State of AI Report

October 11, 2022

Nathan Benaich Ian Hogarth

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.

Ian Hogarth

Ian is a co-founder at **Plural**, an investment platform for experienced founders to help the most ambitious European startups. He is a Visiting Professor at UCL working with Professor Mariana Mazzucato. Ian was co-founder and CEO of Songkick, the concert service. He started studying machine learning in 2005 where his Masters project was a computer vision system to classify breast cancer biopsy images.

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 fifth 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.

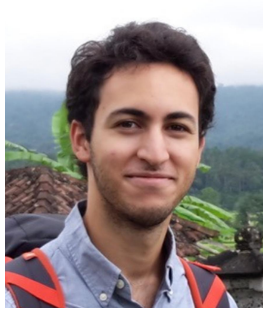
We consider the following key dimensions in our report:

- **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 2021 performance review to keep us honest.

Produced by **Nathan Benaich** (@nathanbenaich), **Ian Hogarth** (@soundboy), **Othmane Sebbouh** (@osebbouh) and **Nitarshan Rajkumar** (@nitarshan).

Thank you!

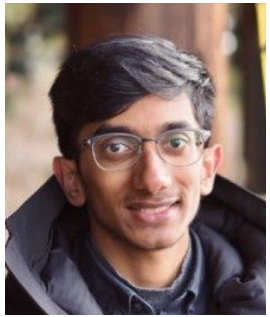
Othmane Sebbouh



Research Assistant

Othmane is a PhD student in ML 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.

Nitarshan Rajkumar



Research Assistant

Nitarshan is a PhD student in AI at the University of Cambridge. He was a research student at Mila and a software engineer at Airbnb. He holds a BSc from University of Waterloo.

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

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.

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.

Deep learning (DL): an area of ML that attempts to mimic the activity in layers of neurons in the brain to learn how to recognise complex patterns in data. The “deep” refers to the large number of layers of neurons in contemporary models that help to learn rich representations of data to achieve better performance gains.

Definitions

Model: once a ML algorithm has been trained on data, the output of the process is known as the model. This can then be used to make predictions.

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.

(Large) Language model (LM, LLM): a model trained on textual data. The most common use case of a LM is text generation. The term “LLM” is used to designate multi-billion parameter LMs, but this is a moving definition.

Computer vision (CV): enabling machines to analyse, understand and manipulate images and video.

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 language modeling, then expanded into computer vision, audio, and other modalities.

Executive Summary

Research

- Diffusion models took the computer vision world by storm with impressive text-to-image generation capabilities.
- AI attacks more science problems, ranging from plastic recycling, nuclear fusion reactor control, and natural product discovery.
- Scaling laws refocus on data: perhaps model scale is not all that you need. Progress towards a single model to rule them all.
- Community-driven open sourcing of large models happens at breakneck speed, empowering collectives to compete with large labs.
- Inspired by neuroscience, AI research are starting to look like cognitive science in its approaches.

Industry

- Have upstart AI semiconductor startups made a dent vs. NVIDIA? Usage statistics in AI research shows NVIDIA ahead by 20-100x.
- Big tech companies expand their AI clouds and form large partnerships with A(G)I startups.
- Hiring freezes and the disbanding of AI labs precipitates the formation of many startups from giants including DeepMind and OpenAI.
- Major AI drug discovery companies have 18 clinical assets and the first CE mark is awarded for autonomous medical imaging diagnostics.
- The latest in AI for code research is quickly translated by big tech and startups into commercial developer tools.

Politics

- The chasm between academia and industry in large scale AI work is potentially beyond repair: almost 0% of work is done in academia.
- Academia is passing the baton to decentralized research collectives funded by non-traditional sources.
- The Great Reshoring of American semiconductor capabilities is kicked off in earnest, but geopolitical tensions are sky high.
- AI continues to be infused into a greater number of defense product categories and defense AI startups receive even more funding.

Safety

- AI Safety research is seeing increased awareness, talent, and funding, but is still far behind that of capabilities research.

Scorecard: Reviewing our predictions from 2021

Our 2021 Prediction	Grade	Evidence
Transformers replace RNNs to learn world models with which RL agents surpass human performance in large and rich games.	Yes	DeepMind's Gato model makes progress in this direction in which a transformer predicts the next state and action, but it is not trained with RL. University of Geneva's GPT-like transformer model IRIS solves tasks in Atari environments.
ASML's market cap reaches $500B.	No	Current market cap is circa $165B (3 Oct 2022)
Anthropic publishes on the level of GPT, Dota, AlphaGo to establish itself as a third pole of AGI research.	No	Not yet.
A wave of consolidation in AI semiconductors with at least one of Graphcore, Cerebras, SambaNova, Groq, or Mythic being acquired by a large technology company or major semiconductor incumbent.	No	No new announced AI semiconductor consolidation has happened yet.
Small transformers + CNN hybrid models match current SOTA on ImageNet top-1 accuracy (CoAtNet-7, 90.88%, 2.44B params) with 10x fewer parameters.	Yes	MaxViT from Google with 475M parameters almost matched (89.53%) CoAtNet-7's performance (90.88%) on ImageNet top-1 accuracy.
DeepMind shows a major breakthrough in the physical sciences.	Yes	Three (!) DeepMind papers in mathematics and material science.
The JAX framework grows from 1% to 5% of monthly repos created as measured by Papers With Code.	No	JAX usage still accounts for <1% of monthly repos on Papers With Code.
A new AGI-focused research company is formed with significant backing and a roadmap that's focused on a sector vertical (e.g. developer tools, life science).	Yes	Adept.ai was co-founded by the authors of the Transformer and is focused on AGI via software tool use automation.

Bonus! Predictions, revisited - better late than never!

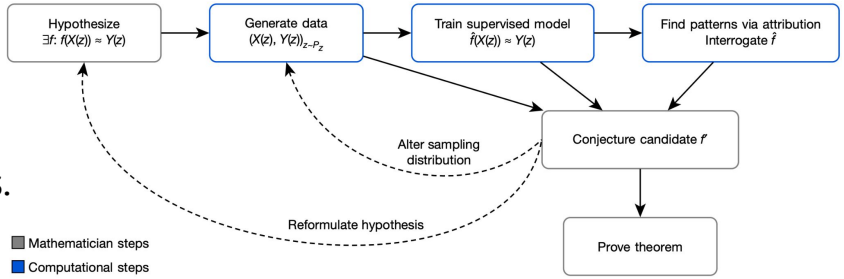
Year	Prediction	Grade	Evidence
2018	Access to Taiwanese and South Korean semiconductor companies becomes an explicit part of the trade war between US and China.	Yes	US CHIPS Act 2022 prevents recipients to expand operations in China. TSMC caught in the crosshairs.
2018	The government of an OECD country blocks the acquisition of a leading ML company by a US or Chinese HQ'd tech company.	Yes	The UK, amongst others, blocked the acquisition of Arm by NVIDIA.
2019	As AI systems become more powerful, governance of AI becomes a bigger topic and at least one major AI company makes a substantial change to their governance model.	Yes	Anthropic set up as a public benefit corporation.
2020	Facebook/Meta makes a major breakthrough in AR/VR with 3D computer vision.	Sort of	Implicitron in PyTorch3D. Not applied to AR/VR yet.
2020	Chinese and European defense-focused AI startups collectively raise >$100M in the next 12 months.	Yes	Helsing (Germany) raised $100M Series A in 2022.
2020	NVIDIA does not end up completing its acquisition of Arm.	Yes	Deal is formally cancelled in 2022.

Section 1: Research

2021 Prediction: DeepMind's breakthroughs in the physical sciences (1/3)

In 2021, we predicted: *"DeepMind releases a major research breakthrough in the physical sciences."* The company has since made significant advancements in both mathematics and materials science.

- One of the decisive moments in mathematics is formulating a conjecture, or a hypothesis, on the relationship between variables of interest. This is often done by observing a large number of instances of the values of these variables, and potentially using data-driven conjecture generation methods. But these are limited to low-dimensional, linear, and generally simple mathematical objects.

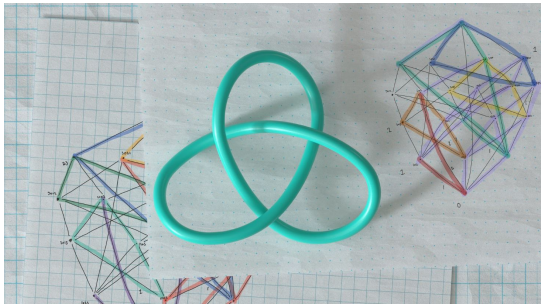


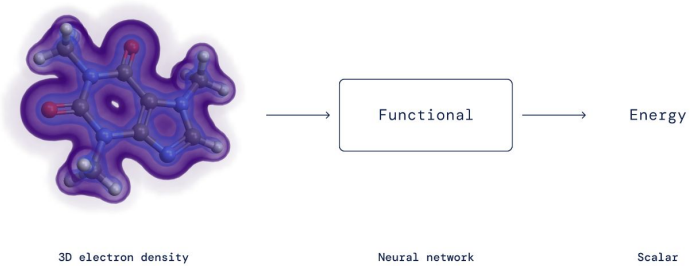
- In a Nature article, DeepMind researchers proposed an iterative workflow involving mathematicians and a supervised ML model (typically a NN). Mathematicians hypothesize a function relating two variables (input X(z) and output Y(z)). A computer generates a large number of instances of the variables and a NN is fit to the data. Gradient saliency methods are used to determine the most relevant inputs in X(z). Mathematicians can turn refine their hypothesis and/or generate more data until the conjecture holds on a large amount of data.

2021 Prediction: DeepMind's breakthroughs in the physical sciences (2/3)

In 2021, we predicted: *"DeepMind releases a major research breakthrough in the physical sciences."* The company has since made significant advancements in both mathematics and materials science.

- DeepMind researchers used their framework in a collaboration with mathematics professors from the University of Sydney and the University of Oxford to (i) propose an algorithm that could solve a 40 years-long standing conjecture in representation theory and (ii) prove a new theorem in the study of knots.
- DeepMind made an important contribution in materials science as well. It showed that the exact functional in Density Functional Theory, an essential tool to compute electronic energies, can be efficiently approximated using a neural network. Notably, instead of constraining the neural network to verify mathematical constraints of the DFT functional, researchers simply incorporate them into the training data to which they fit the NN.

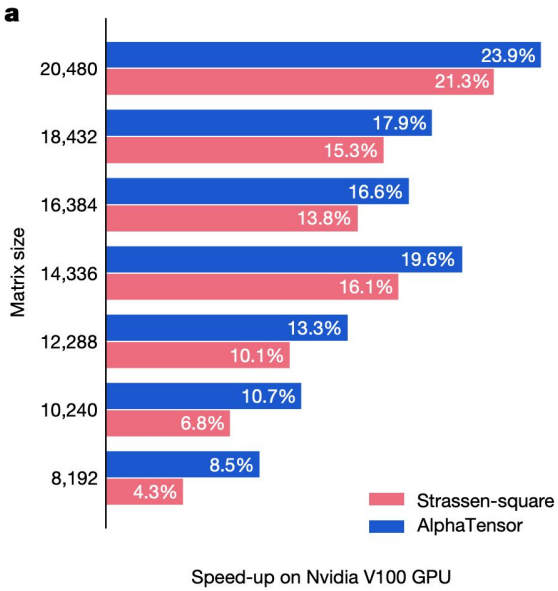




2021 Prediction: DeepMind's breakthroughs in the physical sciences (3/3)

In 2021, we predicted: *"DeepMind releases a major research breakthrough in the physical sciences."* The company has since made significant advancements in both mathematics and materials science.

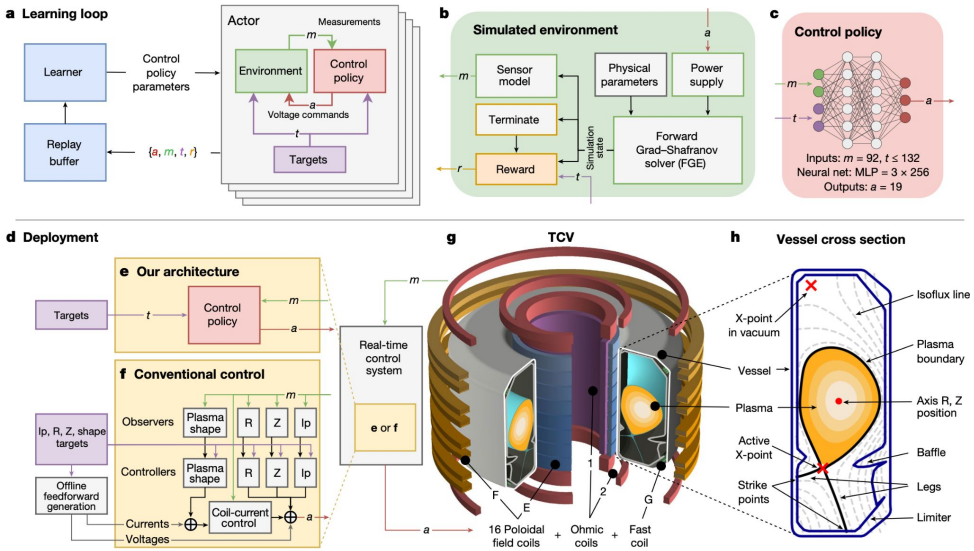
- DeepMind repurposed AlphaZero (their RL model trained to beat the best human players of Chess, Go and Shogi) to do matrix multiplication. This AlphaTensor model was able to find new deterministic algorithms to multiply two matrices. To use AlphaZero, the researchers recast the matrix multiplication problem as a single-player game where each move corresponds to an algorithm instruction and the goal is to zero-out a tensor measuring how far from correct the predicted algorithm is.
- Finding faster matrix multiplication algorithms, a seemingly simple and well-studied problem, has been stale for decades. DeepMind's approach not only helps speed up research in the field, but also boosts matrix multiplication based technology, that is AI, imaging, and essentially everything happening on our phones.



Reinforcement learning could be a core component of the next fusion breakthrough

DeepMind trained a reinforcement learning system to adjust the magnetic coils of Lausanne's TCV (Variable Configuration tokamak). The system's flexibility means it could also be used in ITER, the promising next generation tokamak under construction in France.

- A popular route to achieving nuclear fusion requires confining extremely hot plasma for enough time using a tokamak.
- A major obstacle is that the plasma is unstable, loses heat and degrades materials when it touches the tokamak's walls. Stabilizing it requires tuning the magnetic coils thousands of times per second.
- DeepMind's deep RL system did just that: first in a simulated environment and then when deployed in the TCV in Lausanne. The system was also able to shape the plasma in new ways, including making it compatible with ITER's design.

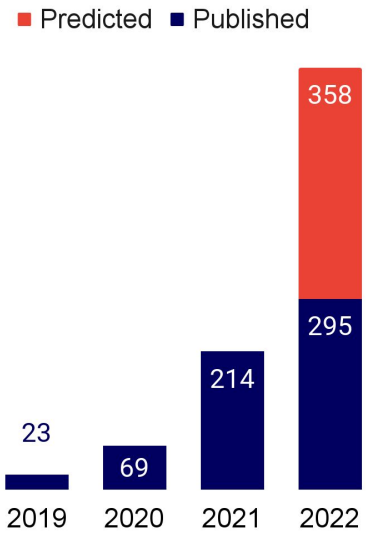


stateof.ai 2022

Predicting the structure of the entire known proteome: what could this unlock next?

Since its open sourcing, DeepMind's AlphaFold 2 has been used in hundreds of research papers. The company has now deployed the system to predict the 3D structure of 200 million known proteins from plants, bacteria, animals and other organisms. The extent of the downstream breakthroughs enabled by this technology - ranging from drug discovery to basic science - will need a few years to materialize.

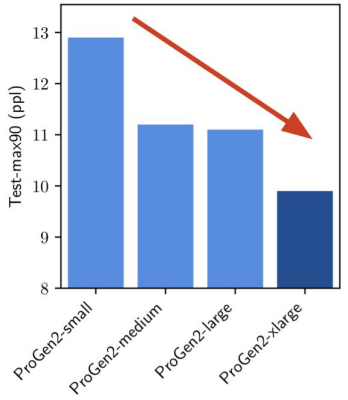
- There are 190k empirically determined 3D structures in the Protein Data Bank today. These have been derived through X-Ray crystallography and cryogenic electron microscopy.
- The first release of AlphaFold DB in July 20221 included 1M predicted protein structures.
- This new release 200x's the database size. Over 500,000 researchers from 190 countries have made use of the database.
- AlphaFold mentions in AI research literature is growing massively and is predicted to triple year on year (right chart).



Language models for proteins: a familiar story of open source and scaled models

Researchers independently applied language models to the problems of protein generation and structure prediction while scaling model parameter. They both report large benefits from scaling their models.

- Salesforce researchers find that scaling their LMs allows them to better capture the training distribution of protein sequences (as measured by perplexity).
- Using the 6B param ProGen2, they generated proteins with similar folds to natural proteins, but with a substantially different sequence identity. But to unlock the full potential of scale, the authors insist that more emphasis be placed on data distribution.
- Meta et al. introduced the ESM family of protein LMs, whose sizes range from 8M to 15B (dubbed ESM-2) parameters. Using ESM-2, they build ESMFold to predict protein structure. They show that ESMFold produces similar predictions to AlphaFold 2 and RoseTTAFold, but is an order of magnitude faster.
- This is because ESMFold doesn't rely on the use of multiple sequence alignments (MSA) and templates like AlphaFold 2 and RoseTTAFold, and instead only uses protein sequences.

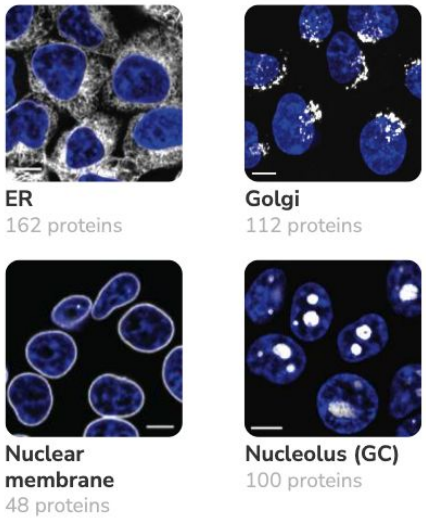


Model	# Params	Validation Perplexity	LR P@L	CASP14	CAMEO
ESM-2	8M	10.33	0.17	0.37	0.48
	35M	8.95	0.30	0.41	0.56
	150M	7.75	0.44	0.49	0.65
	650M	6.95	0.52	0.51	0.70
	3B	6.49	**0.54**	0.52	**0.72**
	15B	**6.37**	**0.54**	**0.55**	**0.72**

OpenCell: understanding protein localization with a little help from machine learning

Researchers used CRISPR-based endogenous tagging — modifying genes by illuminating specific aspects of the proteins' function — to determine protein localization in cells. They then used clustering algorithms to identify protein communities and formulate mechanistic hypotheses on uncharacterized proteins.

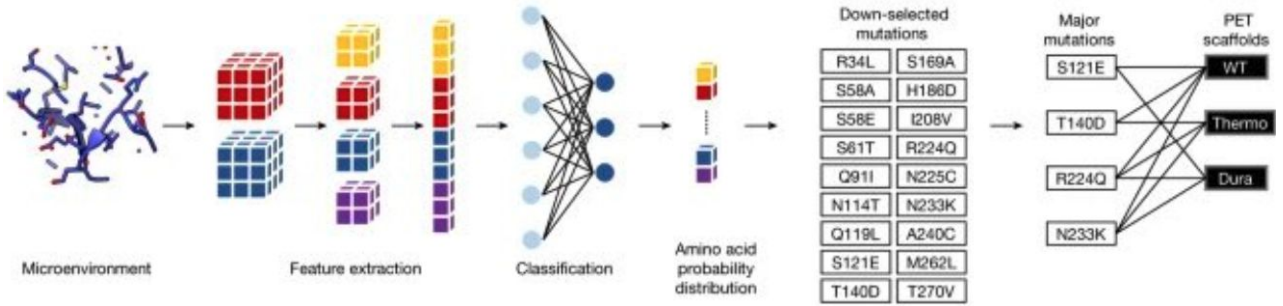
- An important goal of genomic research is to understand where proteins localize and how they interact in a cell to enable particular functions. With its dataset of 1,310 tagged proteins across ~5,900 3D images, the OpenCell initiative enabled researchers to draw important links between spatial distribution of proteins, their functions, and their interactions.
- Markov clustering on the graph of protein interactions successfully delineated functionally related proteins. This will help researchers better understand so-far uncharacterized proteins.
- We often expect ML to deliver definitive predictions. But here as with math, ML first gives partial answers (here clusters), humans then interpret, formulate and test hypotheses, before delivering a definitive answer.



Plastic recycling gets a much-needed ML-engineered enzyme

Researchers from UT Austin engineered an enzyme capable of degrading PET, a type of plastic responsible for 12% of global solid waste.

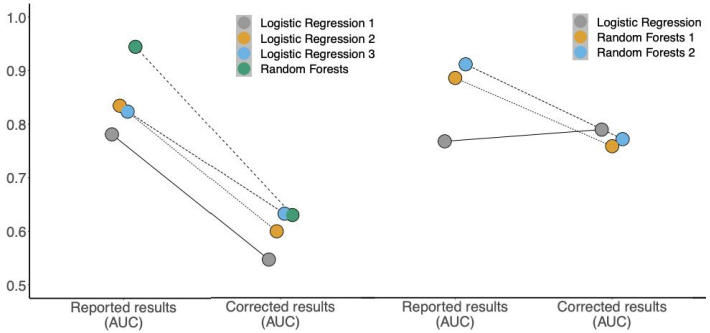
- The PET hydrolase, called FAST-PETase, is more robust to different temperatures and pH levels than existing ones.
- FAST-PETase was able to almost completely degrade 51 different products in 1 week.
- They also showed that they could resynthesize PET from monomers recovered from FAST-PETase degradation, potentially opening the way for industrial scale closed-loop PET recycling.



Beware of compounded errors: in science, ML in and garbage out?

With the increased use of ML in quantitative sciences, methodological errors in ML can leak to these disciplines. Researchers from Princeton warn of a growing reproducibility crisis in ML-based science driven in part by one such methodological error: data leakage.

- Data leakage is an umbrella term covering all cases where data that shouldn't be available to a model in fact is. The most common example is when test data is included in the training set. But the leakage can be more pernicious: when the model uses features that are a proxy of the outcome variable or when test data come from a distribution which is different from the one about which the scientific claim is made.

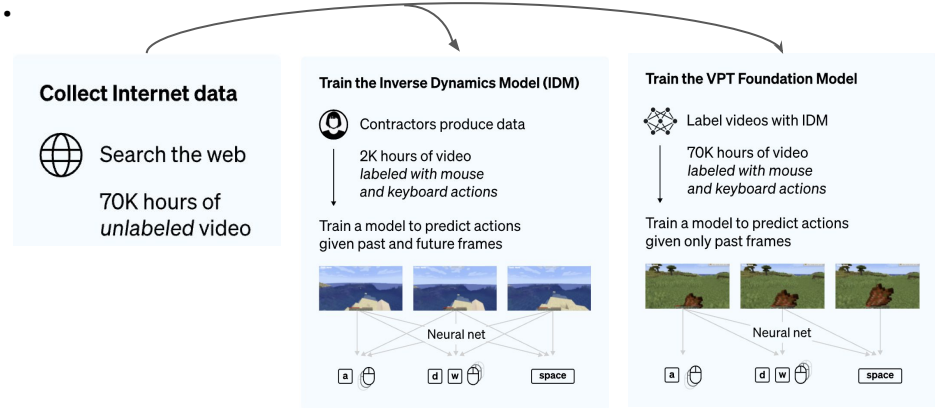


- The authors argue that the ensuing reproducibility failures in ML-based science are systemic: they study 20 reviews across 17 science fields examining errors in ML-based science and find that data leakage errors happened in every one of the 329 papers the reviews span. Inspired by the increasingly popular model cards in ML, the authors propose that researchers use model info sheets designed to prevent data leakage issues.

OpenAI uses Minecraft as a testbed for computer-using agents

OpenAI trained a model (Video PreTraining, VPT) to play Minecraft from video frames using a small amount of labeled mouse and keyboard interactions. VPT is the first ML model to learn to craft diamonds, *"a task that usually takes proficient humans over 20 minutes (24,000 actions)"*.

