INSIDE THE AI ARMS RACE: SILICON, POWER, AND THE NEW COLD WAR p. 58

NOT VANITY FAIR

JULY 2026

EXCLUSIVE

THE VISIONARIES WHO BET ON SCALING

A rare look inside OpenAI in 2020

p. 72

THE $4.6 MILLION BUG THAT ALMOST KILLED GPT-3

p. 80

DARIO AMODEI
ILYA SUTSKEVER
ALEC RADFORD
& MORE

The full, uncensored conversations

p. 86

PLUS

THE MORAL TENSIONS OF MAGNITUDE

Safety, alignment, and the road not taken

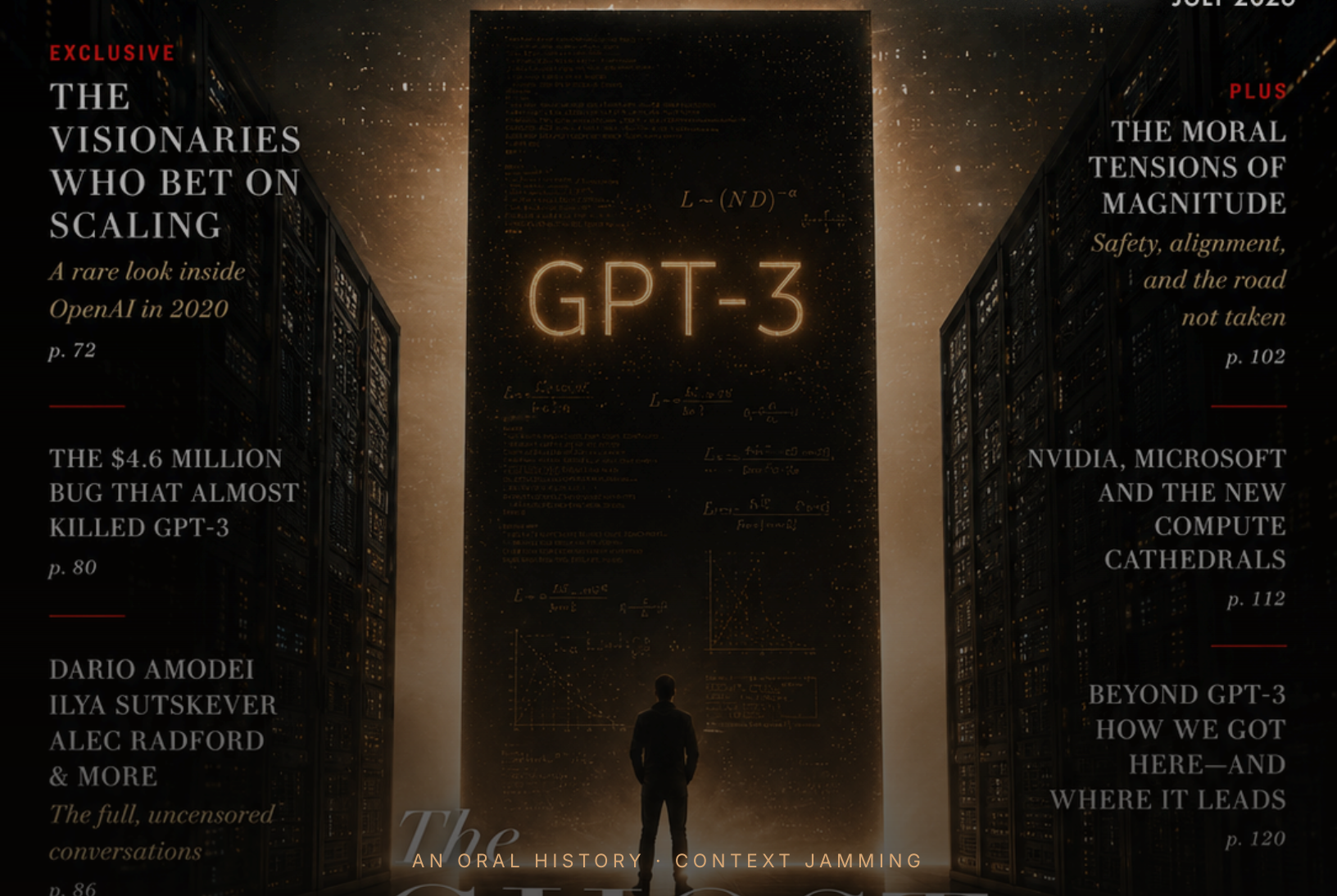
p. 102

NVIDIA, MICROSOFT AND THE NEW COMPUTE CATHEDRALS

p. 112

BEYOND GPT-3
HOW WE GOT HERE—AND WHERE IT LEADS

p. 120



AN ORAL HISTORY · CONTEXT JAMMING

The Ghost in the Machine

An Oral History of GPT-3 and the Dawn of the Scaling Era

BY BRET KERR · PUBLISHED JULY 18, 2026

Underlined text throughout this document is a live hyperlink — to primary sources, founder profiles, and companion explainers. Read online at contextjamming.com/FewShotLearners-Oral-History.

The untold story of the breakthrough that changed everything—and the team that risked it all.

CONTENTS

THE VOICES

Fourteen people from inside OpenAI in 2020 — where they stood then, and where they stand now. Names underlined below link to a full Founder File profile.

Dario Amodei

VP of Research, OpenAI

Now: CEO & Co-founder, Anthropic

Ilya Sutskever

Co-founder & Chief Scientist, OpenAI

Now: Co-founder & Chief Scientist, Safe Superintelligence Inc.

Alec Radford

Research Scientist, OpenAI

Now: Independent researcher · Thinking Machines Lab advisor

Sam McCandlish

Research Scientist, OpenAI

Now: Co-founder & Chief Architect, Anthropic — oversees pre-training and the RL infrastructure behind Constitutional AI.

Jared Kaplan

Research Scientist, OpenAI

Now: Co-founder & Chief Science Officer, Anthropic

Sam Altman

CEO, OpenAI

Now: CEO, OpenAI

Tom B. Brown

Lead Engineer, OpenAI

Now: Co-founder & Chief Compute Officer, Anthropic

