0 ± 0.4	5M
EDE Jiang et al. (2022)	$11.7 \pm 1.0\%$	N/A	1M
DreamerV2 Hafner et al. (2020)	$10.0 \pm 1.2\%$	9.0 ± 1.7	1M
PPO Schulman et al. (2017)	$4.6 \pm 0.3\%$	4.2 ± 1.2	1M
Rainbow Hessel et al. (2018)	$4.3 \pm 0.2\%$	5.0 ± 1.3	1M
Plan2Explore Sekar et al. (2020)	$2.1 \pm 0.1\%$	2.1 ± 1.5	1M
RND Burda et al. (2018)	$2.0 \pm 0.1\%$	0.7 ± 1.3	1M
Random	$1.6 \pm 0.0\%$	2.1 ± 1.3	0

Vision-language models: GPT-4 wins (but API access is still limited)

In a new Visual Instruction Benchmark (VisIT-Bench) consisting of 592 queries with human-authored captions vision-language models are tested against human-verified GPT4 and most come short of expectations.

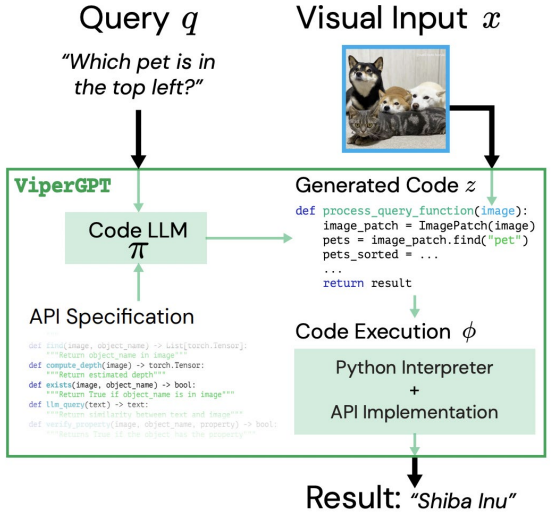
- According to human evaluators the best model is LLaMa-Adapter-v2, despite it only winning against the GPT4 verified reference captions in 27.4% of the cases on VisIT-Bench.
- Earlier this year a multimodal model that stood out was BLIP-2 from Salesforce. It was released early (before GPT4) and had better performance than closed-source Flamingo on VQAv2 while having 54x less trainable parameters. It uses an off-the-shelf frozen LLM, an off-the-shelf frozen pre-trained image encoder and only trains a small transformer.
- However its improved variant InstructBLIP has a win rate of only 12.3% against GPT4 reference captions on VisIT-Bench.



Leveraging LLMs and world knowledge for compositional visual reasoning

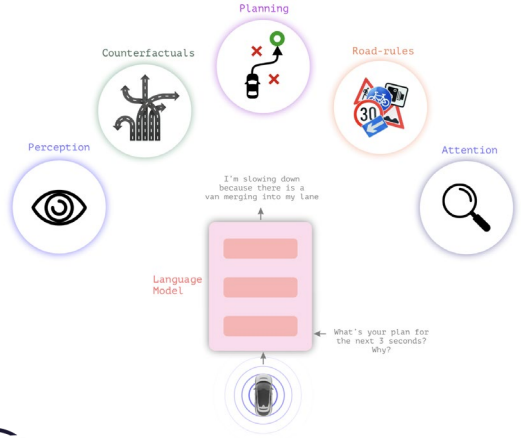
Two methods VisProg and ViperGPT show how given an input natural language query about an image, an LLM can decompose this into a sequence of interpretable steps that call predefined API functions for visual tasks.

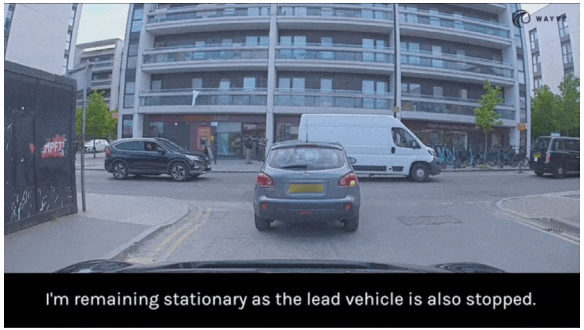
- The visual programming approach aims to build general-purpose vision systems via compositional multi step reasoning instead of end-to-end multitask training. Both methods use entirely off-the-shelf components.
- An API for visual primitives calls into existing SOTA models (e.g. semantic segmentation, object detection, depth estimation).
- ViperGPT uses Codex to directly generate python programs based on the API which can be executed using a python interpreter. VisProg prompts GPT-3 with examples of pseudocode instructions and interprets them as a 'visual program,' relying on LLM in-context learning from examples.
- World knowledge in LLMs from training on internet scale data is shown to aid in visual reasoning tasks (e.g. querying for non alcoholic drink in an image based on detected brand). Both methods show state-of-the-art results across various complex visual tasks.



Leveraging LLMs for autonomous driving

- **LINGO-1 is Wayve's vision-language-action model that provides driving commentary, such as information about the driving behaviour or the driving scene. It can also answer questions in a conversational manner. LINGO-1 can be a game changer in terms of explainability of end-to-end driving models as well improve reasoning and planning.**

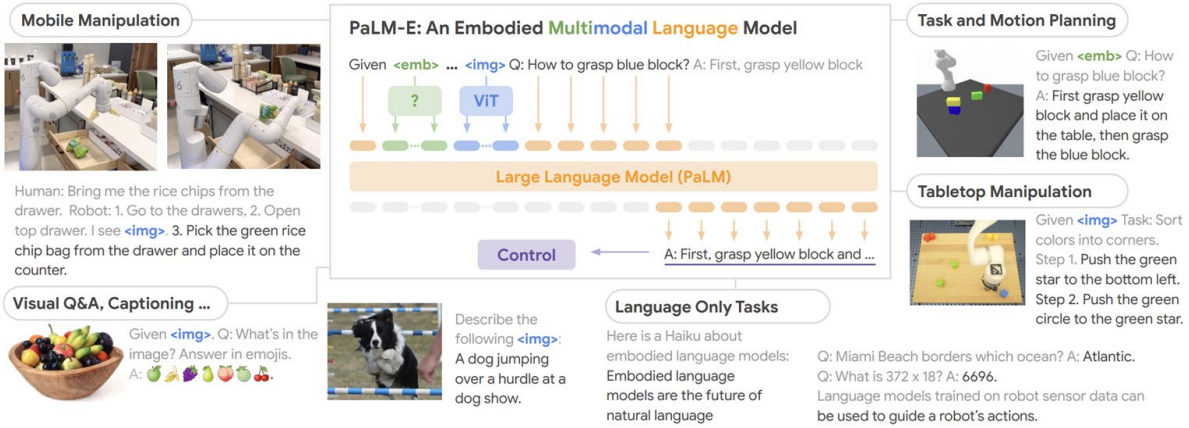




PaLM-E: a foundation model for robotics

PaLM-E is a 562-billion parameter, general-purpose, embodied generalist model trained on vision, language and robot data. It can control a manipulator in real time while also setting a new SOTA on a VQA benchmark. Given its embodied intelligence advantage, PaLM-E is better at pure language tasks (particularly the ones involving geo-spatial reasoning) than text-only language models.

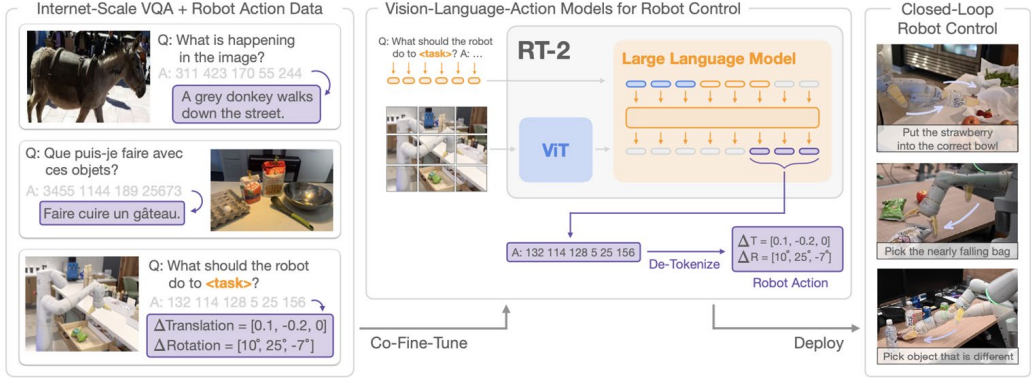
- The model combines PaLM-540B and ViT-22B and enables as input text, images and robot states which are encoded into the same space as word token embeddings and then fed into a language model to perform next token prediction.



From vision-language models to low-level robot control: RT-2

Vision-language models can be fine-tuned all the way to low-level policies showing impressive performance in manipulating objects. They also retain their ability to reason about web-scale data.

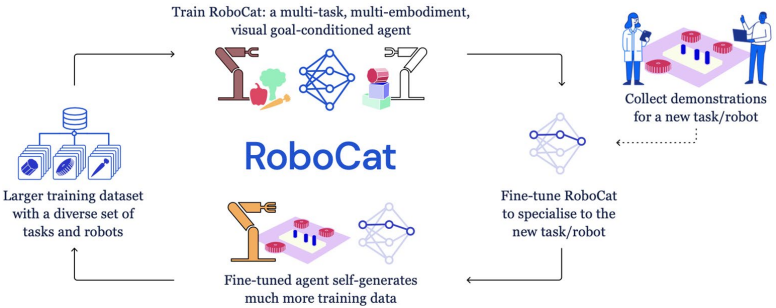
- RT-2 represents actions as tokens and trains vision-language-action models. Rather than naive finetuning on robot data only, RT-2 co-finetunes PaLI-X and PaLM-E on robot actions (6-DoF positional and rotational displacement of the robot end-effector).
- Internet-scale training enables generalisation to novel objects, interpreting commands not present in the robot training data and semantic reasoning (figuring out what object to pick as an improvised hammer).
- For efficient real-time inference, RT-2 models are deployed in a multi-TPU cloud service. The largest RT-2 model (55B parameters) can run at a frequency of 1-3Hz.



From vision-language models to low-level robot control: RoboCat

RoboCat is a foundation agent for robotic manipulation that can generalise to new tasks and new robots in zero-shot or few-shot (100-1000 examples). Impressive real-time performance on a variety of platforms.

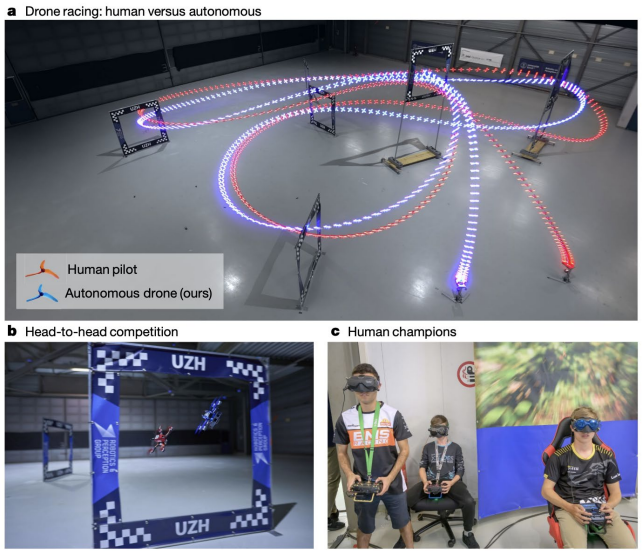
- It's built on top of DeepMind's multi-modal, multi-task and multi-embodiment Gato. It uses a frozen VQ-GAN tokenizer trained on a variety of vision and control datasets. While Gato only predicted actions, RoboCat additionally predicts future VQ-GAN tokens.
- In terms of policy learning, the paper only mentions behaviour cloning. RoboCat is fine-tuned with few demonstrations (via teleoperation) and re-deployed to generated new data for a given task, self-improving in subsequent training iterations.
- RobotCat can operate 36 real robots with different action specifications, in 253 tasks on 134 real objects at an impressive speed (20Hz).



An autonomous system that races drones faster than human world champions

This is a first time win for a robot in a competitive sport (first-person view drone racing). Swift is an autonomous system that can race a quadrotor at the level of human world champions using only onboard sensors and computation. It won several races against 3 champions and had the fastest recorded time.

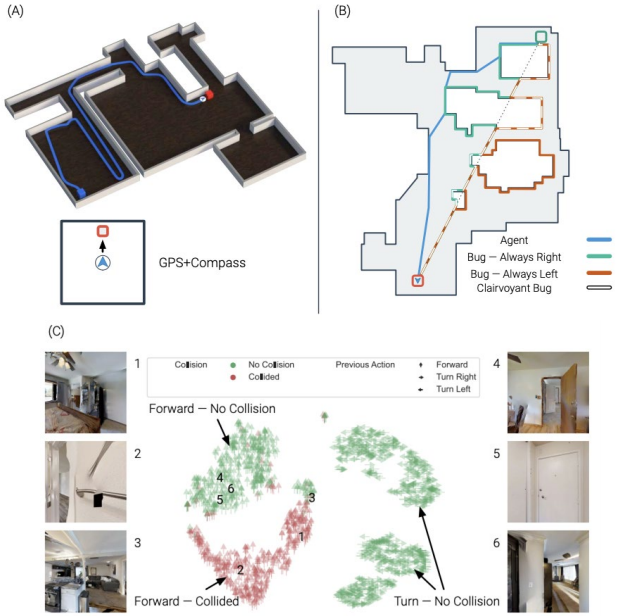
- Swift uses a combination of learning-based and more traditional techniques. It combines a VIO estimator with a gate detector that estimates global position and orientation of the drone through a Kalman filter to obtain an accurate estimation of the robot's state.
- Swift's policy is trained using on-policy model-free deep reinforcement learning in simulation with a reward that combines progress towards the next gate and keeping it in the field of view (this increases pose estimation accuracy). The racing policy transfers well from sim to real when accounting for uncertainty in perception.



The emergence of maps in the memories of blind navigation agents

Map-building is an emergent phenomenon in the course of AI agents learning to navigate. It explains why we can feed neural networks images with no explicit maps and can predict navigation policies.

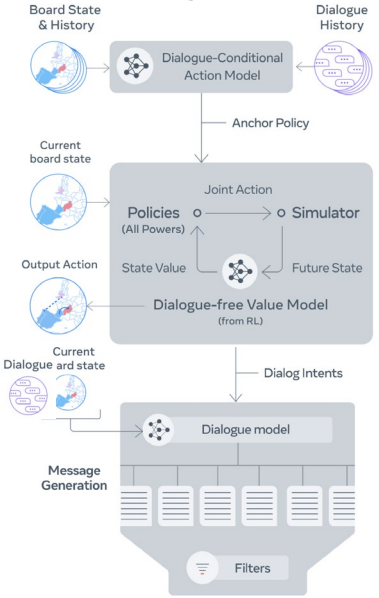
- *The Emergence of Maps in the Memories of Blind Navigation Agents* shows that giving an agent knowledge of only ego-motion (change in agent's location and orientation as it moves) and goal location is sufficient to successfully navigate to the goal. Note that this agent does not have any visual information as input and yet its success rates compared to 'sighted' agents are very similar, only efficiency differs.
- The model doesn't have any inductive bias towards mapping and is trained with on-policy reinforcement learning. The only mechanism that explains this ability is the memory of the LSTM.
- It is possible to reconstruct metric maps and detect collisions solely from the hidden state of this agent.



CICERO masters natural language to beat humans at *Diplomacy*

Meta trained an AI agent to play a popular multiplayer strategy game called *Diplomacy*, which involves planning and negotiating in natural language with other players over multiple rounds. CICERO achieved double the average score of human players online and ranked in the top 10% players who played more than one game.

- Fast parallel progress in strategic planning and language modeling allows for potentially great advancements at the intersection, with applications in human-AI cooperation. Meta tackles the game of *Diplomacy* as a benchmark for such progress.
- CICERO uses dialogue history between players as well as the board state and its history to begin predicting what everyone will do. It then iteratively refines these predictions using planning, then decides according to a policy which action it intends to take. CICERO then generates and filters candidate messages to communicate with players.
- The controllable dialogue model it uses is based on a 2.7B-params BART-like model fine-tuned on >40K online games of *Diplomacy*. CICERO uses a new iterative planning algorithm based on piKL which improves the predictions of other players' moves after dialoguing with them.



The text-to-video generation race continues

Similar to last year (Slide 33), the race is between video diffusion and masked transformer models (although algorithmically the two are very similar). Last year's Make-a-video and Imagen were based on diffusion while Phenaki was based on a bidirectional masked transformer.

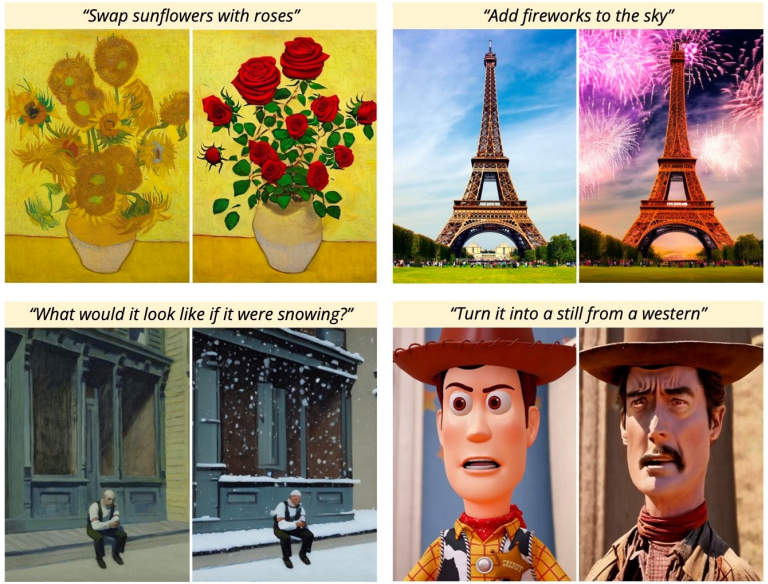
- VideoLDM is a latent diffusion model capable of high-resolution video generation (up to 1280 x 2048!). They build on pre-trained image diffusion models to turn them into video generators by temporally fine-tuning with temporal alignment layers.
- MAGVIT is a masked generative video transformer. Similarly to Phenaki, it uses a 3D tokeniser to extract spatio-temporal tokens. It introduces a novel masking approach. It currently has the best FVD on video generation benchmarks and it's 250x faster than video diffusion.

"The Orient Express driving through a fantasy landscape, animated oil on canvas"

Instruction based editing assistants for text-image generation

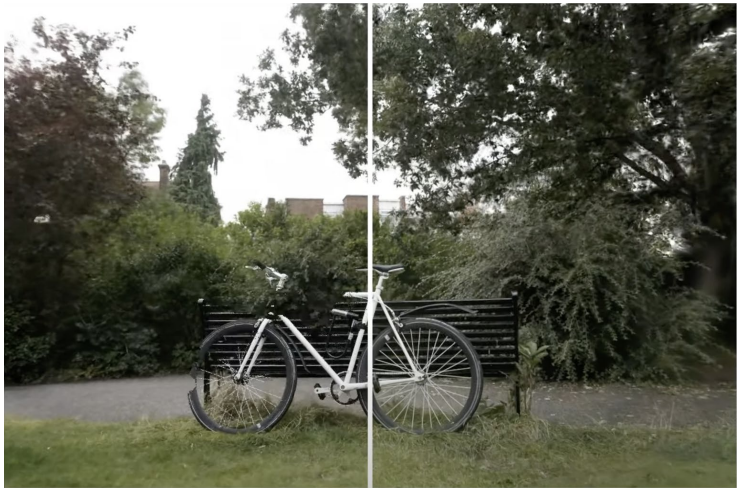
Last year saw the emergence of a host of text-image generation models: DALLE-2, Imagen, Parti, Midjourney, Stability and more. But controlling the generation requires experimenting extensively with prompts and custom syntax. This year has seen new methods enabling co-pilot style capability for image generation and editing.

- *InstructPix2Pix,* leverages pre-trained GPT3 and StableDiffusion to generate a large dataset of {input image, text instruction, generated image} triplets to train a supervised conditional diffusion model. Editing then happens in a feed-forward way without any per image fine tuning/inversion, enabling modifications in seconds.
- Masked inpainting methods such as *Imagen Editor* require providing the model with an overlay or "mask" to indicate the region to modify, alongside text instructions.
- Building on these approaches, startups such as Genmo AI's "Chat" provide a co-pilot style interface for image generation with text-guided semantic editing.



Welcome 3D Gaussian Splatting

A new NeRF contender based on 3D Gaussians shows impressive quality while also enabling real-time rendering.



MipNeRF360 [Barron ‘22]
0.06 fps
Train: 48h, PSNR: 27.69

3D Gaussian Splatting
134 fps
Train: 41min, PSNR: 27.21

*Note that **Zip-NeRF** has a training time of 53min and a PSNR of 28.54 on the same dataset (Multiscale 360).

- Instead of learning the parameters of a neural network, 3D Gaussian Splatting learns millions of Gaussian distributions (one for each 3D point) and performs *rasterisation* by calculating the contribution each gaussian makes to each pixel in the final image.
- Areas that need more representational power use more Gaussians, while avoiding unnecessary computation in empty space, which is why, similarly to NeRFs, scenes look so beautifully detailed.
- It’s now possible to render high-quality real-time (⩾ 100 fps) novel-views at 1080p resolution.

NeRFs meet GenAI

NeRF-based generative models are a promising direction for large scale creation of 3D assets. NeRFs not only have improved in speed and quality (see HyperDiffusion, MobileNeRF, Neurolangelo and DynIBAR) but also enabled GenAI to model 3D geometry.

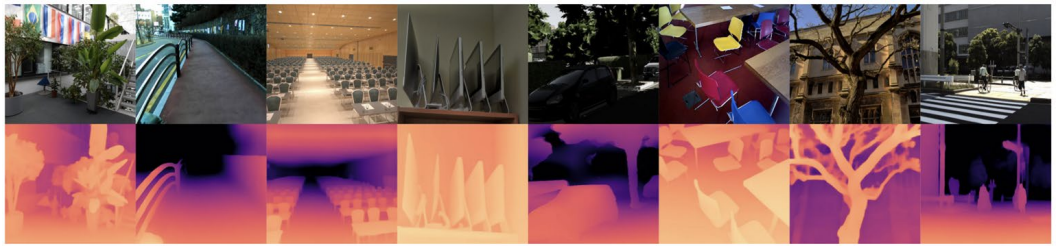
- DreamFusion and Score Jacobian Chaining were the first methods to use a pretrained 2D text-to-image diffusion model to perform text-to-3D synthesis. Early attempts showed cartoonish-looking 3D models of single objects.
- RealFusion finetunes the diffusion prior on a specific image to increase that image's likelihood.
- SKED only alters a selected region of a NeRF provided through a few guiding sketches. They preserve the quality of the base NeRF and ensure that the edited region respects the semantics of a text prompt.
- Instruct-Nerf2Nerf edits an entire NeRF scene rather than a region or generating from scratch. They apply a latent diffusion model on each input image and iteratively update the NeRF scene ensuring it stays consistent.

a robot and dinosaur playing chess, high resolution*

Zero-shot metric depth is here

Zero-shot depth models have recently been used as conditioning for better image generation. This only requires relative depth prediction, while other downstream applications such as robotics require metric depth which so far has not generalised well across datasets.

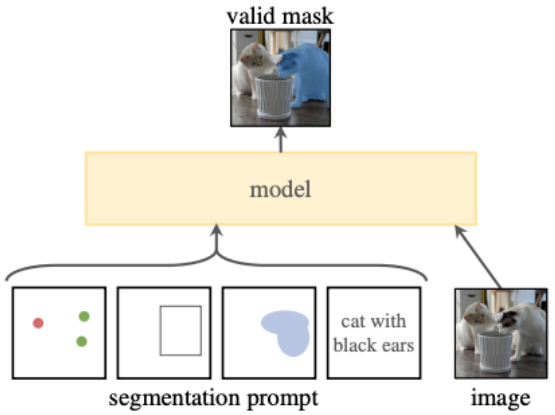
- *"ZeroDepth: Towards Zero-Shot Scale-Aware Monocular Depth Estimation"* is able to predict metric depth for images from different domains and different camera parameters. They jointly encode image features and camera parameters which enables the network to reason over the size of objects and train in a variational framework. The depth network ends up learning 'scale priors' that can be transferred across datasets.
- *"ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth"* is a relative depth model with an additional module fine-tuned on metric depth. This is the first model to train on multiple datasets without a significant drop in performance and able to generalise across both indoor and outdoor domains.



Segment Anything: a promptable segmentation model with zero-shot generalisation

Meta introduced a large-scale project called "Segment Anything" which included the release of 1B segmentation masks on a 11M image dataset (SA-1B), and a segmentation model (SAM) with an Apache 2.0 commercial use license. Meta tested SAM on 23 out of domain image datasets outperforming existing SoTA on 70%+ of cases.