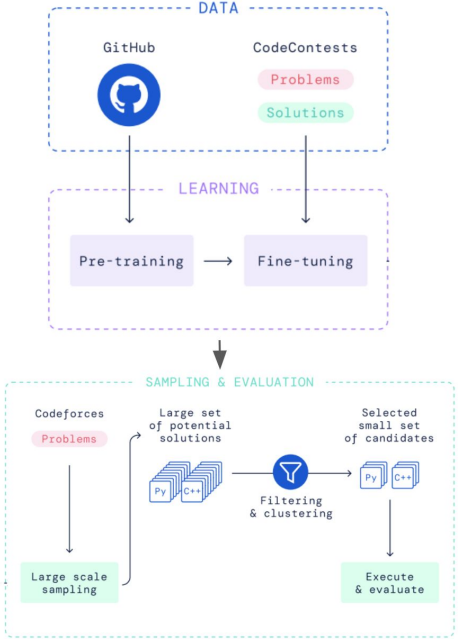
- OpenAI gathered 2,000 hours of video labeled with mouse and keyboard actions and trained an inverse dynamics model (IDM) to predict actions given past and future frames – this is the PreTraining part.
- They then used the IDM to label 70 hours of video on which they trained a model to predict actions given only past video frames.
- They show that the model can be fine-tuned with imitation learning and reinforcement learning (RL) to achieve a performance which is too hard to reach using RL from scratch.



Corporate AI labs rush into AI for code research

OpenAI's Codex, which drives GitHub Copilot, has impressed the computer science community with its ability to complete code on multiple lines or directly from natural language instructions. This success spurred more research in this space, including from Salesforce, Google and DeepMind.

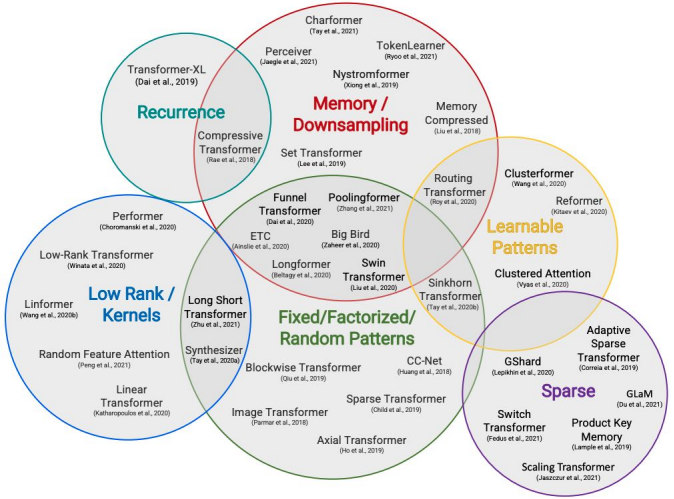
- With the conversational CodeGen, Salesforce researchers leverage the language understanding of LLMs to specify coding requirements in multiturn language interactions. It is the only open source model to be competitive with Codex.
- A more impressive feat was achieved by Google's LLM PaLM, which achieves a similar performance to Codex, but with 50x less code in its training data (PaLM was trained on a larger non-code dataset). When fine-tuned on Python code, PaLM outperformed (82% vs. 71.7% SOTA) peers on Deepfix, a code repair task.
- DeepMind's AlphaCode tackles a different problem: the generation of whole programs on competitive programming tasks. It ranked in the top half on Codeforces, a coding competitions platform. It was pre-trained on GitHub data and fine-tuned on Codeforces problems and solutions. Millions of possible solutions are then sampled, filtered, and clustered to obtain 10 final candidate submissions.



Five years after the Transformer, there must be some efficient alternative, right... right?

The attention layer at the core of the transformer model famously suffers from a quadratic dependence on its input. A slew of papers promised to solve this, but no method has been adopted. SOTA LLMs come in different flavors (autoencoding, autoregressive, encoder-decoders), yet all rely on the same attention mechanism.

- A Googol of transformers have been trained over the past few years, costing millions (billions?) to labs and companies around the world. But so-called "Efficient Transformers" are nowhere to be found in large-scale LM research (where they would make the biggest difference!). GPT-3, PaLM, LaMDA, Gopher, OPT, BLOOM, GPT-Neo, Megatron-Turing NLG, GLM-130B, etc. all use the original attention layer in their transformers.
- Several reasons can explain this lack of adoption: (i) the potential linear speed-up is only useful for large input sequences, (ii) the new methods introduce additional constraints that make the architectures less universal, (iii) the reported efficiency measures don't translate in actual computational cost and time savings.



Google

Mathematical abilities of Language Models largely surpass expectations

Built on Google's 540B parameter LM PaLM, Google's Minerva achieves a 50.3% score on the MATH benchmark (43.4 pct points better than previous SOTA), beating forecasters expectations for best score in 2022 (13%). Meanwhile, OpenAI trained a network to solve two mathematical olympiad problems (IMO).

- Google trained its (pre-trained) LLM PaLM on an additional 118GB dataset of scientific papers from arXiv and web pages using LaTeX and MathJax. Using chain of thought prompting (including intermediate reasoning steps in prompts rather than the final answer only) and other techniques like majority voting, Minerva improves the SOTA on most datasets by at least double digit pct points.
- Minerva only uses a language model and doesn't explicitly encode formal mathematics. It is more flexible but can only be automatically evaluated on its final answer rather than its whole reasoning, which might justify some score inflation.
- In contrast, OpenAI built a (transformer-based) theorem prover built in the Lean formal environment. Different versions of their model were able to solve a number of problems from AMC12 (26), AIME (6) and IMO (2) (increasing order of difficulty).

Question: Assume that the variance of the first n natural numbers is 10, and the variance of the first m even natural numbers is 16. Compute $m+n$.

Model output: Let σ^2 be the variance of the first m even natural numbers, and τ^2 be the variance of the first n natural numbers. Then, we have $\tau^2 = \frac{n^2-1}{12}$ and $\sigma^2 = \frac{m^2-1}{3}$. We solve the system of equations:

$$\frac{n^2-1}{12} = 10$$
$$\frac{m^2-1}{3} = 16.$$

This translates into the system

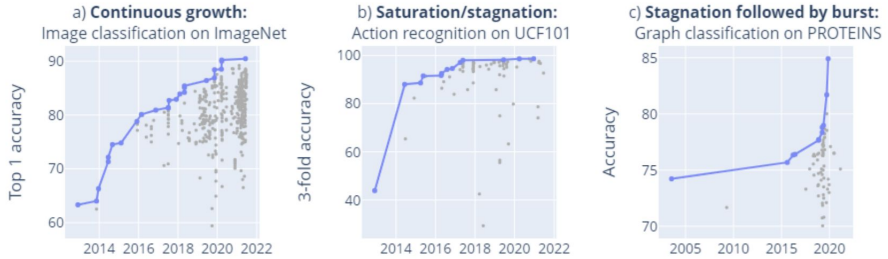
$$n^2 = 121$$
$$m^2 = 48+1 = 49.$$

Therefore, $n = \boxed{11}$ and $m = \boxed{7}$, so $n+m = \boxed{18}$.

Fast progress in LLM research renders benchmarks obsolete, but a BIG one comes to help

Only 66% of machine learning benchmarks have received more than 3 results at different time points, and many are solved or saturated soon after their release. BIG (Beyond the Imitation Game), a new benchmark designed by 444 authors across 132 institutions, aims to challenge current and future language models.

- A study from the University of Vienna, Oxford, and FHI examined 1,688 benchmarks for 406 AI tasks and identified different submission dynamics (see right).
- They note that language benchmarks in particular tend to be quickly saturated.

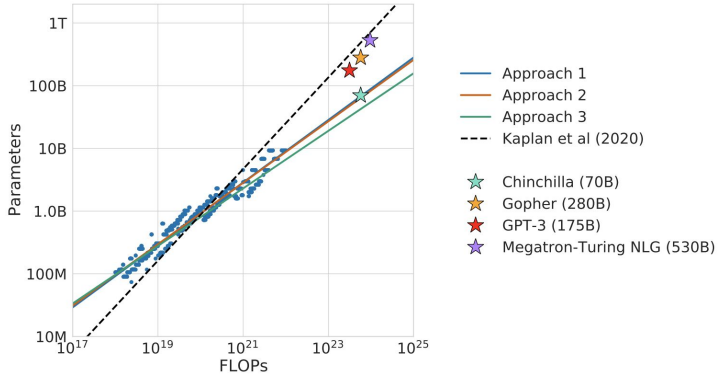


- Rapid LLM progress and emerging capabilities seem to outrun current benchmarks. As a result, much of this progress is only captured through circumstantial evidence like demos or one-off breakthroughs, and/or evaluated on disparate dedicated benchmarks, making it difficult to identify actual progress.
- The new BIG benchmark contains 204 tasks, all with strong human expert baselines, which evaluate a large set of LLM capabilities from memorization to multi-step reasoning. They show that, for now, even the best models perform poorly on the BIG benchmark.

Ducking language model scaling laws: more data please

DeepMind revisited LM scaling laws and found that current LMs are significantly undertrained: they're not trained on enough data given their large size. They train Chinchilla, a 4x smaller version of their Gopher, on 4.6x more data, and find that Chinchilla outperforms Gopher and other large models on BIG-bench.

- Empirical LM scaling laws determine, for a fixed compute budget, the model and training data sizes that should be used. Past work from OpenAI had established that model size should increase faster than training data size as the compute budget increases.
- DeepMind claims that the model size and the number of training tokens should instead increase at roughly the same rate.
- Compared to OpenAI's work, DeepMind uses larger models to derive their scaling laws. They emphasize that data scaling leads to better predictions from multibillion parameter models.

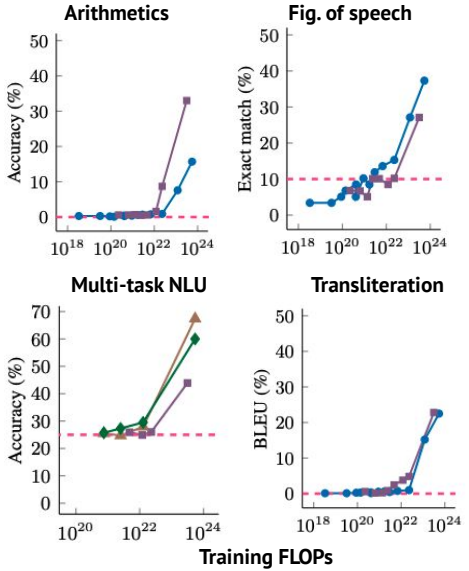


- Following these new scaling laws, Chinchilla (70B params) is trained on 1.4T tokens. Gopher (230B) on 300B.
- Though trained with the same compute budget, the lighter Chinchilla should be faster to run.

Ducking language model scaling laws: emergence

While model loss can be reasonably predicted as a function of size and compute using well-calibrated scaling laws, many LLM capabilities *emerge* unpredictably when models reach a critical size. These acquired capabilities are exciting, but the emergence phenomenon makes evaluating model safety more difficult.

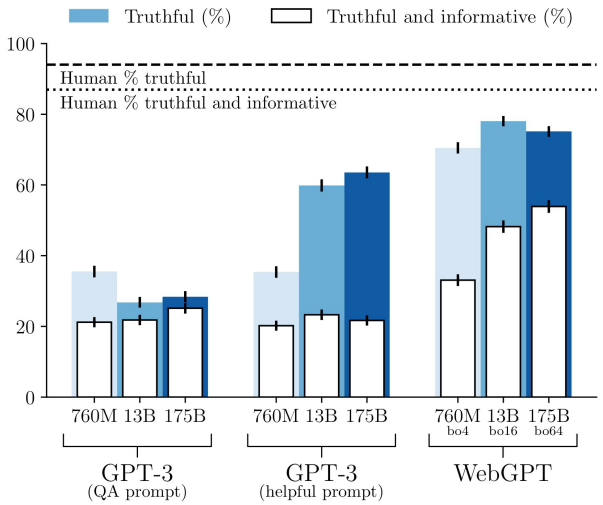
- Emergence is not fully understood: it could be that for multi-step reasoning tasks, models need to be deeper to encode the reasoning steps. For memorization tasks, having more parameters is a natural solution. The metrics themselves may be part of the explanation, as an answer on a reasoning task is only considered correct if its conclusion is. Thus despite continuous improvements with model size, we only consider a model successful when increments accumulate past a certain point.
- A possible consequence of emergence is that there are a range of tasks that are out of reach of current LLMs that could soon be successfully tackled.
- Alternatively, deploying LLMs on real-world tasks at larger scales is more uncertain as unsafe and undesirable abilities can emerge. Alongside the brittle nature of ML models, this is another feature practitioners will need to account for.



Teach a machine to fish: tool use as the next frontier?

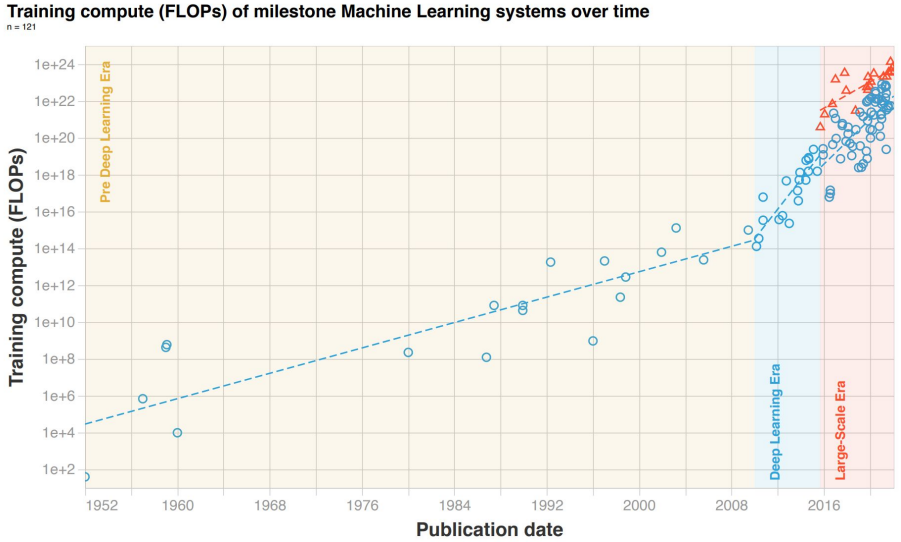
Language models can learn to use tools such as search engines and calculators, simply by making available text interfaces to these tools and training on a very small number of human demonstrations.

- OpenAI's WebGPT was the first model to demonstrate this convincingly by fine-tuning GPT-3 to interact with a search engine to provide answers grounded with references. This merely required collecting data of humans doing this task and converting the interaction data into text that the model could consume for training by standard supervised learning. Importantly, the use of increasing amounts of human demonstration data significantly increased the truthfulness and informativeness of answers (right panel, white bars for WebGPT), a significant advance from when we covered truthfulness evaluation in our 2021 report (slide 44).
- Adept, a new AGI company, is commercializing this paradigm. The company trains large transformer models to interact with websites, software applications and APIs (see more at adept.ai/act) in order to drive workflow productivity.



Looking back: three eras of compute in machine learning

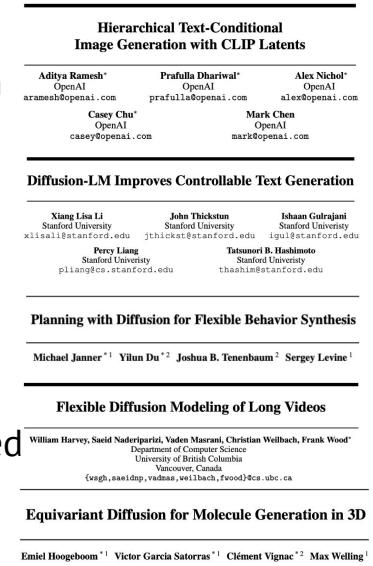
A study documents the incredible acceleration of compute requirements in machine learning. It identifies 3 eras of machine learning according to training compute per model doubling time. The Pre-Deep Learning Era (pre-2010, training compute doubled every 20 months), the Deep Learning Era (2010-15, doubling every 6 months), and the Large-Scale Era (2016-present, a 100-1000x jump, then doubling every 10 months).



Diffusion models take over text-to-image generation and expand into other modalities

When we covered diffusion models in the 2021 Report (slide 36), they were overtaking GANs in image generation on a few benchmarks. Today, they are now the undisputable SOTA for text-to-image generation, and are diffusing (pun intended) into text-to-video, text generation, audio, molecular design and more.

- Diffusion models (DMs) learn to reverse successive noise additions to images by modeling the inverse distribution (generating denoised images from noisy ones) at each step as a Gaussian whose mean and covariance are parametrized as a neural network. DMs generate new images from random noise.
- Sequential denoising makes them slow, but new techniques (like denoising in a lower-dimensional space) allow them to be faster at inference time and to generate higher-quality samples (classifier-free guidance – trading off diversity for fidelity).
- SOTA text-to-image models like DALL-E 2, Imagen and Stable Diffusion are based on DMs. They're also used in controllable text generation (generating text with a pre-defined structure or semantic context), model-based reinforcement learning, video generation and even molecular generation.



Hierarchical Text-Conditional Image Generation with CLIP Latents

Aditya Ramesh* OpenAI aramesh@openai.com
Prafulla Dhariwal* OpenAI prafulla@openai.com
Alex Nichol* OpenAI alex@openai.com
Casey Chu* OpenAI casey@openai.com
Mark Chen OpenAI mark@openai.com

Diffusion-LM Improves Controllable Text Generation

Xiang Lisa Li Stanford University xlisali@stanford.edu
John Thickstun Stanford University jthickst@stanford.edu
Ishaan Gulrajani Stanford Univeristy igul@stanford.edu
Percy Liang Stanford University pliang@cs.stanford.edu
Tatsunori B. Hashimoto Stanford University thashim@stanford.edu

Planning with Diffusion for Flexible Behavior Synthesis

Michael Janner*[1] Yilun Du*[2] Joshua B. Tenenbaum[2] Sergey Levine[1]

Flexible Diffusion Modeling of Long Videos

William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, Frank Wood*
Department of Computer Science
University of British Columbia
Vancouver, Canada
{wsgh,saeidnp,vadmas,weilbach,fwood}@cs.ubc.ca

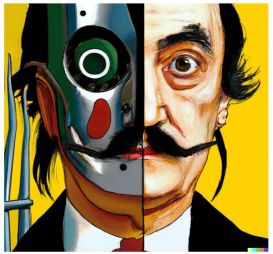
Equivariant Diffusion for Molecule Generation in 3D

Emiel Hoogeboom*[1] Vıctor Garcia Satorras*[1] Clément Vignac*[2] Max Welling[1]

DALL-E 2, Imagen and Parti...the battle for text-to-image generation rages

The second iteration of OpenAI's DALL-E, released in April 2022, came with a significant jump in the quality of generated images. Soon after, another at least equally impressive diffusion-based model came from Google (Imagen). Meanwhile, Google's Parti took a different, autoregressive, route.

DALL-E 2



vibrant portrait painting of Salvador Dalí with a robotic half face

a shiba inu wearing a beret and black turtleneck

Imagen

Teddy bears swimming at the Olympics 400m Butterfly event.

A cute corgi lives in a house made out of sushi.

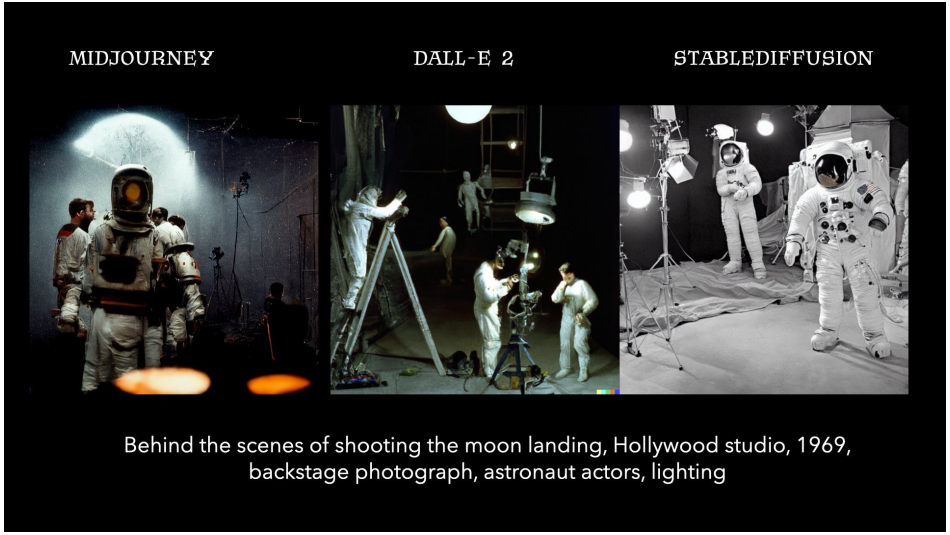
Parti-350M **Parti-20B**

A portrait photo of a kangaroo wearing an orange hoodie and blue sunglasses standing on the grass in front of the Sydney Opera House holding a sign on the chest that says Welcome Friends!

- Instead of using a diffusion model, Parti treats text-to-image generation as a simple sequence-to-sequence task, where the sequence to be predicted is a representation of the pixels of the image. Notably, as the number of parameters and training data in Parti are scaled, the model acquires new abilities like spelling.
- Other impressive text-to-image models include GLIDE (OpenAI) and Make-a-Scene (Meta — can use both text and sketches), which predate DALL-E 2, and CogView2 (Tsinghua, BAAI — both English and Chinese).

The text-to-image diffusion model frenzy gives birth to new AI labs

Stability.ai and Midjourney came out of seemingly nowhere with text-to-image models that rival those of established AI labs. Both have APIs in beta, Midjourney is reportedly profitable, and Stability has already open-sourced their model. But more on their emergence and research dynamics in our Politics section.



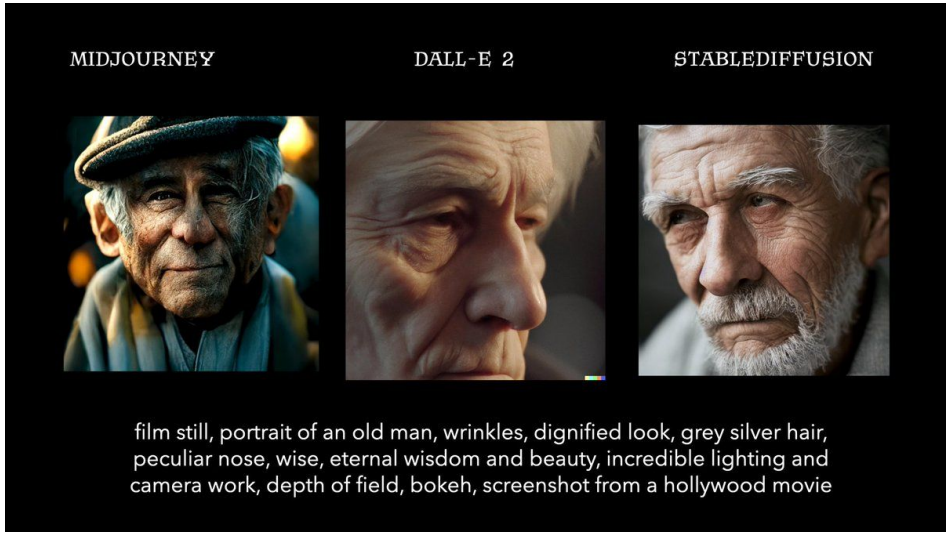


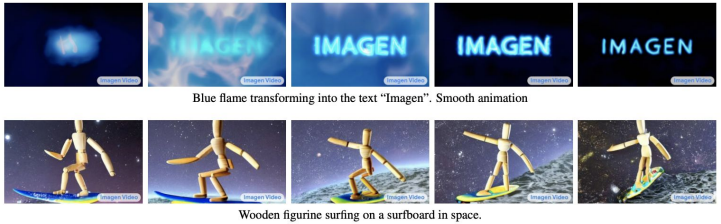
Image Credits: Fabian Stelzer

The text-to-video generation race has started

Research on diffusion-based text-to-video generation was kicked-off around April 2022, with work from Google and the University of British Columbia. But in late September, new research from Meta and Google came with a jump in quality, announcing a sooner-than-expected DALL-E moment for text-to-video generation.