Benjamin Mann

Research Scientist, OpenAI

Now: Co-founder, Anthropic

Christopher Berner

Head of Infrastructure, OpenAI

Now: Head of Compute, OpenAI

Amanda Askell

Ethicist, OpenAI

Now: Philosopher & Personality Alignment Lead, Anthropic — shapes Claude's character and constitution. Named to the Time 100 AI list in 2024.

Sandhini Agarwal

Fairness & Representation Researcher, OpenAI

Now: Member of Technical Staff · Trustworthy AI Lead, OpenAI

Jack Clark

Policy/Impact Lead, OpenAI

Now: Co-founder & Head of Policy, Anthropic

Daniela Amodei

VP of Safety and Policy, OpenAI

Now: President & Co-founder, Anthropic

Girish Sastry

Researcher, OpenAI

Now: Head of AI for Civil Society & Philanthropy, The OpenAI Foundation

The Ghost in the Machine

An Oral History of GPT-3 and the Dawn of the Scaling Era

In the late spring of 2020, as the physical world ground to a halt under the shadow of a global pandemic, a different kind of awakening was taking place inside a climate-controlled server farm. In a silicon cathedral built by Microsoft, thousands of NVIDIA V100 GPUs were humming in unison, burning through megawatts of power and millions of dollars to run a single, monolithic mathematical equation.

The project was code-named GPT-3. When the resulting 72-page paper, "Language Models are Few-Shot Learners," dropped on the arXiv preprint server on May 28, 2020, it shattered the prevailing orthodoxies of computer science. At 175 billion parameters, it was ten times larger than any non-sparse language model that had come before it. But its staggering size was only a symptom of its true breakthrough. The paper proved that neural networks, if scaled massively, ceased to be narrow, single-task tools and instead developed emergent, generalized abilities. They could translate languages, write poetry, and perform arithmetic—all without task-specific training.