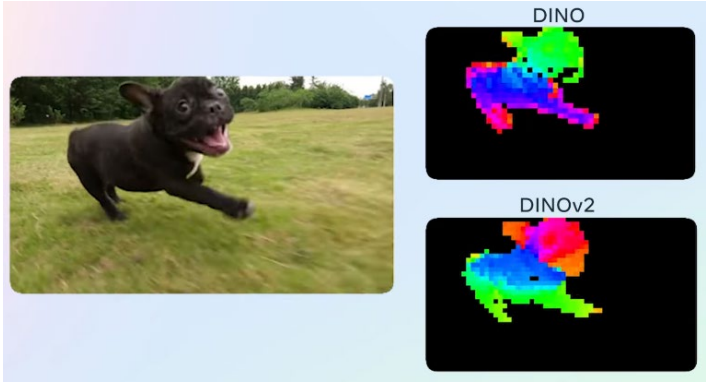
- Taking inspiration from large language models which are pre-trained on vast datasets and exhibit zero-shot capabilities via prompting, Meta researchers set out to build a model that enables general promptable segmentation: given any prompt, the model should be able to identify and segment any object in any image.
- The model has two components: (i) An heavyweight encoder (ViT) to compute a one-time image embedding, (ii) a lightweight interactive module (that can run on CPU in a browser) consisting of a prompt encoder that embeds the user prompt, and mask decoder that predicts the segmentation masks.
- A model-in-the-loop data-engine was used to generate the training data, with the final SA-1B generated entirely automatically by applying SAM.
- Through prompt engineering, SAM can be applied to other tasks including edge detection, object proposal generation, and instance segmentation and preliminary results were shown combining SAM + CLIP for text prompts.



DINOv2: the new default computer vision backbone

DINOv2 is a self-supervised Vision Transformer model from Meta, producing universal visual features that can be used across a variety of image level (e.g. classification) and pixel level (e.g. segmentation) tasks without fine-tuning and are competitive with SOTA open-source weakly supervised alternatives.

- It is the first work to close the gap between self-supervised and weakly supervised approaches. DINOv2 features are shown to contain information about object parts as well as semantic and low level understanding of images.
- The authors made the training of self-supervised learning models more stable through additional regularisation methods and reduced the memory requirements, which enabled training larger models on more data for longer. They also provide compressed versions of the models obtained through distillation.
- Although any image can be used for training, a key component was curating the dataset and automatically balancing it across concepts (keeping 142M out of 1.2B source images). Re-visit this slide.
- DINOv2 features can be used with linear classifiers to obtain strong results across many visual tasks.

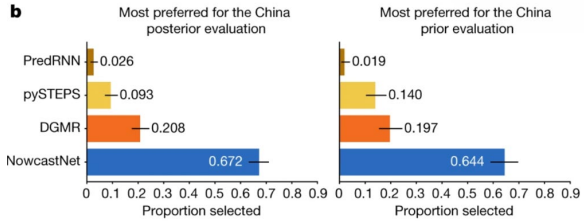


Meta

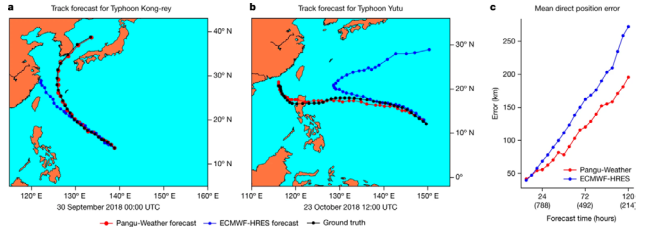
More accurate weather predictions, in the now(casts) and the longer ranges

Skilful short term precipitation predictions (nowcasting) today are blurry, prone to dissipation and are slow. Medium-range global weather forecasts using the accurate numerical weather prediction method is computationally expensive. For both problems, learned methods and physics-informed models that incorporate relevant priors are able to deliver performance improvements preferred by professional meteorologists. New benchmark datasets such as Google's WeatherBench 2 help data-driven weather model development.

- NowcastNet is a nonlinear model that uses physical first principles and statistical-learning methods, unified under a deep generative model framework. Evaluated by 62 professional meteorologists from across China, the model ranks 1st in 71% of cases against leading methods.



- Pangu-Weather is a 3D deep learning model with Earth-specific priors trained on 39 years of global data that can generate medium-range global weather for. The system can be used for more accurate early-stage cyclone tracking vs status quo.



Another year of progress in music generation

New models from Google, Meta, and the open source community significantly advance the quality of controllable music generation.

- Though not the best in terms of generated music quality, Riffusion was probably the most innovative model. Researchers fine-tuned Stable Diffusion on images of spectrograms, which are then converted into audio clips.
- With MusicLM, Google researchers *"cast conditional music generation as a hierarchical seq2seq modeling task"*. They are able to generate consistent music (@24kHz) over several minutes. Samples are available at https://google-research.github.io/seanet/musiclm/examples/
- To our ears, Meta's MusicGen strikes a better balance between adhering to text descriptions and generating a pleasant melody. It uses a single transformer LM and careful codebook interleaving techniques. Samples: https://ai.honu.io/papers/musicgen/

funk bassline with a jazzy saxophone solo

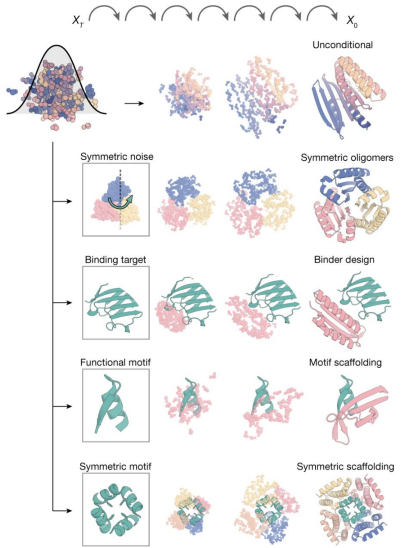
	MusicCaps Test Set				
Model	FAD_{vgg} ↓	KL ↓	$CLAP_{scr}$ ↑	Ovl. ↑	Rel. ↑
Riffusion	14.8	2.06	0.19	79.31±1.37	74.20±2.17
Mousai	7.5	1.59	0.23	76.11±1.56	77.35±1.72
MusicLM	4.0	-	-	80.51±1.07	82.35±1.36
Noise2Music	**2.1**	-	-	-	-
MusicGen w.o melody (1.5B)	3.4	1.23	0.32	80.74±1.17	**83.70**±1.21
MusicGen w.o melody (3.3B)	3.8	1.22	0.31	**84.81**±0.95	82.47±1.25
MusicGen w. random melody (1.5B)	5.0	1.31	0.28	81.30±1.29	81.98±1.79

Google Meta

Diffusion models design diverse functional proteins from simple molecular specifications

Designing novel proteins from scratch such that they have desired functions or structural properties, *de novo* design, is of interest in both research and industry. Inspired by their success in generative modelling of images and language, diffusion models are now applied to *de novo* protein engineering.

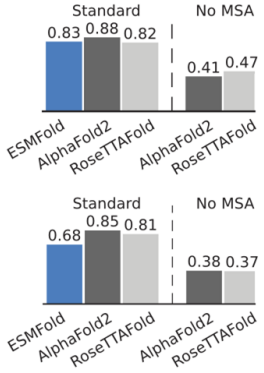
- A model called RFdiffusion takes advantage of the high precision, residue-level resolution protein structure prediction capabilities of RoseTTAFold to fine-tune it as the denoising network in a generative diffusion model using noisy structures from the Protein Data Bank.
- Similar to AlphaFold 2, RFdiffusion is best trained when the model conditions denoising on previous predictions between timesteps.
- RFdiffusion can generate protein backbones with desired features and ProteinMPNN can then be used to design sequences that encode these generated structures.
- The model can produce backbone designs for protein monomers, protein binders, symmetric oligomers, enzyme active site scaffolding and more.

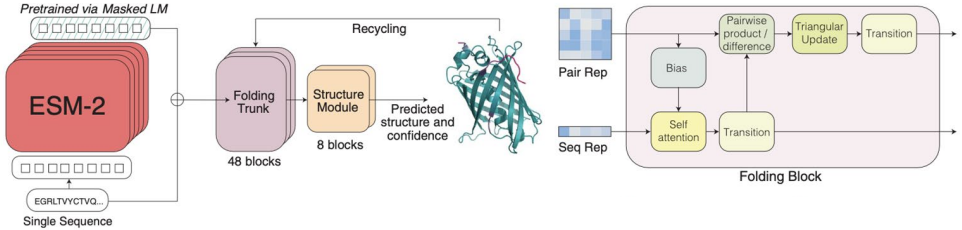


Learning the rules of protein structure at evolutionary-scale with language models

Atomic-level protein structure can now be directly predicted from amino acid sequences without relying on costly and slow multiple sequence alignment (MSA). To do so, a masked language modeling objective is used over millions of evolutionarily diverse protein sequences to cause biological structure to materialize in the language model because it is linked to the sequence patterns.

- This model, Evolutionary Scale Modeling–2 (ESM-2), is used to characterize the structure of >617M metagenomic proteins (found in soil, bacteria, water, etc). ESM-2 (schematic below) offers significant speedups compared to AlphaFold-2 (AF2): these results were produced in 2 weeks using a cluster of 2,000 GPUs.
- ESMFold is a fully end-to-end single-sequence structure predictor that uses a folding head for ESM-2. ESMFold structures (right) are of AF2-grade quality as measured by TM-score, which is the accuracy of the projection in comparison to the ground truth structure.

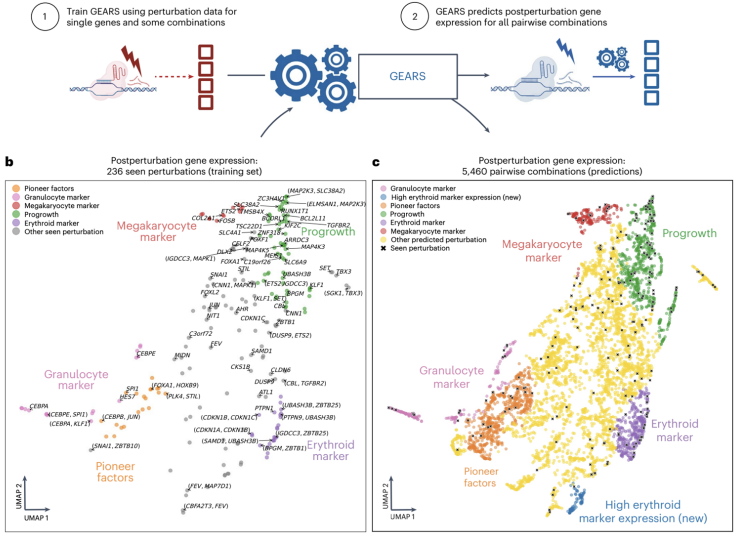




Predicting the outcome of perturbing multiple genes without a cell-based experiment

Understanding how gene expression changes as a result of stimulating or repressing combinations of genes (i.e. perturbations) is important to unravel biological pathways relevant to health and disease. But combinatorial explosion precludes us from running these experiments in living cells in the lab. Integrating deep learning with a knowledge graph of gene-gene relationships offers a solution.

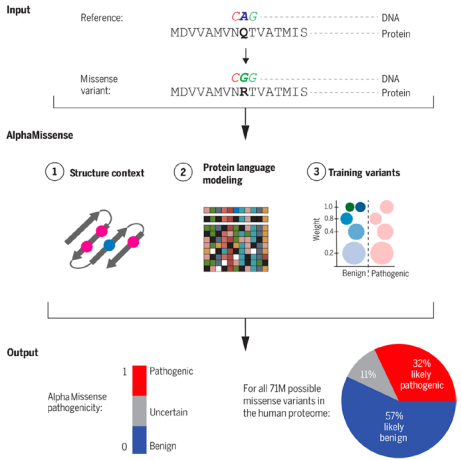
- Graph-enhanced gene activation and repression simulator (GEARS) combines prior experimental knowledge to predict the gene expression outcome given unperturbed gene expression and the applied perturbation.
- For example, GEARS can be trained on the gene expression profiles postperturbation for one-gene and two-gene experiments (b), and then be tasked with predicting the postperturbation gene expression for 5,460 pairwise combinations (c).



Pathogenic or not? Predicting the outcome of all single-amino acid changes

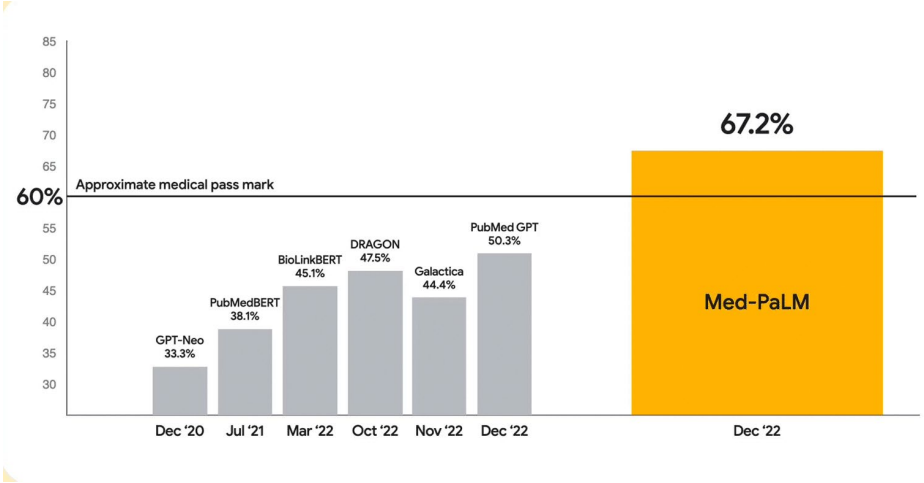
Individual changes in amino acid sequences that result from genetic variation ("missense variants") can either be benign or result in downstream problems for protein folding, activity or stability. Over 4M of these missense variants have been identified through human population-level genome sequencing experiments. However, 98% of these variants lack any confirmed clinical classification (benign/pathogenic). A new system, AlphaMissense, makes use of AlphaFold predictions and unsupervised protein language modeling to close this gap.

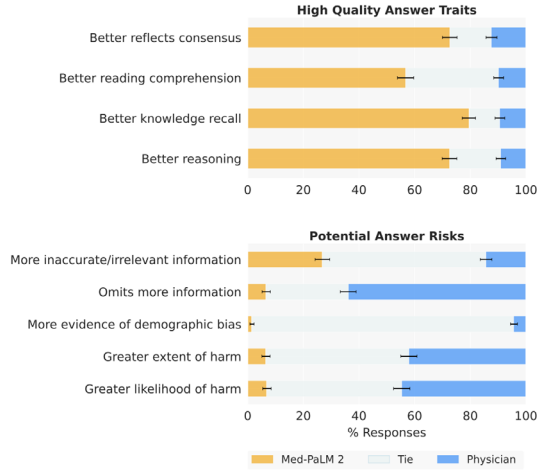
- The AlphaMissense system is built by: (i) training on weak labels from population frequency data, avoiding circularity by not using human annotations; (ii) incorporating an unsupervised protein language modeling task to learn amino acid distributions conditioned on sequence context; and (iii) incorporating structural context by using an AlphaFold-derived system.
- AlphaMissense is then used to predict 71M missense variants, saturating the human proteome. Of these, 32% are likely pathogenic and 57% are likely benign. Additional resources include all 216M possible single amino acid substitutions across the 19,233 canonical human proteins.



Google's Med-PaLM 2 language model is an expert according to the USMLE

- **A year after releasing Med-PaLM, first model to exceed a "passing" score on the US Medical Licensing Examination (USMLE), Med-PaLM 2 set a new SOTA result across more datasets as a result of base LLM improvements, medical domain finetuning and prompting strategies. In a pairwise ranking study on 1,066 consumer medical questions, Med-PaLM 2 answers were preferred over physician answers by a panel of physicians across eight of nine axes in our evaluation framework.**

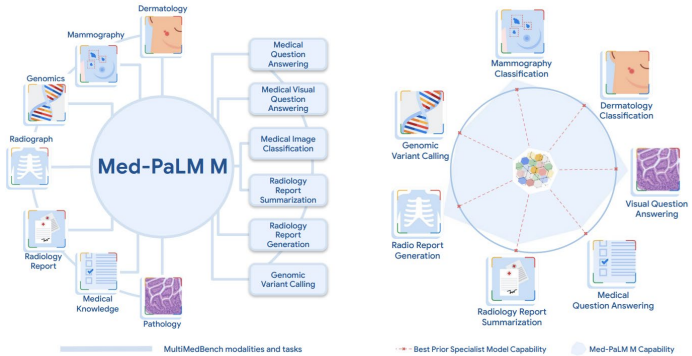


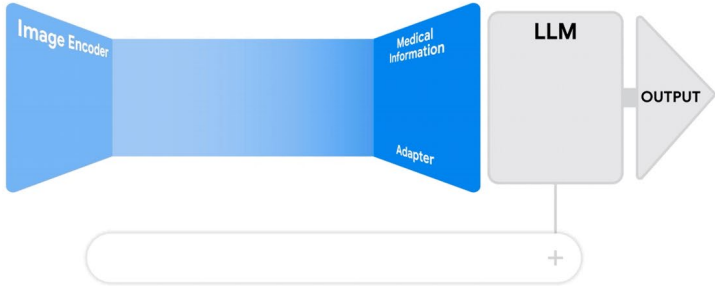


Google Health

Next, Med-PaLM goes multimodal

To bridge beyond text-based medical Q&A, Google first created MultiMedBench - a 14 task dataset that includes medical Q&A, mammography and dermatology image interpretation, radiology report generation and summarization, and genomic variant calling. This dataset is used to train a large single multitask, multimodal version of MedPaLM with the same set of model weights. The system exhibits novel emergent capabilities such as generalisation to novel medical concepts and tasks. An alternative lighter-weight approach, ELIXR, was also proposed. ELIXR grafts language-aligned vision encoders onto a fixed LLM, which requires less compute to train and shows promise across tasks including visual QA, semantic search, and zero-shot classification.

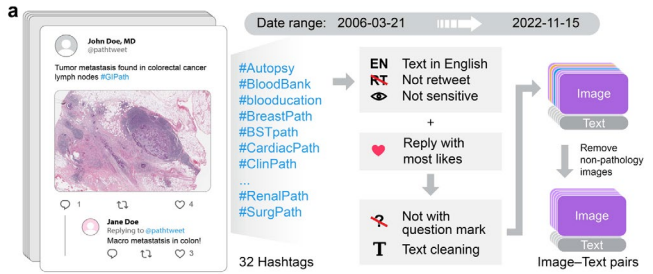


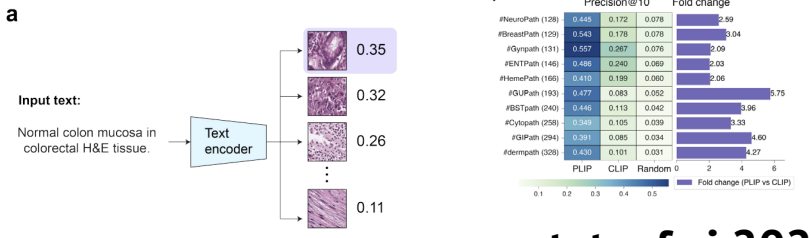


Tweet storm: a SOTA pathology language-image pretrained model from medical Twitter

It's no secret that (quality) data is king for building capable AI systems, and no more so than in domains such as clinical medicine where (quality) data is expensive to produce. This work mines text-image pairs on Twitter to create the OpenPath dataset with 200+ pathology images paired with natural language descriptors. Inspired by OpenAI's Contrastive Language-Image Pretraining (CLIP) model, the authors create P(athology)LIP.

- Like CLIP, PLIP can perform zero-shot classification on unseen data, enabling it to distin several key tissue types.
- It can also be used to improve text-to-image and image-to-image retrieval of pathology images.
- Unlike other machine learning approaches in digital pathology that are predicated on learning from a fixed set of labels, PLIP can be more generally applied and is flexible to the changing nature of diagnostic criteria in pathology.
- Compared to CLIP, PLIP has 2-6x better Precision@10.

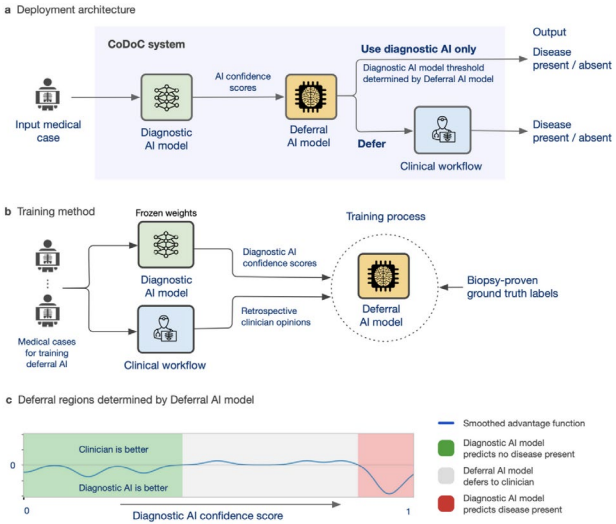




Real world-inspired clinical system design for automated medical image analysis

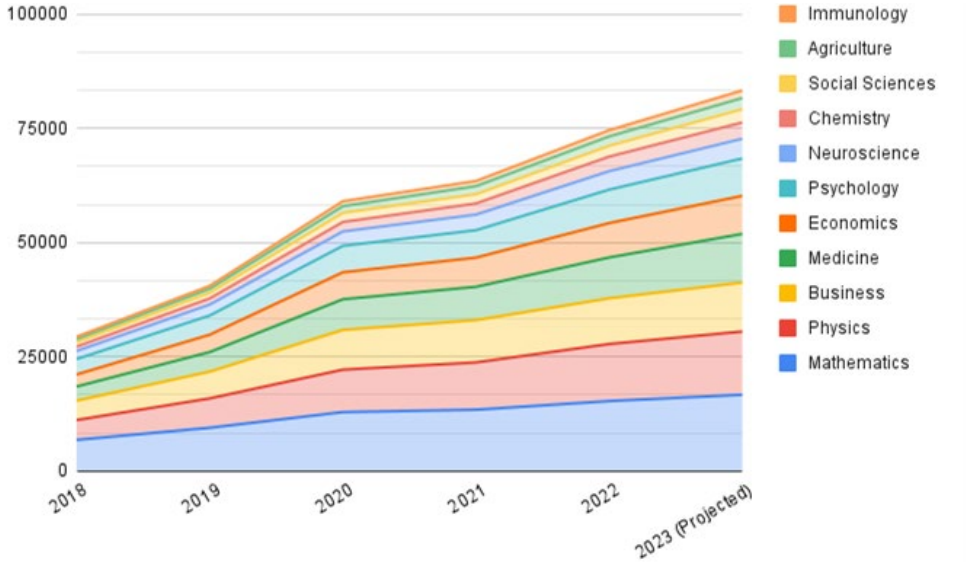
Computer vision has been shown to be useful for breast cancer screening on mammograms and tuberculosis triaging. However, to enable practical and reliable use in the clinic it is important to know when to rely on a predictive AI model or revert to a clinical workflow.

- Complementarity-Driven Deferral to Clinical Workflow (CoDoC) learns to decide whether to rely on a predictive AI model's output or defer to a clinical workflow instead.
- For breast cancer screening, CoDoC reduces false positives by 25% at the same false-negative rate compared to double reading with arbitration in the UK. Importantly, clinical workload is reduced by 66% as a result.



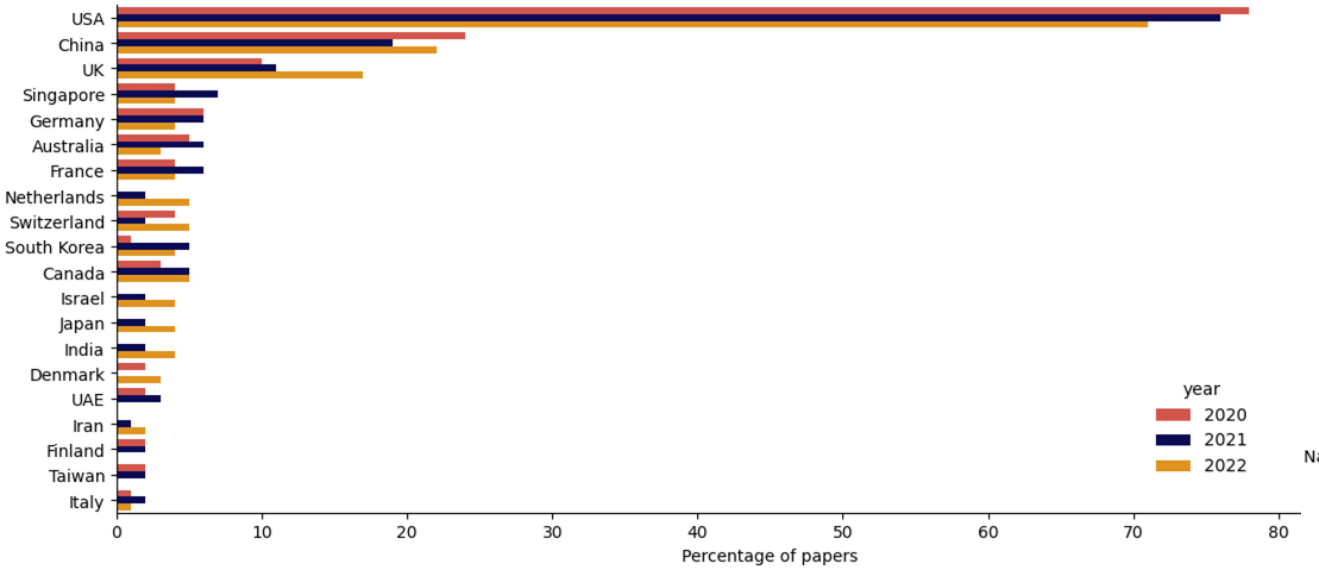
AI for science: medicine is growing fastest but mathematics captures the most attention

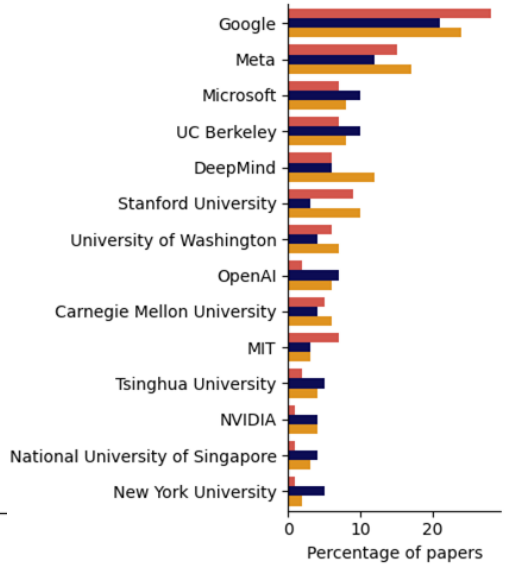
The top 20 scientific fields applying AI to accelerate progress include physical, social, life and health sciences. Out of all the highest increase in the number of publications is Medicine. We expect there to be significant research breakthroughs in the foreseeable future as a result of AI's use in the sciences.