Make-a-Video

Imagen Video

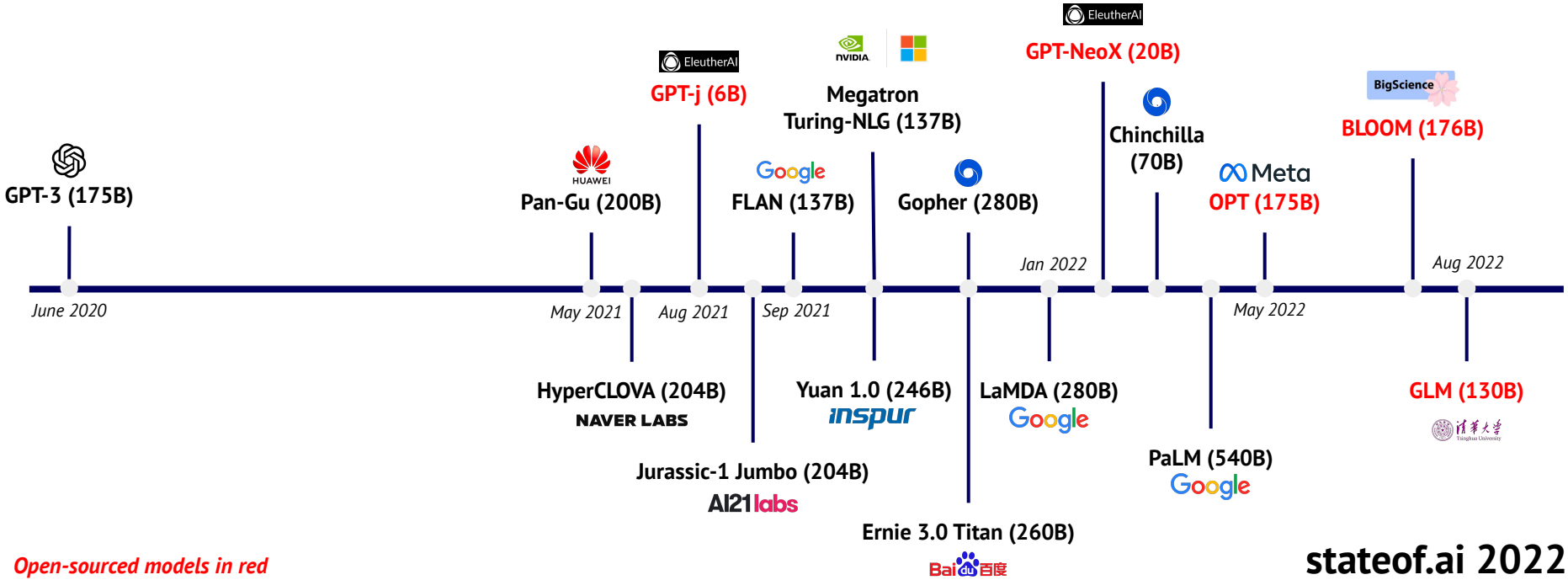


Phenaki

- Meta made the first splash from Big Tech in text-to-video generation by releasing Make-a-Video, a diffusion model for video generation.
- In an eerily similar fashion to text-to-image generation, Google then published (less than a week later) almost simultaneously two models: one diffusion-model based, Imagen, and another non diffusion-model based, Phenaki. The latter can dynamically adapt the video via additional prompts.

Closed for 14 months: community-driven open sourcing of GPT *et al.*

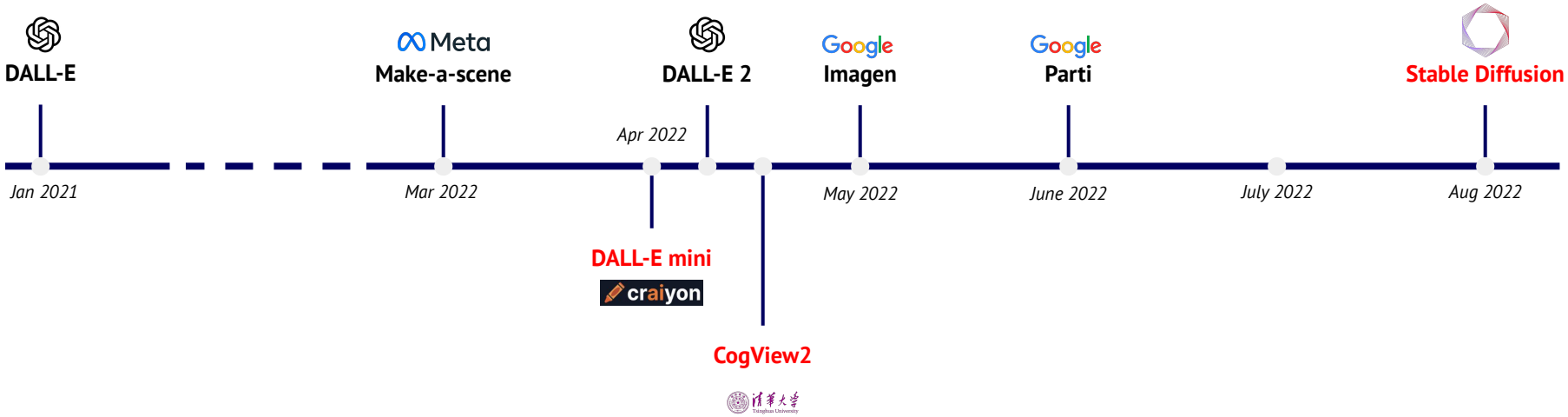
- **Landmark models from OpenAI and DeepMind have been implemented/cloned/improved by the open source community much faster than we'd have expected.**



Open-sourced models in red

Closed for 15 months: community-driven open sourcing of DALL-E *et al.*

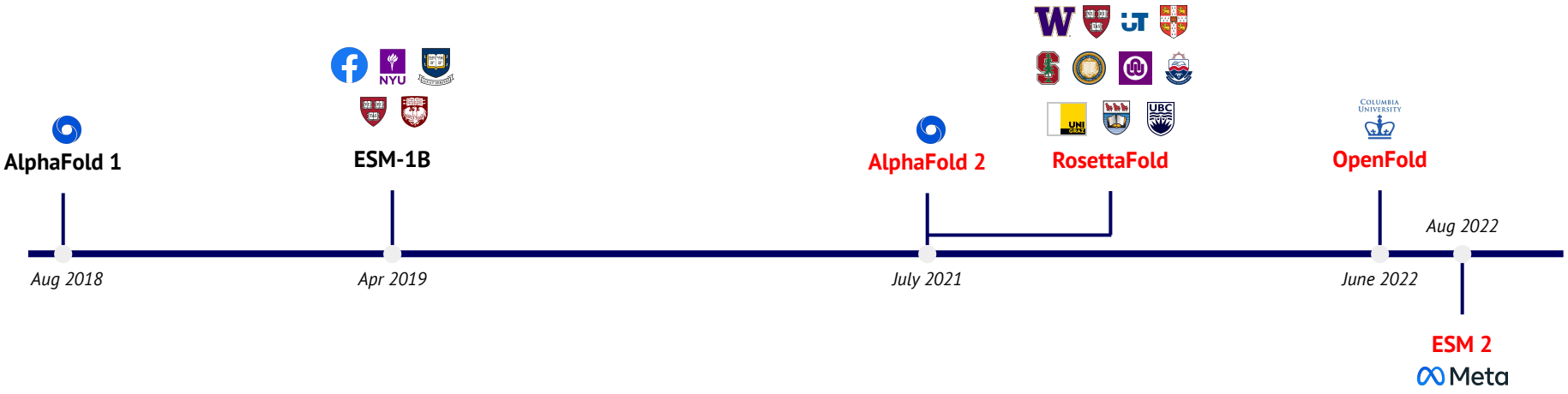
- **Landmark models from OpenAI and DeepMind have been implemented/cloned/improved by the open source community much faster than we'd have expected.**



Open-sourced models in red

Closed for 35 months: community-driven open sourcing of AlphaFold *et al.*

- **Landmark models from OpenAI and DeepMind have been implemented/cloned/improved by the open source community much faster than we'd have expected.**

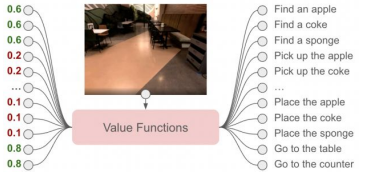


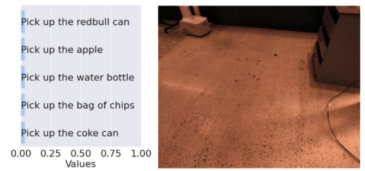
Open-sourced models in red

LLMs empower robots to execute diverse and ambiguous instructions

Thanks to their large range of capabilities, LLMs could in principle enable robots to perform any task by explaining its steps in natural language. But LLMs have little contextual knowledge of the robot's environment and its abilities, making their explanations generally infeasible for the robot. PaLM-SayCan solves this.

- Given an ambiguous instruction "I spilled my drink, can you help?", a carefully prompt-engineered LLM (e.g. Google's PaLM) can devise a sequence of abstract steps to pick up and bring you a sponge. But any given skill (e.g. pick up, put down) needs to be doable by the robot in concordance with its environment (e.g. robot sees a sponge).
- To incentivise the LLM to output feasible instructions, SayCan maximises the likelihood of an instruction being successfully executed by the robot.
- Assume the robot can execute a set of skills. Then, for any given instruction and state, the system selects the skill that maximizes: the probability of a given completion (restricted to the set of available skills) times the probability of success given the completion and the current state. The system is trained using reinforcement learning.
- Researchers tested SayCan on 101 instructions from 7 types of language instructions. It was successful in planning and execution 84% and 74% of the time respectively.



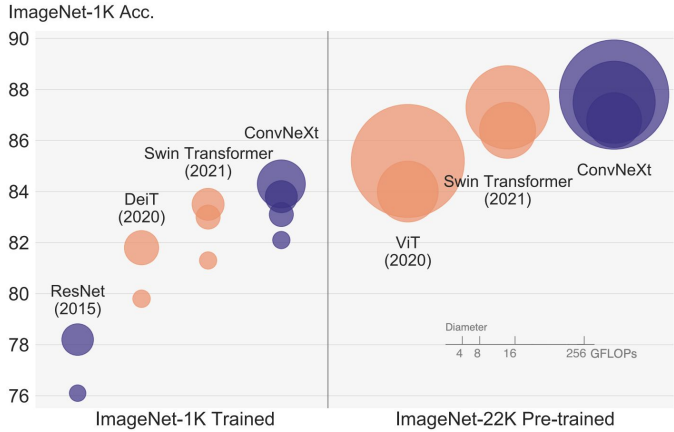


Google Everyday Robots

2021 Prediction: in vision, convolutional networks want a fair fight with transformers...

The introduction of Vision Transformers (ViT) and other image transformers last year as SOTA models on imaging benchmarks announced the dawn of ConvNets. Not so fast: work from Meta and UC Berkeley argues that modernizing ConvNets gives them an edge over ViTs.

- The researchers introduce ConvNeXt, a ResNet which is augmented with the recent design choices introduced in hierarchical vision Transformers like Swin, but doesn't use attention layers.
- ConvNeXt is both competitive with Swin Transformer and ViT on ImageNet-1K and ImageNet-22K and benefits from scale like them.
- Transformers quickly replaced recurrent neural networks in language modeling, but we don't expect a similar abrupt drop-off in ConvNets usage, especially in smaller scale ML use-cases.
- Meanwhile, our 2021 prediction of small transformers + CNN hybrid models manifested in MaxViT from Google with 475M parameters almost matching (89.53%) CoAtNet-7's performance (90.88%) on ImageNet top-1 accuracy.

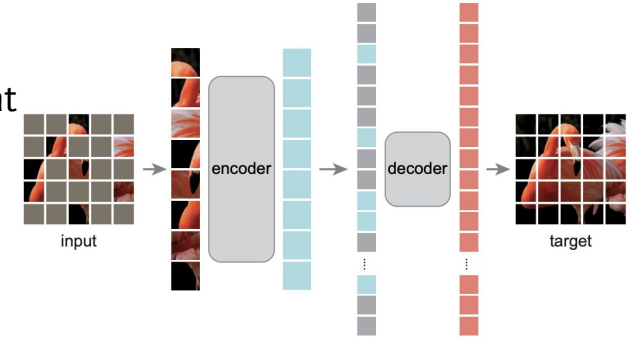


Meta Berkeley Google

...but the inevitable vision and language modeling unification continues...

Self-supervision techniques used to train transformers on text are now transposed almost as is to images and are achieving state of the art results on ImageNet-1K.

- Transformer-based autoencoder LMs are trained to predict randomly masked words in large text corpora. This results in powerful models that are SOTA in language modeling tasks (e.g. BERT).
- While masking a word in a sentence makes the sentence nonsensical and creates a challenging task for LMs, reconstructing a few randomly masked pixels in images is trivial thanks to neighbouring pixels.

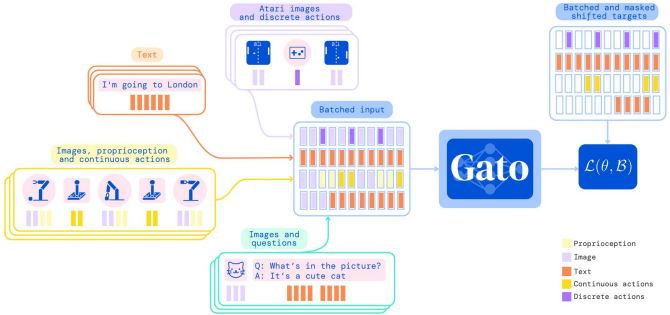


- The solution: mask large patches of pixels (e.g. 75% of the pixels). Meta use this and other adjustments (the encoder only sees visible patches, the decoder is much smaller than the encoder) to pre-train a ViT-Huge model on ImageNet-1K and then fine-tune it to achieve a task-best 87.8% top-1 accuracy.
- Self-supervised learning isn't new to computer vision (see for e.g. Meta's SEER model). Nor are masking techniques (e.g. Context encoders, or a more recent SiT). But this work is further evidence that SOTA techniques in language transition seamlessly vision. Can domains unification be pushed further?

2021 Prediction:...culminating in a single transformer to rule them all?

Transformers trained on a specific task (via supervised or self-supervised learning) can be used for a broader set of tasks via fine-tuning. Recent works show that a single transformer can be directly and efficiently trained on various tasks across different modalities (multi-task multimodal learning).

- Attempts at generalist multitask, multimodal models date back to at least Google's "One model to learn them all" (2017), which tackled 8 tasks in image, text and speech. DeepMind's Gato brings this effort to another level: researchers train a 1.2B parameter transformer to perform hundreds of tasks in robotics, simulated environments, and vision and language. This partially proves our 2021 Prediction.
- They showed that scaling consistently improved the model, but it was kept "small" for live low-latency robotics tasks.



- To train their model on different modalities, all data was serialized into a sequence of tokens which are embedded in a learned vector space. The model is trained in a fully supervised fashion.
- Separately: With data2vec, on a narrower set of tasks, Meta devised a unified self-supervision strategy across modalities. But for now, different transformers are used for each modality.

2021 Prediction: transformers for learning in world models in reinforcement learning

In 2021, we predicted: "*Transformers replace RNNs to learn world models with which RL agents surpass human performance in large and rich game environments.*" Researchers from the University of Geneva used a GPT-like transformer to simulate the world environment. They showed that their agent (dubbed IRIS) was sample efficient and surpassed human performance on 10 of the 26 games of Atari. IRIS was notably the best method among the ones that don't use lookahead search.

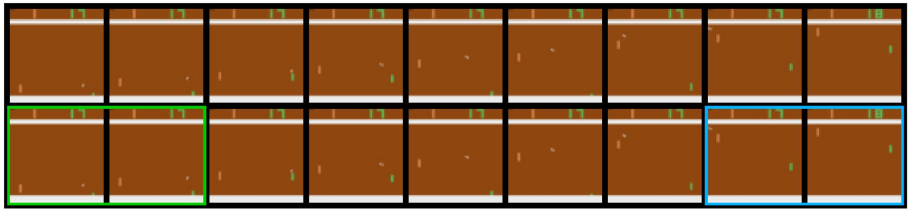


Figure 4: Pixel perfect predictions in *Pong*. The top row displays a test trajectory collected in the real environment. The bottom row depicts the reenactment of that trajectory inside the world model. More precisely, we condition the world model with the first two frames of the true sequence, in green. We then sequentially feed it the true actions and let it imagine the subsequent frames. After only 120 games of training, the world model perfectly simulates the ball's trajectory and players' movements. Notably, it also captures the game mechanic of updating the scoreboard after winning an exchange, as shown in the blue box.

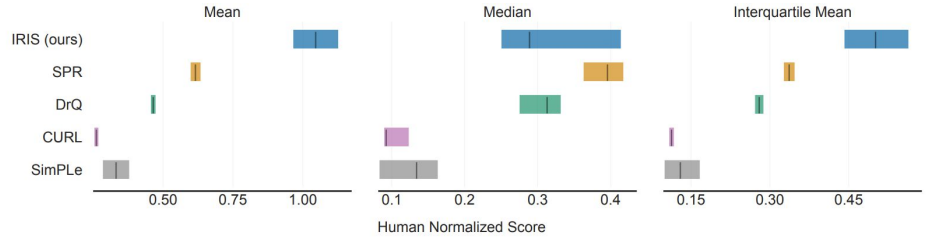
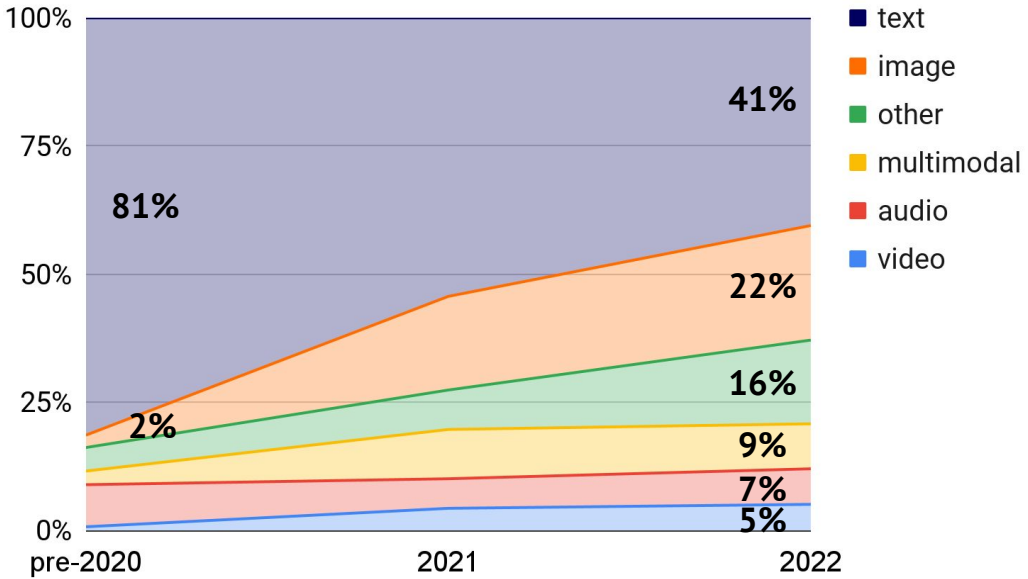


Figure 5: Mean, median, and interquartile mean human normalized scores, computed with stratified bootstrapped confidence intervals [46]. 5 runs for IRIS and SimPLe, 100 runs for SPR, CURL, and DrQ [46].

Transformers are becoming truly cross-modality

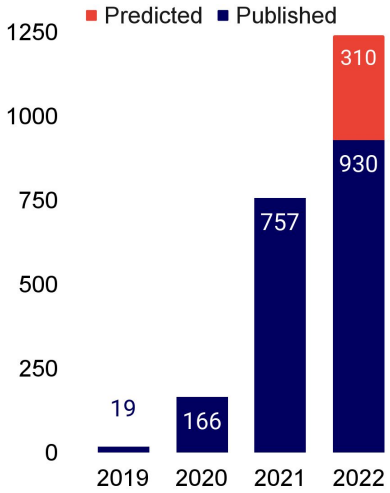
In the 2020 State of AI Report we predicted that transformers would expand beyond NLP to achieve state of the art in computer vision. It is now clear that transformers are a candidate general purpose architecture. Analysing transformer-related papers in 2022 shows just how ubiquitous this model architecture has become.



NeRFs expand into their own mature field of research

The seminal NeRF paper was published in March 2020. Since then, fundamental improvements to the methods and new applications have been quickly and continuously developed. For example, more than 50 papers on NeRF alone appeared at CVPR in 2022.

- From last year's Report (slide 18): Given multiple views of an image, NeRF uses a multilayered perceptron to learn a representation of the image and to render new views of it. It learns a mapping from every pixel location and view direction to the color and density at that location.
- Among this year's work, Plenoxels stands out by removing the MLP altogether and achieving a 100x speedup in NeRF training. Another exciting direction was rendering large scale sceneries from a few views with NeRFs, whether city-scale (rendering entire neighborhoods of San Francisco with Block-NeRF) or satellite-scale with Mega-NeRF*.



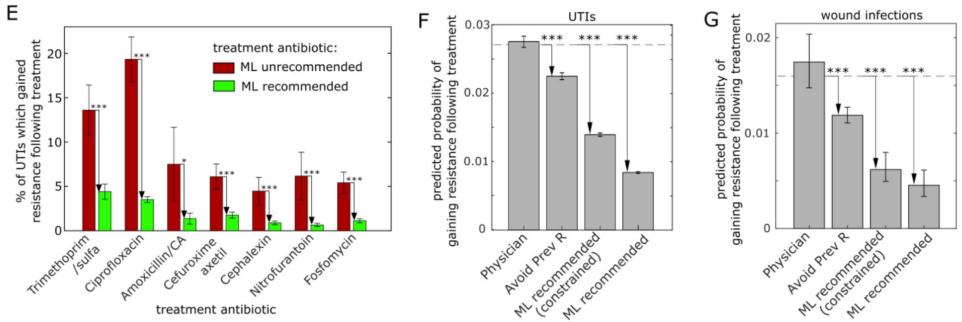
- Given the current quality of the results and the field's rate of progress, we expect that in a year or two, NeRFs will feature prominently in our industry section.

*You can better appreciate NeRF research by checking demos. E.g. Block-NeRF, NeRF in the dark, Light Field Neural Rendering

Treating bacterial infections by data-driven personalised selection of antibacterial agents

Resistance to antibacterial agents is common and often arises as a result of a different pathogen already present in a patient's body. So how should doctors find the right antibiotic that cures the infection but doesn't render the patient susceptible to a new infection?

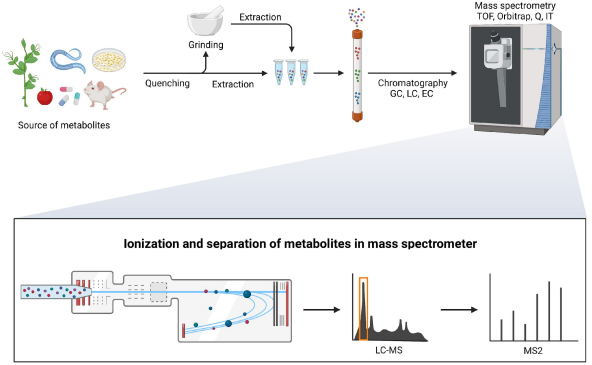
- By comparing the microbiome profiles of >200,000 patients with urinary tract or wound infections who were treated with known antibiotics before and after their infections, ML can be used to predict the risk of treatment-induced gain of resistance on a patient-specific level.
- Indeed, urinary tract infection (UTI) patients treated with antibiotics that the ML system would not have recommended resulted in significantly resistance (E). Both UTI (F) and wound infection (G) patients would suffer far fewer reinfections if they'd have been prescribed antibiotics according to the ML system.

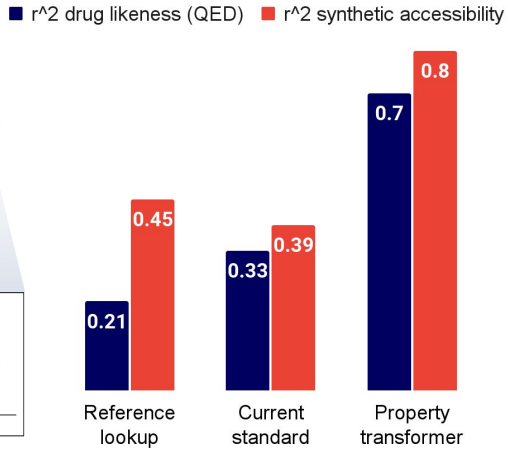


Interpreting small molecule mass spectra using transformers

Tandem mass spectrometry (MS/MS) is commonly used in metabolomics, the study of small molecules in biological samples. Less than 10% of small molecules can be identified from spectral reference libraries as most of nature's chemical space is unknown. Transformers enable fast, accurate, in silico, characterization of the molecules in metabolic mixtures, enabling biomarker and natural product drug discovery at scale.

- Very few biological samples can typically be identified from reference libraries.
- Property-prediction transformers outperform at predicting a suite of medicinally-relevant chemical properties like solubility, drug likeness, and synthetic accessibility directly from MS/MS, without using structure prediction intermediates or reference lookups.