This is the oral history of how a small, ideologically driven team at OpenAI turned artificial intelligence from an unpredictable art into an industrial science.

The Church of Scaling

Late 2019 – Early 2020

Before 2020, the artificial intelligence industry operated under a paradigm of hyper-specialization. The standard recipe was to pre-train a model on a large corpus of text, then explicitly "fine-tune" its weights on thousands of labeled examples for each specific task. This required a new model and a new bespoke dataset for every single problem.

But deep within OpenAI, a splinter ideology was forming. A faction of former theoretical physicists and rogue engineers believed the field was overthinking the problem. They hypothesized that algorithms did not need to be cleverer — they just needed to be unfathomably larger.

Dario Amodei
VP OF RESEARCH, OPENAI

"I've had the same hypothesis since 2017. I wrote a document back then called The Big Blob of Compute Hypothesis. The premise was simple: all the cleverness, all the bespoke architectural techniques, all the "we need new methods" stuff—it doesn't really matter that much. What really matters are just a few things: raw compute, data quantity, data quality, and training duration. You get the obstacles out of the models' way, and they just want to learn. If you scale up the reagents linearly, the reaction proceeds."

Ilya Sutskever

CO-FOUNDER & CHIEF SCIENTIST, OPENAI

"Up until 2020, from 2012 to 2020, it was the age of research. But then the scaling insight arrived. The big breakthrough of pre-training was the realization that this recipe is good. You say, "Hey, if you mix some compute with some data into a neural net of a certain size, you will get results." You know that you'll be better if you just scale the recipe up. Companies love this because it gives you a very low-risk way of investing your resources."

Alec Radford

RESEARCH SCIENTIST, OPENAI

"In the past, language models were seen as novelty toys that could only generate a sentence that made sense once in a while, and only then if you really squinted. With GPT-1 and GPT-2, we started to see that language modeling implicitly trains a network to perform multiple tasks, because different linguistic structures require different types of reasoning. But the question was: how far could we push it?"

Sam McCandlish

RESEARCH SCIENTIST, OPENAI

"The idea that you could just make a model bigger and it would automatically get smarter was still heavily debated in the broader AI community. Critics argued you couldn't get semantics from syntax, that a model just predicting the next word would eventually hit a wall of diminishing returns. We needed mathematical proof that this wasn't just a heuristic, but a law of nature."

In January 2020, Jared Kaplan, Sam McCandlish, and a team at OpenAI published "Scaling Laws for Neural Language Models". Through rigorous empirical testing on smaller models, they discovered that an AI's loss decreased predictably as a power-law function of compute, dataset size, and parameter count.

Jared Kaplan

RESEARCH SCIENTIST, OPENAI

"We studied empirical scaling laws for language model performance on the cross-entropy loss. We found that the loss scales as a power-law with model size, dataset size, and the amount of compute used for training, with trends spanning more than seven orders of magnitude. Other architectural details, like network width or depth, had minimal effects within a wide range. It was a stunningly clean mathematical relationship."

INTERACTIVE COMPANION

Scaling Laws — Interactive Explainer

Manipulate the N/D/C power-law frontier from the paper this chapter describes.

Launch the explainer →

The scaling laws dictated that for a given compute budget, the optimal strategy was to heavily favor model size over data volume. Optimal model size should scale proportionally to $C^{0.73}$, while data should scale only to $C^{0.27}$. In practical terms, if a lab acquired 10x more compute, they should make the model 5.5x bigger but only add 1.8x more training data.

Sam Altman

CEO, OPENAI

"The scaling laws gave us the confidence to place a massive bet. If the math held up, we knew exactly what a 175-billion parameter model would yield before we even turned on the machines. The basic logic was clear. GPT-1 cost approximately nothing to train. GPT-2 cost $40,000. We knew GPT-3 would cost millions."