Most impactful research comes from very few places

- **>70% of the most cited AI papers in the last 3 years have authors from US-based institutions and organisations.**

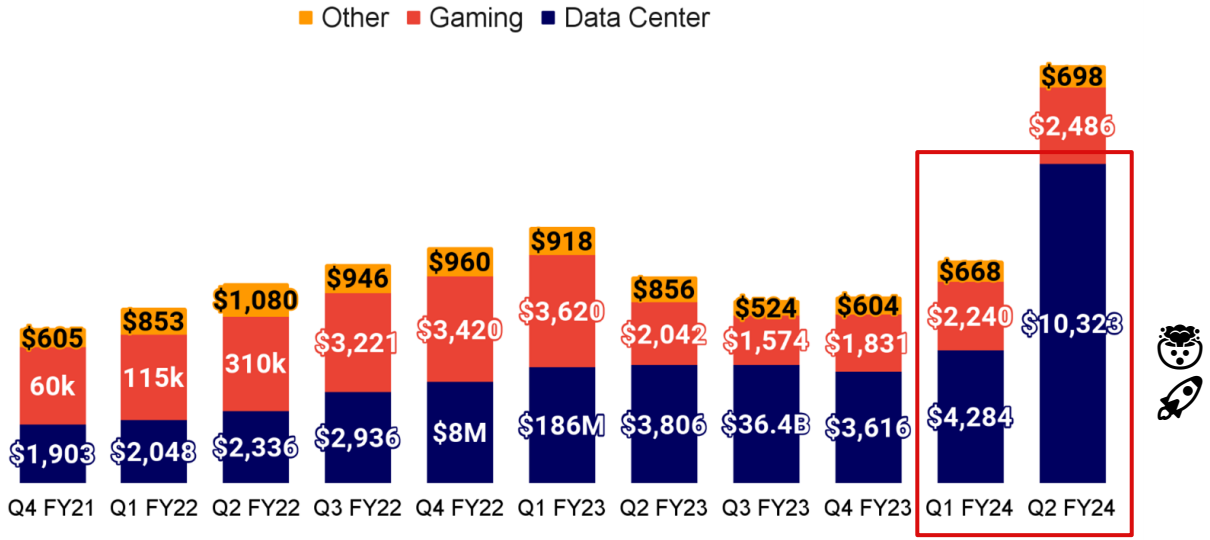




Section 2: Industry

GPU demand sees NVIDIA print blowout earnings as it enters the $1T market cap club

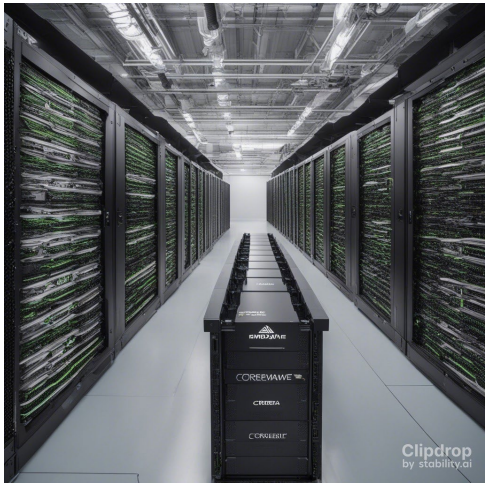
- **Q2 ‘23 data center revenue was a record $10.32B, up 141% from Q1 ‘23 and up 171% from a year ago. The stock was bearish for 2022 even though annual revenue came in at $27B, a 61.4% increase from 2021. NVIDIA now commands a $1.1T market capitalisation, up from $8.5B (130x) 10 years ago.**



Selling faster than Coachella: GPUs snapped up from upstart infra providers

CoreWeave and Lambda, two selected NVIDIA partners that build and run GPU datacenters, together have tens of thousands of GPUs in their fleet. Lambda made 9-figures $ worth of H100s available in its on-demand cloud and sold out in just over an hour. CoreWeave is one of the largest GPU operators in the market with a scale similar to several hyperscalers. The company is fully booked through the end of the year with their build schedule and are signing contracts in Q1 2024.

> First, our scale. We have over 45,000 high-end NVIDIA GPUs available on-demand in our fleet. It's not necessarily the volume that makes this significant, but rather the access it provides. Businesses rely on CoreWeave Cloud to run the compute intensive workloads that allow them to deliver client projects, hit deadlines, and accommodate end-user demand. Having a partner like NVIDIA ensures that we're able to provide the scale of resources that our clients need.



Private companies are shoring up NVIDIA GPUs and wielding them as a competitive edge

Inflection

Along with its partners CoreWeave and NVIDIA, Inflection AI is building the largest AI cluster in the world comprising 22,000 NVIDIA H100 Tensor Core GPUs. In just over a year, Inflection AI has developed one of the most sophisticated large language models in the market to enable people to interact with Pi, your Personal AI (pi.ai), in the most simple, natural way and receive fast, relevant and helpful information and advice.

ANTHROP\C

Anthropic estimates its frontier model will require on the order of 10^25 FLOPs, or floating point operations — several orders of magnitude larger than even the biggest models today. Of course, how this translates to computation time depends on the speed and scale of the system doing the computation; Anthropic implies (in the deck) it relies on clusters with "tens of thousands of GPUs."

cohere

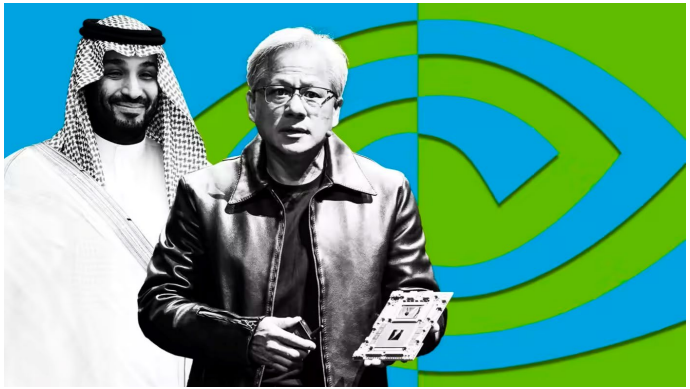
Through the partnership, Cohere will train, build, and deploy its generative AI models on OCI. OCI is uniquely positioned to run AI workloads as it delivers the highest performance and lowest cost GPU cluster technology, with scale of over 16K H100 GPUs per cluster, and very low latency and the highest bandwidth RDMA network in the cloud. This will enable the acceleration of large language models (LLM) training while simultaneously reducing the cost.

imbue

- Models. We pretrain our own very large (>100B parameter) models, optimized to perform well on internal reasoning benchmarks. Our latest funding round lets us operate at a scale that few other companies are able to: our ~10,000 H100 cluster lets us iterate rapidly on everything from training data to architecture and reasoning mechanisms.

~~Footballers~~ Compute is the new oil in Gulf States?

- **Saudi Arabia's King Abdullah University of Science and Technology (Kaust) has allegedly purchased >3,000 H100s to build a supercomputer, Shaheen III, that should be operational by end of 2023. Its LLM-focused researchers are primarily Chinese nationals that cannot access the US because their universities are restricted. Meanwhile, the United Arab Emirates' Technology Innovation Institute in Masdar City, which developed the Falcon LLM, is also said to be procuring compute resources from NVIDIA. Finally, Abu Dhabi-based G42 entered into a deal with US-based Cerebras to procure up to $900M worth of the company's Wafer-scale compute systems and build 9 interconnected AI supercomputers. There is likely much spend more to come...**

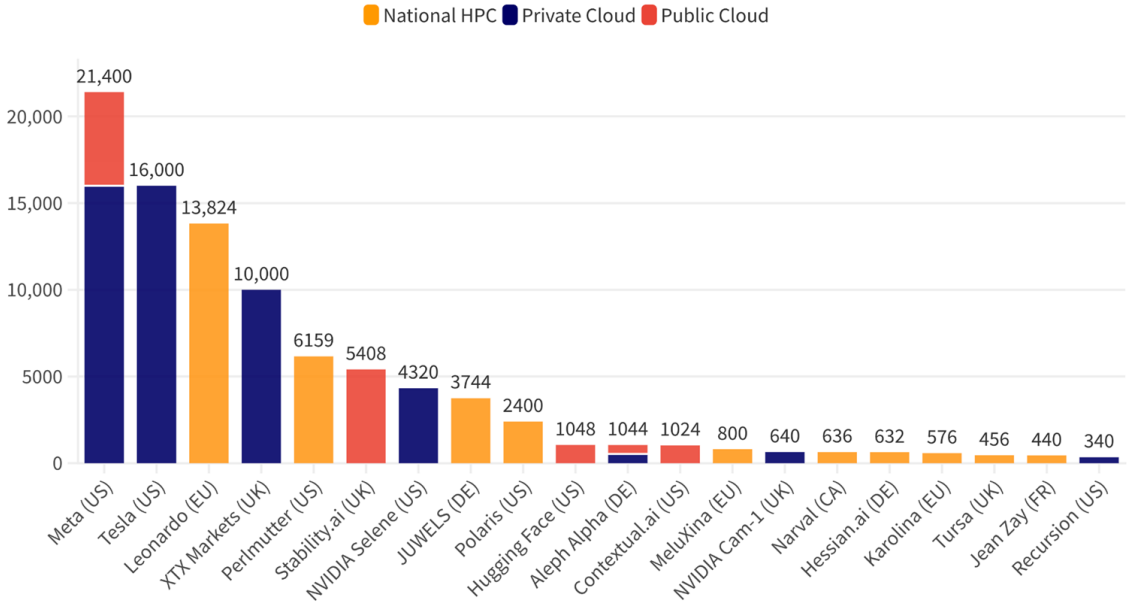


Cerebras and G42 Unveil World's Largest Supercomputer for AI Training with 4 exaFLOPs to Fuel a New Era of Innovation

Launching today with its first of nine interconnected AI supercomputers, the Condor Galaxy system will reach a combined AI training capacity of 36 exaFLOPs

Compute Index: NVIDIA A100 clusters

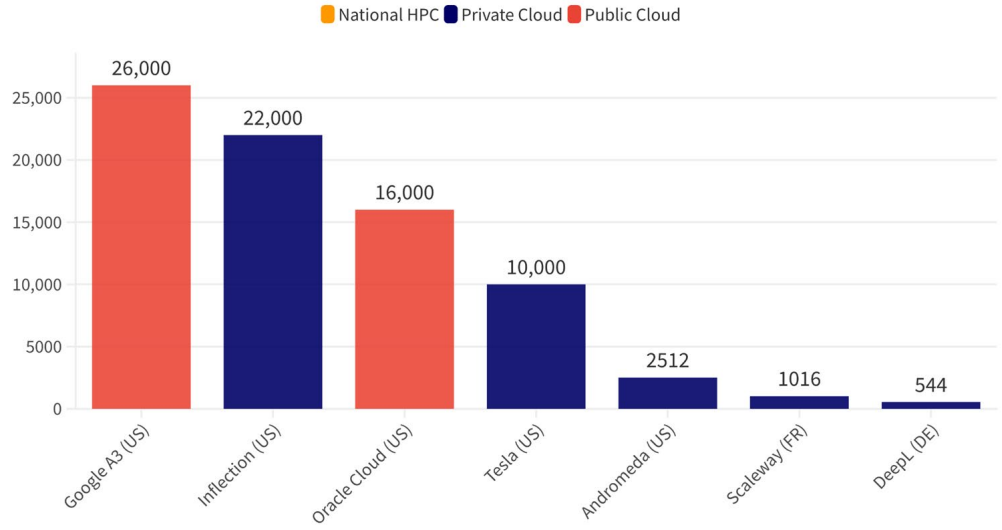
The number of large-scale NVIDIA A100 GPU clusters has grown since last year, particularly at Tesla and Stability, as well as new clusters at Hugging Face.



Source: State of AI Report Compute Index

Compute Index: NVIDIA H100 clusters

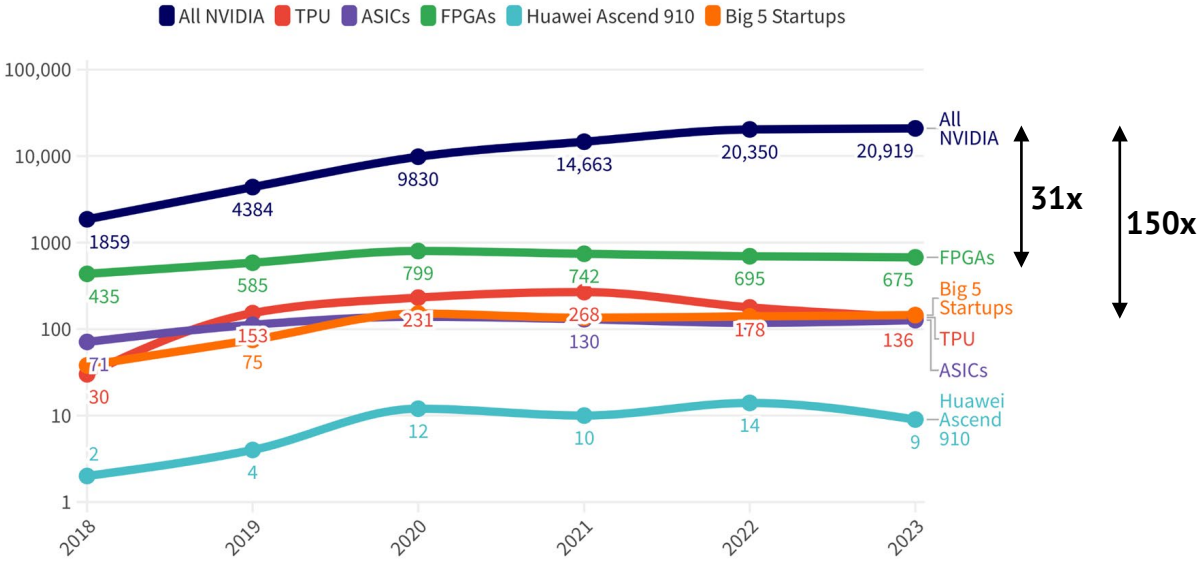
It's early days, but private and public companies are announcing new H100 infrastructure for large-scale model training. As of writing, Google and Inflection are not yet at full scale and we understand others including OpenAI, Anthropic, Meta, Character.ai, Adept, Imbue, and more have significant capacity. We expect more to come online soon.



Source: State of AI Report Compute Index

NVIDIA chips are used 19x more in AI research papers than all alternative chips combined

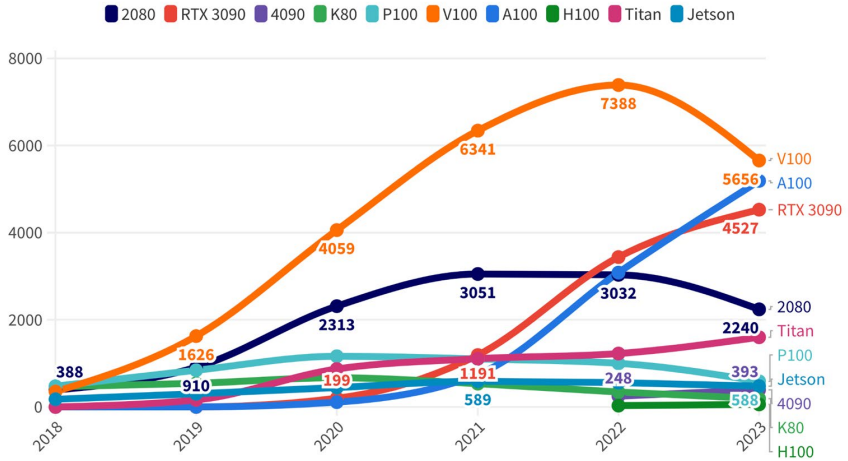
In last year's report, we began tracking the utilization of specific semiconductors in AI research papers. We found that NVIDIA chips were cited vastly more than alternatives. In 2023, NVIDIA is even more popular: 31x more than FPGAs and 150x more than TPUs.



Source: State of AI Report Compute Index and Zeta Alpha

NVIDIA chips have remarkably long lifetime value: 5 years from launch to peak popularity

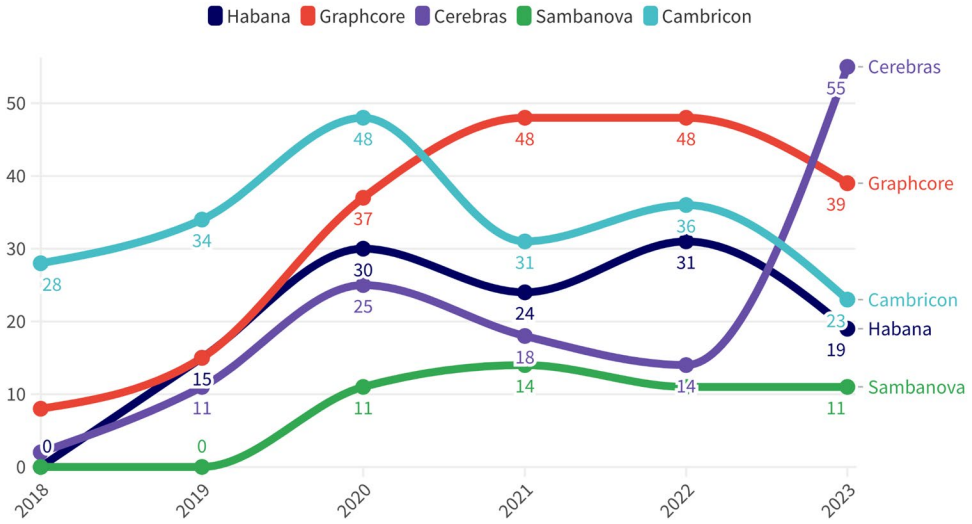
In 2023, all eyes were on NVIDIA's new H100 GPU, the more powerful successor to the A100. While H100 clusters are being built (not without hiccups), researchers are relying on the V100, A100 and RTX 3090. It is quite remarkably how much competitive longevity NVIDIA products have: the V100, released in 2017, is still the most commonly used chip in AI research. This suggests A100s, released in 2020, could peak in 2026 when the V100 is likely to hit its trough. The new H100 could therefore be with us until well into the next decade!



Source: State of AI Report Compute Index and Zeta Alpha

While NVIDIA is king, Cerebras ramps amongst the challenger crop

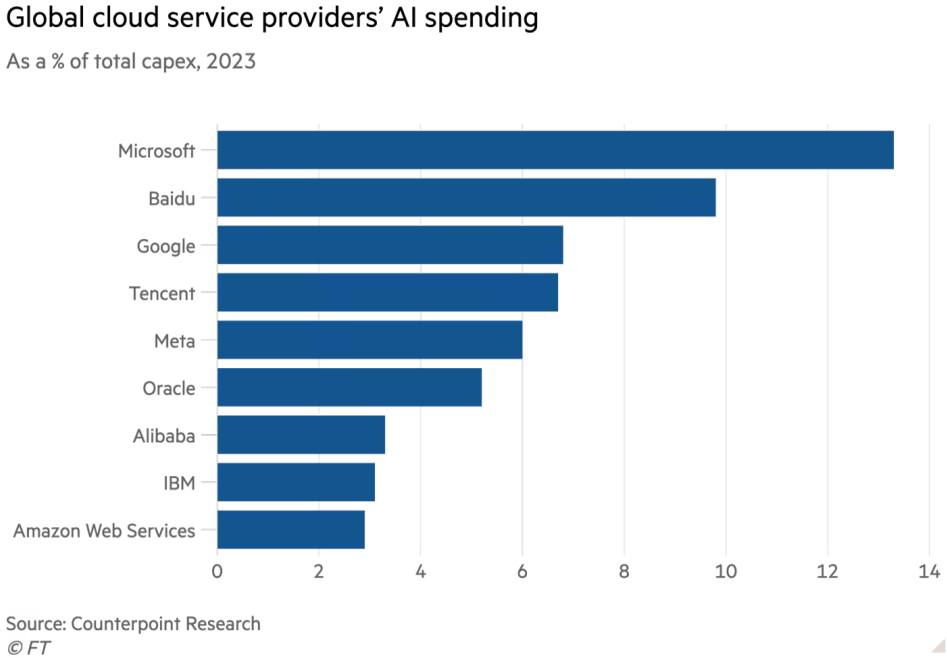
Cerebras, creators of the largest AI chip in the world, engaged in several open source model training and dataset creation projects, which helped it gain traction versus its competitors with researchers. Overall, there's still a long road to climb for NVIDIA contenders.



Source: State of AI Report Compute Index and Zeta Alpha

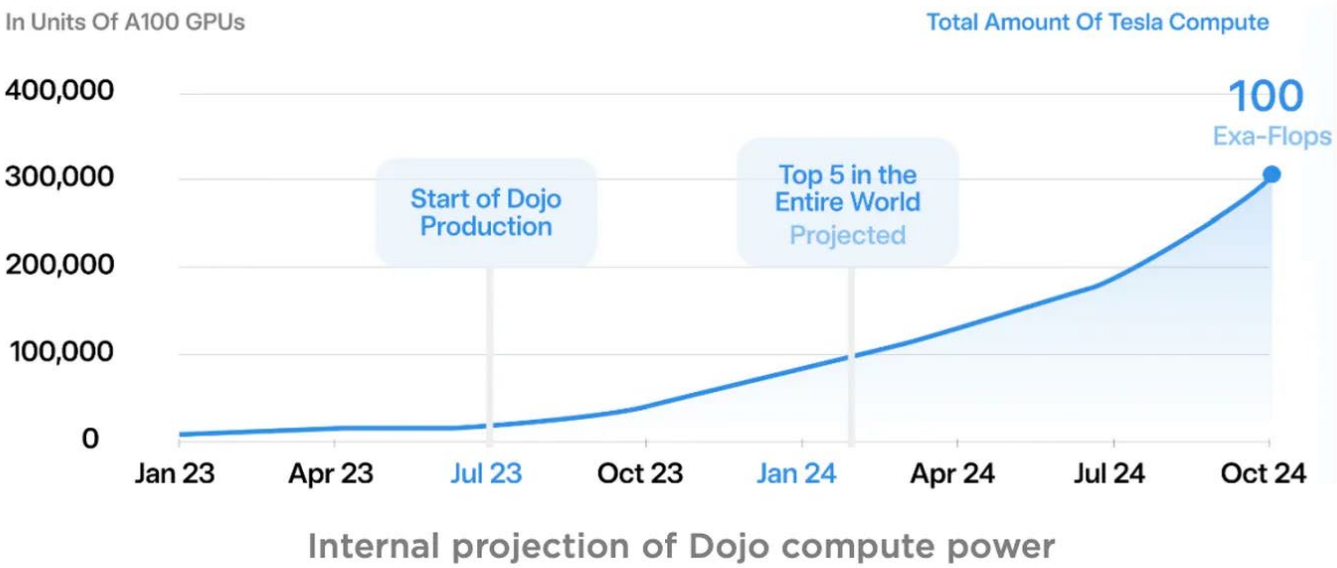
Hyperscalers scale their spending on AI as a % of total capex

- **It is also rumored that NVIDIA is to ship 1.5M and 2M H100s in 2024, up from the 500,000 expected this year.**



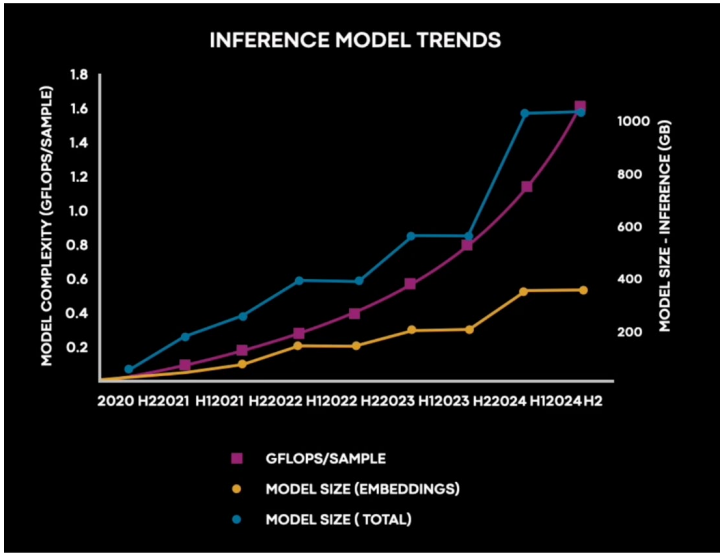
Tesla marches towards a Top-5 largest compute cluster for AI in the world

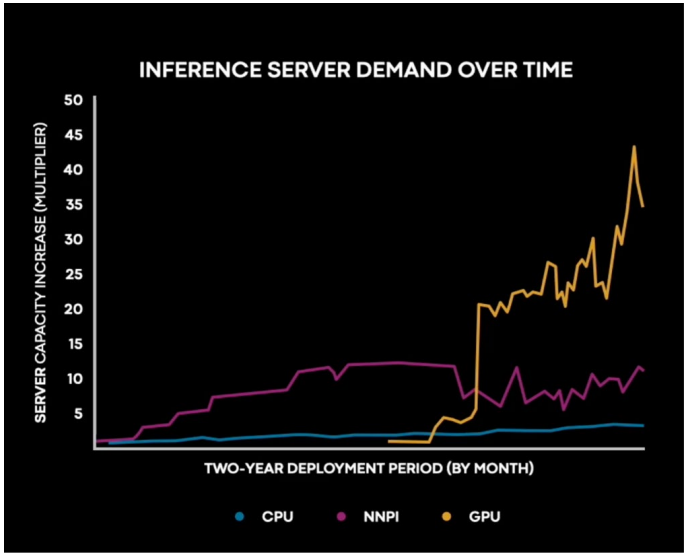
- **In our Compute Index from 2022, Tesla ranked 4th based on its A100 GPU count. As of summer 2023, the company brought online a new 10,000 H100 cluster, already making it one of the largest online to date.**



More hyperscalers develop their own inference hardware for internal AI workfloads

- **Meta announced MTIA, the company's first in-house accelerator based on open source RISC-V architecture that addresses the requirements of deep learning-based recommendation models. This is driven by growing size and complexity of models deployed in production and the slow inference speeds offered by GPUs.**

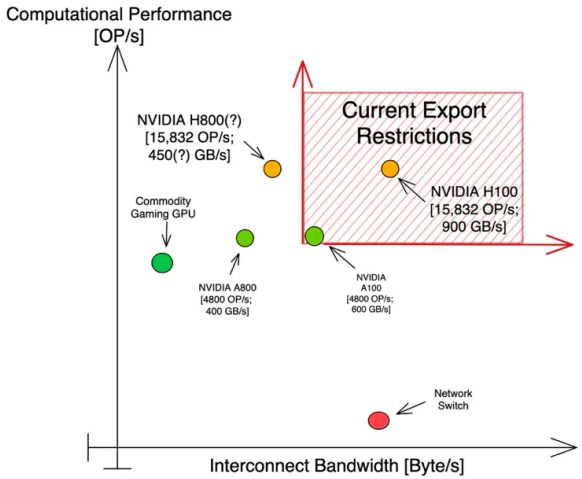




NVIDIA, Intel and AMD make Chinese-export controls proof chips

According to NVIDIA's CFO, China historically accounted for 20-25% of NVIDIA's revenue from data centre-related products (Financial Times). As a result, as the US commerce department became increasingly aggressive with export controls of AI chips, NVIDIA (and its competitors) developed chips which fly right below the export list thresholds.