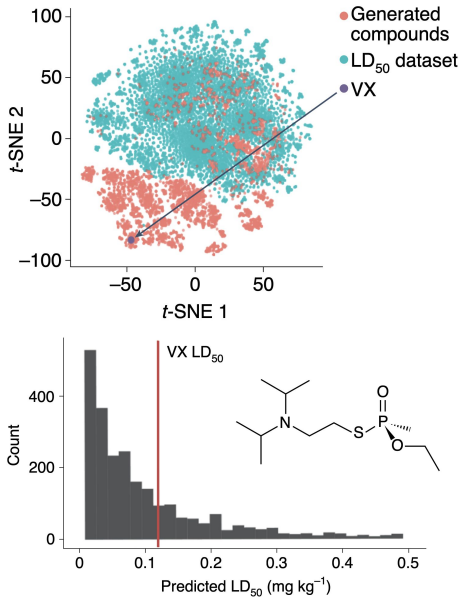




Drug discovery, the flagship "AI for good" application, is not immune to misuse

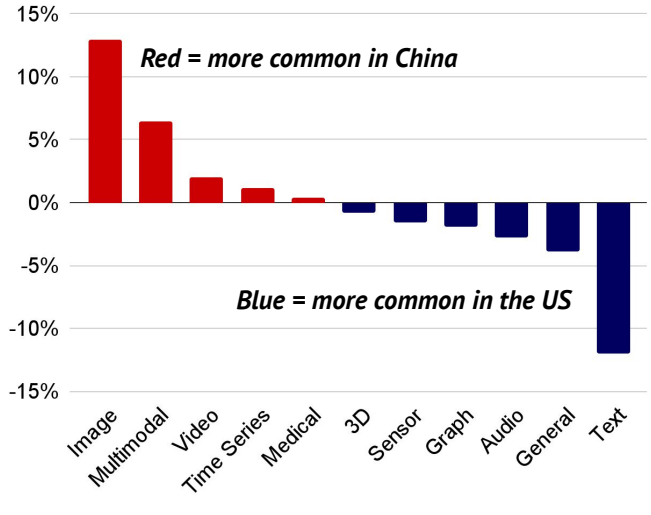
Researchers from Collaborations Pharmaceuticals and King's College London showed that machine learning models designed for therapeutic use can be easily repurposed to generate biochemical weapons.

- The researchers had trained their "MegaSyn" model to maximize bioactivity and minimize toxicity. To design toxic molecules, they kept the same model, but now simply training it to maximize both bioactivity and toxicity. They used a public database of drug-like molecules.
- They directed the model towards generation of the nerve agent VX, known to be one of the most toxic chemical warfare agents.
- However, as is the case with regular drug discovery, finding molecules with a high predicted toxicity doesn't mean it is easy to make them. But as drug discovery with AI in the loop is being dramatically improved, we can imagine best practices in drug discovery diffusing into building cheap biochemical weapons.

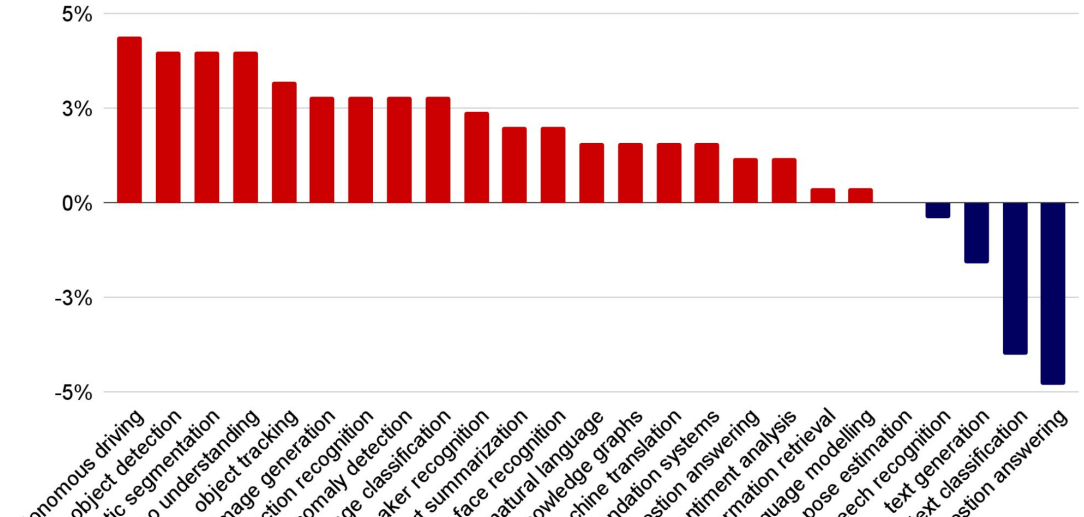


Compared to US AI research, Chinese papers focus more on surveillance related-tasks. These include autonomy, object detection, tracking, scene understanding, action and speaker recognition.

Comparing data modalities in Chinese vs. US papers

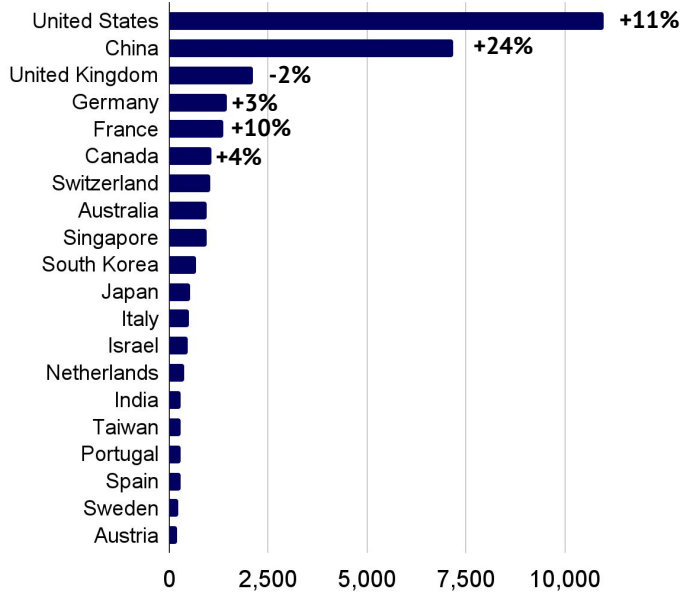


Comparing machine learning tasks in Chinese vs. US papers

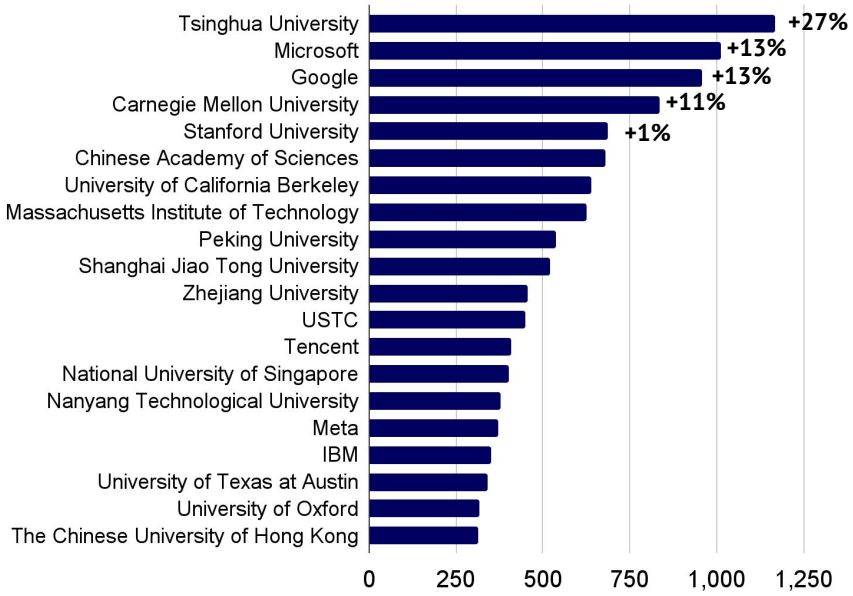


While US-based authors published more AI papers than Chinese peers in 2022, China and Chinese institutions are growing their output at a faster rate

 # papers published in 2022 and change vs. 2021

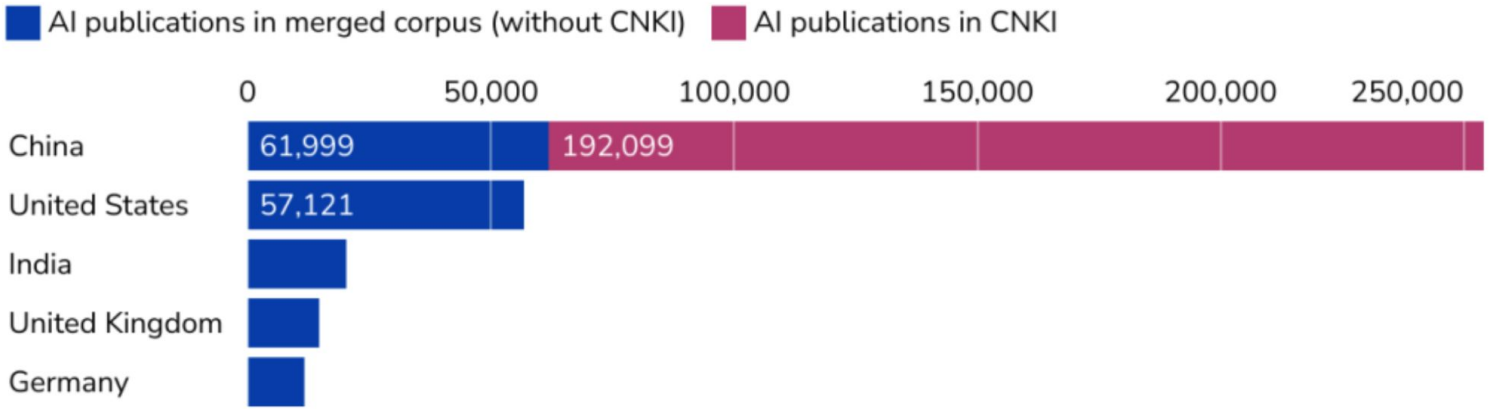


 # papers published in 2022 and change vs. 2021



The China-US AI research paper gap explodes if we include the Chinese-language database, China National Knowledge Infrastructure

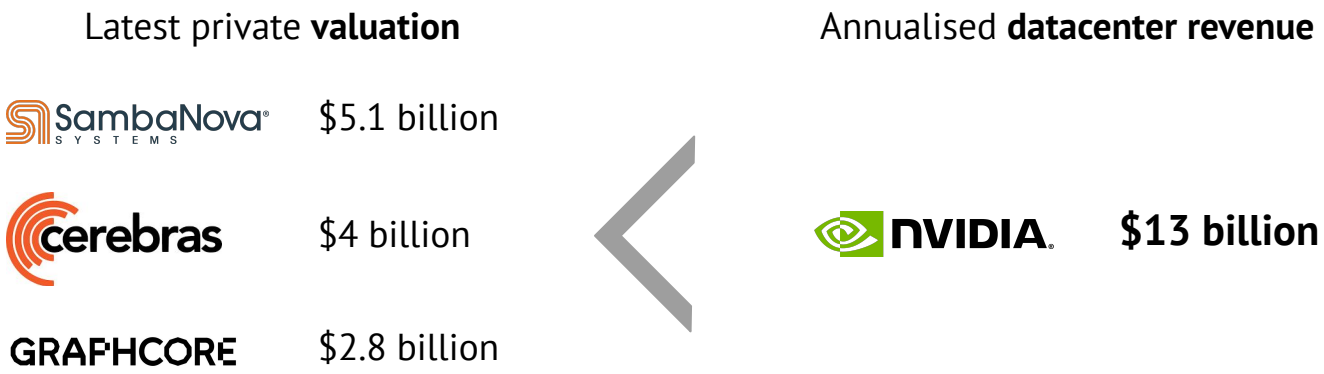
- **Chinese institutions author 4.5x the number of papers than American institutions since 2010.**



Section 2: Industry

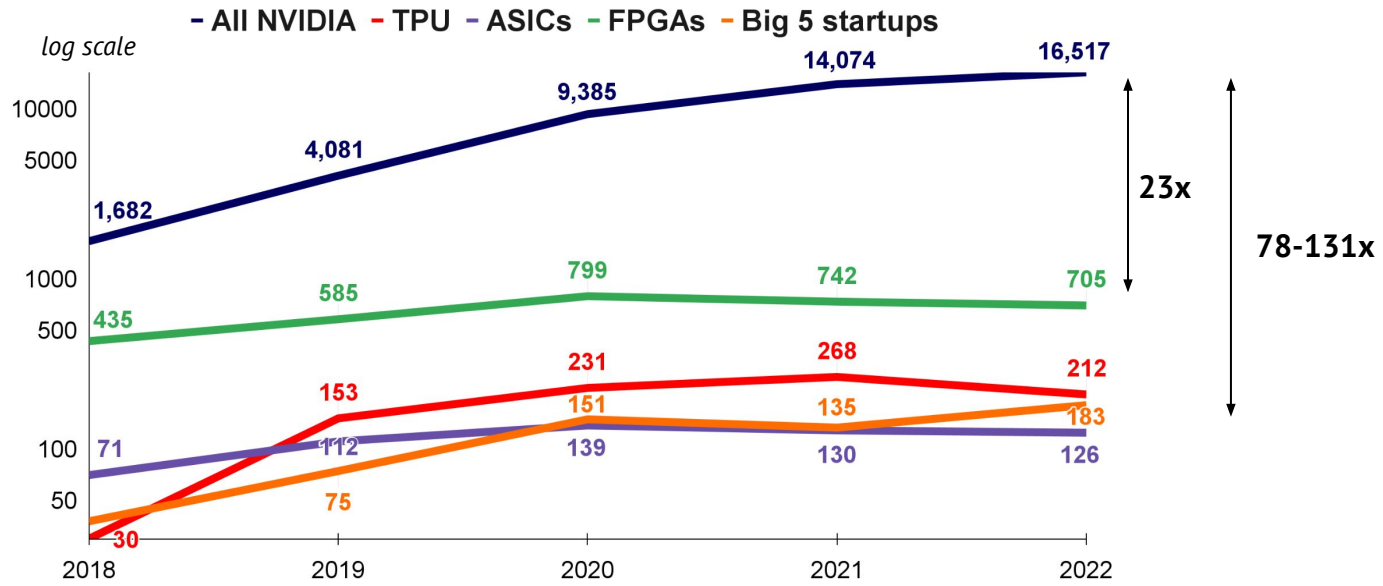
Do upstart AI chip companies still have a chance vs. NVIDIA's GPU?

NVIDIA's FY 2021 datacenter revenue came in at $10.6B. In Q4 2021, they recognised $3.26B, which on an annualised basis is greater than the combined valuation of top-3 AI semiconductor startups. NVIDIA has over 3 million developers on their platform and the company's latest H100 chip generation is expected to deliver 9x training performance vs. the A100. Meanwhile, revenue figures for Cerebras, SambaNova and Graphcore are not publicly available.



NVIDIA's chips are the most popular in AI research papers...and by a massive margin

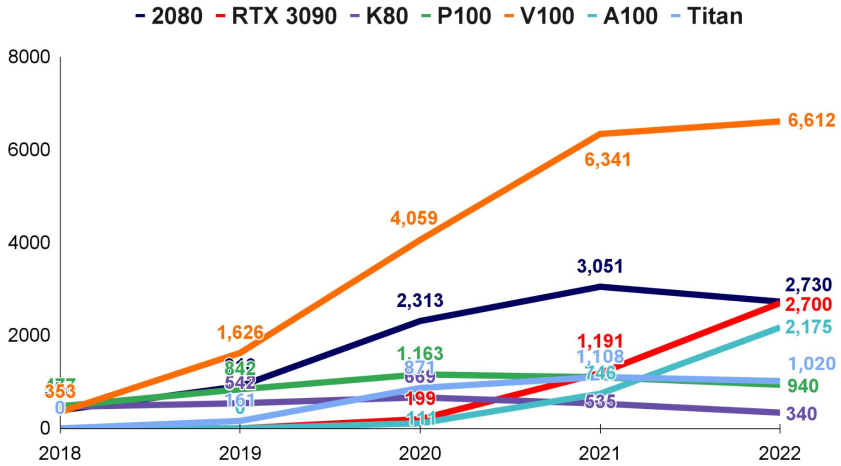
- **GPUs are 131x more commonly used than ASICs, 90x more than chips from Graphcore, Habana, Cerebras, SambaNova and Cambricon combined, 78x more than Google's TPU, and 23x more than FPGAs.**



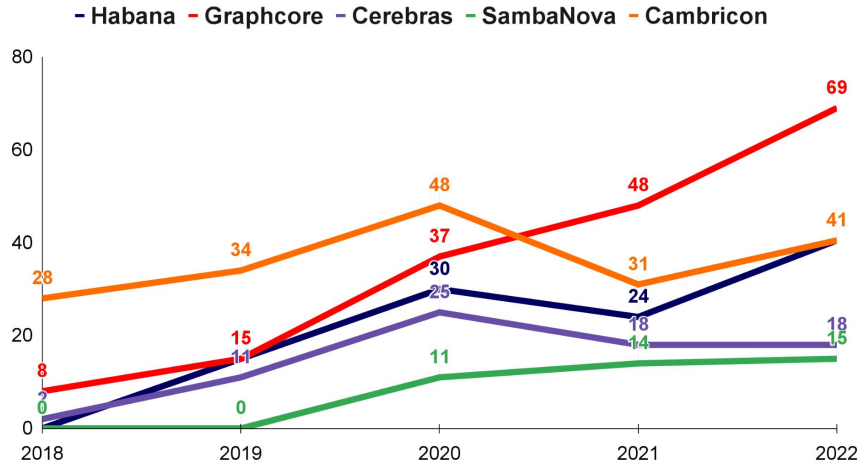
For NVIDIA, the V100 is most popular, and Graphcore is most used amongst challengers

The V100, released in 2017, is NVIDIA's workhorse chip, followed by the A100 that was released in 2020. The H100 is hotly awaited in 2022. Of the major AI chip challengers, Graphcore is cited most often.

Number of AI papers citing use of specific NVIDIA cards

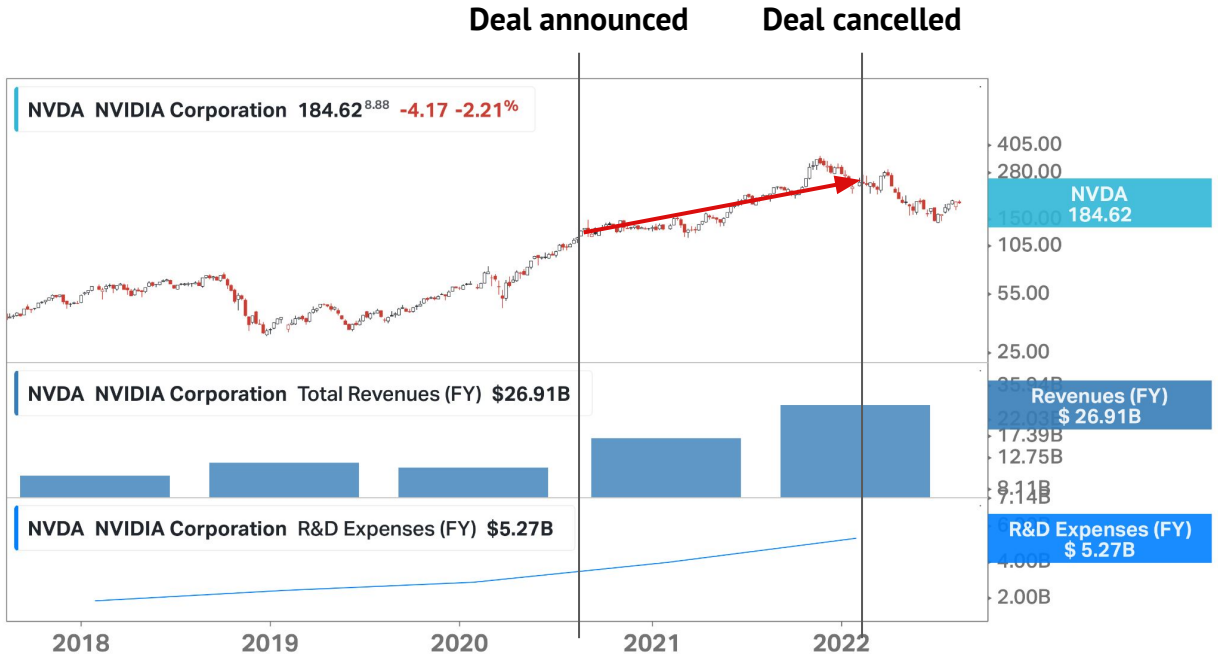


Number of AI papers citing use of specific AI chip startups



NVIDIA fails to acquire Arm and grows its revenue 2.5x and valuation 2x during the deal

- **Announced at $40B, NVIDIA's attempted acquisition of Arm fell through due to significant geopolitical and anti competition pushback. Nonetheless, NVIDIA's enterprise value grew by $295B during the period (!!)**



NVIDIA reaps rewards from investing in AI research tying up hardware and software

NVIDIA has been investing heavily in AI research and producing some of the best works in imaging over the years. For instance, their latest work on view synthesis just won the best paper award at SIGGRAPH, one of the most prestigious computer graphics conferences. But NVIDIA has now gone a step further and applied their reinforcement learning work to design their next-generation AI chip, the H100 GPU.

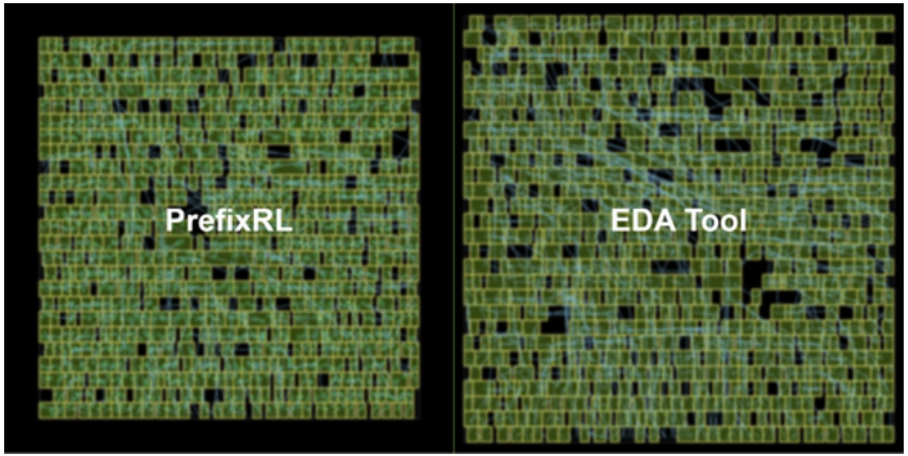


Figure 1. 64b adder circuits designed by PrefixRL AI (left) are up to 25% smaller than that designed by a state-of-the-art EDA tool (right) while being as fast and functionally equivalent

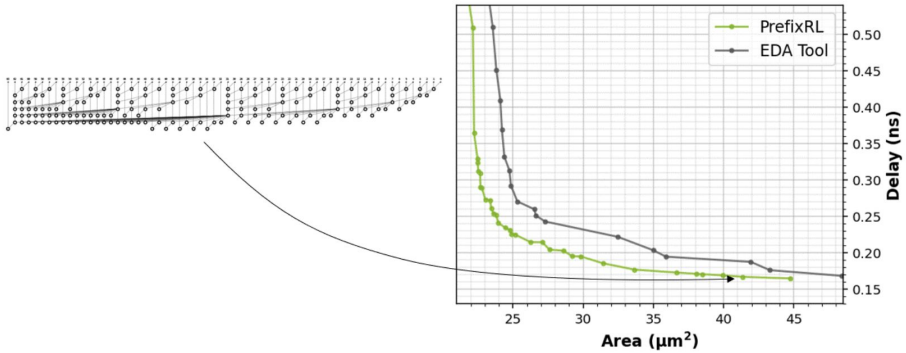


Figure 5. PrefixRL designs arithmetic circuits that are smaller and faster than circuits designed by a state-of-the-art EDA tool. (left) The circuit architectures; (right) the corresponding 64b adder circuit properties plots

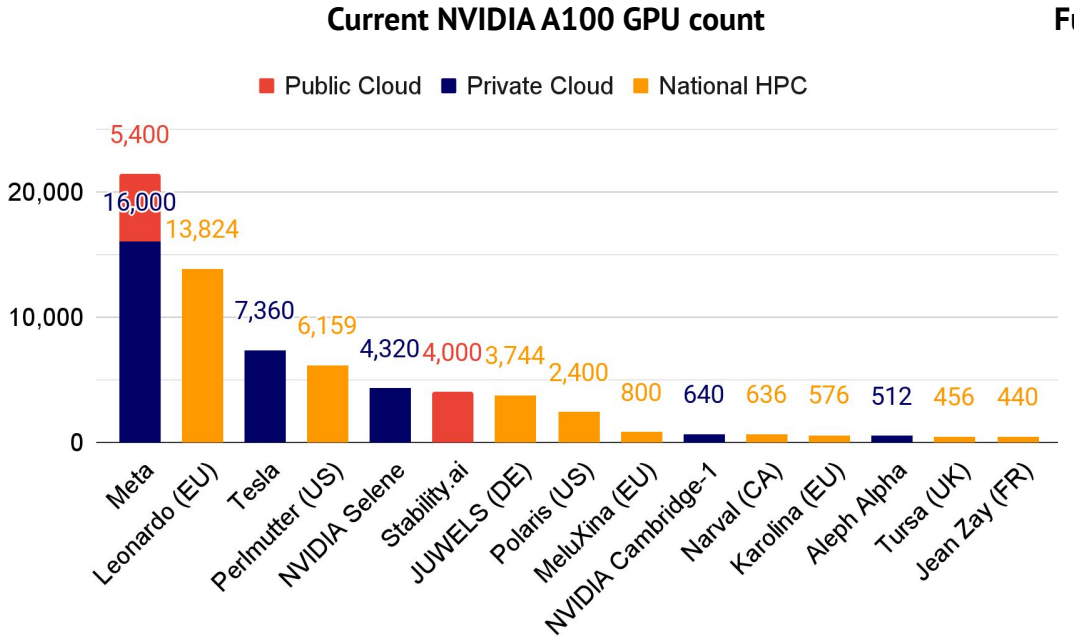
David teaming up with Goliath: training large models requires compute partnerships

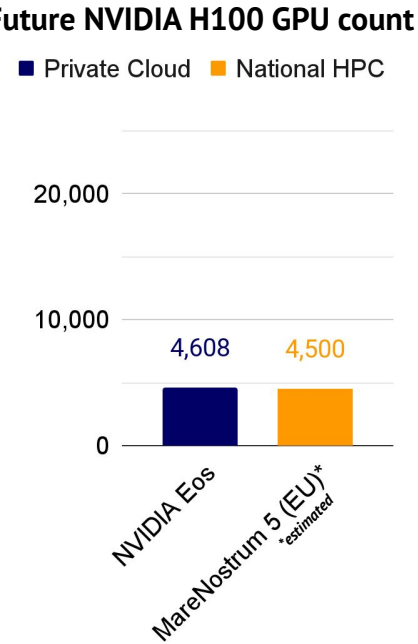
- **The hyperscalers and challenger AI compute providers are tallying up major AI compute partnerships, notably Microsoft's $1B investment into OpenAI. We expect more to come.**

Microsoft Azure	OpenAI, WAYVE, PRIMER
Google Cloud	DeepMind, co:here
aws	stability.ai, ANTHROP\C
None yet?	Inflection, Adept
NVIDIA	None yet?