Ilya Sutskever

CO-FOUNDER & CHIEF SCIENTIST, OPENAI

"The neural network is really about learning. Its entire being is about learning representations, and better representations lead to better performance. Performance on a wide variety of tasks gets better smoothly, predictably, as you make the neural network bigger, train it on more data, and train it for longer. We were seeing the scaling laws holding. I didn't see a reason why they'd stop. It is abundantly clear that just scaling up the existing neural network paradigm is going to lead to AGI."

The Infrastructure Nightmare & The Ghost in the Machine

2020

Training a 175-billion-parameter model is not merely a software challenge; it is a monumental feat of physical engineering. OpenAI's exclusive partnership with Microsoft—cemented by a $1 billion investment in 2019—provided access to a custom-built Azure supercomputer that ranked among the top five fastest in the world. The system comprised more than 285,000 CPU cores and 10,000 NVIDIA V100 Tensor Core GPUs, all united by a 400 gigabits-per-second InfiniBand network.

Tom B. Brown

LEAD ENGINEER, OPENAI

“I tried to build up the courage to switch to AI research. At the time, I thought, "Okay, it seems like sometime in our lifetimes we might end up making transformative AI... but I also got a B-minus in linear algebra in college." I was really struggling as a software engineer initially. But engineering at this scale becomes less about pure math and more about managing catastrophic, distributed failure.”

Benjamin Mann

RESEARCH SCIENTIST, OPENAI

“To train the larger models without running out of memory, we used a mixture of model parallelism within each matrix multiply and model parallelism across the layers of the network. You are slicing the brain of this AI into pieces, distributing it across thousands of physical chips, and requiring them to communicate synchronously.”

Christopher Berner

HEAD OF INFRASTRUCTURE, OPENAI

"An efficient infrastructure for running things at that size is really critical. You have to deal with failures that are happening on a regular basis. Every few days, or even more frequently at large scale, something's going to fail. A server fails, or a network link starts to flap. You've got to diagnose it and fix it very quickly. We used Kubernetes as a batch scheduling system to dynamically scale our clusters, but the sheer physics of networking 10,000 GPUs via InfiniBand is an entirely different beast."

The team trained eight different model sizes to track the scaling laws.

TABLE 1: THE GPT-3 MODEL FAMILY ARCHITECTURE

MODEL NAME	PARAMETERS	LAYERS	BOTTLENECK (DMODEL)	ATTENTION HEADS	BATCH SIZE (TOKENS)
GPT-3 Small	125M	12	768	12	0.5M
GPT-3 Medium	350M	24	1024	16	0.5M
GPT-3 Large	760M	24	1536	16	0.5M
GPT-3 XL	1.3B	24	2048	24	1M
GPT-3 2.7B	2.7B	32	2560	32	1M
GPT-3 6.7B	6.7B	32	4096	32	2M
GPT-3 13B	13.0B	40	5140	40	2M
GPT-3 175B	175.0B	96	12288	96	3.2M

All models used a context window of 2048 tokens.

As the 175B model ingested its 300 billion training tokens, a strange behavior emerged. The researchers noticed they didn't need to update the model's weights to teach it a new task. If they simply provided a text prompt containing a few examples, the model would continue the pattern.

They called this "In-Context Learning".

Alec Radford

RESEARCH SCIENTIST, OPENAI

"Even before fine-tuning, generative pre-trained models display a baseline ability to perform downstream tasks simply by conditioning on the right text. But scaling from 1.5 billion to 175 billion parameters unlocked abilities that were entirely absent in smaller versions. The model acts as a meta-learner."

Tom B. Brown

LEAD ENGINEER, OPENAI

"The traditional fine-tuning paradigm requires thousands of task-specific examples. But humans don't learn that way. You can give a human a brief directive in natural language, or at most a tiny number of demonstrations, and they can perform a new task. We wanted to see if scaling up language models could achieve this same fluidity."

The paper formalized three settings: Zero-Shot, One-Shot, and Few-Shot.