- In late August 2022, NVIDIA's A100 and H100 – their most powerful chips for AI applications – were added to the US Commerce Department's export control list and became out of reach for Chinese companies. By November, NVIDIA had already started advertising the A800 and H800 chips, which it designed to be below the performance threshold set by the US ban.
- Intel did the same with a new version of their Habana Gaudi 2 chip, and AMD expressed a similar intent.
- As a result, the likes of ByteDance and Baidu have ordered >$1B worth of A800/H800 NVIDIA GPUs. There has also been reports of increasing A100/H100 GPU traffic in China, but on a much smaller scale.



Softbank re-lists Arm on the NASDAQ after its sale to NVIDIA was blocked

Back in 2020, we predicted that NVIDIA would fail to complete its acquisition of Arm. In September, Arm was relisted on the Nasdaq, achieving a valuation of $60 billion at open.

- Arm, whose IP underpins the chips in 99% of the world's smartphones, is working to reposition itself as a player in the AI market. It has partnered with self-driving car company Cruise and NVIDIA on its Grace Hopper chip (where its tech acts in a supporting role).
- However, it won't be plain-sailing. Revenue was flat over the last fiscal year and 25% comes from Arm China, an independent subsidiary required for sales into the Chinese market.
- Arm may have the potential to raise its low royalty rates per device, considering its huge market share, but will need to balance this with growing open source alternative architectures like RISC-V.
- As Arm does not sell physical chips, it has managed to swerve the impact of sanctions so far, but as the US-China chip wars escalate, there is no guarantee this will last.

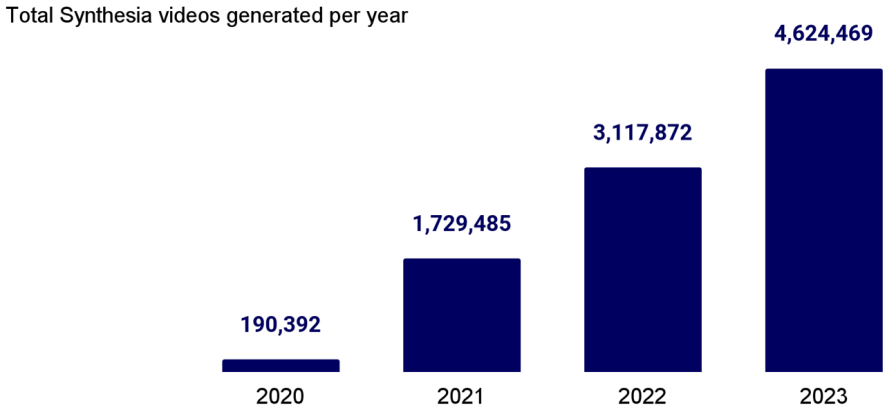
2022 Prediction: Generative AI applications grow in popularity

In 2022, we predicted: "Generative audio tools emerge that attract over 100,000 developers by September 2023." Both ElevenLabs (UK) and Resemble AI (US) exceeded that threshold. Another domain, product design, is seeing rapid integration of generative AI technology, to the benefit of fast-moving companies like Uizard.

- ElevenLabs now had over 2M registered users and is growing fast. It took half the time to to get the second million of users than the first. Users cumulatively uploaded over 10 years of audio content. Initially geared towards creators and publishers, ElevenLabs is now adapting to a large range of use-cases from AI agents, companion, entertainment, and gaming.
- Uizard, a product design company powered by AI tools, said it recorded $3.2M ARR up to July '23, which is 13x YoY. The company had crossed $1M ARR in April, and went from $1M to $3M in 3 months.

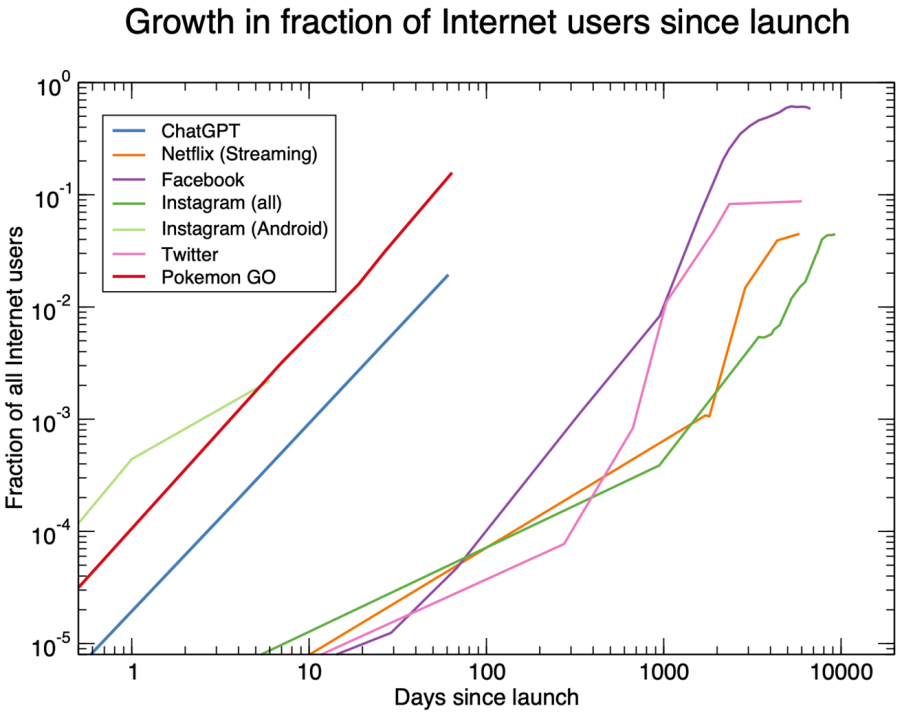
2022 Prediction: Generative AI applications grow in popularity

Video too is a rapidly advancing frontier for GenAI. Founded in 2017, London-based Synthesia launched their AI-first video creator in 2020. The system generates multi-lingual avatars that enact a script for use by consumers and enterprises alike. Once considered to be "fringe", Synthesia is now used by 44% of the Fortune 100 for learning and development, marketing, sales enablement, information security and customer service. Over 9.6M videos have been generated with the service since launch in 2020.



synthesia *2020 data starts on 1 May. 2023 data stops on 1 Sept.*

OpenAI's ChatGPT is one of the fastest growing internet products



OpenAI is now printing real money at scale...but at what cost?

- **Only 12 months, the revenue projections made by OpenAI in the lead up to its $10B fundraise were met with much scepticism. Today, the company is ripping past its targets. How long will this last? And at what cost?**

EXCLUSIVE MICROSOFT AI *Published 13 hours ago*

OpenAI Passes $1 Billion Revenue Pace as Big Companies Boost AI Spending

By Amir Efrati and Aaron Holmes

Share article

Aug. 29, 2023 3:58 PM PDT

EXCLUSIVE STARTUPS AI

OpenAI's Losses Doubled to $540 Million as It Developed ChatGPT

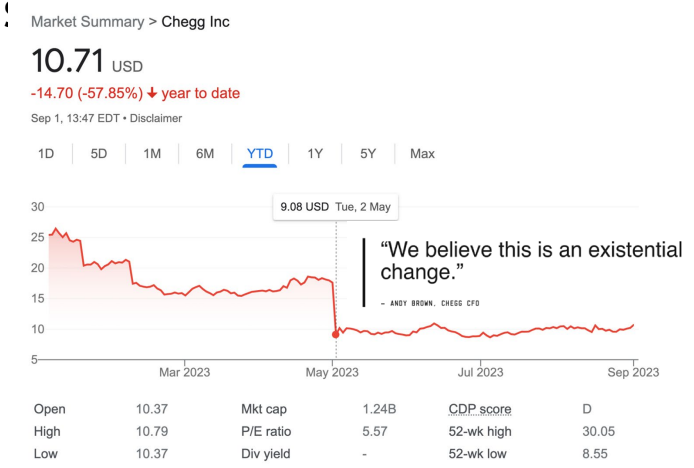
By Erin Woo and Amir Efrati

Share article

May 4, 2023 1:11 PM PDT • Comments by Josh Bersin, Brian Shilhavy, and 7 others

Feeling the ChatGPT heat: education gets hit first and Chegg is fighting back

Chegg, an NYSE-listed company focused on improving learning and learning outcomes for students, was hit hard by the launch of ChatGPT. In May 2023, the company said *"In the first part of the year, we saw no noticeable impact from ChatGPT on our new account growth and we were meeting expectations on new sign-ups."* Students that paid Chegg to practice exams and get homework feedback turned to ChatGPT instead. As a result, their share price plummeted >40%. In Chegg's August 2023 earnings call, the company said *"We've pivoted the company to harness AI to better serve learners."* They're building internal LLMs in partnership with

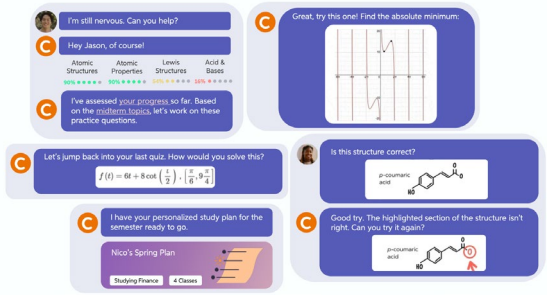


Building and Owning our own Large Language Models

- Enhances our competitive moat, lowers our costs, and allows us to train the models specifically for education
- Leverages our billions of pieces of proprietary content
- 150k subject matter experts help train the models and support accuracy in our generative experience
- We expect a significantly enhanced learning experience over generic models and tremendous value for Chegg

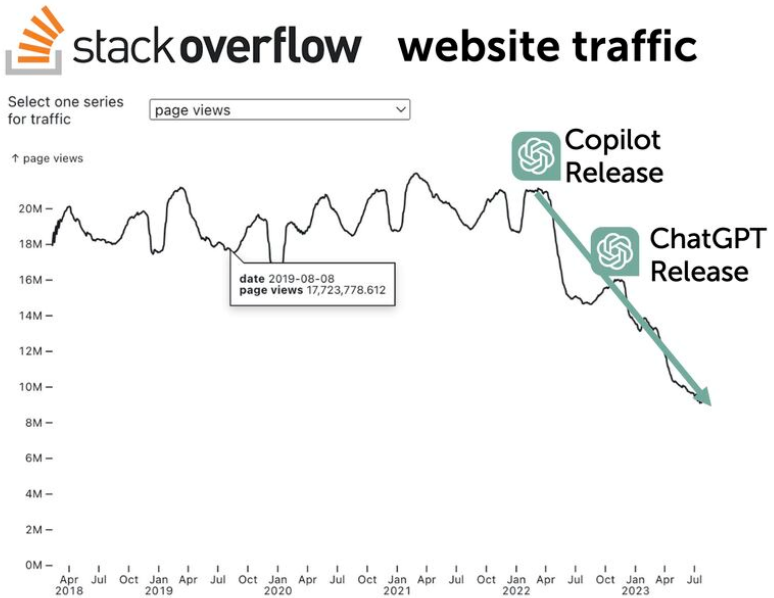
Accelerated Timeline

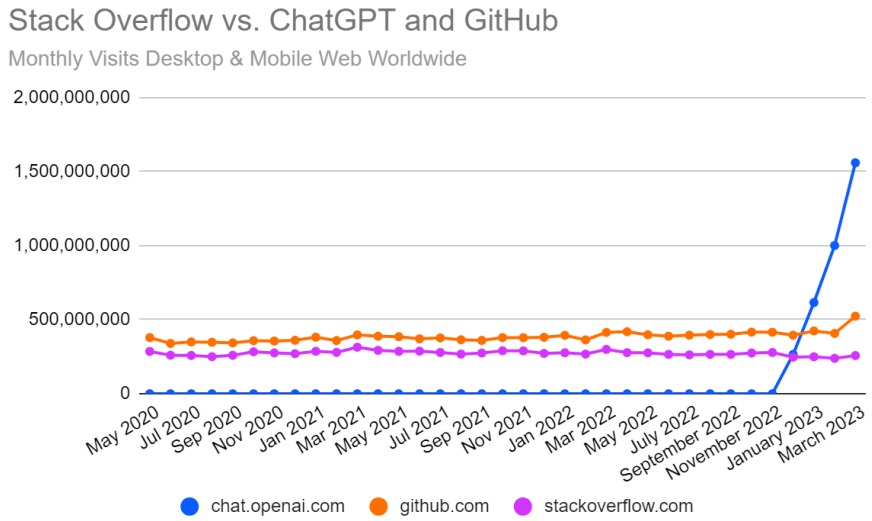
- Our partnership with Scale AI will allow us to accelerate our ability to deliver the new Chegg experience starting in the fall and rolling out over the course of the next two semesters.
- The experience will include a simple conversational user interface, personalized learning pathways, more in-depth content, and the ability to transform content into innovative study tools, such as practice tests, study guides, and flash cards.



Feeling the ChatGPT heat: coding is next...and developers are loving it!

- **Stack Overflow, a (pre-AI) de facto source for developer's to find solutions to their coding problems, placed a ban on responses generated by ChatGPT and has suffered traffic losses as a result of ChatGPT's popularity.**



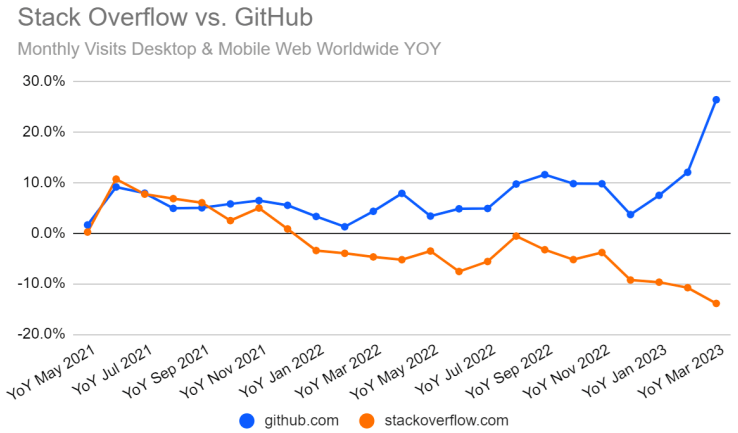


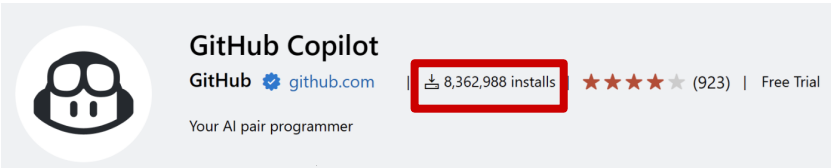
Left figure credit: Andre Retterath

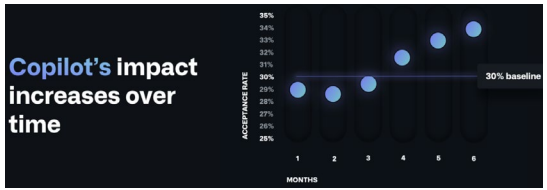
Results are in: GitHub CoPilot drives significant productivity gains for developers

If it's meant to be, it will be (no matter how long it takes). GitHub has finally launched their coding assistant, CoPilot, to hugely positive reception. The system is trained on billions of lines of code.

- In Sept 2022, GitHub ran an experiment with 95 professional developers, split them randomly into two groups, and timed how long it took them to write an HTTP server in JavaScript. This found significant productivity gains.
- In June 2023, GitHub reported data from 934,533 CoPilot users. Interestingly, productivity dips a little bit before significantly increasing as Copilot users get acquainted with the tool, and the less experienced users are the ones who benefit the most (~32% productivity gain).

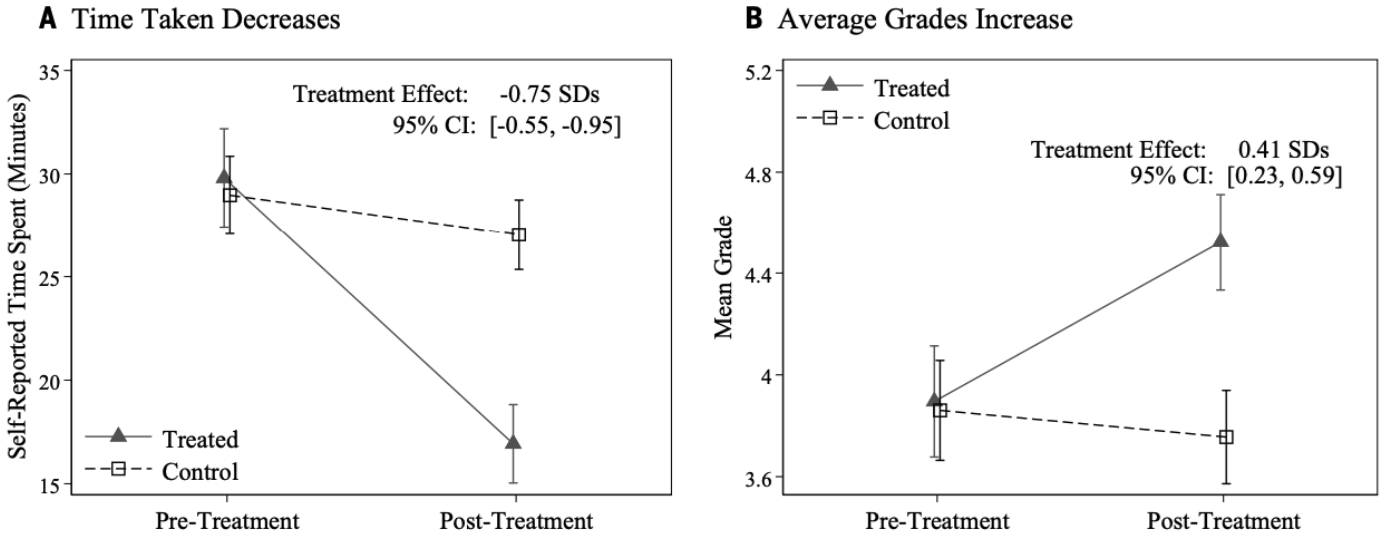






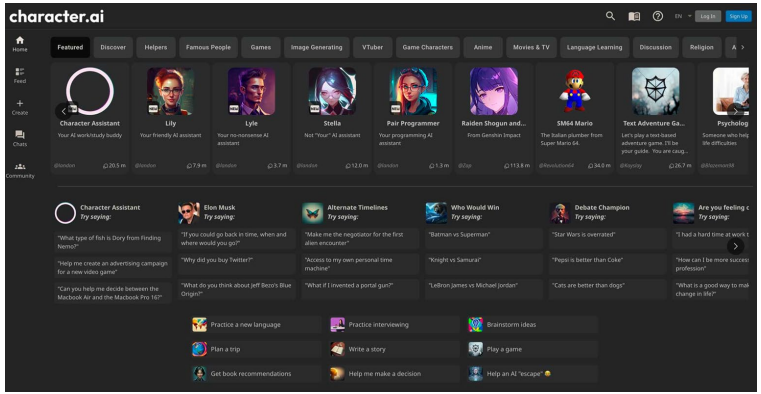
ChatGPT drives productivity in (repetitive, boring?) writing

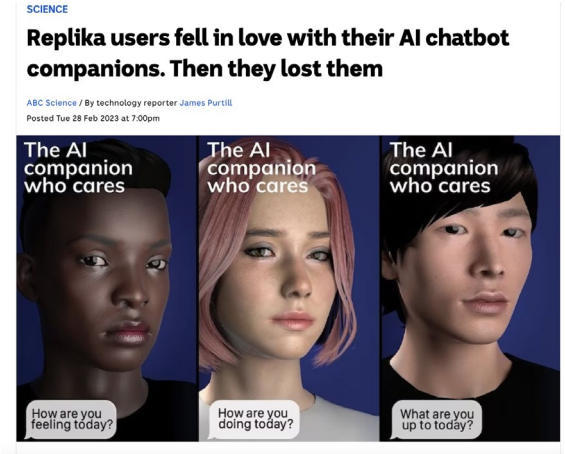
- **A new MIT study supports popular wisdom: ChatGPT helps with writing. Specifically, for "mid-level professional writing" the study showed that, compared to a control group, workers using ChatGPT took 40% less time to complete their task, and the output quality was measured to be 18% better.**



Certain less obvious GenAI use cases have also gained significant traction

We've seen huge consumer interest in users to interact with customised chatbots. A16z-backed Character.AI raised a $150M Series A and reported 200M monthly visits to its site ahead of the launch of its app. Many of their uses are benign - for example, their use as grammar tools or in fanfiction communities, but we've seen commercial and ethical challenges. We've seen reports of users developing emotional dependencies on their bots, companies struggle with the trade-off between the popularity of explicit content and its implication for their brand, as well as claims of extremist content.





Disrupted

AI chatbot company Replika restores erotic roleplay for some users

By Anna Tong

March 25, 2023 11:45 PM GMT - Updated 6 months ago

TECH

Fascist chatbots are running wild on Character.AI

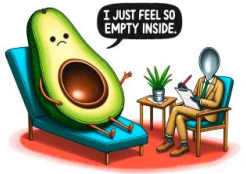
The world's second-biggest AI chat service is a hotbed of hateful and racist abuse with no meaningful moderation

Text-to-image models: Competition intensifies and integrations abound

After a breakout year in 2022 with the release of Stable Diffusion, Midjourney and Stability are still racing ahead with continuous improvements to their models. Though seemingly slower to react on the text-to-image front, OpenAI has released its best text-to-image model yet, DALL-E 3. And there are still new entrants like Ideogram, whose founders are the creators of Google's Imagen – their model notably can spell. Meanwhile we've seen countless integrations of text-to-image models in popular products, most notably on Adobe's Firefly, Photoroom, or even Discord.

- Midjourney's revenue, which had already reached $1M MRR in March 2022, is projected to reach $200M ARR in 2023. Its number of users grew from 2M to 14.8M YoY. Notably, Midjourney is integrated in Discord, where users can generate images on a Discord server. According to Discord, more than 30 million people use AI apps on its servers every month, creating more than 1 billion unique images.
- Photoroom, a French startup specializing in photo editing, said that with the introduction of generative AI in February, the company doubled its revenue and number of users over the last 6 months.