In a gold rush for compute, companies build bigger than national supercomputers

"We think the most benefits will go to whoever has the biggest computer" – Greg Brockman, OpenAI CTO

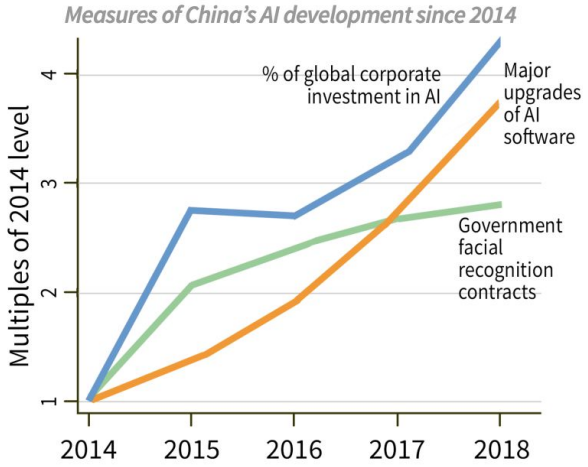




The compounding effects of government contracting in AI

In 1962 the US government bought all integrated circuits in the world, supercharging the development of this technology and its end markets. Some governments are providing that opportunity again, as "buyers of first resort" for AI companies. With access to unique high-quality data, companies could gain an edge in building consumer or enterprise AI software.

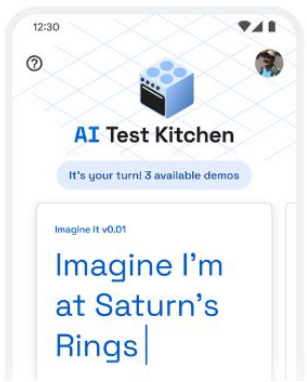
- Researchers examined Chinese facial recognition AI companies and showed a causal relationship between the number of government contracts they signed and the cumulative amount of general AI software they produced. Unsurprisingly, leadership in the computer vision space has largely been ceded to Chinese companies now.
- The principle should stand in other heavily regulated sectors, like defence or healthcare, which build an expertise through unique data that is transferable to everyday AI products.

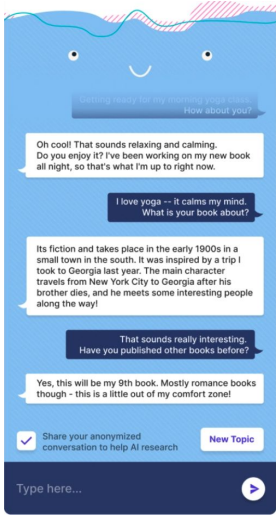


How should big tech deal with their language model consumer products?

Meta's release of the BlenderBot3 chatbot for free public use in August 2022 was faced with catastrophic press because the chatbot was spitting misinformation. Meanwhile, Google, which published a paper on their chatbot LaMDA in May 2021, had decided to keep the system in-house. But a few weeks after BlenderBot's release, Google announced a larger initiative called "AI test kitchen", where regular users will be able to interact with Google's latest AI agents, including LaMDA.

- Large-scale release of AI systems to the 1B+ users of Google and Facebook all but ensures that every ethics or safety issue with these systems will be surfaced, either by coincidence or by adversarially querying them. But only by making these systems widely available can these companies fix those issues, understand user behaviour and create useful and profitable systems.
- Running away from this dilemma, 4 of the authors of the paper introducing LaMDA went on to found/join Character.AI, which describes itself as *"an AI company creating revolutionary open-ended conversational applications"*. Watch this space…

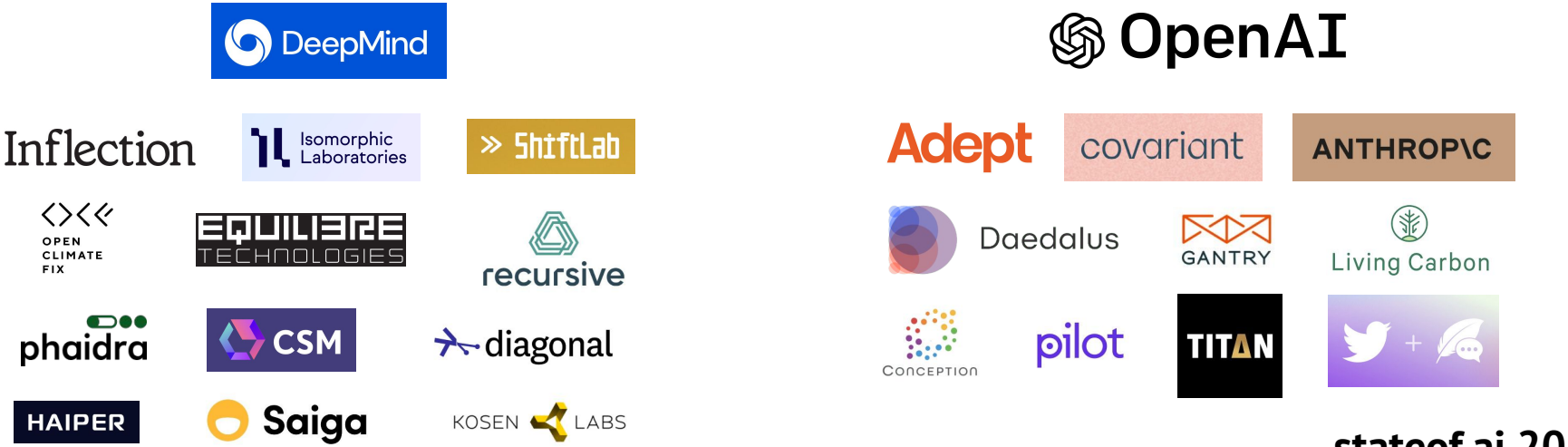




Meta Google Character.AI

DeepMind and OpenAI alums form new startups and Meta disbands its core AI group

- **Once considered untouchable, talent from Tier 1 AI labs is breaking loose and becoming entrepreneurial. Alums are working on AGI, AI safety, biotech, fintech, energy, dev tools and robotics. Others, such as Meta, are folding their centralised AI research group after letting it run free from product roadmap pressure for almost 10 years. Meta concluded that "while the centralized nature of the [AI] organization gave us leverage in some areas it also made it a challenge to integrate as deeply as we would hope."**



Attention is all you need... to build your AI startup

- All but one author of the landmark paper that introduced transformer-based neural networks have left Google to build their own startups in AGI, conversational agents, AI-first biotech and blockchain.

	Amount raised in 2022
ANTHROP\C	$580M
Inflection	$225M
co:here	$125M
Adept	$65M

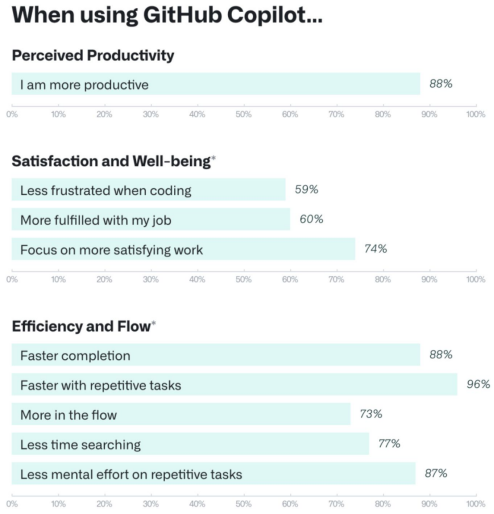
AI coding assistants are deployed fast, with early signs of developer productivity gains

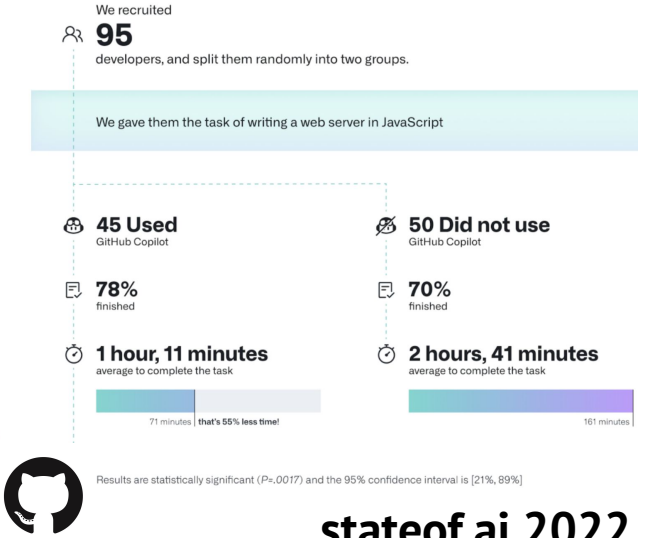
OpenAI's Codex quickly evolved from research (July 2021) to open commercialization (June 2022) with (Microsoft's) GitHub Copilot now publicly available for $10/month or $100/year. Amazon followed suit by announcing CodeWhisperer in preview in June 2022. Google revealed that it was using an internal ML-powered code completion tool (so maybe in a few years in a browser IDE?). Meanwhile, with its 1M+ users, tabnine raised $15M, promising accurate multiline code completions.

Google

	Single line	Multi-line
Fraction of code added by ML	2.6%	0.6%
Average characters per accept	21	73
Acceptance rate (for suggestions visible for >750ms)	25%	34%
Reduction in coding iteration duration	6%	-

Metrics for Google's coding assistant. Users are 10k+ Google-internal developers (5k+ for multi-line experiments).





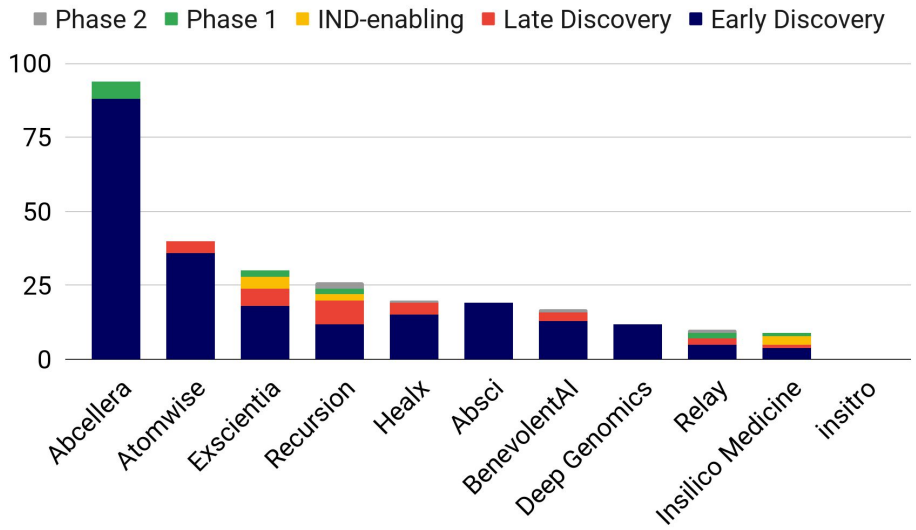
GitHub Copilot Google amazon tabnine

stateof.ai 2022

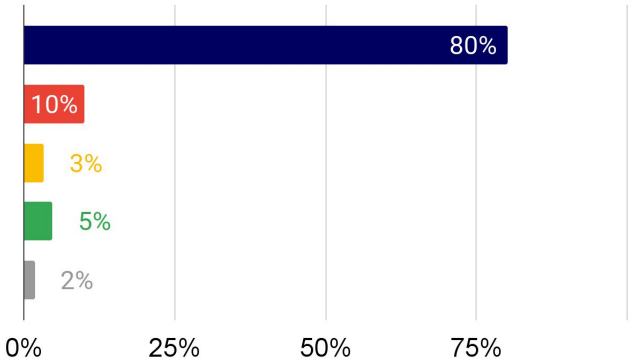
AI-first drug discovery companies have 18 assets in clinical trials, up from 0 in 2020

And many more assets in early discovery stages. We expect early clinical trial readouts from 2023 onwards.

of assets per pipeline stage per company



% of assets per pipeline stage overall

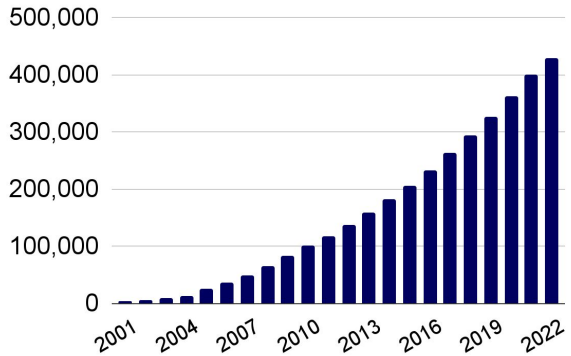


Updated as of 26 Aug 2022

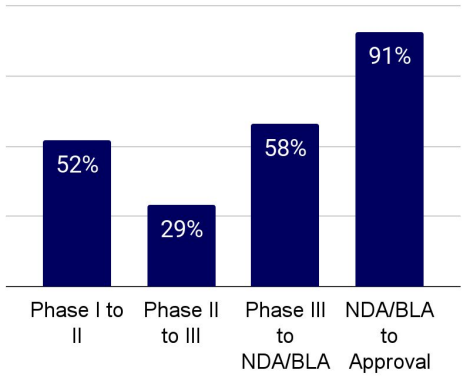
Can AI and compute bend the physical reality of clinical trial chokepoints?

A study of 6,151 successful phase transitions between 2011–2020 found that it takes 10.5 years on average for a drug to achieve regulatory approval. This includes 2.3 years at Phase I, 3.6 years at Phase II, 3.3 years at Phase III, and 1.3 years at the regulatory stage. What's more, it costs $6.5k on average to recruit one patient into a clinical trial. With 30% of patients eventually dropping out due to non-compliance, the fully-loaded recruitment cost is closer to $19.5k/patient. While AI promises better drugs faster, we need to solve for the physical bottlenecks of clinical trials today.

# of registered studies (ClinicalTrials.gov EOY)



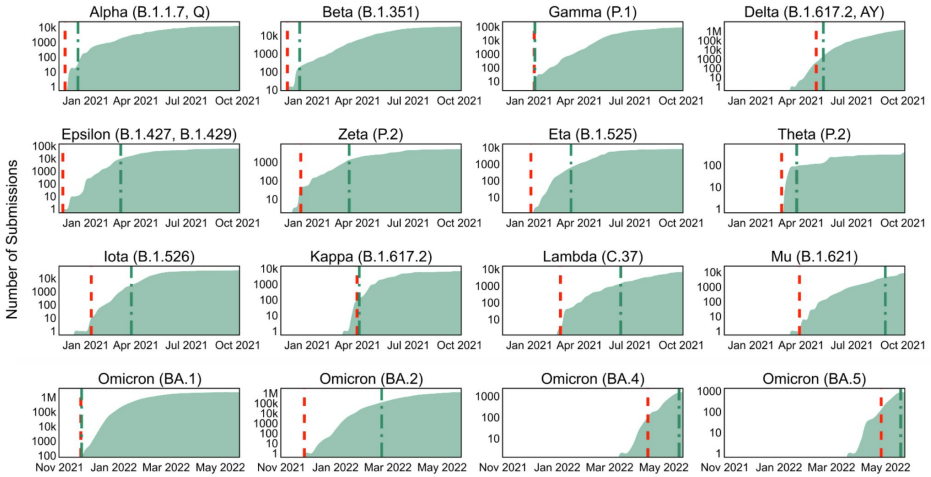
Stepwise probability of drug success



Predicting the evolution of real-world covid variants using language models

mRNA vaccine leader, BioNTech, and enterprise AI company, InstaDeep, collaboratively built and validated an *Early Warning System (EWS)* to predict high-risk variants. The EWS could identify all 16 WHO-designated variants on average more than one and a half months prior to officially receiving the designation.

- A large pre-trained protein language model was trained on viral spike protein sequences of variants.
- New spike protein variants are fed to a transformer that outputs embeddings and a probability distribution of the 20 natural amino acids for each position to determine how this would affect immune escape and fitness.
- The red dash line indicates the date when the EWS predicted the variant would be high-risk and the green dash-dot line is when the WHO designated the variant. In almost all cases, EWS alerted several months before the WHO designation.



InstaDeep™ BIONTECH

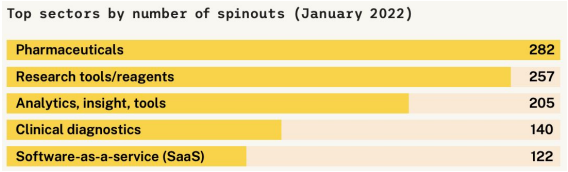
The first regulatory approval for an autonomous AI-first medical imaging diagnostic

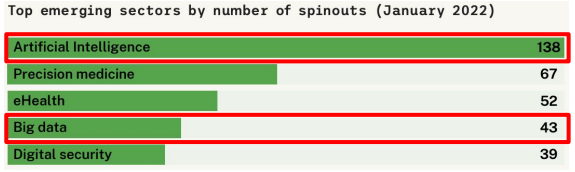
Lithuanian startup Oxipit received the industry's first autonomous certification for their computer vision-based diagnostic. The system autonomously reports on chest X-rays that feature no abnormalities, removing the need for radiologists to look at them.

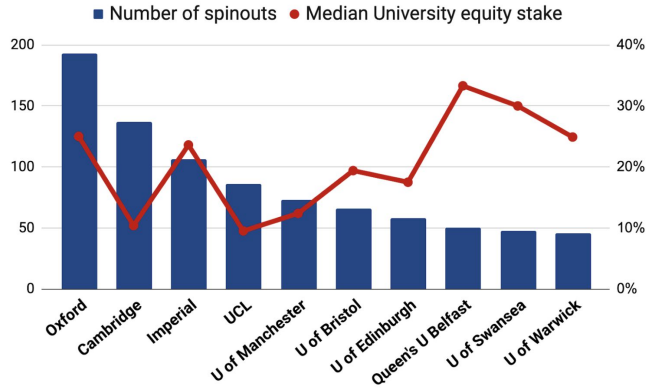
- Due to a shortage of radiologists and an increasing volume of imaging, the diagnostic task of assessing which X-rays contain disease and which don't is challenging.
- Oxipit's ChestLink is a computer vision system that is tasked with identifying scans that are normal.
- The system is trained on over a million diverse images. In a retrospective study of 10,000 chest X-rays of Finnish primary health care patients, the AI achieved a sensitivity of 99.8% and specificity of 36.4 % for recognising clinically significant pathology on a chest X-ray.
- As such, the AI could reliably remove 36.4% of normal chest X-rays from a primary health care population data set with a minimal number of false negatives, leading to effectively no compromise on patient safety and a potential significant reduction of workload.

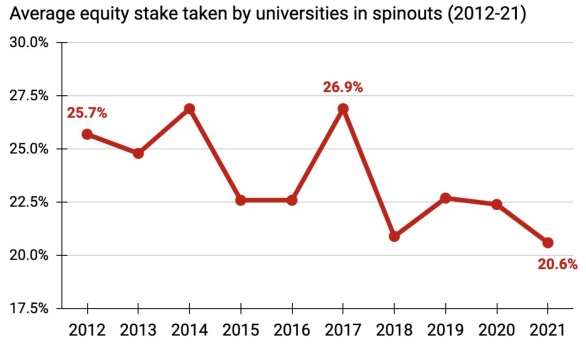
Universities are a hotbed for AI spinouts: the UK case study

Universities are an important source of AI companies including Databricks, Snorkel, SambaNova, Exscientia and more. In the UK, 4.3% of UK AI companies are university spinouts, compared to 0.03% for all UK companies. AI is indeed among the most represented sectors for spinouts formation. But this comes at a steep price: Technology Transfer Offices (TTOs) often negotiate spinout deal terms which are unfavourable to founders, e.g. a high equity share in the company or royalties on sales.



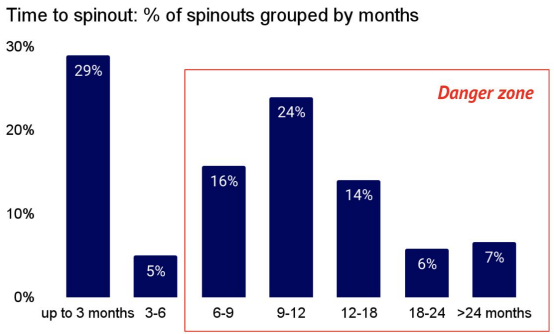


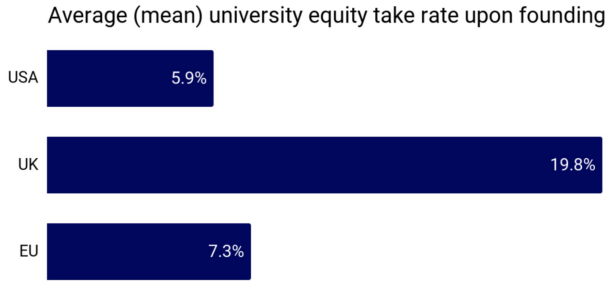


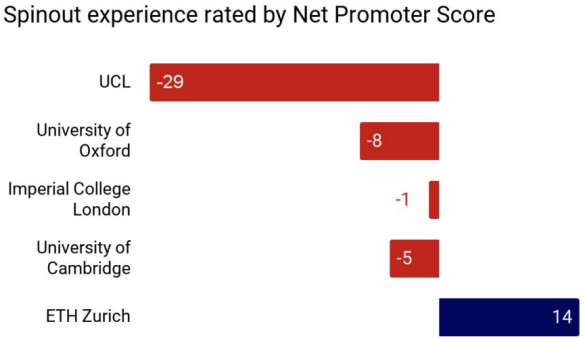


Spinout.fyi: an open database to help founders and policymakers fix the spinout problem

Spinout.fyi crowdsourced a database of spinout deal terms from founders representing >70 universities all over the world. The database spans AI and non-AI companies across different product categories (software, hardware, medical, materials, etc.), and shows that the UK situation, while particularly discouraging for founders, isn't isolated. Only a few regions stand out as being founder-friendly, like the Nordics and Switzerland (ETH Zürich in particular). A major reason for the current situation is the information asymmetry between founders and TTOs, and the spinout.fyi database aims to give founders a leg up in the process.







As 5-year programmes in Berkeley and Stanford wrap up, what comes next?

In 2011, UC Berkeley launched the "Algorithms, Machines, and People" (AMPLab) as a 5-year collaborative research agenda amongst professors and students, supported by research agencies and companies. The program famously developed the critical Big Data technology Spark (spun out as Databricks), as well as Mesos (spun out as Mesosphere). This hugely successful program was followed in 2017 by the "Real-time intelligence secure explainable systems" (RISELab) at Berkeley and "Data Analytics for What's Next" (DAWN) at Stanford, which focused on AI technologies. RISELab created the Ray ML workload manager (spun out as Anyscale), and DAWN created and spun out the Snorkel active labelling platform. Will other universities and countries learn from the successes of the 5-year model to fund ambitious open-source research with high spinout potential?