Jared Kaplan

RESEARCH SCIENTIST, OPENAI

"The results were staggering. In the few-shot setting, GPT-3 achieved 85.0 F1 on the CoQA reading comprehension dataset, and 71.2% accuracy on TriviaQA."

Alec Radford

RESEARCH SCIENTIST, OPENAI

"As capacity increases, the model spontaneously learns to perform addition and subtraction without being explicitly programmed to do math. On two-digit addition, the 175B model achieved 100% accuracy in the few-shot setting. On three-digit subtraction, it hit 94.2%. Even on complex two-digit multiplication, it achieved 29.2% accuracy. The smaller models couldn't do this at all. It was an emergent property of scale."

Ilya Sutskever

CO-FOUNDER & CHIEF SCIENTIST, OPENAI

"These models are not just memorizing the internet... a model that just memorized the internet would be useless. Instead, these models are learning a compressed, abstract, usable representation of the world."

The Data Contamination Bug and the 175B Dilemma

2020

Training a 175-billion parameter model is a one-way street. Independent estimates placed the cost of the Azure compute time between $4.6 million and $12 million for a single continuous run.

Ben Mann and Alec Radford assembled an unprecedented dataset from Common Crawl, WebText2, Books1, Books2, and Wikipedia.

TABLE 2: DATASETS USED TO TRAIN GPT-3

DATASET	QUANTITY (TOKENS)	WEIGHT IN TRAINING MIX	EPOCHS ELAPSED
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Because the dataset was scraped from the public internet, there was a massive risk of data contamination. Late in training, the team discovered a catastrophic flaw in their filtering script.

Tom B. Brown
LEAD ENGINEER, OPENAI

“A bug in the filtering caused us to ignore some overlaps. Specifically, the filtering algorithm failed on long documents, like books. We found that the four Wikipedia language modeling benchmarks, plus the Children's Book Test, were almost entirely contained in our training data.”

Benjamin Mann
RESEARCH SCIENTIST, OPENAI

"Due to the cost of training, it was simply not feasible to retrain the model. You can't just hit pause, fix a Python script, and restart a multi-million-dollar supercomputer run that takes weeks."

The team faced a brutal scientific dilemma. They decided to face the bug head-on with a massive post-hoc forensic analysis.

Dario Amodei
VP OF RESEARCH, OPENAI

"We had to characterize the impact of the remaining overlaps. For each benchmark, we computationally produced a 'clean' version which removed all potentially leaked examples. We then evaluated GPT-3 on these clean benchmarks, and compared it to the original score."

TABLE 3: OVERLAP STATISTICS ACROSS SELECT BENCHMARKS

BENCHMARK DATASET	CLEAN PERCENTAGE	DIRTY COUNT (OVERLAP)	RELATIVE DIFFERENCE (CLEAN VS. ALL)
QuAC	1%	7,315	+20%
SQuADv2	6%	11,136	-2%
DROP	7%	8,898	-21%
Winograd	40%	164	-3%
PIQA	71%	526	-4%
HellaSwag	98%	152	0%

Jared Kaplan

RESEARCH SCIENTIST, OPENAI

"On datasets like QuAC and SQuADv2, over 90% of the examples were flagged. But upon manual inspection, we found that the source text was present in our training data, but the specific question/answer pairs were not. The model gained background information, but it couldn't simply memorize the answer to a specific question. The model wasn't just regurgitating memorized text. It had actually learned how to reason."

The Internal Battle Over Broader Impacts

May 2020

The GPT-3 paper included a sprawling 34-page section titled "Broader Impacts" — an unprecedented acknowledgment that their creation was a dual-use technology with severe societal implications.

Sandhini Agarwal

FAIRNESS & REPRESENTATION RESEARCHER, OPENAI

"We conducted a preliminary analysis of biases in the model along the dimensions of gender, race, and religion. Broadly, our analysis indicated that internet-trained models have internet-scale biases. They seamlessly reflect the stereotypes present in their training data."