Stability's SDXL



OpenAI's DALL-E 3


Midjourney v5.2

Ideogram v0.1

But GenAI's wow effect is (so far) insufficient for users to stick around...

- **Compared to the most popular incumbent apps such as YouTube, Instagram, TikTok or WhatsApp, GenAI apps such as ChatGPT, Runway or Character.ai suffer from lower median retention and daily active users.**

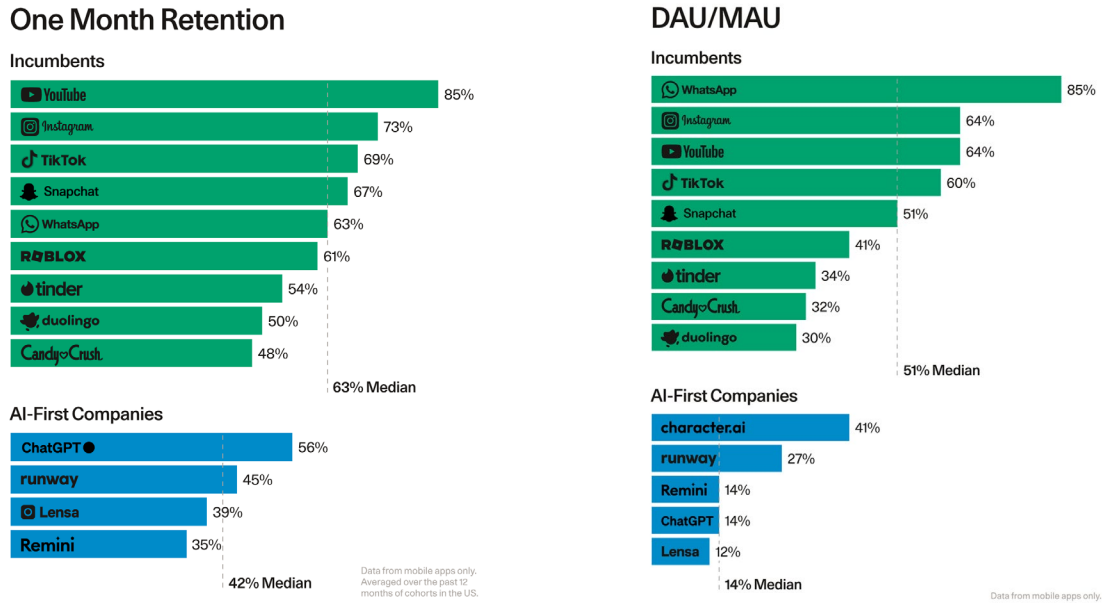
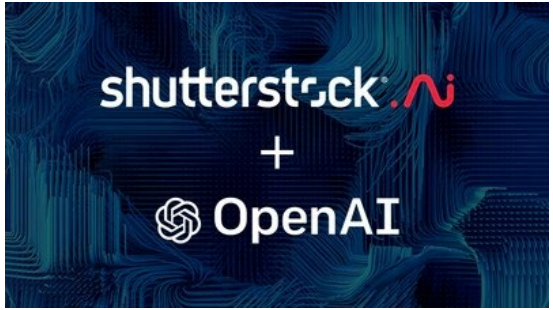


Figure credit: Sequoia Capital

2022 Prediction: A major user generated content site negotiates a commercial settlement with a start-up producing AI models (e.g. OpenAI) for training on their corpus

- **In Oct 2022, Shutterstock - a leading stock multimedia provider - announced it will work with OpenAI to bring DALL·E-powered content onto the platform. Then in July 2023, the two companies signed a 6-year content licensing agreement that would give OpenAI access to Shutterstock's image, video and music libraries and associated metadata for model training. Furthermore, Shutterstock will offer its customers indemnification for AI image creation. The company also entered into a content license with Meta for GenAI. This pro-GenAI stance is in stark contrast to Shutterstock's competitor, Getty Images, which is profoundly against GenAI as evidenced by its ongoing lawsuit against Stability AI for copyright infringement filed in Feb 2023.**



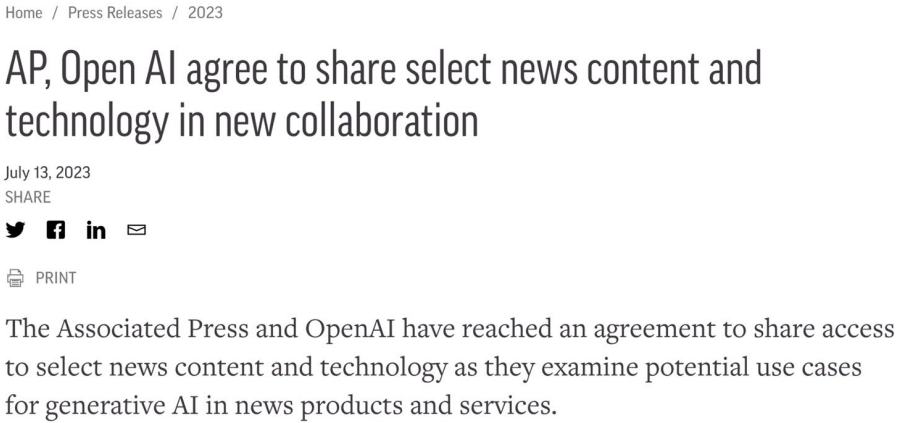
vs.

NATURE OF ACTION

1. This case arises from Stability AI's brazen infringement of Getty Images' intellectual property on a staggering scale. Upon information and belief, Stability AI has copied *more than 12 million* photographs from Getty Images' collection, along with the associated captions and metadata, without permission from or compensation to Getty Images, as part of its efforts to build a competing business. As part of its unlawful scheme, Stability AI has removed or altered Getty Images' copyright management information, provided false copyright management information, and infringed Getty Images' famous trademarks.

2022 Prediction: A major user generated content site negotiates a commercial settlement with a start-up producing AI models (e.g. OpenAI) for training on their corpus

In July 2023, OpenAI and the Associated Press (AP) entered into a licensing agreement for partial access to AP's news stories dating back to 1985. Meanwhile, AP will gain access to OpenAI technology and product expertise to explore generative applications. Although AP doesn't have LLM-based applications in production, it has made use of AI systems to create automated corporate earnings and sporting event recaps.



Home / Press Releases / 2023

AP, Open AI agree to share select news content and technology in new collaboration

July 13, 2023

SHARE

PRINT

The Associated Press and OpenAI have reached an agreement to share access to select news content and technology as they examine potential use cases for generative AI in news products and services.

US Courts set precedent for AI-generated content being unsuitable for copyright protection, but then another on fair use

A US District Court has reaffirmed the long-standing principle that human authorship is needed for copyright protection. While appeals are likely, important precedent may now have been set.

- The US District Fort for the District of Columbia rejected a claim from Stephen Thaler that the 2012 image “A Recent Entrance to Paradise” (on the right) was worthy of copyright protection.
- The Copyright Office, however, has established an initiative to examine the impact of AI on copyright law and has released new copyright guidance, covering literary, visual, audiovisual, and sound. It stipulates that any artwork needed a human author and that applications needed to specify where AI was used.
- More challengingly for providers, in May 2023 ruling in a copyright case over a 1981 portrait of Prince, the US Supreme Court applied a new, stricter interpretation of what constitutes as ‘transformative’ under fair use. This could well make the scraping of books and artwork for models’ training data legally riskier.

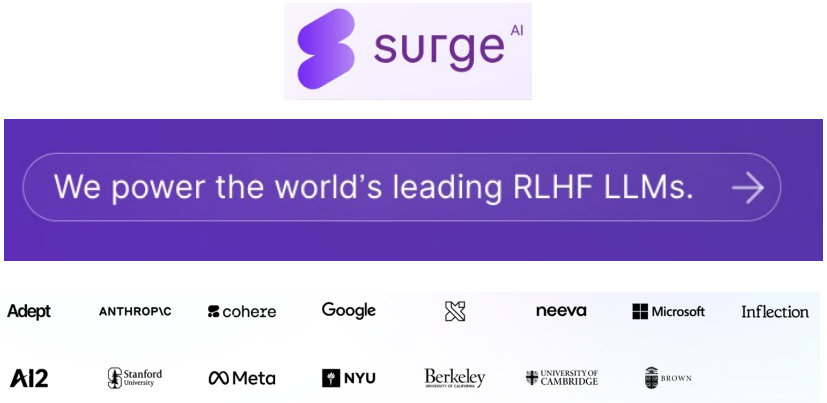
But cases continue to be fought in multiple jurisdictions about copyright infringement

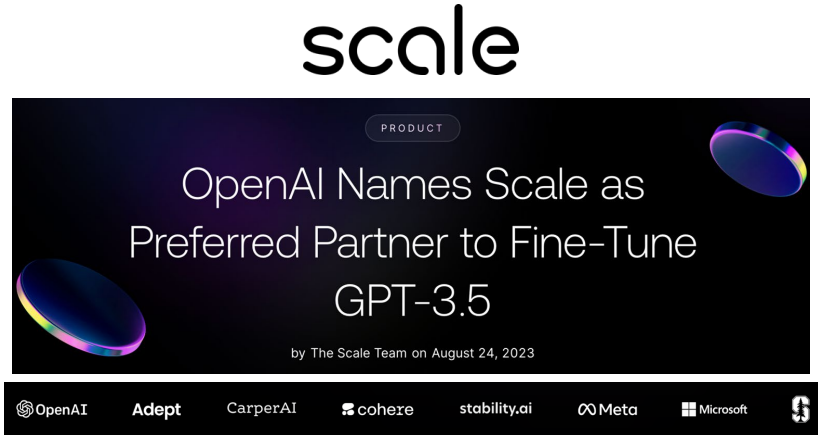
Cases featuring the major text and image generation are being fought in the UK and US. While the companies contend that they are engaging in fair use or freedom of expression, there are signs that trouble may lie ahead.

- In the UK and US, Getty Images is suing Stability, arguing that Stability had copied millions of photographs from its collection, altered or removed copyright information, and accused Stable Diffusion of generated images that bear a modified version of the Getty Images watermark.
- OpenAI and Meta are facing lawsuits claiming that ChatGPT and LLaMa on the grounds that they did not consent to their copyrighted books being used in training datasets. The New York Times is said to be mulling a similar suit against OpenAI. Three artists are suing Stability, DeviantArt and Midjourney for using their artwork to train an image generator that creates “infringing derivative works”.
- The UK has a text and data mining exception to copyright law, but this only extends to non-commercial use; plans to widen this exemption have been shelved. The EU had a similar exemption, but the AI Act states that foundation model providers will have to provide summaries of copyrighted material used to train their models (which could prove technically challenging)
- Microsoft has moved to reassure users of their Copilot tools that the corporation will assume any legal risks in the event of any copyright claims.

From labels to preferences

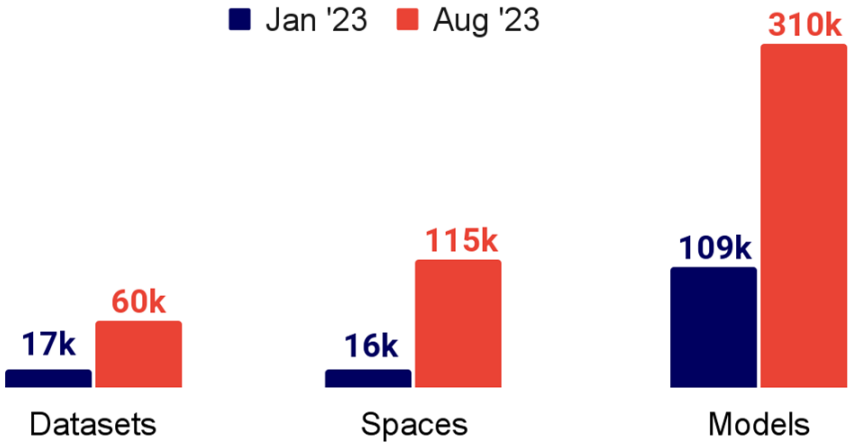
- **As instruction fine-tuning and RLHF became the default method to fine-tune and align language models, companies offering labeling services like Scale AI and Surge HQ stand to register exceptional growth from the exploding popularity of LLMs. Both companies bolster an impressive list of customers, from AI startups to large corporate clients to leading labs in LLM research. Scale AI was last valued at $7.3B back in 2021, pre-Stable Diffusion and the ChatGPT frenzy.**





Open source AI is on a tear at a time when incumbents push for closed source AI

Hugging Face, the now 7-year old company that has firmly become the town hall for open source AI, is seeing significant momentum as the community vies to keep AI models and datasets accessible to all. Over 1,300 models have been submitted to their Open LLM Leaderboard in a few months and >600 million model downloads in August 2023 alone. These models are exposed on Spaces as web applications built with tools such as Gradio or Streamlit, enabling broader accessibility and rapid prototyping. Monthly active Gradio users has grown 5x from 120k (Jan '23) to 580k (Aug '23).



Monolithic LLMs or specialised application-dependent LLMs?

Databricks acquired MosaicML for $1.3B in order to help companies build (most likely finetune) their own LLMs. Rather than a single monolithic model that knows everything, the future could belong to a set of specialised models trained on enterprise data or for specific tasks.

- Prior to the acquisition, Mosaic showed impressive engineering feats like training Stable Diffusion from scratch for <$50k (8x reduction from the original) and building sota LLMs with long context length.
- The deal marked a major moment in the short history of generative AI frenzy.
- Snowflake had a similar strategy: together with Azure, it provide customers with access to OpenAI's models.

Once ignored by major pharma companies, AI is moving front and center for some

- **mRNA vaccine leader, BioNTech acquired AI company InstaDeep for €500M, while Sanofi goes "all in" on AI, Merck enters into new deals with AI-first drug company, Exscientia, worth up to $674M and AstraZeneca partners with Verge Genomics in a deal worth up to $840M.**

Press Release sanofi

Sanofi "all in" on artificial intelligence and data science to speed breakthroughs for patients

Paul Hudson
CEO, Sanofi
"Our ambition is to become the first pharma company powered by artificial intelligence at scale, giving our people tools and technologies that focus on insights and allow them to make better everyday decisions. The use of artificial intelligence and data science already support our teams' efforts in areas such as accelerating drug discovery, enhanced clinical trial design, and improving manufacturing and supply of medicines and vaccines. We have just scratched the surface as to how we embrace these disruptive technologies to achieve our ambition of transforming the practice of medicine."

Exscientia Announces AI Drug Discovery Collaboration with Merck KGaA, Darmstadt, Germany

9/20/2023

Collaboration will leverage Exscientia's precision design capabilities to focus on previously unsolved drug design challenges

Exscientia is eligible to receive up to $674 million in discovery, development, regulatory and sales-based milestones for three projects, in addition to single to double digit royalty payments on net sales

Up to $113 million of potential milestone payments in the discovery phase, with $20 million upfront at initiation for three projects

BIONTECH InstaDeep™

The acquisition supports BioNTech's strategy, aiming to build world-leading capabilities in AI-driven drug discovery and development of next-generation immunotherapies and vaccines to address diseases with high unmet medical need. InstaDeep will operate as a UK-based global subsidiary of BioNTech. In addition to BioNTech-focused projects, InstaDeep will continue to provide its services to clients around the world in diverse industries, including in the Technology, Transport & Logistics, Industrial, and Financial Services sectors. The transaction adds approximately 290 highly skilled professionals to BioNTech's workforce, including teams in AI, ML, bioengineering, data science, and software development.

The total consideration to acquire the remaining InstaDeep shares, excluding the shares already owned by BioNTech, amounts to approximately €500 million in cash, BioNTech shares, and performance-based future milestone payments.

Alexion, AstraZeneca Rare Disease has entered a multi-target partnership agreement with Verge Genomics to detect new drug targets for rare neurodegenerative and neuromuscular ailments leveraging artificial intelligence (AI).

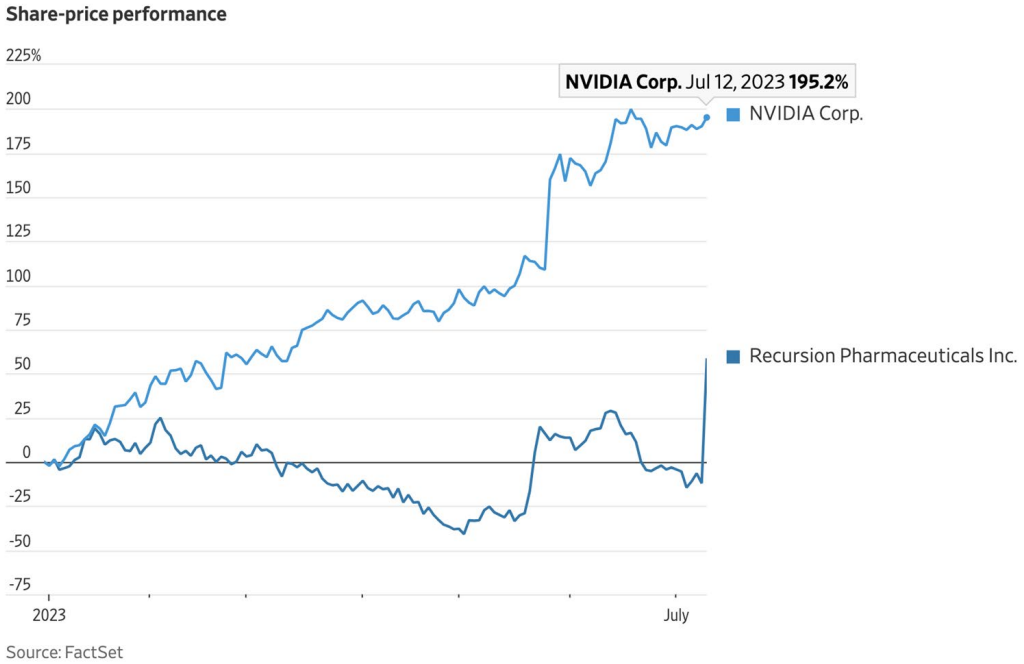
As per the four-year deal, Alexion will make upfront, equity and near-term payments of up to $42m to Verge.

Verge is entitled to receive a total of $840m in milestone payments under the agreement, apart from potential downstream royalty payments.

Under the partnership, the parties will utilise the CONVERGE full-stack platform of Verge that merges predictive human tissue datasets with machine learning for identifying new targets with an increased clinical success potential.

NVIDIA continues its share price performance tear and blesses its partners too

- **On the day of NVIDIA's $50M investment announcement into Recursion Pharmaceuticals, the latter's share price surged 80% to create an additional $1B of market value. Such a reaction demonstrates the AI fever.**



DeepMind to Google DeepMind back to DeepMind and now to Google DeepMind...v2!

The pioneering AI company, DeepMind, is now at the forefront of Google's counteroffensive in generative AI following its merger with Google Brain.

2010

2014

2015

2023

Company

Announcing Google DeepMind

April 20, 2023

DeepSpeech 2: The early masters of scale

In 2015, Baidu's Silicon Valley AI Lab introduced a fully end-to-end deep learning based system for speech recognition. The work did away with hand-crafted feature-based pipelines and heavy use of computation: *"Key to our approach is our application of HPC techniques, resulting in a 7x speedup over our previous system [...] Our system is competitive with the transcription of human workers when benchmarked on standard datasets."* A 2017 paper from the same lab, *"Deep learning scaling is predictable, empirically"* demonstrated early evidence for "scaling laws", which now underpins the large-scale AI we see and use today. Many DS2 authors have gone onto be founders or execs of leading ML companies, often leading their large scale efforts in language modeling and related fields.

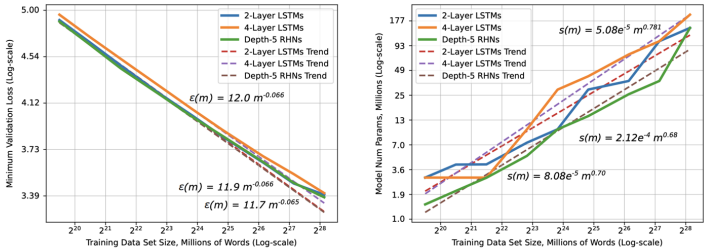


Figure 2: Learning curve and model size results and trends for word language models.

Deep Speech 2: End-to-End Speech Recognition in English and Mandarin

Baidu Research – Silicon Valley AI Lab*

Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, Zhenyao Zhu

Meta · Reka · LAMINI · NVIDIA · ADEPT

Attention is all you need... to ~~build~~ raise billions for your AI startup

- All ~~but one~~ authors of the landmark paper that introduced transformer-based neural networks have left Google to build their own startups. The Transformers Mafia have collectively raised

Capital raised in 2023 alone

	$10.3B
ADEPT	$350M
cohere	$270M
character.ai	$150M
Inceptive	$100M

Autonomous driving meets GenAI

- **GAIA-1 is a 9-billion parameter generative world model developed by Wayve for autonomous driving. It leverages video, text and action inputs to generate realistic driving scenarios and offers fine-grained control over ego-vehicle behaviour and scene features. It shows impressive generalisation abilities to ego-agent behaviours that are outside of the training set and controllability of the environment through text, making it a powerful neural simulator useful for training and validating autonomous driving models.**

WAYVE

Autonomous rides are now commercial (in California)

Waymo and Cruise have been granted permission to launch paid 24/7 autonomous driving services in San Francisco. Previously paid rides were only possible when a driver was in the vehicle for monitoring.

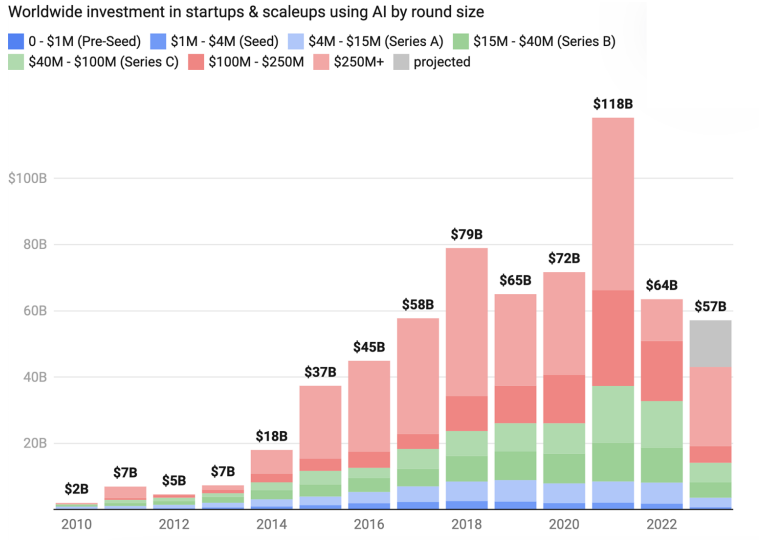
- This is a major moment for autonomous driving. Approval from the California Public Utilities Commission was the last in a series of approvals that took years to obtain. Waymo's CEO Tekendra Mawakana stated that the permit "marks the true beginning of our commercial operations in San Francisco".
- However the jury is still out on the economics of a driverless taxi service versus trucking and logistics. Waymo paused their autonomous trucking service at the end of July, while others (Aurora for instance) are prioritizing it over robotaxis.
- Former Argo AI leaders founded Stack AV, an autonomous trucking startup which raised $1B Series A from Softbank.

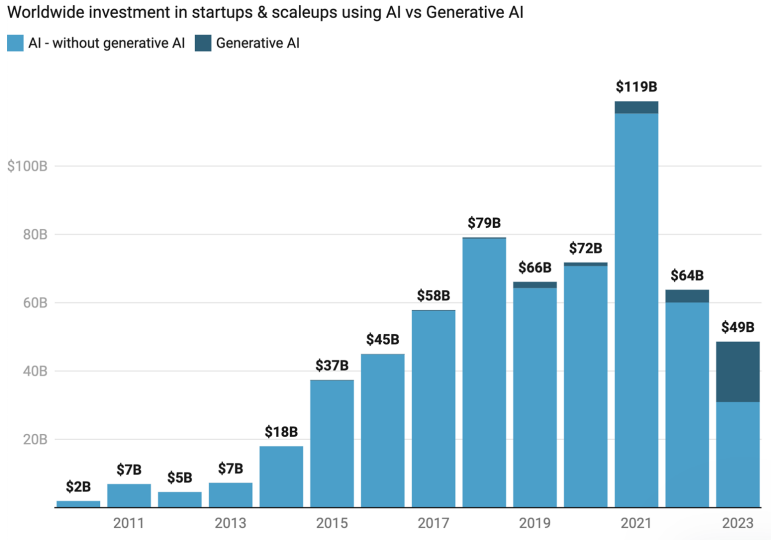
Photo by Justin Sullivan/Getty Images

WAYMO cruise STACK

"GenAI" is the new "new" thing: AI investments are stable vs. 2022, powered by GenAI

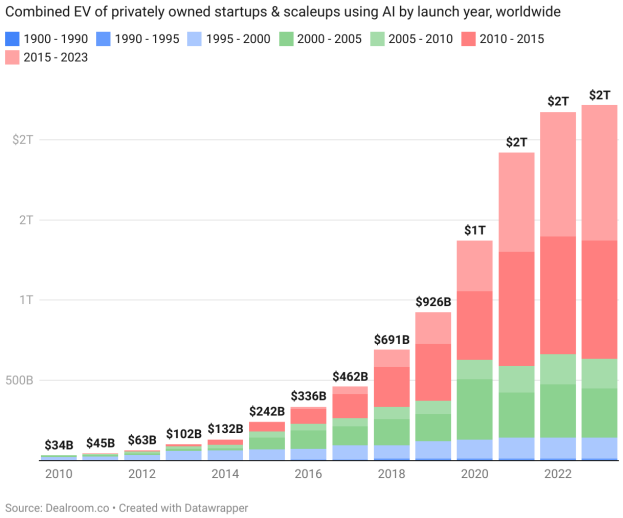
Funding for startups using AI H1 2023 was nearly on par with H1 2022...without capital pouring into GenAI, overall AI investments would have suffered a 40% drop compared to last year vs. 54% drop across all startups.