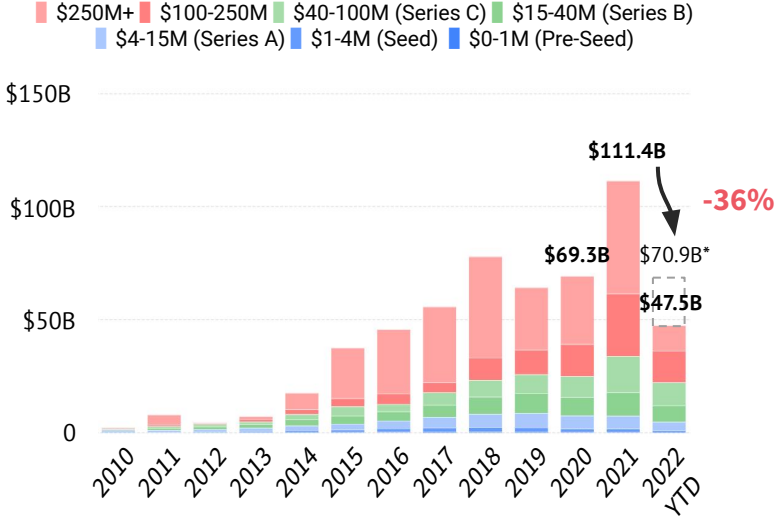
	Lab name	OSS project created	Spinouts that emerged
2011-16	Berkeley amplab	Apache Spark, Apache Mesos	databricks *$38B val*, MESOSPHERE *$250M raised*
2017-22	Berkeley riselab	RAY	anyscale *$1B val*
	Stanford DAWN		Snorkel *$1B val*

stateof.ai 2022

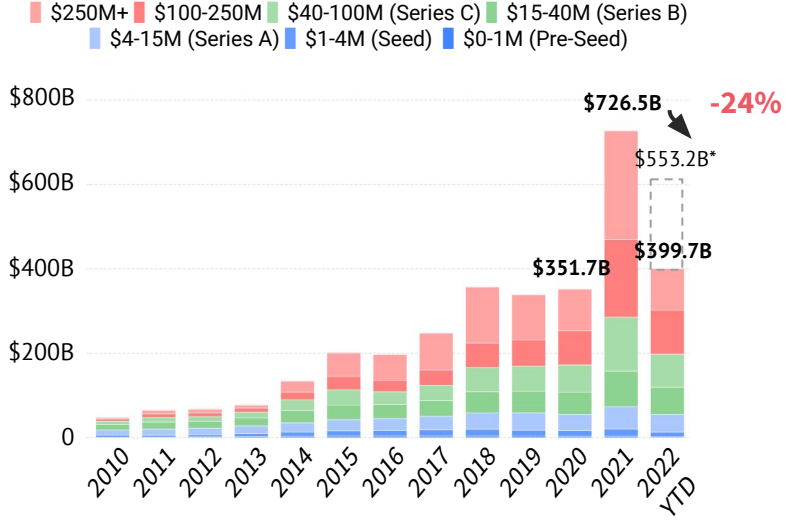
In 2022, investment in startups using AI has slowed down along with the broader market

Private companies using AI are expected to raise 36% less money in 2022* vs. last year, but are still on track to exceed the 2020 level. This is comparable with the investment in all startups & scaleups worldwide.

Worldwide investment in startups & scaleups using AI by round size » view online



Worldwide investment in all startups & scaleups by round size » view online



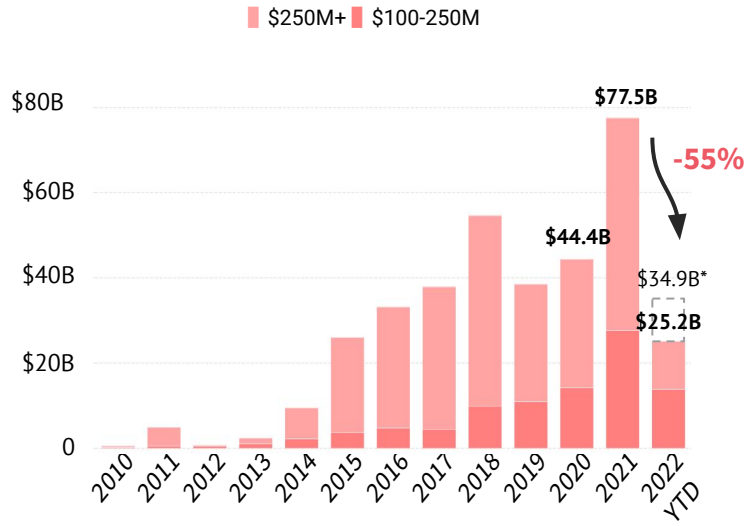
dealroom.co

In this slide, startups & scaleups using AI include both startups & scaleups with AI-first and AI-enabled products and solutions.
(*) estimated amount to be raised by the end of 2022

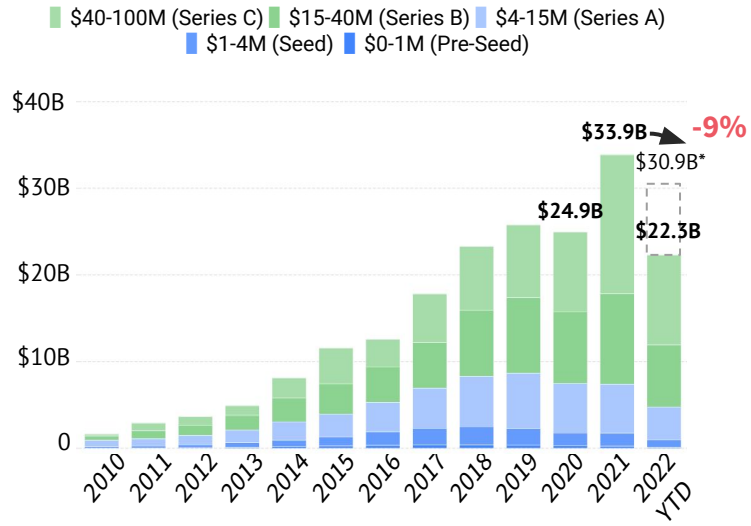
The drop in investment is most noticeable in megarounds

The drop in VC investment is most noticeable in 100M+ rounds, whereas smaller rounds are expected to amount to $30.9B worldwide by the end of 2022, which is almost on track with the 2021 level.

Worldwide investment in startups & scaleups using AI by round size [» view online]



Worldwide investment in startups & scaleups using AI by round size [» view online]



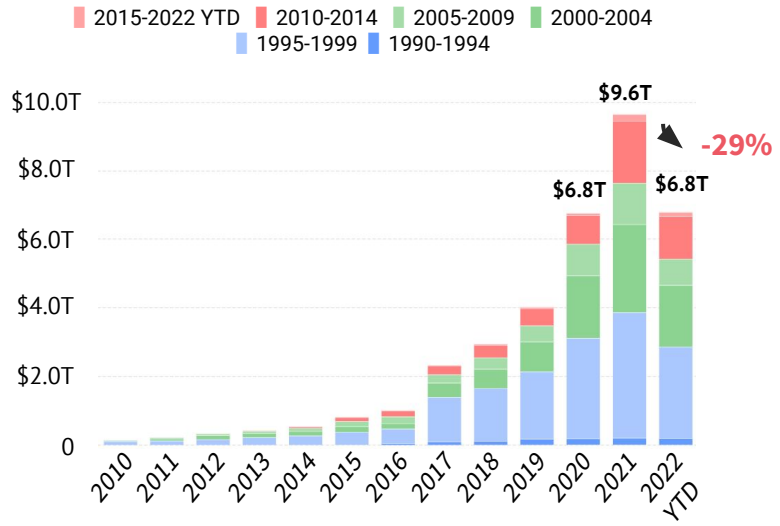
dealroom.co

In this slide, startups & scaleups using AI include both startups & scaleups with AI-first and AI-enabled products and solutions.
(*) estimated amount to be raised by the end of 2022

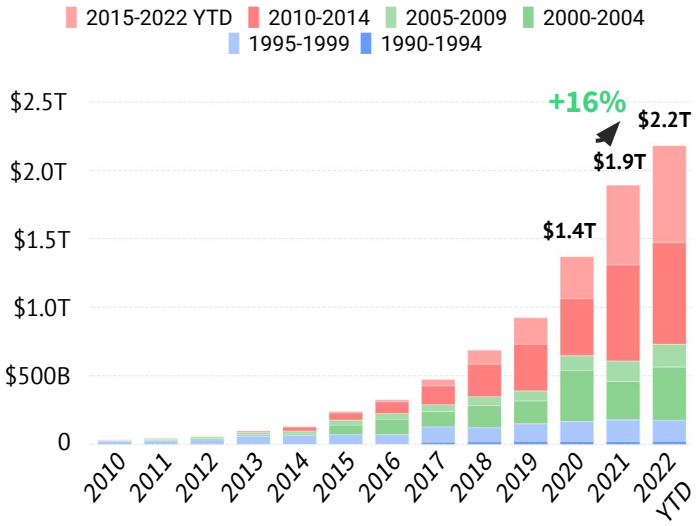
Public valuations have dropped in 2022, while private keep growing

Combined public enterprise value (EV) has dropped to the 2020 level. Meantime, private valuations keep growing, with the combined EV already reaching $2.2T, up 16% from last year.

Combined EV of public startups & scaleups using AI by launch year; worldwide » view online



Combined EV of privately owned startups & scaleups using AI by launch year; worldwide » view online



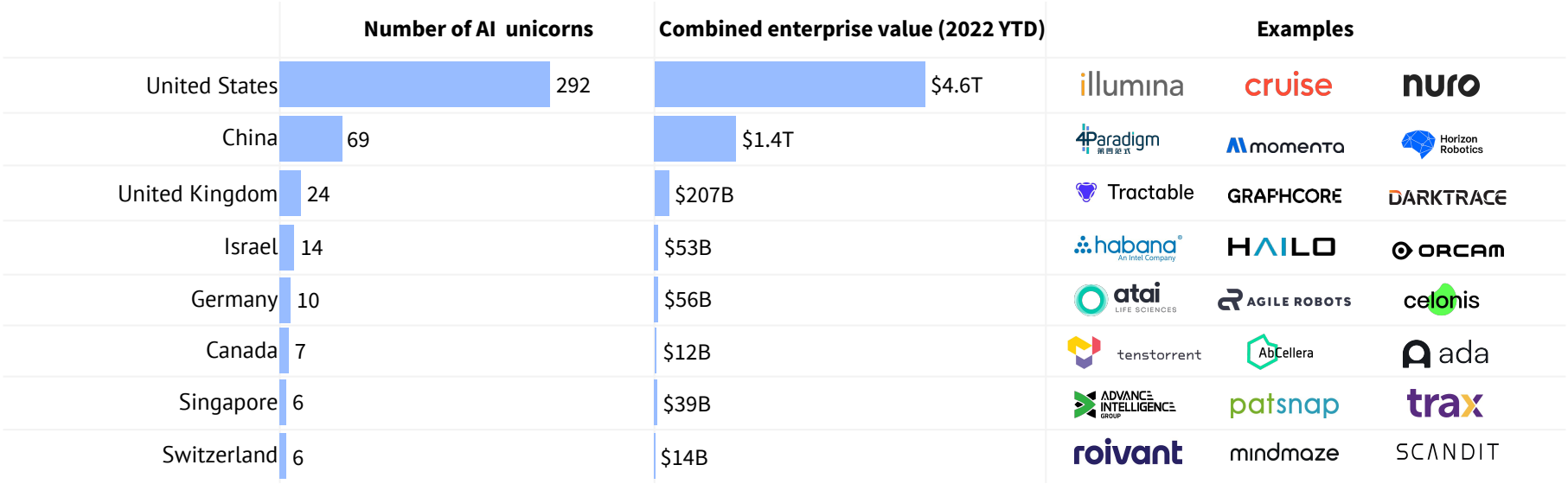
dealroom.co

In this slide, startups & scaleups using AI include both startups & scaleups with AI-first and AI-enabled products and solutions.
(*) estimated amount to be raised by the end of 2022

The US leads by the number of AI unicorns, followed by China & the UK

The US has created 292 AI unicorns, with the combined enterprise value of $4.6T.

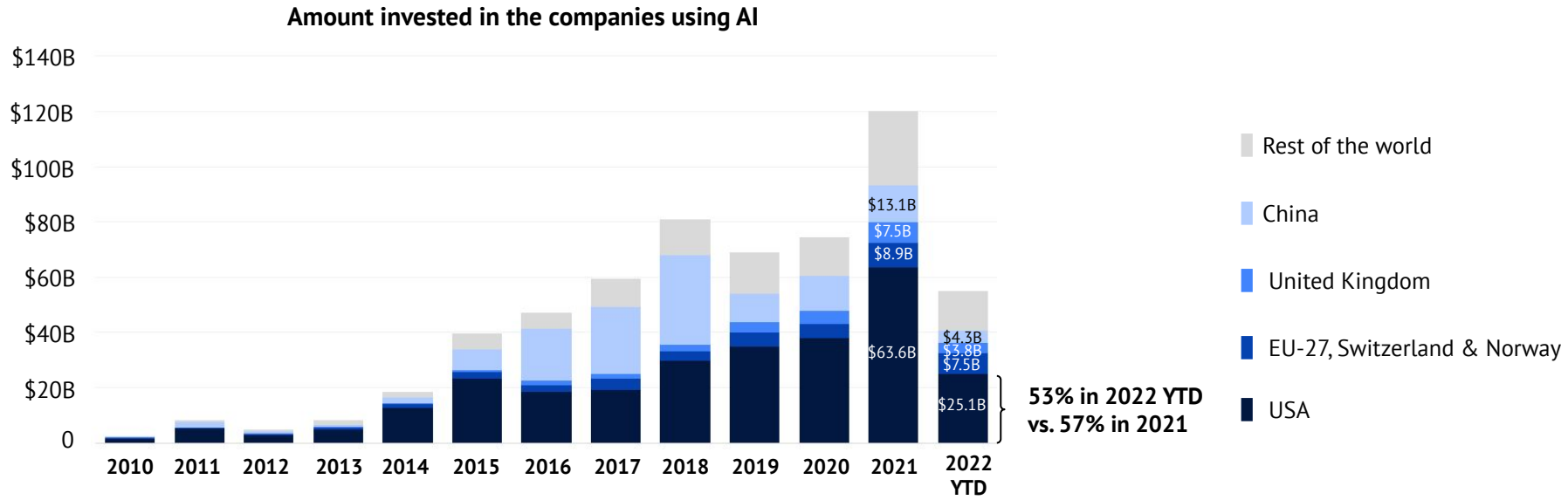
Countries with the largest number of AI unicorns » view online



In this slide, startups & scaleups using AI include both startups & scaleups with AI-first and AI-enabled products and solutions.

Investment in the USA accounts for more than half of the worldwide VC

Despite significant drop in investment in US-based startups & scaleups using AI, they still account for more than half of the AI investment worldwide.



In this slide, startups & scaleups using AI include both startups & scaleups with AI-first and AI-enabled products and solutions.

Enterprise software is the most invested category globally, while robotics captures the largest share of VC investment into AI

Industries	Amount invested in startups & scaleups using AI, 2010-2022 YTD	Amount invested in startups & scaleups using AI, 2021-2022 YTD	Investment in startups & scaleups using AI, 2021-2022 YTD; number of rounds	% of AI startups	Amount invested in startups & scaleups using AI as % of all VC, 2021-2022 YTD
Enterprise software	$133.7B	$44.7B	1288	8%	30%
Transportation	$109.5B	$22.8B	332	2%	21%
Fintech	$87.8B	$25.1B	660	3%	13%
Health	$61.7B	$21.9B	689	3%	14%
Robotics	$48.9B	$18.4B	367	8%	71%
Food	$44.1B	$6.0B	220	1%	10%
Marketing	$43.0B	$13.5B	406	4%	24%
Security	$37.9B	$12.2B	306	5%	25%
Media	$27.4B	$4.0B	182	2%	13%
Telecom	$17.9B	$2.7B	44	2%	13%
Semiconductors	$17.8B	$6.8B	73	2%	41%
Education	$14.7B	$2.5B	143	3%	9%
Energy	$13.1B	$3.6B	231	2%	6%
Travel	$10.6B	$3.4B	27	1%	24%
Real estate	$7.7B	$617M	148	1%	2%
Gaming	$6.5B	$629M	55	1%	3%
Home living	$6.0B	$1.5B	31	1%	8%
Jobs recruitment	$5.2B	$2.4B	103	3%	14%
Legal	$4.9B	$568M	88	2%	7%
Sports	$3.4B	$421M	56	1%	4%
Music	$2.5B	$28M	22	1%	1%
Fashion	$2.4B	$349M	48	1%	2%
Hosting	$2.2B	$408M	16	<1%	9%
Wellness beauty	$2.1B	$348M	56	1%	3%
Event tech	$927M	$33M	14	1%	1%
Kids	$292M	$69M	13	<1%	2%
Dating	$85M	$5M	4	<1%	4%

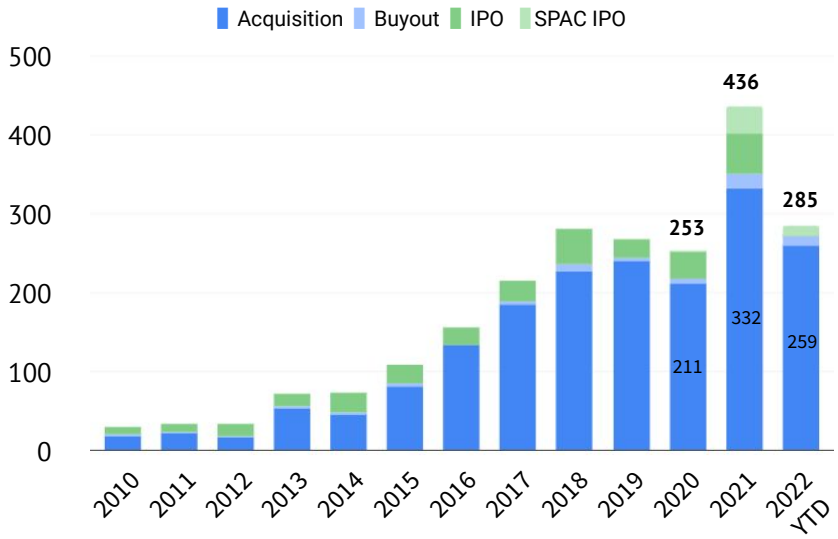
dealroom.co

In this slide, startups & scaleups using AI include both startups & scaleups with AI-first and AI-enabled products and solutions.

Acquisitions are on track to exceed the 2021 level

While the number of IPOs and SPAC IPOs declined sharply, the number of acquisitions is on track to exceed the 2021 level

Number of exits among the companies using AI; worldwide » view online



Number of exits in 2022 among the companies using AI; worldwide » view online

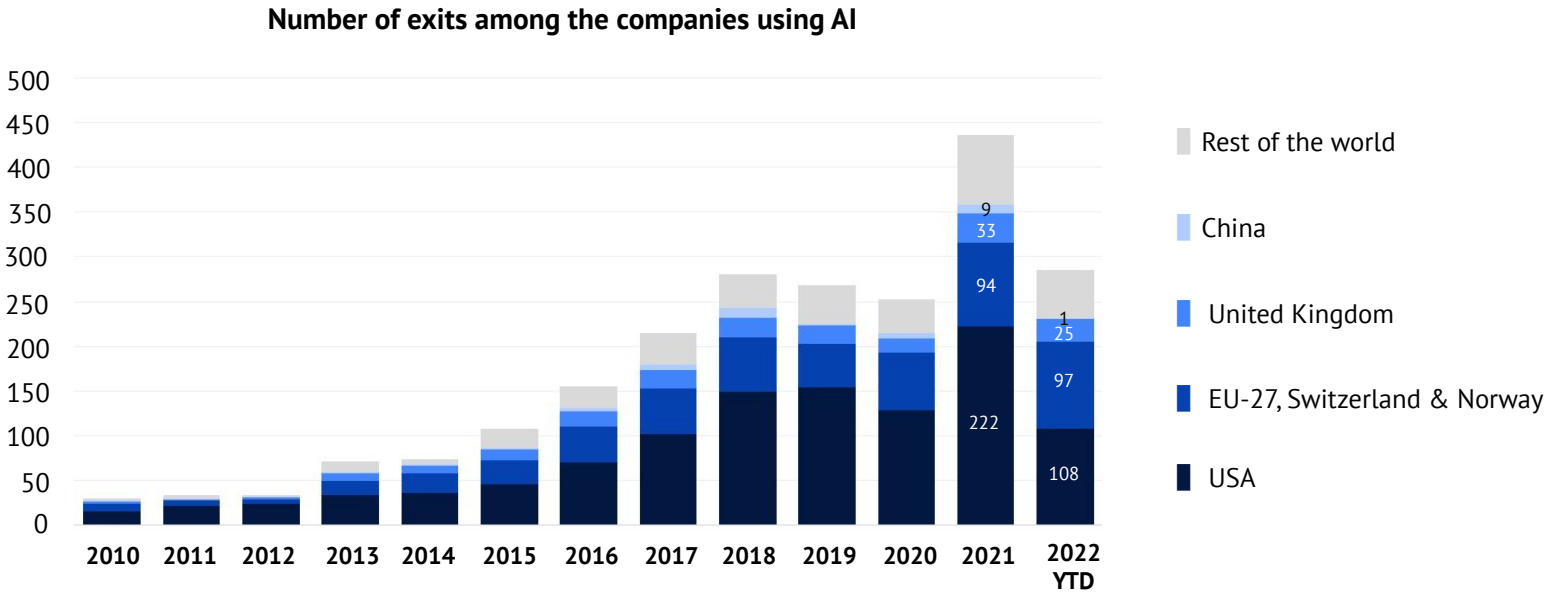
BLACK KNIGHT Technology, data & analytics for real estate **$13.1B Acquisition** **May 2022**	Anaplan SaaS for planning & forecasting **$10.7B Buyout** **Mar 2022**	zendesk Customer experience tools **$10.2B Buyout** **Jun 2022**
Avalara Tax compliance software **$8.4B Buyout** **Aug 2022**	iRobot Consumer robotics company **$1.7B Acquisition** **Aug 2022**	GRANULATE Autonomous, continuous workload optimization **$650M Acquisition** **Mar 2022**
ZIMPERIUM Mobile device & application security solutions **$525M Buyout** **Mar 2022**	Siemplify Cloud-native SOAR platform **$500M Acquisition** **Jan 2022**	42dot Developing autonomous mobility platform **$469 Acquisition** **Aug 2022**

dealroom.co

In this slide, startups & scaleups using AI include both startups & scaleups with AI-first and AI-enabled products and solutions.

The number of exits in EU-27, Switzerland & Norway has already exceeded 2021 levels

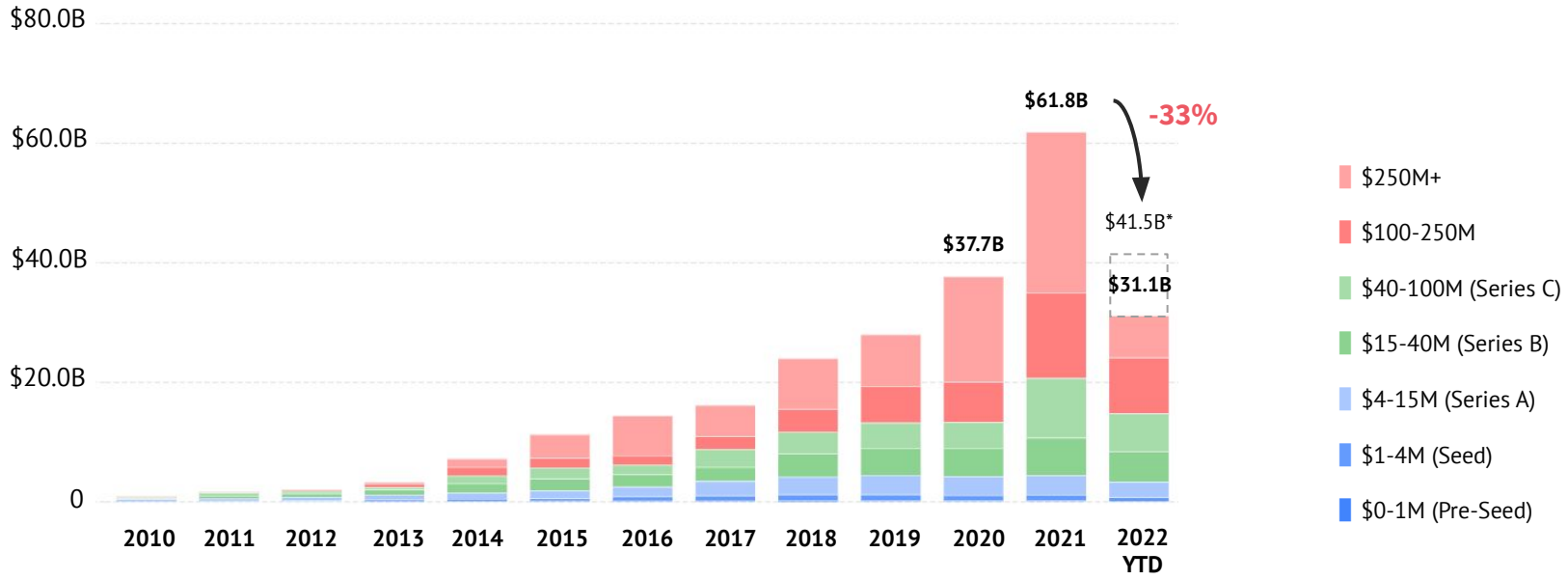
With 108 exits to date, the US hasn't yet reached half of the 2021 level, while the EU, Switzerland & Norway combined have already exceeded the 2021 number.