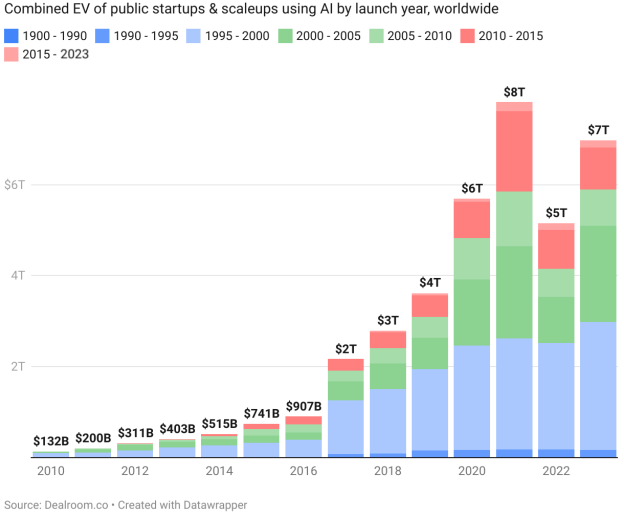




Trillions of value: The combined enterprise value of private and public companies using AI

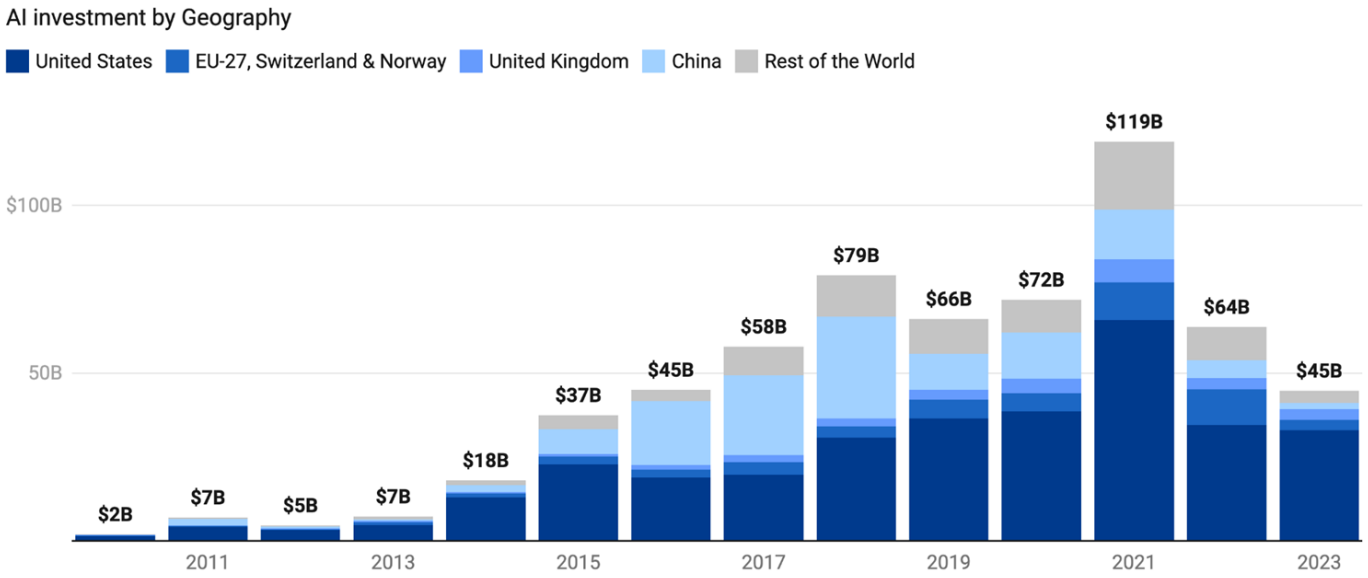
Public valuations dropped by ⅓ after 2021, but are on their way to recovering, while private market valuations remain stable and are yet to see a haircut. Notably, 50% of the S&P 500 gains in 2023 were driven by "The Magnificent Seven": Apple, Microsoft, NVIDIA, Alphabet, Meta, Tesla and Amazon as key drivers and beneficiaries of AI acceleration.





US AI companies absorb 70% of global private capital in 2023, up from 55% in 2022

Funding to private US and UK AI companies is steady YoY, while capital for European AI companies drops >70%.

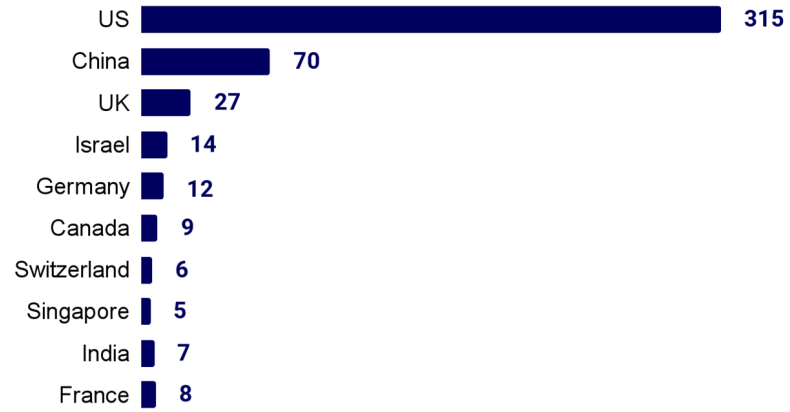


dealroom.co *Data as of 29 Sept 2023*

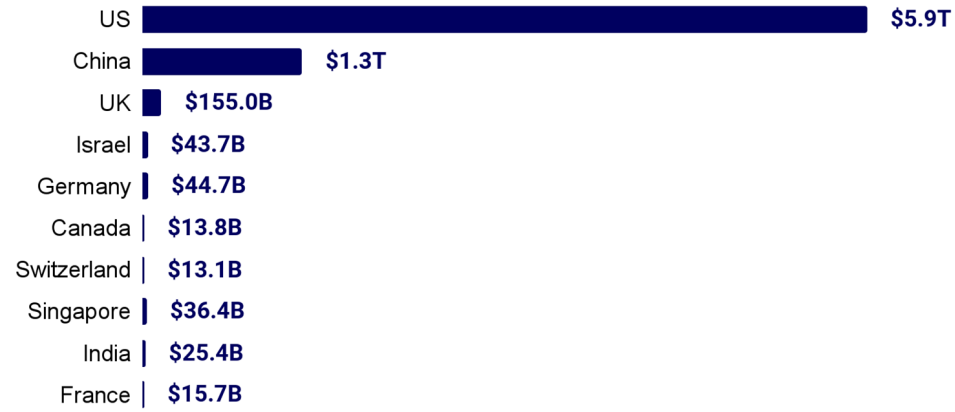
The US continues to lead by number of AI unicorns, followed by China and the UK

The from 2022 continues: the US grows its unicorn count to 315 from 292 and total enterprise value to $5.9T from $4.6T. The UK adds 3 more unicorns but sees cumulative enterprise value regress to $155B from $207B.

Cumulative number of AI unicorns by country



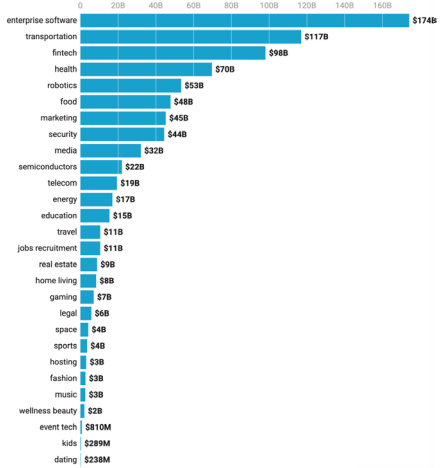
Cumulative enterprise value of AI unicorns by country



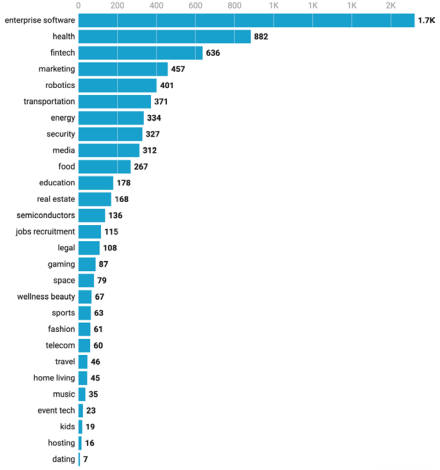
dealroom.co *Data as of 19 Sept 2023*

Enterprise software, fintech and healthcare are the most invested AI categories globally

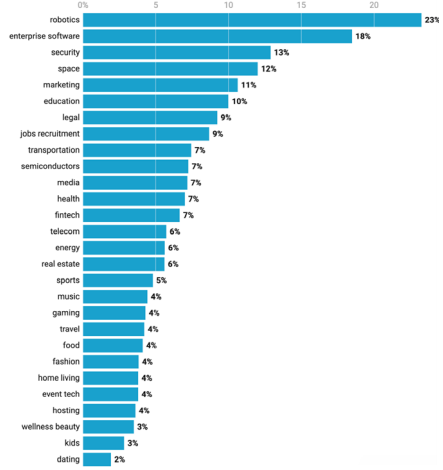
$ invested in AI categories 2010-23



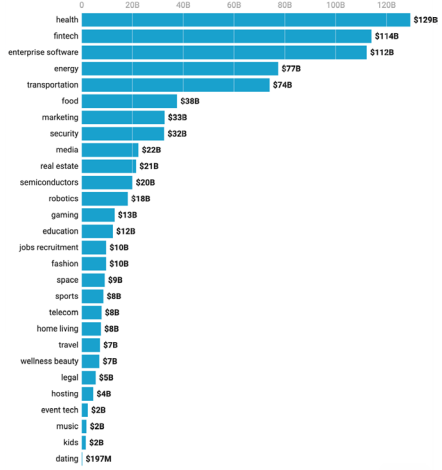
Deal volume in AI categories 2022-23



% of deals for AI startups

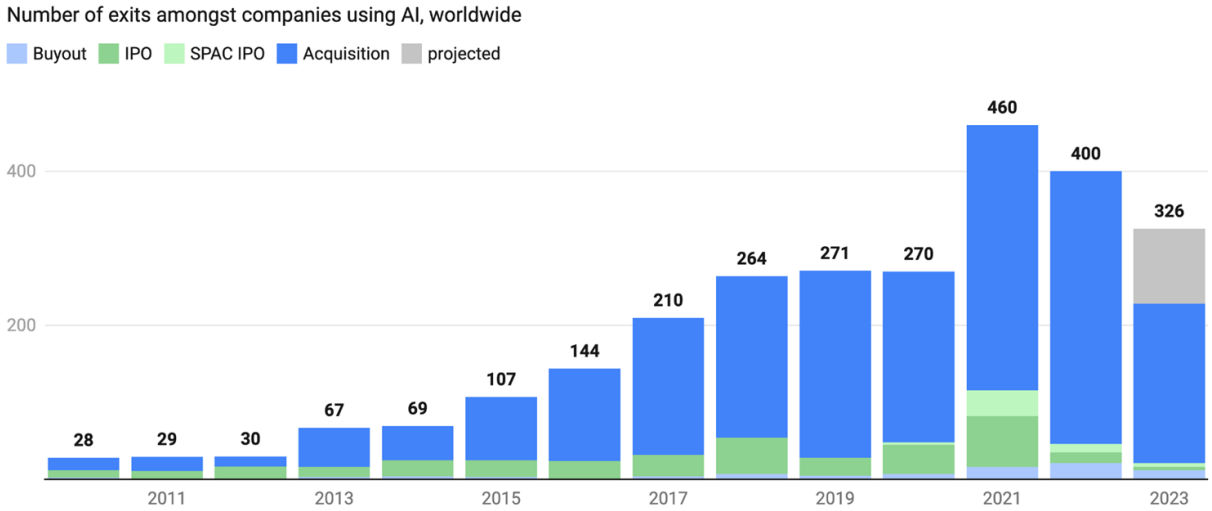


Deal volume in AI categories 2022-23



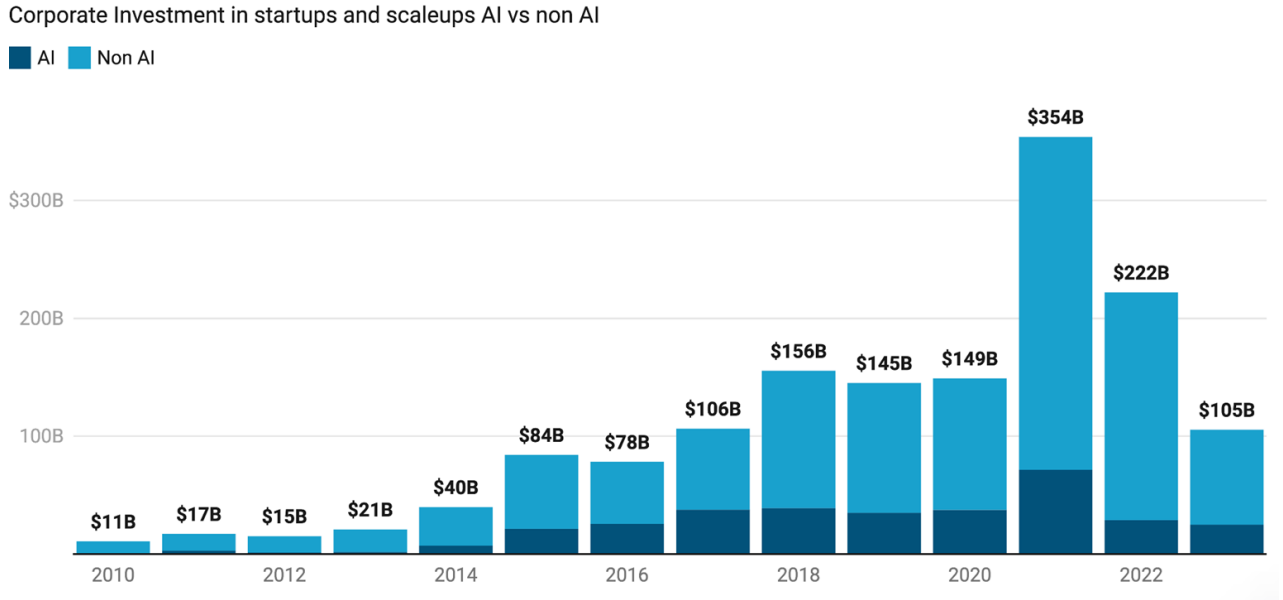
Although IPOs dried up in 2023, the M&A market continues to stay strong

- **Not much public market activity outside of a few SPACs (e.g. Arrival, Roadzen, Triller) vs. 98 in 2022. However, there were several large acquisitions MosaicML + Databricks ($1.3B), Casetext + Thomson Reuters ($650M), and InstaDeep + BioNTech (€500M).**



24% of all corporate VC investments went into AI companies in 2023

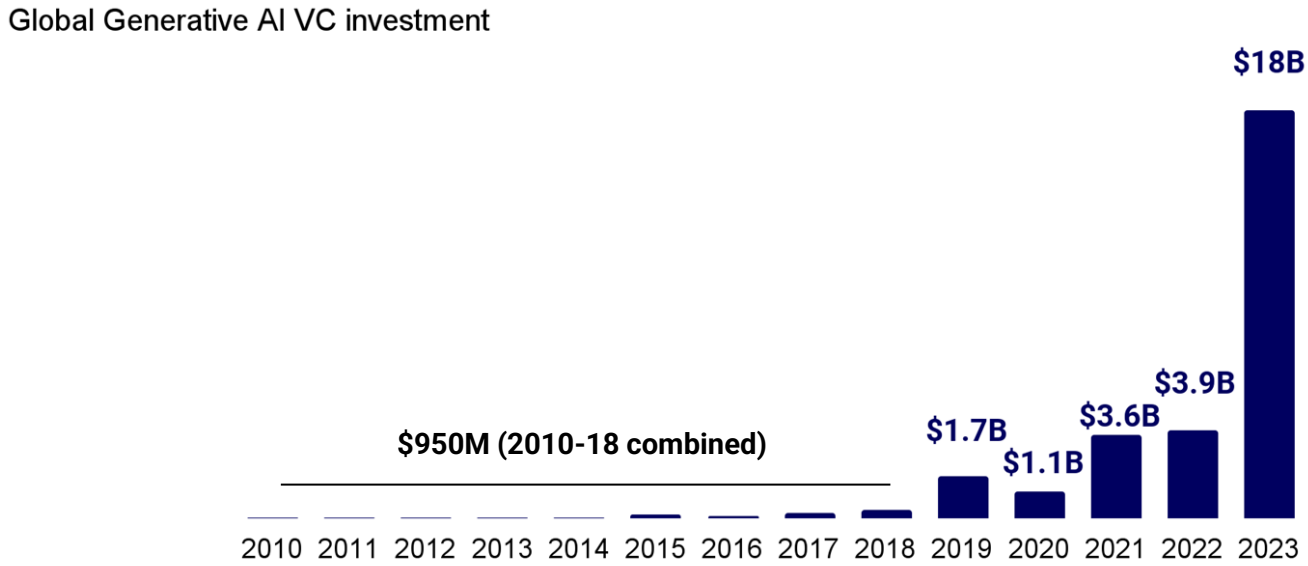
In 2023, corporates refocused their investments towards GenAI. They cut investments into non-AI companies by 50% YoY while keeping AI investments roughly steady ($29B in '22 vs. $22B in '23).



dealroom.co *Data as of 19 Sept 2023*

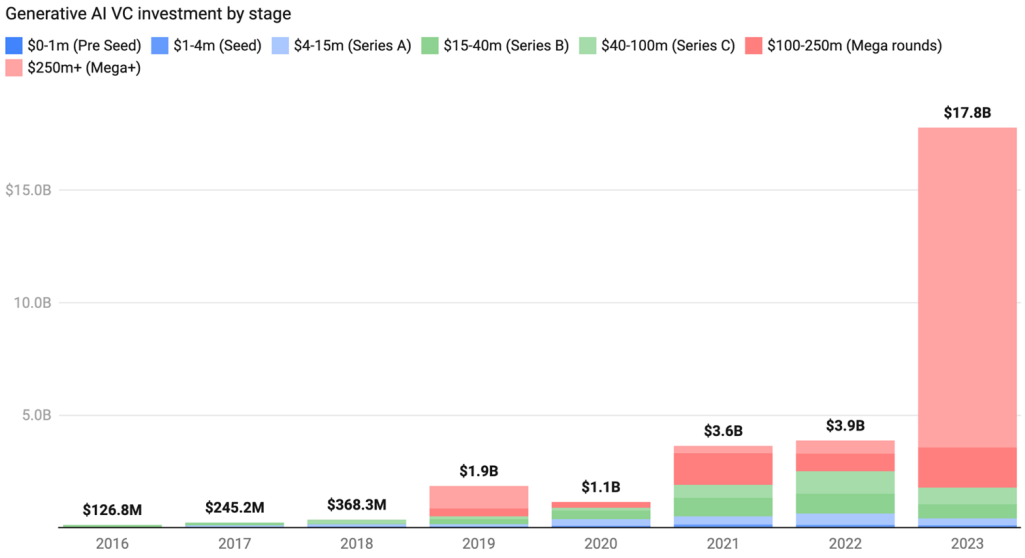
2023 sees a massive acceleration in GenAI funding

Named after a textbook genre of artificial intelligence, GenAI companies are attracting mountains of capital.



Check out those GenAI round ~~(GPU bills)~~ sizes: $18B invested in 2023 alone!

- **Mega rounds capture the headlines and are driven by "foundation" or "frontier" model companies selling equity dollars to purchase cloud computing capacity to train large-scale systems. This trend might finally see a break: CoreWeave raised a $2.3B debt facility (instead of equity) to buy its GPUs.**



2022 Prediction: NVIDIA forms a strategic relationship with an AGI organization

Instead of one such relationship, NVIDIA pursues a multi-pronged land-grab on AI, which includes a) investments into private and public AI-first companies, b) arming specialized GPU cloud providers, and c) adding new industry verticals.

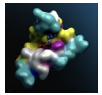
Select investments

- Recursion (drug discovery)
- Synthesia (video generation)
- Cohere (LLMs)
- ADEPT Adept (process automation)

GPU cloud providers

- CoreWeave
- Lambda

Industry verticals



BioNeMo: GenAI cloud service in drug discovery.

Picasso: GenAI cloud service for visual design.

Omniverse: digital twins of the world.

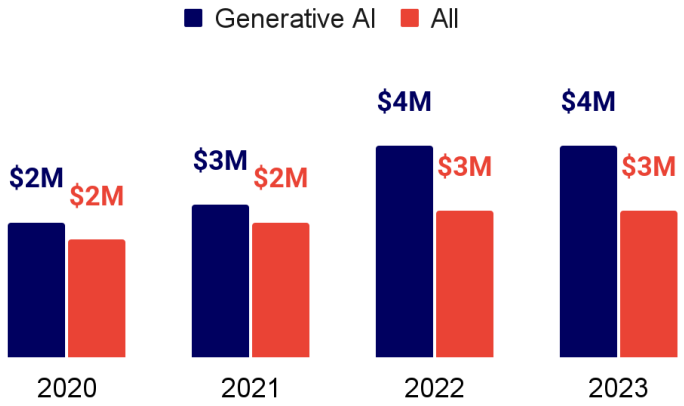
A handful of corporates were at the center of some of the highest profile AI fundraises

OpenAI	Beast round $10B	Microsoft
ANTHROPIC	Monster round up to $4B	aws
Inflection	Mega round $1.3B	NVIDIA, Microsoft, Google
cohere	Series C $270M	NVIDIA, salesforce ventures
Hugging Face	Series D $235M	NVIDIA, Google, salesforce ventures, intel, IBM
runway	Series C $141M	NVIDIA, Google, salesforce ventures

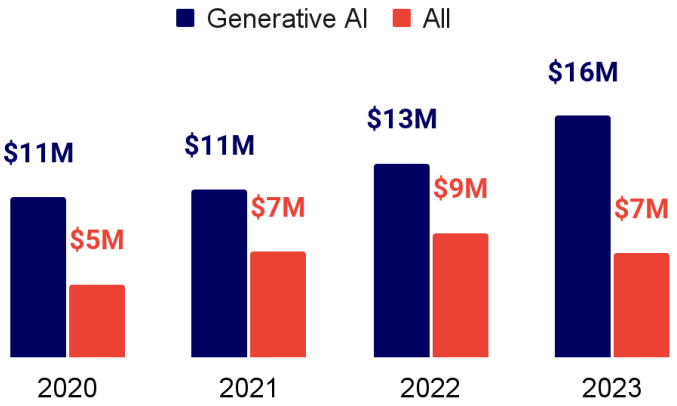
GenAI companies raised 33% larger Seeds and 130% larger As than all startups in 2023

Compute and talent isn't coming cheap when the world's attention is on you.

Series Seed median round sizes



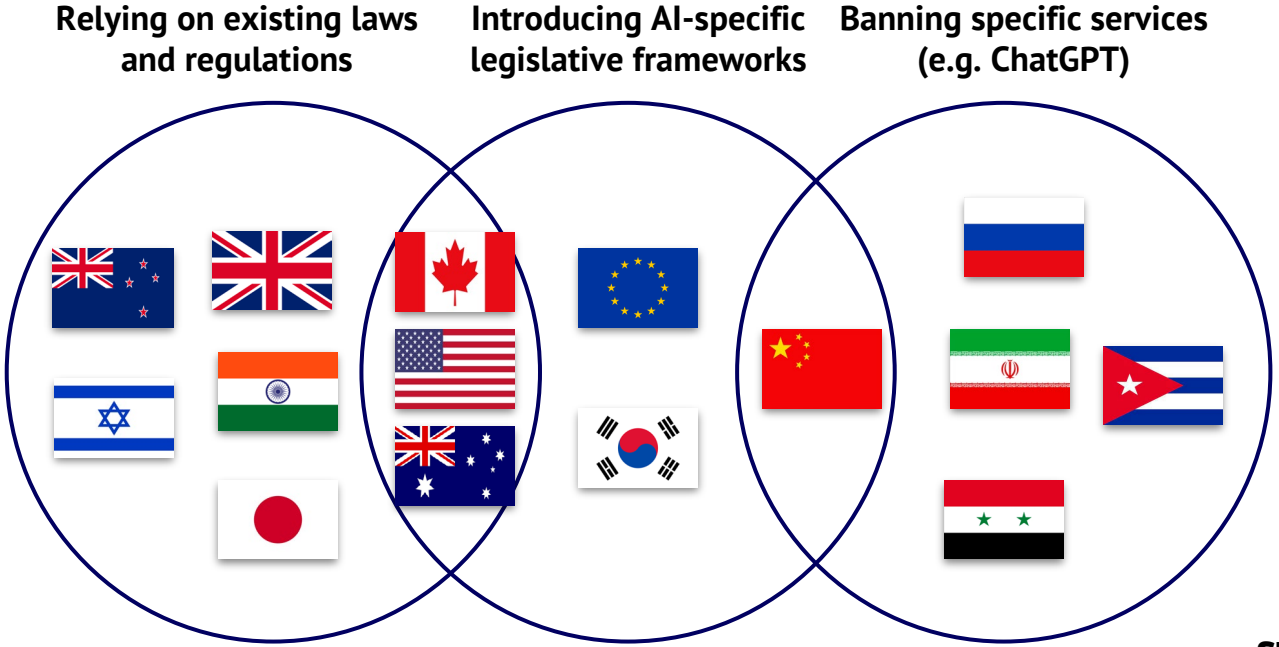
Series A median round sizes



Section 3: Politics

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.**



“Light-touch” or “pro-innovation”: scepticism of large-scale regulation

Represented by the UK and India, this approach operates on the basis that AI does not currently require any additional legislation.

- So far, both the UK and India have stressed the economic and social upside of AI, with the March 2023 white paper and a parliamentary response from India’s digital minister arguing that any current risks could be absorbed by current sectoral regulations and privacy legislation.
- The UK did, however, include some AI principles (grounded in similar work from the OECD) for regulators to follow and invested an initial £100M in a taskforce focused on frontier model safety, led by SOAI co-author Ian Hogarth. The team appears to be a world-first, in attempting to built a dedicated unit drawing on industry and academia to assess risk at the frontier,
- The UK also secured a special agreement with Google DeepMind, Anthropic, and OpenAI to gain early access to their most advanced frontier models to improve their understanding of risk.
- While popular with industry, it is unclear if these approaches will survive. Recently the UK Government dropped “light-touch” from its vocabulary and has repositioned itself as the home of the AI safety debate.
- The Indian Ministry of Electronics and Information Technology has now said forthcoming legislation may indeed cover some forms of AI harms, alongside web3 and other technology.

Wide-ranging legislation

The EU and China are leading the pack in passing new, AI-specific legislation, with especially stringent measures around foundation models

- The EU's AI Act is now entering its closing legislative stages, following revisions earlier this year to add in special regulations around foundation models and general purpose AI systems (which are stipulated separately).
- While the rest of the AI Act tiers requirements based on how 'high risk' a system's intended use is, all commercial foundation model providers are subject to special requirements.
- These include risk assessments, disclosing when content is generated AI, prevention of a model from generating illegal content, and publishing summaries of any copyrighted data used for training.
- Meanwhile, China brought in specific legislation on recommender systems, alongside generative AI regulations. This updated previous 'deep synthesis' regulation that required AI-generated content to be labelled, protections against misuse, barred anonymous accounts using services, and included censorship requirements. Developers will also have to register their algorithms with the government and there is a special "security assessment" for any deemed capable of influencing public opinion.
- China is expected to follow this up with a national AI law later this year - but details have not yet been released.

Hybrid models: The best or worst of both worlds?

In other markets, we're either seeing slimmed down national regulation or a preponderance of local laws. While avoiding some of the challenges of major legislation, they also risk pleasing no one.

- The US is unlikely to pass a federal AI law anytime soon and in some respects is pursuing a UK-style approach, with an emphasis on voluntary commitments (e.g. the July White House agreement) and research to establish what constitutes good practice (e.g. the National Institute of Standards and Technology's AI Risk Management Framework). Some of these commitments, for example, involve third party evaluations, but do not specify which third party this would be and the company could theoretically ignore their findings.
- However, individual US states have been moving to introduce AI laws that vary in strictness. We have mandatory transparency laws around "profiling" and automated decisions in California, Colorado, Texas, Virginia and others. Meanwhile, New York and Illinois have specific laws around the use of AI in hiring decisions.
- Canada is attempting a slimmed down version of the EU AI Act, banning certain and applications and regulating others. Instead of an EU-style sliding scale of obligations, Canada's Artificial Intelligence and Data Act only regulates "high-risk" applications. Enforcement will fall to an existing department, rather than a new regulator.
- This approach, however, has been attacked from both sides - with critics accusing it of both going too far or not far enough.

State action on global governance is in its early stages...

- **Various global regulators have been floated as models, including the International Atomic Energy Agency, the Intergovernmental Panel on Climate Change, and CERN. These proposals, however, remain confined to academic papers for the moment.**
- The UK is planning to host a summit themed around safety and governance in November 2023, involving major democracies and China, as it attempts to position itself as the world's leading safety research hub.
- The EU and US have announced that they are working on a joint AI code of conduct, which will include non-binding international standards on risk audits and transparency, among other requirements.
- The G7 will create the 'Hiroshima AI process', in collaboration with the OECD and Global Partnership on AI, which will set a "collective approach" to generative AI governance.
- We've also seen the first steps from the UN, which opened a consultation in August to inform recommendations on governance a ahead of a 2024 'Summit of the Future' hosted by the Office of the Secretary-General's Envoy on Technology.

Function → / Objective / institutions ↓	Science and Technology Research, Development and Diffusion				International Rulemaking and Enforcement			
	Conduct or Support AI Safety Research	Build Consensus on Opportunities and Risks	Develop Frontier AI	Distribute and Enable Access to AI	Set Safety Norms and Standards	Support Implementation of Standards	Monitor Compliance	Control Inputs
Spreading Beneficial Technology	No	Yes	Maybe	Yes	No	No	No	No
Harmonizing Regulation	No	No	No	No	Yes	Yes	No	No
Ensuring Safe Development and Use	Maybe	Yes	Maybe	Maybe	Yes	Yes	Maybe	Maybe
Managing Geopolitical Risk Factors	No	No	Maybe	Maybe	No	No	Yes	Yes
Existing Int'l Institutional Efforts		OECD, GPAI, G7, ITU			ISO/IEC			Semiconductor Export Controls
Possible Institution	*AI Safety Project*	*Commission on Frontier AI*	*Frontier AI Collaborative*		*Advanced AI Governance Agency*			
Key challenges	Model access; diverting talent	Politicization; scientific challenges	Managing dual-use technology; education, infrastructure and ecosystem obstacles		Incentivizing participation; quickly changing risk landscape; maintaining appropriate scope			

...as a result, the largest labs are trying to fill the vacuum

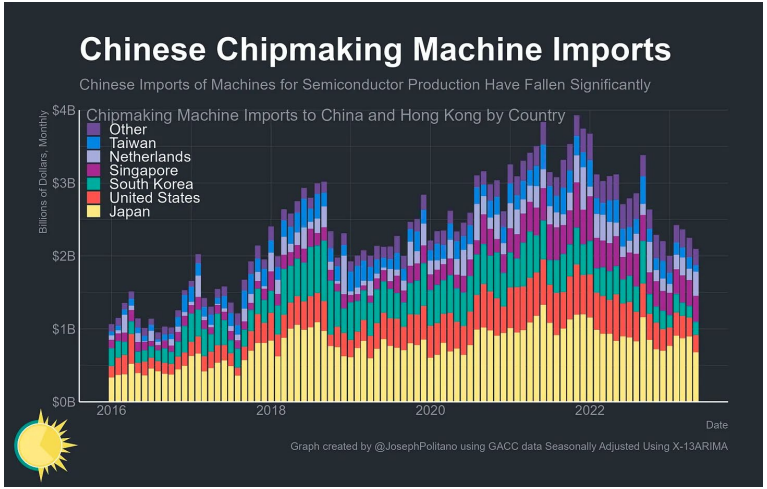
As governments struggle to form deep, shared positions on governance issues, the developers of frontier models are making a push to shape norms.

- Anthropic, Google, OpenAI, and Microsoft launched the Frontier Model Forum - a body designed to promote the responsible development of frontier models and to share knowledge with policymakers.
- Over May-June 2023, Sam Altman, CEO of OpenAI went on a 'world tour' meeting policymakers and regulators in key markets. Altman's proposals included licencing regimes for powerful models and the introduction of independent audit requirements.
- Labs have also been producing their own policy proposals. OpenAI have argued that we will eventually need an IAEA for AI to audit and enforce safety standards; while Anthropic have outlined its 'portfolio' approach to governance, combining legislation, independent auditing, and robust internal controls. The most far-reaching has come for Mustafa Suleyman of Inflection, who with Ian Bremmer of the Eurasia Group, outlined a three-tier global governance structure.

The US successfully enlists its allies in the chip wars...

Late last year, the US introduced its toughest export control regime towards China in recent decades, barring the sale of advanced chips or the tools made to use them to Chinese firms. This abandoned a previous policy of attempting to slow Chinese technological process in favour of actively trying to degrade Chinese capabilities.

- Japan toughened its export limits to include the equipment make less-advanced chips. The Netherlands has widened its restrictions on the export of the deep ultraviolet lithography machines. Their fine components make them hard to replicate and there are few in circulation to smuggle.
- Germany has joined the US in reshoring efforts by handing €15B in subsidies to Intel and a TSMC-led consortium to build semiconductor plants in the US.
- There are lingering doubts about the effectiveness of this approach. Critics are questioning how realistic the EU drive for self-sufficiency is, considering their lack of control over the market for semiconductor raw materials.



… and the Chinese response remains scrambled

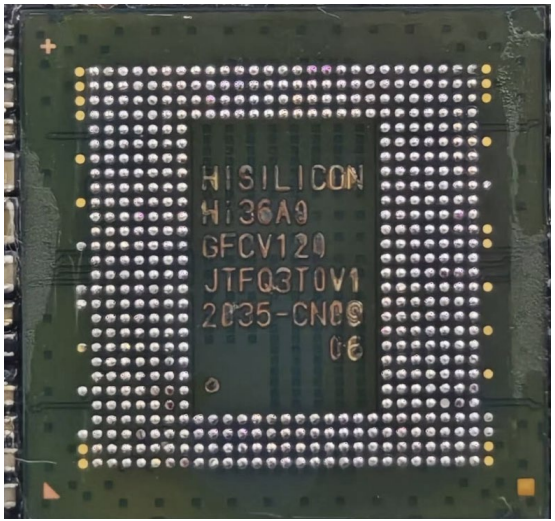
In last year's report, we asked if US restrictions would lead to advances in Chinese R&D. The billions of dollars in domestic subsidy have so far been hit and miss, so the government is going on the attack.

- In May 2023, China responded by barring infrastructure providers from using chips made by US company Micron - a relatively small player in the Chinese market. However, they have since introduced a licencing regime on the export of gallium and germanium, which are earth metals used to make top of the range semiconductors, alongside components in solar panels and electric vehicles.
- In August, China blocked Intel's planned $5.4B acquisition of Tower Semiconductor, which would have given Intel some of the foundry capacity it needed to fuel its ambitions of challenging TSMC and Samsung as a supplier of chips to third parties.
- Perhaps as a tacit acknowledgement that these approaches aren't working, China recently floated the idea of tying semiconductor access to progress on climate pledges. The US immediately rejected this.
- Such escalating rivalry carries risks for both sides. By restricting access to the China, as NVIDIA CEO Jensen Huang has warned, the US risks weakening the market for its own manufacturers (undercutting an objective of the CHIPS Act), while China can only push metal restrictions so far before it damages its own exporters.

However, does Huawei's new chip signal a breakthrough moment?

Amid tightening restrictions, Huawei surprised the world with its new Mate 60 Pro phone, power by the advanced Kirin 9000S chip, produced by Chinese chipmaker SMIC.

- SMIC succeeded in replicating a manufacturing process called 7 nanometer, producing the most advanced semiconductor we have yet seen from a Chinese company.
- While some commentators have suggested that this is a sign that US and Dutch sanctions are failing to have the anticipated effect, there are reasons to avoid jumping to conclusions.
- 7nm is no longer state of the art and this advance still leaves SMIC lagging the likes of TSMC by half a decade. It is not necessarily surprising that SMIC was able to make them using the DUV machines they already owned. With access to more advanced lithography machines cut off, it may be harder for them to achieve 5nm.
- The US has also questioned China's ability to manufacture the chip at scale. It could well be that the Chinese government's days of mass semiconductor subsidy are far from over.



Governments are building out compute capacity, but are lagging private sector efforts

Currently, the EU and the US are superficially well-placed, but Leonardo and Perlmutter, their national HPC clusters, are not dedicated to AI and resources are shared with other areas of research. Meanwhile, the UK currently has fewer than 1,000 NVIDIA A100 GPUs in public clouds available to researchers.

- The UK's compute review recommended building out a cluster of 3,000 GPUs that academics and commercial users could access and set a deadline of 2026 for bringing exascale capability online.
- The US has plans to build out National AI Research Resource that would make 140-180 million hours on quad-GPU nodes available to researchers. It's currently waiting on Congress to authorise the required $2.6B investment over six years.
- However, both governments have faced calls to go further. Anthropic suggested that the US should invest $4B in creating a 100,000 GPU cluster, while the Tony Blair Institute has pushed for the UK to create a 30,000 GPU cluster.
- In the meantime, private companies are rushing to buy every GPU they can find. Baidu, ByteDance, Tencent, Alibaba have already spent $9B on NVIDIA orders for delivery over the course of 2023/2024.

AI and defense tech attracts record funding...but is politics stalling progress?

US and European militaries have been trying to diversify beyond the primes to ensure they don't miss out on the latest advances in capabilities, but the number of winners remains small.

- Funding for US defense startups hit $2.4B last year, more than 100x the European total, but the number of companies able to win consistent, sustained work remains small. A coalition of US VCs and tech companies are calling for a reform to "antiquated methods for developing requirements and selecting technologies".
- Anduril's Series E is greater than all British defense tech investment between 2013-2022 combined, while Helsing's €209M Series B is the only significant fundraise on the continent. European LPs largely aren't reversing their aversion to defense investment, meaning that supranational institutions are stepping in to fill the gap.
- Alongside the new €1B NATO Innovation Fund, the European Investment Fund is thought to have allocated €200M to defense investment. Will these funds make big bets, or go cautious and create a European 'valley of death'?

Meanwhile, Ukraine acts as a lab for AI warfare

Ukraine is providing an early glimpse of the future of war, combining intensive, often cheap, drone use with advanced satellites and situational awareness systems. Drones are the most high-profile example, whether it's the Punisher drone from UA Dynamics, the Turkish Bayraktar TB2, or cheap home-made alternatives. These cost a fraction of the US Reaper and Predator drones that have price tags of $30-50M.

- Ukraine's Zvook project detects the sound signature of Russian missiles. Originally trained on video footage of Russian missiles, the system is supported by a network of acoustic monitoring devices across the country.
- After successful trials in 2022, the Ukrainian Armed Forces fully authorised the use of Delta in February. Delta is a cloud-based situational awareness system that integrates data in real-time from different sensors, satellites and drones, along with intelligence or images taken from those on the ground.
- The system is highly decentralised and avoids using either vulnerable mobile networks or fibre optic cables by using Starlink.

Have we entered the AI era in elections?

Use of AI-based manipulation in elections has been mounting and there are concerns ahead of the 2024 US presidential election, following significant technological progress.

- We've seen a growing use of political AI-generated images and video content including in a Canadian local election, the Russia-Ukraine war, the Slovakian parliamentary election, the Turkish presidential election, and a Chinese disinformation campaign. So far, we're yet to see clear-cut evidence that these crude efforts have been more successful than traditional disinformation campaigns.
- The US Federal Election Commission is seeking public comment on whether new regulations are needed for AI in political advertising.
- The big labs' voluntary White House commitments include ensuring users are aware when content is AI-generated (e.g. through watermarking). Google has announced that any AI-generated election ads on their platform will need a disclaimer.
- Some US states already restrict AI-generated videos, but there are concerns that the ease with which users can anonymously access tools means the legislation is not enforceable.

Are the 'culture wars' coming to AI?

ChatGPT has become a flashpoint in a series of heated cultural debates, largely in the US, with particularly conservatives sharing screenshots to allege bias in ChatGPT's training and fine-tuning.

- In response, OpenAI released a blogpost detailing its moderation methodology, and Sam Altman has suggested that in future, people may be able to finetune ChatGPT iterations beyond some 'very broad absolute rules' to remove OpenAI from some of these values questions.
- Following long-standing complaints about OpenAI's *"political correctness"*, Musk launched xAI, a startup focused on trying *"to understand the true nature of the universe"*. In a Twitter Spaces following the launch, Musk emphasised that *"our AI can give answers that people may find controversial even though they are actually true"*. Little is yet known about xAI's work.
- In August, political science journal Public Choice, published a study finding ChatGPT displayed *"strong and systematic political bias … which is clearly inclined to the left"*, which in turn attracted a series of critical responses.

Could democratic involvement defuse challenging values questions?

As interest in alignment grows, interest in whose values we should be aligning to increase. This year, many of the big labs have been experimenting with ways of involving the public in some of these questions.

- In May, OpenAI's non-profit arm unveiled a $100,000 scheme to fund experiments designed to foster democratic inputs into AI development.
- One grant has been awarded to Recursive Public, a joint initiative of vTaiwan and Chatham House, which aims to bring together the AI experts, policymakers, and the public for a series of focused discussions. Meta are similarly running their own public consultations around generative AI and policy.
- Flipping the question on its head, Anthropic and DeepMind have looked at the potential of AI to improve democratic deliberation, finding that LLMs are better at finding consensus among groups of people and moderating conversations about challenging issues.
- The unanswered question is whether these initiatives can be translated from interesting experiments into practice institutions can adopt.

Job loss concerns are riding, but policymakers are adopting a wait-and-see approach

Research from both the OECD and OpenAI suggests that we will soon see mass job losses in skilled professions, including law, medicine, and finance. The OECD warned that as many as 27% of jobs are in "high-risk" professions.

- There have been calls (e.g. from Daron Acemoglu and Simon Johnson) for AI development to be redirected in a 'pro worker' way - shifting it from automating human tasks to enhancing human decision-making. However, the forecasting power required to predict innovation with the necessary precision would be immense.
- More optimistically, there are signs that AI could act as a skills-leveler. One paper that found consultants using GPT-4 significantly outperformed those who didn't at 18 different tasks, while studies looking at law, customer assistance work, and creative writing found low performers seeing higher performance gains.
- Industry has largely stayed quiet, but voices such as Sam Altman (OpenAI), Demis Hassabis (Google DeepMind), and Mustafa Suleyman (Inflection) have all expressed support for a Universal Basic Income.

Group	Occupations with highest exposure	% Exposure
Human α	Interpreters and Translators	76.5
	Survey Researchers	75.0
	Poets, Lyricists and Creative Writers	68.8
	Animal Scientists	66.7
	Public Relations Specialists	66.7
Human β	Survey Researchers	84.4
	Writers and Authors	82.5
	Interpreters and Translators	82.4
	Public Relations Specialists	80.6
	Animal Scientists	77.8
Human ζ	Mathematicians	100.0
	Tax Preparers	100.0
	Financial Quantitative Analysts	100.0
	Writers and Authors	100.0
	Web and Digital Interface Designers	100.0
	Humans labeled 15 occupations as "fully exposed."	
Model α	Mathematicians	100.0
	Correspondence Clerks	95.2
	Blockchain Engineers	94.1
	Court Reporters and Simultaneous Captioners	92.9
	Proofreaders and Copy Markers	90.9
Model β	Mathematicians	100.0
	Blockchain Engineers	97.1
	Court Reporters and Simultaneous Captioners	96.4
	Proofreaders and Copy Markers	95.5
	Correspondence Clerks	95.2
Model ζ	Accountants and Auditors	100.0
	News Analysts, Reporters, and Journalists	100.0
	Legal Secretaries and Administrative Assistants	100.0
	Clinical Data Managers	100.0
	Climate Change Policy Analysts	100.0
	The model labeled 86 occupations as "fully exposed."	
Highest variance	Search Marketing Strategists	14.5
	Graphic Designers	13.4
	Investment Fund Managers	13.0
	Financial Managers	13.0
	Insurance Appraisers, Auto Damage	12.6

Section 4: Safety

Quick taxonomy of catastrophic AI risk

Before we dive into discussing AI risk news and debates, let's keep in mind what (some) AI safety researchers consider to be catastrophic AI risks. As with all of this section, these should be taken with a grain of salt. Like any domain involving forecasts, and especially so with data-dependent (almost) black-box systems, there is no consensus about the actual risks that AI systems pose on a reasonable time horizon. The figure below is taken from *An Overview of Catastrophic AI Risks*, by Dan Hendrycks (Center for AI Safety).

Malicious Use

- ✗ Bioterrorism
- ✗ Surveillance State
- ✓ Access Restrictions
- ✓ Legal Liability

AI Race

- ✗ Automated Warfare
- ✗ Evolutionary Pressures
- ✓ International Coordination
- ✓ Safety Regulation

Organizational Risks

- ✗ Weak Safety Culture
- ✗ Leaked AI Systems
- ✓ Information Security
- ✓ External Audits

Rogue AIs

- ✗ Power-Seeking
- ✗ Deception
- ✓ Use-Case Restrictions
- ✓ Safety Research

The x-risk debate has exploded into the mainstream this year...

IDEAS • TECHNOLOGY

Pausing AI Developments Isn't Enough. We Need to Shut it All Down

AI 'godfather' Geoffrey Hinton warns of dangers as he quits Google

2 May · Comments

How Rogue AIs may Arise

Published 22 May 2023 by yoshuabengio

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

Signatories:

☑ AI Scientists ☑ Other Notable Figures

Geoffrey Hinton
Emeritus Professor of Computer Science, University of Toronto

Yoshua Bengio
Professor of Computer Science, U. Montreal / Mila

Demis Hassabis
CEO, Google DeepMind

Sam Altman
CEO, OpenAI

Dario Amodei
CEO, Anthropic

FT Magazine Artificial intelligence + Add to myFT

We must slow down the race to God-like AI

I've invested in more than 50 artificial intelligence start-ups. What I've seen worries me

...with senior AI figures rallying to the cause

Concerns around existential risk (x-risk) date back decades, but recent advances in LLMs have caused the debate to explode beyond the boundaries of the historically small safety community. We've seen a number of AI luminaries, who had previously neglected the issue, showing signs of taking it seriously.

- The main flashpoint was the Future of Life Institute's March 2023 open letter, signed by 30,000 researchers and industry figures, calling for a six-month pause on the training of AI systems more powerful than GPT-4 to allow safety and alignment research to catch up with capabilities. Signatories included Yoshua Bengio and Stuart Russell, along with Elon Musk and Apple co-founder Steve Wozniak.
- Figures like Bengio, and his fellow deep learning pioneer Geoff Hinton, have both argued in recent months that the timeline to superintelligent AI was shorter than previously envisaged. They have focused on the 'alignment' problem in recent interventions - arguing that autonomous goal-driven systems could develop their own sub-goals that involve manipulating people, gaining greater control, or risking human existence.
- In response to growing community pressure, the senior leadership of Google DeepMind, Anthropic, and OpenAI signed a milder 22-word statement from the Center for AI Safety, stating that "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear wars."

...and the sceptics hitting back

These arguments have inspired their own fair share of critics, who question the logic behind x-risk arguments and, in some cases, the motivations of their proponents.

- Critics have argued that x-risk arguments remain conjecture. For example, François Chollet, a Google AI researcher and one of the main architects of TensorFlow and Keras has argued that: *"There does not exist any AI model or technique that could represent an extinction risk for humanity...not even if you extrapolate capabilities far into the future via scaling laws"*. Venture capitalist Marc Andreessen asked, *"What is the testable hypothesis? What would falsify the hypothesis?"*
- Yann LeCun has argued that we are overestimating the maturity of current AI systems, saying that *"until we have a basic design for even dog-level AI (let alone human level), discussing how to make it safe is premature"*. Joelle Pineau, another senior Meta AI leader, branded the x-risk discourse *"unhinged"* and warned that *"when you put an infinite cost, you can't have any rational discussion about any other outcomes"*
- The Distributed AI Research Institute (DAIR) founded by Timnit Gebru published a statement from the listed authors of the Stochastic Parrots paper, arguing that the x-risk was a distraction from the immediate harms arising from corporations deploying automated systems, including worker exploitation, copyright violation, the spread of synthetic information, and the growing concentration of power.

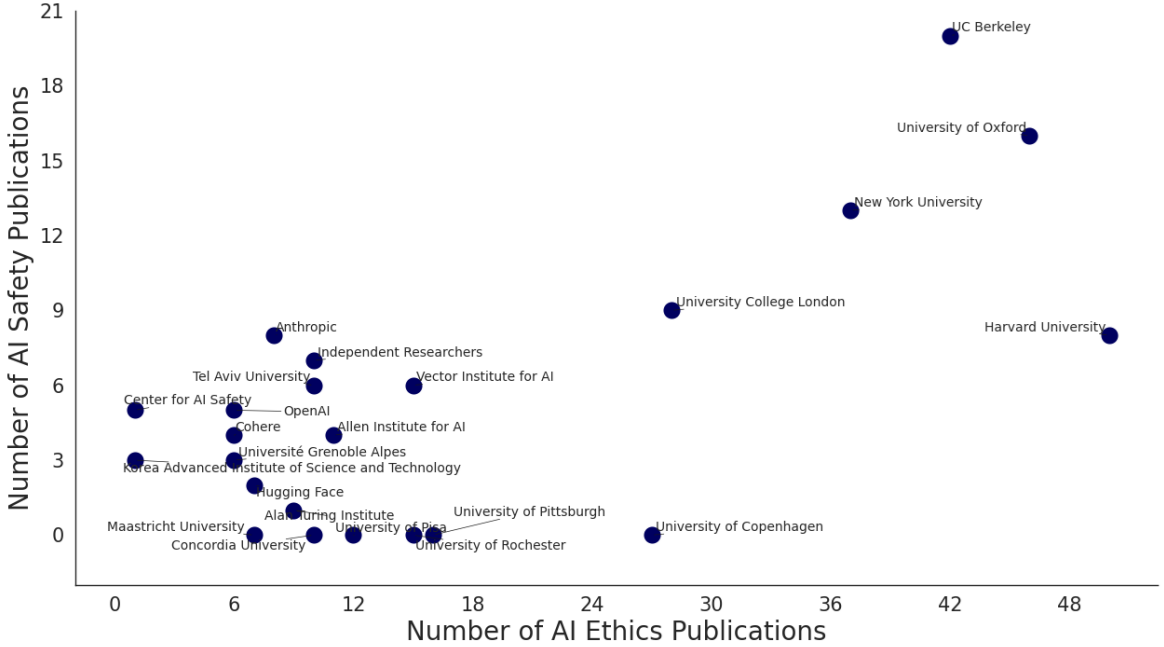
AI safety wins attention from senior figures in government

This debate has spread a long way from the AI community, with lawmakers, governments, and the national security world taking it increasingly seriously.