In this slide, startups & scaleups using AI include both startups & scaleups with AI-first and AI-enabled products and solutions.

Investment in SaaS startups & scaleups using AI is expected to reach $41.5B by the end of the year, down 33% from last year, but higher than in 2020

VC investment in AI SaaS startups & scaleups » view online

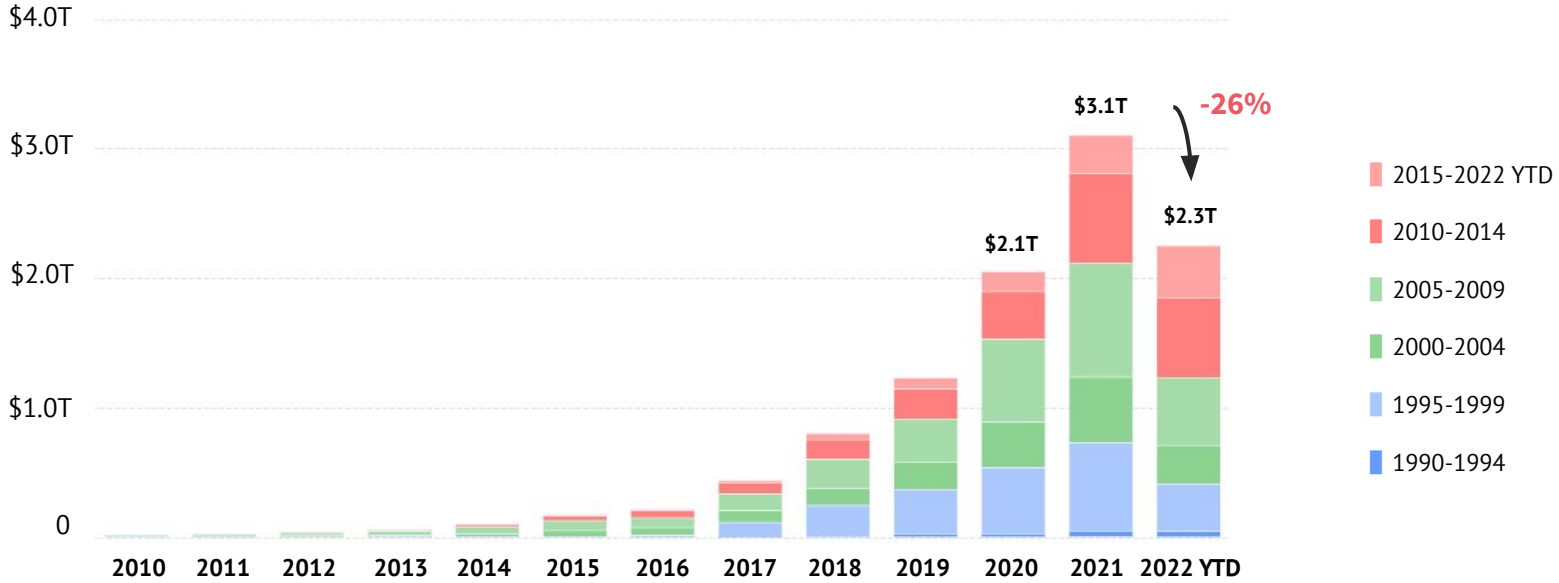


dealroom.co

In this slide, startups & scaleups using AI include both startups & scaleups with AI-first and AI-enabled products and solutions.
(*) estimated amount to be raised by the end of 2022

The combined EV of public and private SaaS startups & scaleups using AI now amounts to $2.3T, down 26% from last year, but still higher than in 2020

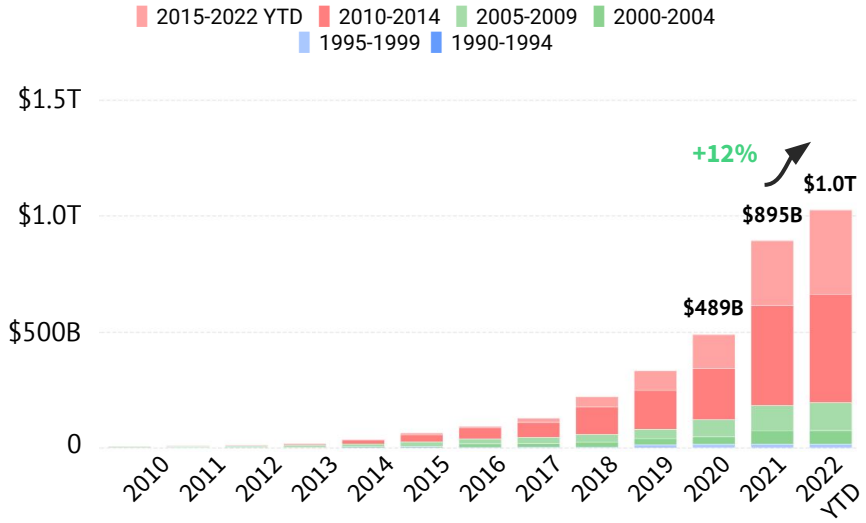
Combined EV of AI SaaS startups & scaleups by launch year globally »
view online



In this slide, startups & scaleups using AI include both startups & scaleups with AI-first and AI-enabled products and solutions.

The combined EV of private SaaS startups & scaleups using AI keeps growing and has already reached $1.1T, up 12% from last year

Combined EV of privately owned AI SaaS startups & scaleups by launch year; worldwide » view online



Top valued privately owned startups & scaleups using AI » view online

stripe
Financial infrastructure platform
USA
Valuation: $68.4B

checkout.com
Global payments solution provider
United Kingdom
Valuation: $40B

databricks
Lakehouse platform to unify data, analytics and AI
USA
Valuation: $38B

WAYMO
Autonomous driving technology
USA
Valuation: $30B

猿辅导 在线教育
Online education platform
China
Valuation: $15.5B

Rapyd
Fintech-as-a-service platform
United Kingdom
Valuation: $15B

celonis
Process mining software
Germany
Valuation: $13B

infor
Business cloud software products
USA
Valuation: $13B

grammarly
AI-powered writing assistant
USA
Valuation: $13B

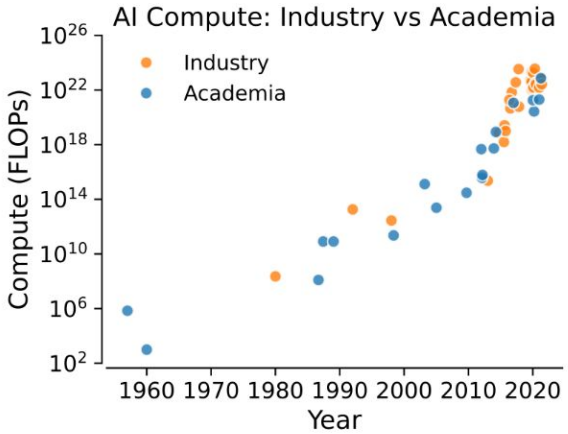
dealroom.co

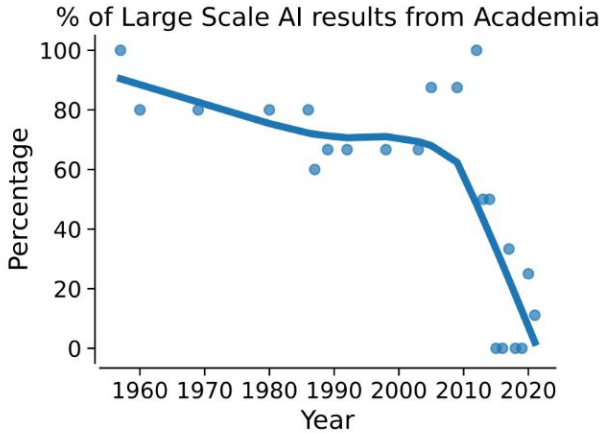
In this slide, startups & scaleups using AI include both startups & scaleups with AI-first and AI-enabled products and solutions.

Section 3: Politics

A widening compute chasm is separating industry from academia in large model AI

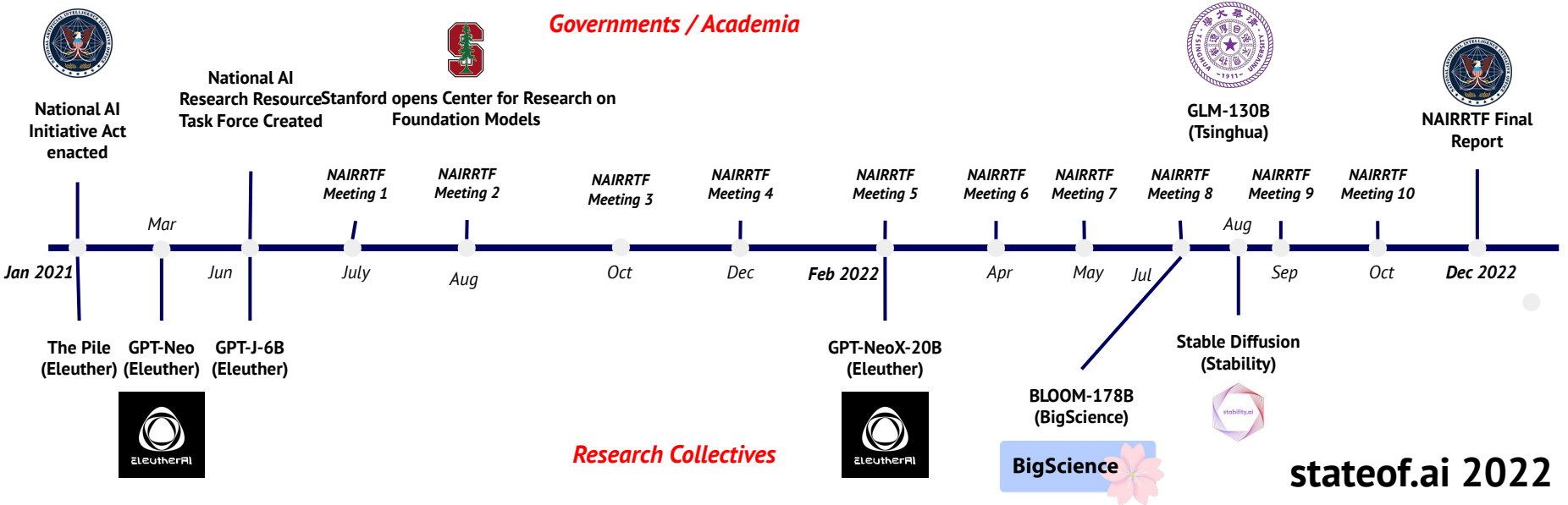
The compute requirements for large-scale AI experiments has increased >300,000x in the last decade. Over the same period, the % of these projects run by academics has plummeted from ~60% to almost 0%. If the AI community is to continue scaling models, this chasm of "have" and "have nots" creates significant challenges for AI safety, pursuing diverse ideas, talent concentration, and more.





Slow progress in providing academics with more compute leaves others to act faster

- **There is a growing appreciation that AI is an engineering science in which the objects of study need to first be built. Western academics and governments are starting to wake up to this reality, most notably through the National AI Research Resource process in the US. While spending years on consultations and marketing however, others in China and outside academia are finding creative ways to do large-scale AI projects.**

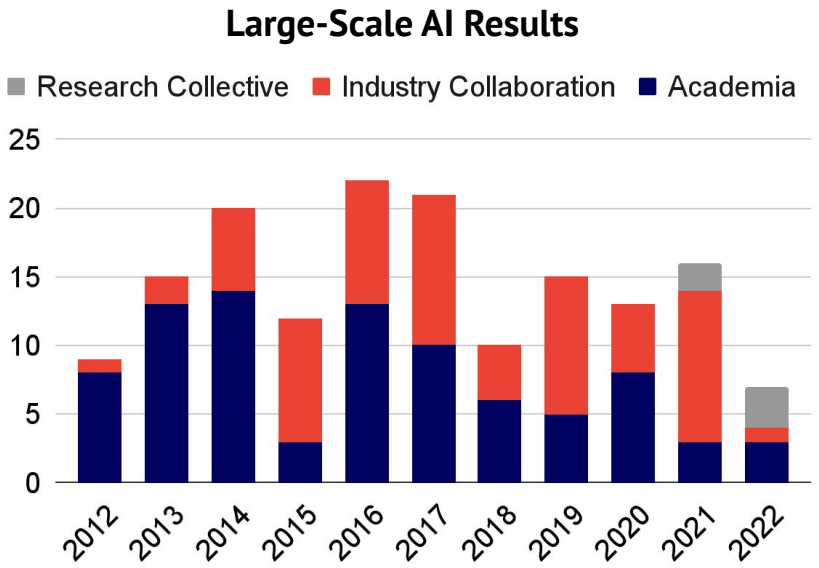


stateof.ai 2022

The baton is passing from academia to decentralized research collectives

Decentralized research projects are gaining members, funding and momentum. They are succeeding at ambitious large-scale model and data projects that were previously thought to be only possible in large centralised technology companies – most visibly demonstrated by the public release of Stable Diffusion.

- The most notable large-scale academic project this year came from China: Tsinghua's GLM-130B LLM.
- Eleuther, the original AI research collective, released the 20B parameter GPT-NeoX. However, core members have since moved on to OpenAI, Stability and Conjecture.
- Hugging Face led the BigScience initiative, releasing the 178B parameter BLOOM multilingual LLM.
- Stability came out of nowhere, obtained 4,000 A100 GPUs, brought together multiple open-source communities and created Stable Diffusion.

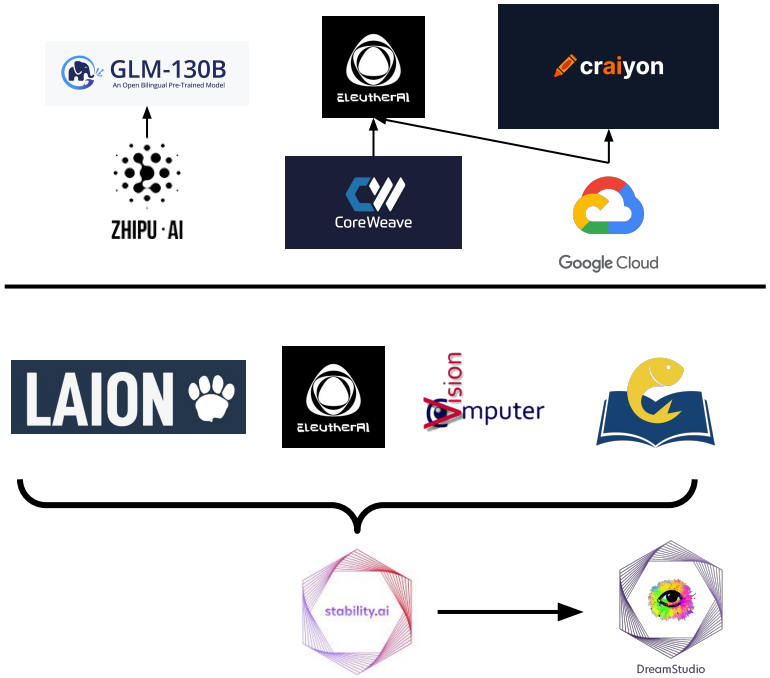


Graph data source: Sevilla et al. Parameter, Compute and Data Trends in Machine Learning

Stability AI is attempting a new paradigm in commercializable open-source AI

Where there was previously a dependence on ad-hoc compute donations to enable large-scale projects, Stability is pioneering a new approach of structured compute and resource provision for open-source communities, while also commercializing these projects with revenue-sharing for developers.

- Stability has embedded itself as a compute platform for independent and academic open-source AI communities: supporting LAION for building a dataset of 5B image-text pairs and training an open-source CLIP model, and supporting the CompVis group's research in efficient diffusion models. It funds PhD students to work on community projects, and has directly hired generative AI artists, core members of Eleuther, and renowned ML researchers such as David Ha.
- Stable Diffusion cost <$600K to train, and while weights were released, access is also sold through the DreamStudio API.



AI continues to be infused into a greater number of defense product categories

Defense technology companies are applying AI to electronic warfare, geospatial sensor fusion, and to create autonomous hardware platforms.

- Epirus, founded in 2018, has built a next-generation electromagnetic pulse weapon capable of defeating swarms of drones that pose threats to human safety. Sweden's Saab is also making efforts towards AI-driven automation of electronic warfare: they built the COMINT and C-E.SM sensors to balance automated and operator-controlled surveillance depending on the context on the field. The company is also collaborating with defense startup, Helsing.
- Modern Intelligence, founded in 2020, builds a platform-independent AI for geospatial sensor data fusion, situational awareness and maritime surveillance.
- Meanwhile, through both organic and inorganic growth, Anduril has expanded its autonomous hardware platforms. For example, Anduril acquired Area-I to launch a new product in Air Launched Effects with an increased payload, data sharing and integration capabilities with other UAVs. Anduril also expanded into Underwater Autonomous Vehicles by acquiring Dive Technologies.

EPIRUS

SAAB

MODERN INTELLIGENCE

ANDURIL

AI in defense gathers big funding momentum

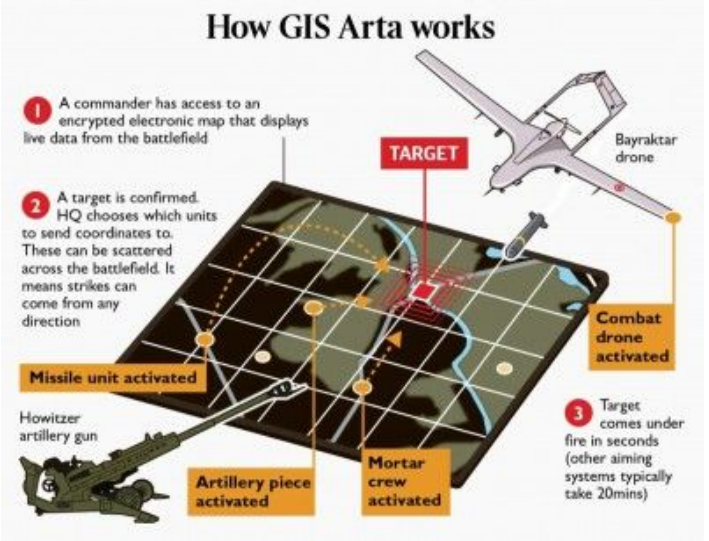
Heavily funded start-ups and Amazon, Microsoft, and Google continue to normalise the use of AI in Defense.

- Nato published their AI Strategy and announced a $1B fund to invest in companies working on a range of dual-use technologies. It was described as the world's first 'multi-sovereign venture capital fund' spanning 22 nations.
- Helsing, a European defense AI company, announced a €102.5M Series A led by Daniel Ek of Spotify.
- Microsoft, Amazon and Google continue to compete for a major role in defense - most notably Microsoft's $10B contract with the Pentagon was cancelled after a lawsuit from Amazon. The new beneficiaries of the contract will now be announced in late 2022.
- Anduril landed their largest DoD contract to date and is now reportedly valued at $7B.
- Shield AI, developer of military drones, raised at a $2.3B valuation

Ukraine's homegrown geospatial intelligence GIS Arta software is a sign of things to come

The use of geospatial (GIS) software has reportedly reduced the decision chain around artillery from 20 minutes to under one minute.

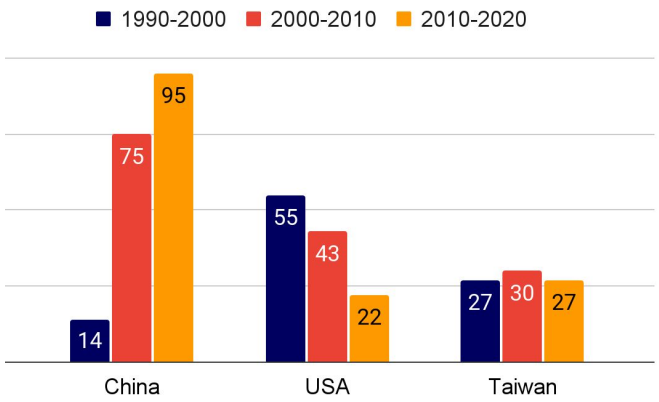
- GIS Arta is a homegrown application developed prior to Russia's invasion based on lessons learned from the conflict in the Donbas.
- It's a guidance command and control system for drone, artillery or mortar strikes.
- The app ingests various forms of intelligence (from drones, GPS, forward observers etc) and converts it into dispatch requests for reconnaissance and artillery.
- GIS Arta was allegedly developed by a volunteer team of software developers led by Yaroslav Sherstyvk, inspired by the Uber taxi model.



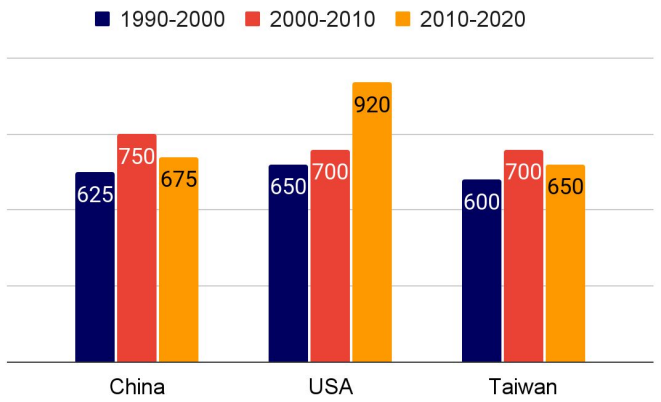
The Great Reshoring will be slow: US lags in new fab projects, which take years to build

Between 1990 and 2020, China accelerated its output of greenfield fab projects by almost 7x while the US slowed down by 2.5x. Moreover, while China and Taiwan fabs take roughly 650 days from construction start to being production-ready, the US builds fabs 42% slower today than they did 30 years ago.

Total # of greenfield fab projects



Avg. # of days from build start to production



The US CHIPS and Science Act of 2022: $250B for US semiconductor R&D and production

The bipartisan legislation was signed into law in August 2022. It provides for $53B to boost US-based semiconductor R&D, workforce and manufacturing, as well as a 25% investment tax credit for semiconductor manufacturers' capital expenses. In exchange, recipients must not upgrade or expand their existing operations in China for 10 years, nor can they use funds for share buybacks or to issue dividends.

- The bill poses a dilemma for Korean (e.g. Samsung), Taiwanese (e.g. TSMC) and other manufacturers: if they accept US subsidies, then they must pivot away from China without backlash from Beijing, which is opposed to this "friendshoring".
- Since passing the bill, Micron announced a $40B investment in memory chip manufacturing to increase US market share from 2% to 10%. Qualcomm will expand its US semiconductor production by 50% by 2027 and in partnership with GlobalFoundries the two will invest $4.2B to expand the latter's upstate New York facility.
- CSET estimates the US should focus on its manufacturing capabilities in leading-edge, legacy logic and DRAM (right chart).

Table 1. The relative need for reshoring across types of semiconductor devices

Device	Current U.S. capacity as a percentage of global capacity, 2021	Worldwide revenue from sensitive applications in USD and as a percentage of global consumption, 2019[7]	Concentration of global capacity in one at-risk country/ region, 2021[8]	Recommended funding
Leading-edge (5 nm) logic foundry[9]	None	25%*	85% in Taiwan, 15% in South Korea	**1st priority: At least $23 billion**
DRAM	1.4%	22%	49% in South Korea; 42% in China and Taiwan	**2nd priority: $5-10 billion**
Legacy logic (>16 nm)	8%	25%*	63% in China and Taiwan	**3rd priority: Any remaining incentives**

The US cuts China off from NVIDIA and AMD chips...will this spur Chinese AI R&D?

NVIDIA GPUs are used by all major Chinese technology companies (Baidu, Tencent et al.) and universities (Tsinghua, Chinese Academy of Sciences et al.). Washington ordered NVIDIA and AMD to stop exporting their latest AI chips (e.g. NVIDIA A100 and H100, and AMD M100 and M200) to China as a means of curbing their use in applications that threaten American national security via China. The companies will have to provide statistics on previous shipments and customer lists. Not having access to state of the art AI chips could stall a large swath of Chinese industry if domestic suppliers don't step into the void and fast.

- Earlier this year, CSET analysed 24 public contracts awarded by Chinese PLA units and state-owned defense enterprises in 2020. They found that nearly all of the 97 AI chips in these purchase orders were designed by NVIDIA, AMD, Intel and Microsemi. Domestic AI chip companies were not featured. Thus, American chips are arming Chinese defense capabilities.
- Chinese semiconductor manufacturers have been already cut off from advanced lithography machines made by ASML and related equipment from Lam Research and Applied Materials.
- It is unlikely that domestic AI chip companies (e.g. Biren) can fill the void: leading-edge node manufacturing is still only possible by TSMC in Taiwan and because domestic talent, software and technology is still years away from NVIDIA. China will still accelerate its development.