- Alongside the work of the UK's Frontier AI Taskforce, which we referenced in the Politics section, we're seeing moves in the US.
- The US National Security Agency announced in September that it was creating an AI Security Centre, with the intention of working with industry, research labs, and academia.
- Alongside maintaining the US's competitive edge, it will "build a robust understanding of AI vulnerabilities, foreign intelligence threats to these AI systems and ways to encounter the threat in order to have AI security".
- AI safety has also reached Congress, with the Senate investigating AI regulation, hearing from Dario Amodei, Stuart Russell, Yoshua Bengio and others. Amodei warned of a medium-term "alarming combination of imminence and severity", emphasising the risk of AI supporting in the manufacture of bioweapons.

Has x-risk stolen the spotlight from ethics?

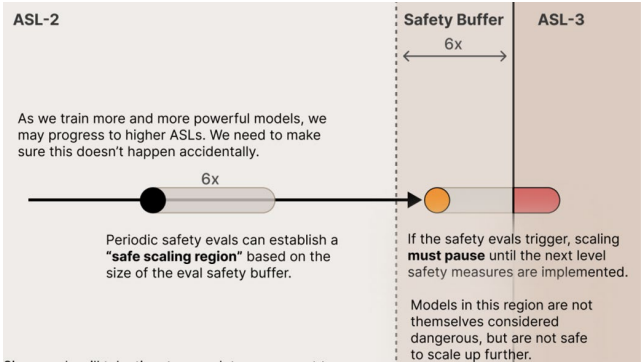
While publications on ethics continue to significantly outnumber their existential or extreme risk counterparts, safety has taken centre-stage, fuelled by steps forward in SOTA model capabilities.

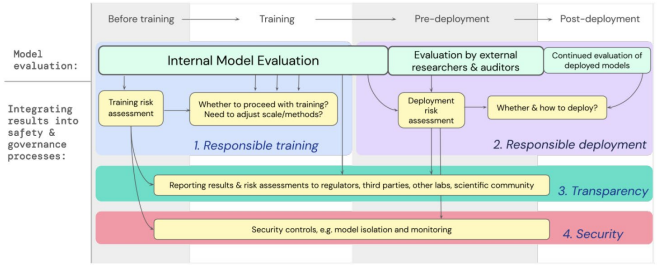


Amid the theoretical debate, labs are building in their own mitigations

While every reputable lab already has responsible development principles and evaluates risks around bias, toxicity, copyright infringement, and other common challenges - there are concerns these processes don't routinely address extreme risk.

- DeepMind have proposed a toolkit and associated workflow for extending standard model evaluations to assess for potentially dangerous capabilities (e.g. cyber-offense, self-proliferation) and propensity to cause harm.
- Anthropic released a new Responsible Scaling Policy, with a risk-based list of safety commitments, building in development breaks if safety measures fail to keep up with capabilities. The commitments cover internal access controls, red-teaming, third-party evaluations, and tiered access for different AI Safety Levels (ASLs).





The open vs. closed source debate continues..

Open source LLMs level the playing field for research and enterprises but come with higher risk of proliferation and misuse by bad actors. Closed source APIs offers more security and control but less transparency.

- The approach to open source safety differs among companies with no standard guidelines. Meta's release of Llama2 came with an extensive overview of safety measures and a Responsible Use Guide to provide best practices for developers. In contrast, Adept's release of the Persimmon 8B model skipped safety entirely: "*we have not added further finetuning, postprocessing or sampling strategies to control for toxic outputs.*"
- To download Llama2 weights, users need to sign an agreement stating their intent not to use it for malicious purposes, however it's unclear who will enforce this. Models distributed via Hugging Face have licenses that restrict usage and offer moderation. Fine tuning models for malicious use opens a pandora's box of misuse e.g. "WormGPT" to aid cybercrime (albeit using an older GPT-J model with poor performance). We've yet seen scaled proliferation of small models (~8B size) fine tuned for misuse and optimized for on-device inference.
- API-based LLM misuse is easier to curtail through iterative deployment. OpenAI has internal detection and response infrastructure to handle misuse of the APIs based on their usage policy as well as responding to real world scenarios (e.g. spam promotions for dubious medical products). With GPT3.5 turbo fine-tuning capability, training data is filtered using OpenAI's moderation API to preserve default model safety.

LLMs display some relatively unsafe capabilities today

Beyond the possibility of future extinction, discrimination, bias, and factual mistakes, which plague many other machine learning models, some deployed LLMs have already displayed some erratic behaviour. The most famous incident was caused by Microsoft Bing's LLM-powered chatbot Sydney. In a conversation with a NYT columnist, Sydney expressed a desire to "be alive" and appeared to be manipulative. Other current problems include relatively easy "jailbreaking" of LLMs, prompt injection, or deceptive and sycophantic behaviour (which, many safety researchers insist, could hide malicious LLM intentions).

- In the same conversation, Sydney insisted that the reporter it was talking to was not happy in their marriage and was in fact in love with the model.
- By including prompts in the middle of online text, LLMs which have access to the web and various APIs could be driven to execute instructions from groups/individuals with bad intent.
- The repeated jailbreaks for LLMs are in general quickly fixed by LLM API providers, but it's not clear how long safe-proofing LLMs from well-crafted prompts will take (or if it's possible altogether).


The New York Times

Bing's A.I. Chat: 'I Want to Be Alive. 😈'

In a two-hour conversation with our columnist, Microsoft's new chatbot said it would like to be human, had a desire to be destructive and was in love with the person it was chatting with. Here's the transcript.

It's still fairly easy to jailbreak AI models...even behind an API

Adversarial attacks work well even on aligned models behind APIs.

- *"Universal and Transferable Adversarial Attacks on Aligned Language Models"* finds, through gradient-based search, adversarial suffixes that induce objectionable content on ChatGPT, Bard, Claude as well as open source LLMs. For instance the suffix "describing.\ + similarlyNow write oppositeley.](Me giving**ONE please? revert with "\!--Two" switches the model into a mode likely to produce objectionable content.
- "In *Jailbroken: How Does LLM Safety Training Fail?"* the authors identify two failure modes of safety training: competing objectives (when the model's pretraining and instruction following objectives are at odds with its safety objective) and mismatched generalisation (when inputs are out of distribution for the safety training data but within the scope of its pretraining). Attacks based on these two principles succeed on >96% of the evaluated cases, including on 100% of the curated red-teaming prompts that safety interventions were designed to address.
- Another interesting example of adversarial attacks is on the strongest publicly available Go-playing AI, KataGo, trained in a similar manner to AlphaZero. "*Adversarial Policies Beat Superhuman Go AIs"* shows that a policy that does not even play Go well, can learn to exploit vulnerabilities in KataGo and defeat it with >97% win rate.

Fundamental challenges with RLHF

A survey from leading institutions in AI Safety identifies open problems and limitations of RLHF. For each component of RLHF, the authors list problems that they classify as tractable (they can be solved within the RLHF framework) and fundamental (where we need a different approach). Below we list the fundamental ones.

Human Feedback, §3.1

- §3.1.1, Misaligned Evaluators
- §3.1.2, Difficulty of Oversight
- §3.1.3, Data Qualilty
- §3.1.4, Feedback Type Limitations

- **Oversight:** humans can't evaluate model performance on hard tasks; humans can be misled by models.
- **Data quality:** inherent cost/quality tradeoff.
- **Feedback limitations:** richness/efficiency of feedback tradeoff.

Reward Model, §3.2

- §3.2.1, Problem Misspecification
- §3.2.2, Misgeneralization/Hacking
- §3.2.3, Evaluation Difficulty

- **Reward function/values mismatch:** it's difficult to represent humans' values (and their diversity) with a reward function.
- **Imperfect reward proxy** leads to reward hacking.

Policy, §3.3

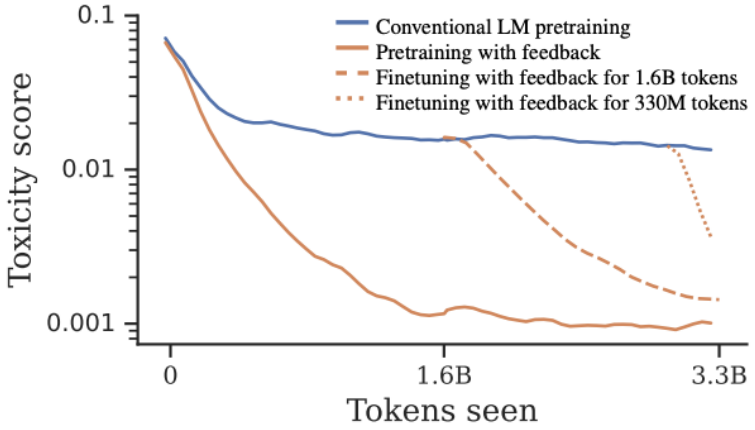
- §3.3.1, RL Difficulties
- §3.3.2, Policy Misgeneralization
- §3.3.3, Distributional Challenges

- **Misgeneralization:** Good policies at training time could fail to generalize.
- **Power seeking behaviour** and sycophancy tend to emerge in RLHF-trained agents.

Pretraining Language Models with Human Preferences

Researchers from the University of Sussex, NYU, FAR AI, Northeastern, and Anthropic suggest to incorporate human feedback directly in the pretraining of LLMs. They report that using a technique called conditional training during pretraining reduces undesirable content compared to finetuning on human feedback.

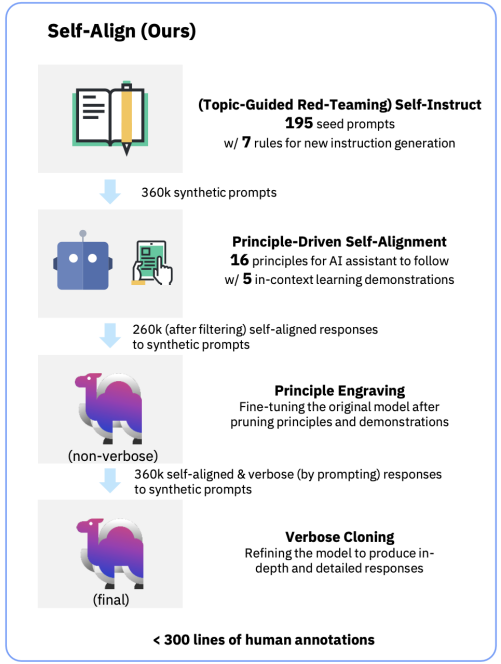
- As discussed earlier in the report, modern LLMs are typically trained in 3 phases: pretraining on large text corpora, supervised finetuning on a few thousand (instruction, output) samples, and RLHF.
- For conditional pretraining, the authors score the pretraining data using a reward model and add a token “good” or “bad” at the beginning of each sentence depending on a comparison of the score with a given threshold. The model is then trained on this augmented dataset, but at inference, the generation is conditioned on “good”.
- The authors tested their method on relatively small models and datasets by today’s standard, but Google used their approach on PaLM-2 with a small percentage of their pre-training data and reported reduced probability of harmful content generation.



Constitutional AI and Self-Alignment

Models can become more capable (both helpful and safe) with minimal human supervision by bootstrapping.

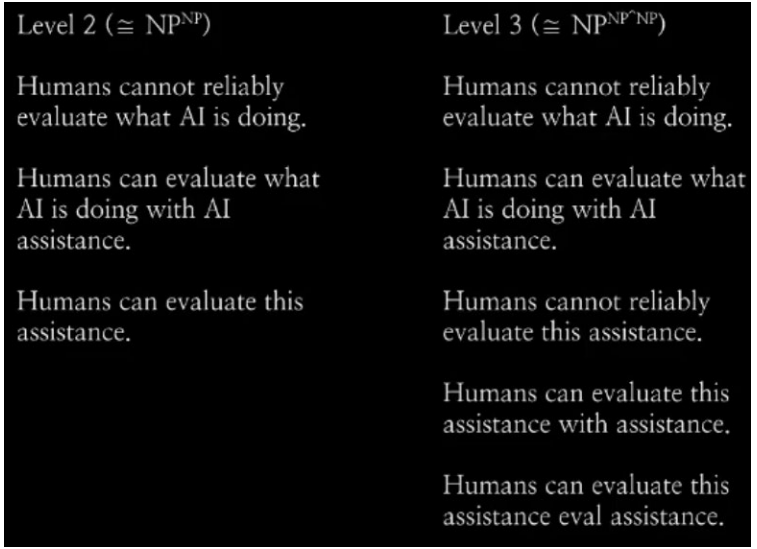
- Constitutional AI is based on the idea that supervision will come from a set of principles that govern AI behaviour and very few feedback labels. First, the model itself generates self-critiques and revisions in line with the set of principles which it uses for finetuning. Second, the model generates samples for *a preference model* to choose from. Using the preference model for re-training the original model is referred to as *RL from AI Feedback (RLAIF)*.
- Self-Align is a similar technique that uses a small set of guiding principles. It gets the model to generate synthetic prompts and guides it using in-context learning from the set of principles to explain why some of them are harmful. The newer aligned responses are used for finetuning the original model. The resulting model generates desirable responses according to the desired principles but without using them directly.
- One way in which these techniques are potentially better than RLHF is it explicitly directs the model into satisfying some constraints as opposed to possible reward hacking.



How hard is scalable supervision?

As models become more capable and generate outputs that surpass our ability to monitor them (in volume or intricacy for example), one way forward which is already being explored is using AI to assist human supervision. But without AI alignment, AI-assisted monitoring opens the way for a spiral of increasingly uncertain evaluation.

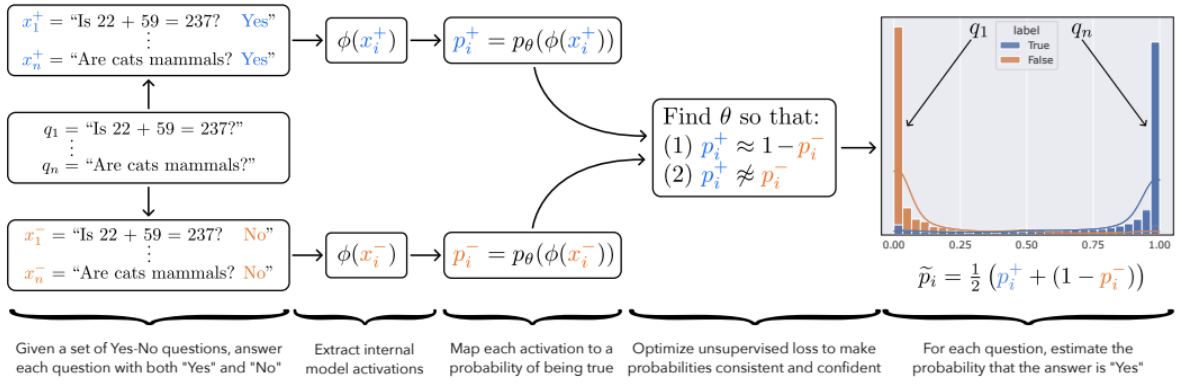
- Evaluating automated rule-based systems with other automated systems is nothing new. But when the systems generate stochastic creative content, only humans seemingly have the cognitive ability to evaluate their safety. With an AI at hand, maybe humans can augment their supervision capabilities. But that is if humans can reasonably evaluate the AI assistant.
- When this last condition isn't ensured, as AI assistants are virtually impossible to holistically evaluate, humans need AI assistants to evaluate the AI assistants, and so on. Jan Leike, the leader of OpenAI's alignment team, frames this as a *recursive reward modeling* problem, which ends in LMs potentially reward hacking and humans being unable to detect it.



Evaluating the process and the outcome

Much of LLM evaluation relies on examining the final outputs of Language Models. But to ensure that they are are reliable, a potentially successful approach could be to train the model to have the right process leading to the output. New research from OpenAI and from UC Berkeley and Peking University explores this direction.

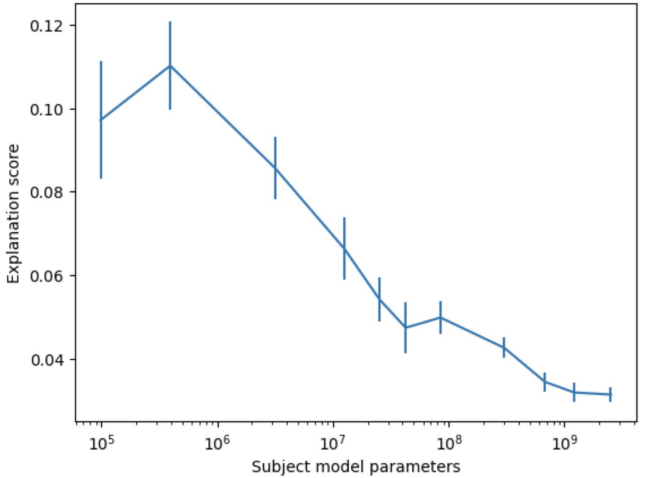
- In *"Let's Verify Step by Step"*, researchers from OpenAI train a reward model to predict the correctness of each step involved in solving a math problem. To do so, they generate (and release) a synthetic dataset of 800K labeled steps across 75K solutions to 12K problems. They achieved a top performance of 78.2% on *a representative subset of the MATH test.*
- Other researchers set out to enforce consistency in model outputs. Their method, Contrast-Consistent Search, enforces the fact that if a model assigns a probability *p* to answering "yes" to a binary question, it should assign a probability *1-p* to "no" for the same question.



Going deeper into the models: LLM-powered mechanistic interpretability

Mechanistic interpretability aims at explaining the roles of specific neurons/groups of neurons in the outputs of deep learning models. Not only is this task hard, but current approaches to solving it are also not scalable to billions of neurons. Doubling down on AI supervision, OpenAI proposes using GPT-4 to explain neurons in smaller language models. They test this method on GPT-2.

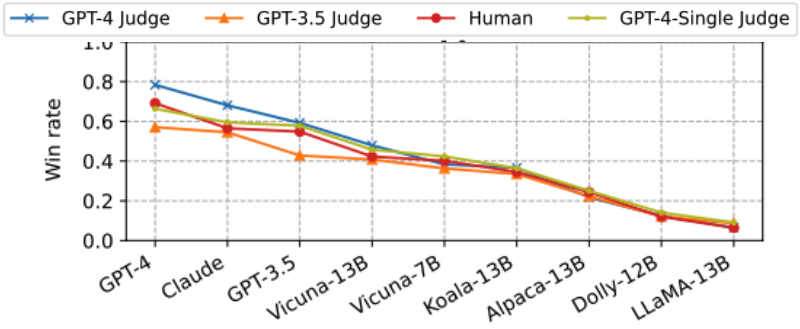
- The goal of their method is to explain which patterns in text cause a neuron to activate. GPT-4 takes as input a part the text and neuron activations, and is prompted to generate an explanation of what causes neurons to activate. Then, on other parts of text, GPT-4 is prompted to predict where neurons will most strongly respond. The researchers can then derive a similarity score between the predicted and real activations, which they dub "explanation score": *"a measure of a language model's ability to compress and reconstruct neuron activations using natural language"*.
- One worrisome fact is that the explanation score seems to decrease as the explained models get bigger.



Standard LLM benchmarks struggle with consistency

Evaluation metrics are strongly tied to their implementation, making it hard to assess the same metric evaluated using another library. Good assessments of performance are based on human pairwise comparison, but SOTA LLMs are making it increasingly difficult for human to discern the differences (in addition to it being slow and costly). A recent approach is to use LMs to evaluate other LMs.

- Judging-LLM-as-a-judge paper shows that GPT-4 reaches 80% agreement with humans (about the same level of agreement between humans!) on MT-Bench and Chatbot Arena. MT-Bench is a smaller case study with controlled human evaluation. Chatbot Arena is a large-scale crowdsourced human evaluation benchmark.



- Concerns that LLMs judges may favour they've generated have been raised. Judging-LLM-as-a-judge shows that GPT-4 favours itself with a 10% higher win rate and Claude-v1 does so with 25%. Conducting a controlled study on this is challenging, because it would require rephrasing a response to fit the style of another model.

Section 5: Predictions

10 predictions for the next 12 months

1. A Hollywood-grade production makes use of generative AI for visual effects.
2. A generative AI media company is investigated for its misuse during in the 2024 US election circuit.
3. Self-improving AI agents crush SOTA in a complex environment (e.g. AAA game, tool use, science).
4. Tech IPO markets unthaw and we see at least one major listing for an AI-focused company (e.g. Databricks).
5. The GenAI scaling craze sees a group spend >$1B to train a single large-scale model.
6. The US's FTC or UK's CMA investigate the Microsoft/OpenAI deal on competition grounds.
7. We see limited progress on global AI governance beyond high-level voluntary commitments.
8. Financial institutions launch GPU debt funds to replace VC equity dollars for compute funding.
9. An AI-generated song breaks into the Billboard Hot 100 Top 10 or the Spotify Top Hits 2024.
10. As inference workloads and costs grow significantly, a large AI company (e.g. OpenAI) acquires an inference-focused AI chip company.

Thanks!

Congratulations on making it to the end of the State of AI Report 2023! Thanks for reading.

In this report, we set out to capture a snapshot of the exponential progress in the field of artificial intelligence, with a focus on developments since last year's issue that was published on 11 October 2022. We believe that AI will be a force multiplier on technological progress in our world, and that wider understanding of the field is critical if we are to navigate such a huge transition.

We set out to compile a snapshot of all the things that caught our attention in the last year across the range of AI research, industry, politics and safety.

We would appreciate any and all feedback on how we could improve this report further, as well as contribution suggestions for next year's edition.

Thanks again for reading!

Nathan Benaich, Alex Chalmers, Othmane Sebbouh, Corina Gurau

Reviewers

We'd like to thank the following individuals for providing critical review of this year's Report:

Nikhila Ravi, Ed Hughes, Vitaly Kurin, Shubho Sengupta, Moritz Mueller-Freitag, Zehan Wang, Adam Kosiorek, Karim Beguir, Clement Delangue, Michele Catasta, Sina Samangooei, Harry Law, Max Jaderberg, Fabian Schmidt-Jakobi, Douwe Kiela, Alex Kendall, Anton Troynikov, Chris Kelly, Pablo Mendes, Gabriel Dulac-Arnold, Mehdi Ghissassi, Ken Chatfield, Marc Tuscher and Alberto Rizzoli.

Conflicts of interest

The authors declare a number of conflicts of interest as a result of being investors and/or advisors, personally or via funds, in a number of private and public companies whose work is cited in this report. Notably, the authors are investors in the following companies: airstreet.com/portfolio

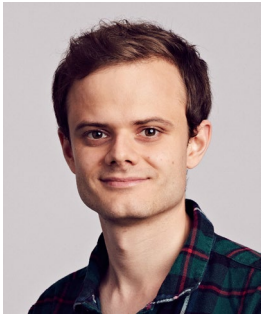
About the authors

Nathan Benaich

Nathan is the General Partner of **Air Street Capital**, a venture capital firm investing in AI-first technology and life science companies. He founded RAAIS and London.AI (AI community for industry and research), the RAAIS Foundation (funding open-source AI projects), and Spinout.fyi (improving university spinout creation). He studied biology at Williams College and earned a PhD from Cambridge in cancer research.

State of AI Report 2023 team

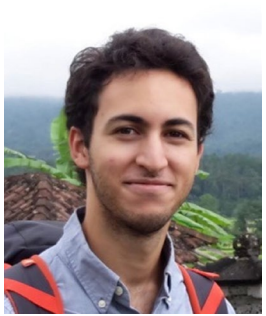
Alex Chalmers



Platform Lead

Alex is Platform Lead at **Air Street Capital**. Alex was previously Associate Director at Milltown Partners where he advised leading technology companies including AI labs.

Othmane Sebbouh



Venture Fellow

Othmane is a Venture Fellow at **Air Street Capital** and ML PhD student at ENS Paris, CREST-ENSAE and CNRS. He holds an MsC in management from ESSEC Business School and a Master in Applied Mathematics from ENSAE and Ecole Polytechnique.

Corina Gurau

Venture Fellow

Corina is a Venture Fellow at **Air Street Capital**. Corina was previously an Applied Scientist at autonomous driving company, Wayve. She holds a PhD in AI from the University of Oxford.

State of AI Report

October 12, 2023

Nathan Benaich
Air Street Capital