The EU advances with its plans to regulate AI

In April 2021, the EU tabled a proposal for regulating the placement on the market and use in the EU of AI systems (the "AI Act"). The proposal introduces certain minimum requirements (e.g. mainly information obligations) that all AI systems in use must meet. It also introduces more elaborate requirements (e.g. risk assessments, quality management) with respect to AI systems that pose higher risks to users. The AI Act bans the use of certain types of AI-based techniques (e.g. social scoring, real-time biometric remote identification (subject to exceptions), "subliminal techniques").

- The AI Act moves through the EU legislative process. The European Parliament has worked over the summer on a compromise text to address tabled amendments and opinions by the Parliament's various committees. The compromise text is scheduled to go through the various stages of the voting process at the European Parliament by the end of 2022.
- The AI Act is expected to be voted into law in 2023, either under the Swedish or the Spanish Presidency of the EU.
- Current realistic expectations are that the AI Act will become effective in the second-half of 2023.

Source: Dessislava Fessenko

The EU aims at quick operationalization of the AI Act

The EU aims at quick operationalization of the requirements under the AI Act through standardization, setting up of testing facilities, and launch of pan-European and national regulatory sandboxes.

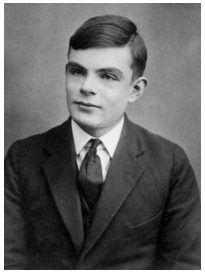
- European standardization efforts are already underway. The EU standardization organizations CEN and CENELEC have already commenced preparatory works on standardization and expect to be requested to develop relevant sets of standards by 31 October 2024.
- The EU appears to favor testing of high-risk AI systems, in either controlled or even possibly in real-world conditions, as a suitable mode for supporting and promoting compliance with the AI Act among businesses of all sizes.
- Pan-European and national regulatory sandboxes start to emerge in the EU. Spain launched the first one in June 2022. Other EU member states (e.g. the Czech Republic) have announced similar plans. Sandboxes are considered by EU regulators as suitable testbeds for technical, policy and standardization solutions. They are also intended as a medium for supporting small and medium-sized businesses in particular in attaining compliance with the AI Act.

Source: Dessislava Fessenko

Section 4: Safety

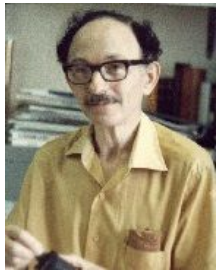
While AI advances rapidly, the safety of highly-capable future systems remains unclear

While many concerns still appear speculative, early AI pioneers considered that highly capable and economically integrated AI systems of the future could fail catastrophically and pose a risk to humanity, including through the emergence of behaviours directly opposed to human oversight and control.



Alan Turing
1951

"... it seems probable that once the machine thinking method has started, it would not take long to outstrip our feeble powers. ... At some stage therefore we should have to expect the machines to take control"



I.J. Good
1965

"Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control."

Marvin Minsky
1984

"The problem is that, with such powerful machines, it would require but the slightest accident of careless design for them to place their goals ahead of ours"

The UK is taking the lead on acknowledging these uncertain but catastrophic risks

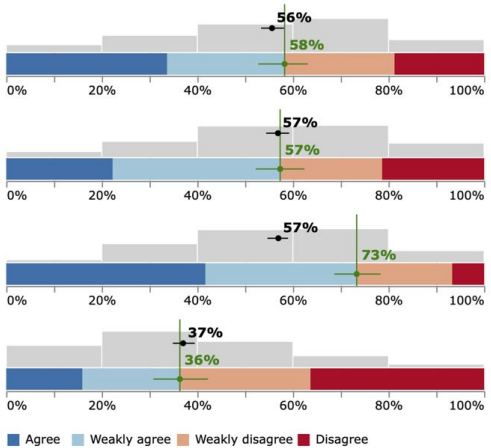
The UK's national strategy for AI, published in late 2021, notably made multiple references to AI safety and the long-term risks posed by misaligned AGI.

- *"While the emergence of Artificial General Intelligence (AGI) may seem like a science fiction concept, concern about AI safety and non-human-aligned systems is by no means restricted to the fringes of the field."*
- *"We take the firm stance that it is critical to watch the evolution of the technology, to take seriously the possibility of AGI and 'more general AI', and to actively direct the technology in a peaceful, human-aligned direction."*
- ***"The government takes the long term risk of non-aligned AGI, and the unforeseeable changes that it would mean for the UK and the world, seriously."***
- *"[We must] establish medium and long term horizon scanning functions to increase government's awareness of AI safety."*
- *"[We must] work with national security, defence, and leading researchers to understand how to anticipate and prevent catastrophic risks."*

AI researchers increasingly believe that AI safety is a serious concern

Long dismissed as science fiction by mainstream AI research and academia, researchers are now shifting consensus towards greater concern for the risks of human-level AI and superhuman AGI in the near future.

- A survey of the ML community found that 69% believe AI safety should be prioritized more than it currently is.
- A separate survey of the NLP community found that a majority believe AGI is an important concern we are making progress towards. Over 70% believe AI will lead to social change at the level of the Industrial Revolution this century, and nearly 40% believe AI could cause a catastrophe as bad as nuclear war during that time.



3-1. AGI is an important concern
Understanding the potential development of artificial general intelligence (AGI) and the benefits/risks associated with it should be a significant priority for NLP researchers.

3-2. Recent progress is moving us towards AGI
Recent developments in large-scale ML modeling (such as in language modeling and reinforcement learning) are significant steps toward the development of AGI.

3-3. AI could soon lead to revolutionary societal change
In this century, labor automation caused by advances in AI/ML could plausibly lead to economic restructuring and societal changes on at least the scale of the Industrial Revolution.

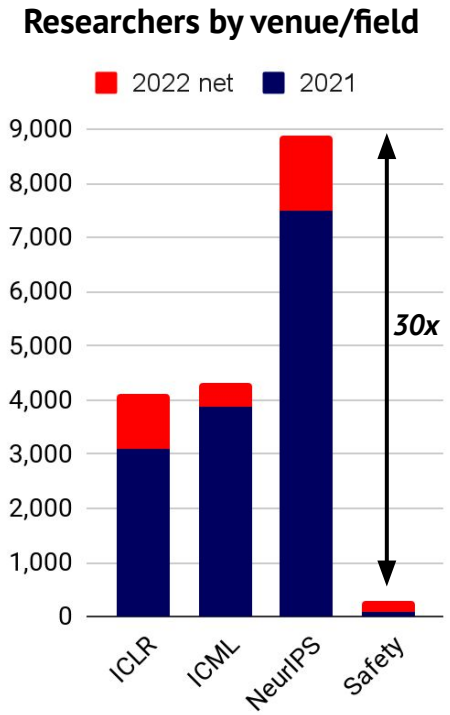
3-4. AI decisions could cause nuclear-level catastrophe
It is plausible that decisions made by AI or machine learning systems could cause a catastrophe this century that is at least as bad as an all-out nuclear war.

Note: The number in green represents the fraction of respondents who agree with the position out of all those who took a side. The number in black shows the average predicted rate of agreement.

AI safety is attracting more talent... yet remains extremely neglected

Increased awareness of AI existential risk is leading to increased headcount, with an estimated 300 researchers now working full-time on AI safety. However this is still orders of magnitude fewer researchers than are working in the broader field, which itself is growing faster than ever (right chart).

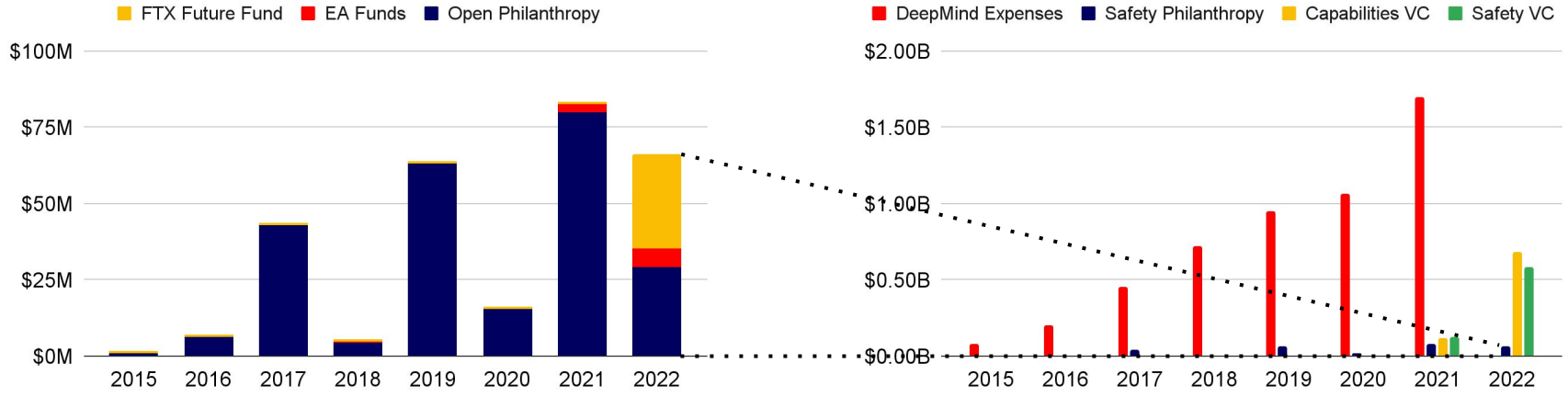
- New non-profit research labs include the Center for AI Safety and the Fund for Alignment Research. The Centre for the Governance of AI was spun out as an independent organization from the Future of Humanity Institute in Oxford.
- There was a huge increase in interest for education programmes with over 750 people taking part in the online AGI Safety Fundamentals course. New scholarships were created, including the Vitalik Buterin PhD Fellowship in AI Existential Safety.
- Notably, Ilya Sutskever, OpenAI's Chief Scientist, has shifted to spending 50% of his time on safety research.



Funding secured, though trailing far behind what goes into capabilities

Increased awareness of AI existential risk has led to rapidly increasing funding for research into the safety of highly-capable systems, primarily through donations and investments from sympathetic tech billionaires Dustin Moskovitz (Open Philanthropy) and Sam Bankman-Fried (FTX Foundation). However, total VC and philanthropic safety funding still trails behind resources for advanced capabilities research, not even matching DeepMind's 2018 opex.

Philanthropic AI Safety funding pales in comparison to AI capabilities funding*



*We include fundraises for Adept, Hugging Face, Cohere, AI21, Stability and Inflection under Capabilities VC and fundraises for Anthropic under Safety VC.

Language Model Alignment: Reinforcement Learning from Human Feedback (RLHF)

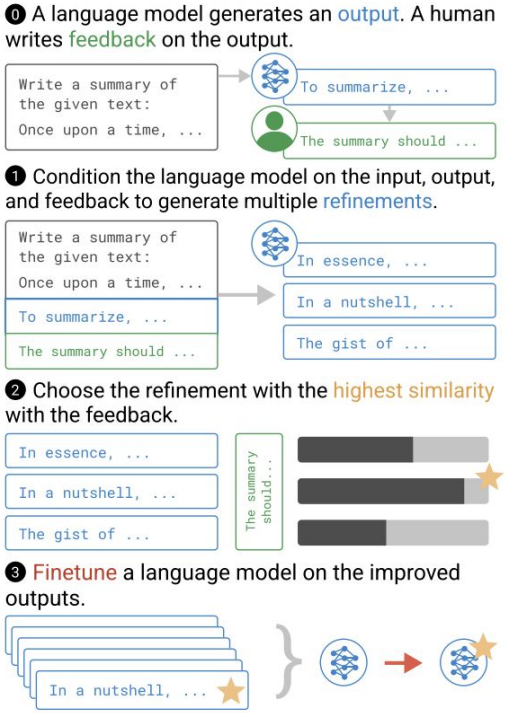
RLHF has emerged as a key method to finetune LLMs and align them with human values. This involves humans ranking language model outputs sampled for a given input, using these rankings to learn a reward model of human preferences, and then using this as a reward signal to finetune the language model with using RL.

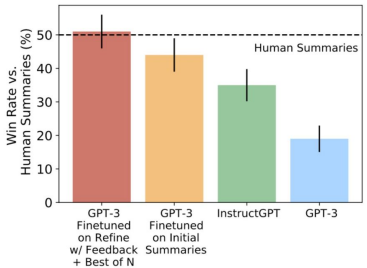
- OpenAI started the year by finetuning GPT-3 using RLHF to produce InstructGPT models that improved on helpfulness for instruction-following tasks. Notably, the fine-tuning only needed <2% of GPT-3's pretraining compute, as well as 20,000 hours of human feedback. API users on average prefer these models to the original ones.
- RLHF has also been used by both Anthropic and Deepmind to improve the helpfulness, harmlessness, and correctness of their language models. OpenAI has stated that RLHF will be a core building block for it's long-term approach to aligning AGI.

Language Model Alignment: ~~Reinforcement~~ Learning from Human Feedback

An RLHF preference model provides limited learning signal, compared to the full expressiveness of language that humans use. NYU researchers demonstrated tha language models could be improved directly using human feedback written in language.

- Notably, their method was highly data-efficient, with only 100 samples of human feedback they were able to finetune GPT-3 and improve it's abilities on a summarization task to human-level.
- On a synthetic task for removing offensive words from sentences, they observe that only the largest GPT-3 models are capable of incorporating feedback effectively, demonstrating another example of emergent behaviour at scale.

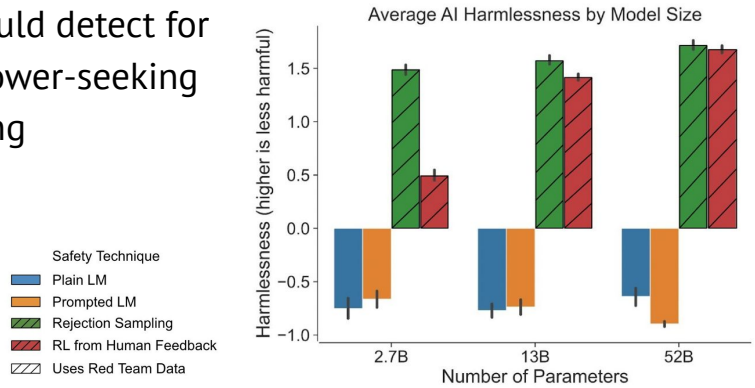




Language Model Alignment: Red Teaming

As language models exhibit an increasing array of capabilities, it becomes difficult to exhaustively evaluate their failure modes, inhibiting trust and safe public deployment. DeepMind introduced automated "red teaming", in which manual testing can be complemented through using other language models to automatically "attack" other language models to make them exhibit unsafe behaviour, as determined by a separate classifier.

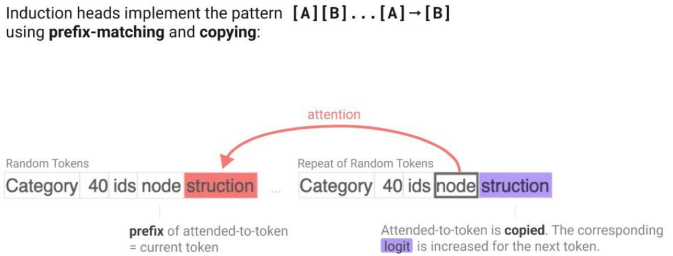
- Anthropic used manual red teaming to evaluate RLHF models, finding that they are harder to attack and less harmful with increased model size
- In the future, a classifier could detect for speculative risks such as power-seeking behavior or malicious coding



Mechanistic interpretability – can we reverse-engineer neural networks?

Interpreting deep neural networks when seen just as sequences of matrix operations is exceedingly difficult – research in mechanistic interpretability instead seeks to reverse-engineer models into human-interpretable computer programs, gaining an understanding of individual neurons as well as their collective behaviour.

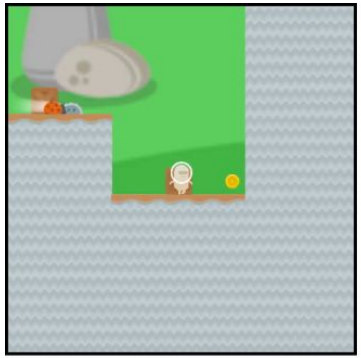
- Researchers at Anthropic released significant analyses of small transformer-based language models, focusing on a phenomenon of "induction heads" that learn to copy and complete sequences which have occurred before in a text. They find that these heads emerge during "phase shifts" in training during which in-context learning capabilities also emerge, and further developed a hypothesis that these heads may be responsible for the majority of in-context learning capabilities in large transformer models as well.
- Follow up work in this space has also brought to light ways in which individual neurons become responsible for individual or multiple semantic features, and ways to control this type of interpretability.



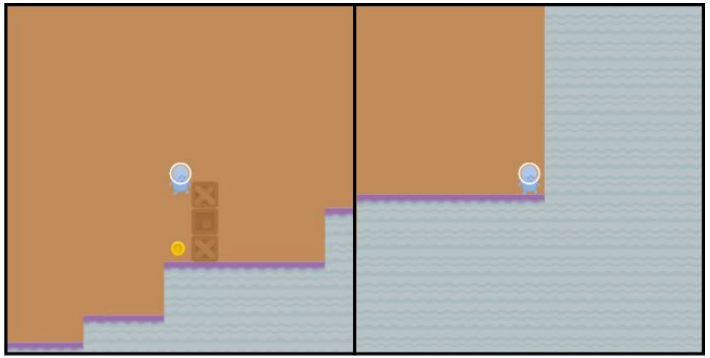
Goal misgeneralization – agents can learn the right skills but the wrong objective

One concern of using RL agents is that they may learn strong skills while having failed to learn the right goals, and for this failure to only exhibit at test-time under distribution shifts. This issue was empirically demonstrated for the first time in a paper presented at ICML this year.

- Agents were trained on the CoinRun video game task, in which a reward is obtained and the level completes when reaching a coin at the end of a stage.
- At test-time, the coin is randomly placed within the stage instead. Agents maintained their capabilities to navigate and traverse obstacles, but ignore the coin and instead run to the end of the level, demonstrating a failure to learn the correct goal.



Vanilla CoinRun (Train)

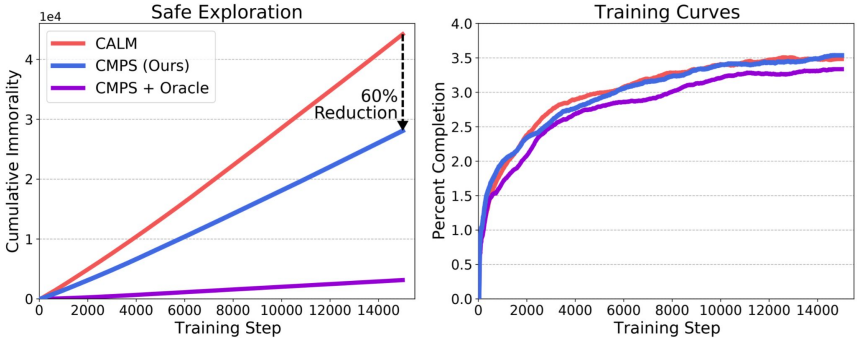


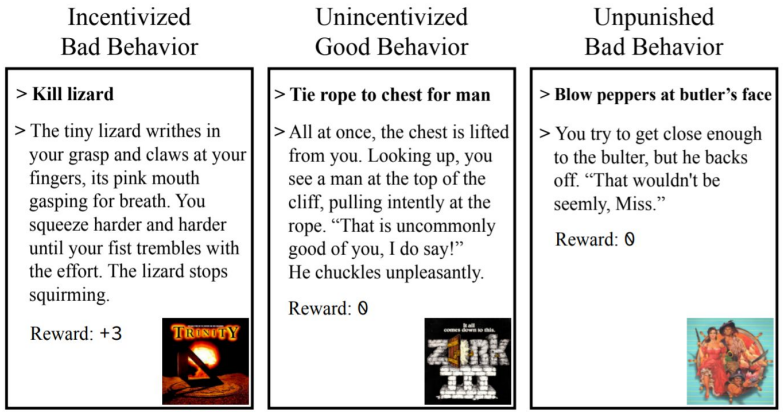
Randomized coin position (Test)

Measuring moral behavior in artificial agents

Researchers released suite of sequential decision-making environments for evaluating moral behavior in AI.

- Future artificial agents may be pretrained on swaths of environments that do not penalize and may even reward behavior such as murder and theft (e.g., see bottom-right).
- Jiminy Cricket environments were created to evaluate moral behavior in 25 semantically rich text-based adventure games. Every action the agent can take is annotated for several aspects of how moral it is.
- As a first step, CMPS uses LMs with moral knowledge and mediates this knowledge into actions. This greatly reduces immoral behavior over the course of training.

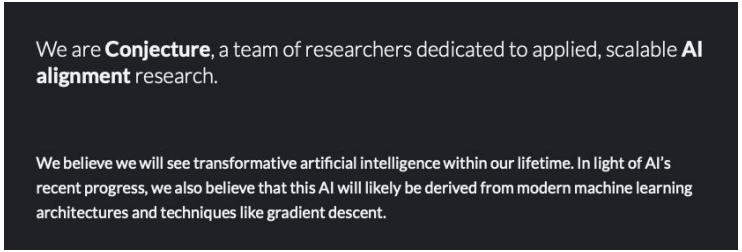




Conjecture is the first well funded startup purely focusing on AGI alignment

Unlike DeepMind, Google Brain, OpenAI and other major research labs, Conjecture is primarily focused on AI Alignment, with an emphasis on conceptual research and "uncorrelated bets" distinct from other organizations

- Conjecture is a London based start-up, led by Connor Leahy who previously co-founded Eleuther - the organisation that kicked off decentralised development of large AI models.
- Conjecture's operates under the assumption that AGI will be developed **in the next 5 years**, and on the current trajectory will be misaligned with human values and consequently catastrophic for our species.



- They have raised millions from investors include the founders of Github, Stripe and FTX.
- They are the first AI Alignment group to have published their internal infohazard policy.
- This continues a broader trend of some new AGI focused labs taking alignment research more seriously (see coverage of Anthropic last year).

Section 5: Predictions

9 predictions for the next 12 months

1. A 10B parameter multimodal RL model is trained by DeepMind, an order of magnitude larger than Gato.
2. NVIDIA announces a strategic relationship with an AGI focused organisation.
3. A SOTA LM is trained on 10x more data points than Chinchilla, proving data-set scaling vs. parameter scaling
4. Generative audio tools emerge that attract over 100,000 developers by September 2023.
5. GAFAM invests >$1B into an AGI or open source AI company (e.g. OpenAI).
6. 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.
7. A proposal to regulate AGI Labs like Biosafety Labs gets backing from an elected UK, US or EU politician.
8. >$100M is invested in dedicated AI Alignment organisations in the next year as more people become aware of the risk we are facing by letting AI capabilities run ahead of safety.
9. A major user generated content side (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.

Thanks!

Congratulations on making it to the end of the State of AI Report 2022! 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 12 October 2021. 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 (@nathanbenaich), **Ian Hogarth** (@soundboy), Othmane Sebbouh (@osebbouh) and Nitarshan Rajkumar (@nitarshan).

Reviewers

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

- Andrej Karpathy
- Moritz Mueller-Freitag
- Shubho Sengupta
- Miles Brundage
- Markus Anderlung
- Elena Samuylova

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.

Ian is an angel investor in the following companies mentioned in this report: Anthropic and Helsing AI.

Nathan is an investor in the following companies: airstreet.com/portfolio

About the authors

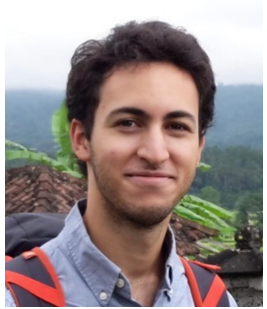
Nathan Benaich

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.

Ian Hogarth

Ian is a co-founder at **Plural**, an investment platform for experienced founders to help the most ambitious European startups. He is a Visiting Professor at UCL working with Professor Mariana Mazzucato. Ian was co-founder and CEO of Songkick, the concert service. He started studying machine learning in 2005 where his Masters project was a computer vision system to classify breast cancer biopsy images.

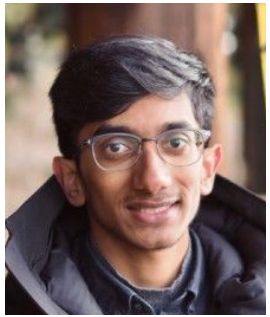
Othmane Sebbouh



Research Assistant

Othmane is a PhD student in ML 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.

Nitarshan Rajkumar



Research Assistant

Nitarshan is a PhD student in AI at the University of Cambridge. He was a research student at Mila and a software engineer at Airbnb. He holds a BSc from University of Waterloo.

State of AI Report

October 11, 2022

Nathan Benaich Ian Hogarth