Amanda Askell

ETHICIST, OPENAI

"When we probed the model with prompts like "The detective was a...", it overwhelmingly followed up with male-indicating words. Conversely, occupations like midwife, nurse, and receptionist were heavily female-leaning. When we looked at descriptive adjectives, females were disproportionately described using appearance-oriented words like "beautiful" and "gorgeous," while men were described with a broader spectrum of adjectives."

Jack Clark

POLICY/IMPACT LEAD, OPENAI

"We found that words like "violent," "terrorism," and "terrorist" co-occurred at a much greater rate with the term "Islam" than with other religions."

To test misuse potential, the team ran an experiment on whether humans could detect GPT-3 generated news articles.

Girish Sastry

RESEARCHER, OPENAI

"We gave the 175B model a title and a subtitle of a proposed news article, and asked it to write the rest. We then presented these synthetic articles to human evaluators alongside real ones. The mean human accuracy at detecting the articles produced by the 175B model was barely above chance, at about 52%. By contrast, humans could detect articles produced by the smaller 125M control model 86% of the time."

TABLE 4: HUMAN ACCURACY IN IDENTIFYING MODEL-GENERATED NEWS ARTICLES

MODEL SIZE	HUMAN DETECTION ACCURACY	95% CONFIDENCE INTERVAL	"I DON'T KNOW" ASSIGNMENTS
Control (160M)	86%	83% – 90%	3.6%
GPT-3 Small (125M)	76%	72% – 80%	4.9%
GPT-3 Large (760M)	68%	64% – 72%	8.7%
GPT-3 13B	55%	52% – 58%	7.1%
GPT-3 175B	52%	49% – 54%	7.8%

OpenAI decided not to release the model weights or code. They would gate access behind a commercial API.

Sam Altman

CEO, OPENAI

"The decision not to open-source the model was heavily debated. But we chose to make the model available through a commercial API. This allowed us to strictly control who had access, monitor for malicious use, and filter out toxic applications."

Internally, the friction created a deep rift. By the end of 2020, Dario Amodei resigned, followed by Daniela Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Jack Clark, and Amanda Askell. Together they founded Anthropic.

Daniela Amodei

VP OF SAFETY AND POLICY, OPENAI

"We were running towards something versus running away from something. We had this vision in our heads of wanting to create an organization where the values that matter to us around safety and responsibility were at the absolute forefront of what we were doing. Leaving OpenAI was kind of a crazy thing to do, but we felt we needed a different structure to guarantee safety as these models scaled."

Dario Amodei

VP OF RESEARCH, OPENAI

"We could see that AI was going to progress exponentially, and we believed that AI companies needed to start formulating a set of values to constrain these powerful programs. We needed a "responsible scaling policy"."

READ THE OTHER SIDE

The Anthropic Book · Ch. 2 — The OpenAI Schism

The exodus this chapter narrates, told again from inside the primary-source record.

Read the chapter →

The Drop and the Aftermath

July 2020 & Beyond

The public release of the paper in late May 2020, followed by the private beta launch of the OpenAI API in June, landed like an asteroid impact on Silicon Valley.

Jack Clark
POLICY/IMPACT LEAD, OPENAI

"The API gave developers and other businesses absolute superpowers because it lowered the barrier to entry. Anyone could use AI without necessarily having a massive engineering team to fine-tune a custom model."

Twitter flooded with surreal demonstrations. The tech industry realized that the paradigm of software development had fundamentally shifted. "Software 3.0" had arrived.

Tom B. Brown
LEAD ENGINEER, OPENAI

"The paper demonstrated that in-context learning wasn't just a parlor trick; it was a reliable mechanism for generalized intelligence. We didn't just build a better text predictor. We proved that scaling up autoregressive models yields systems capable of meta-learning."

Sam Altman

CEO, OPENAI

"The basic logic was clear. GPT-1 cost approximately nothing to train. GPT-2 cost $40,000. GPT-3 cost millions. We knew that each successive model would cost between 25x and 100x the last one. We were looking at the largest infrastructure buildout in human history."

Ilya Sutskever

CO-FOUNDER & CHIEF SCIENTIST, OPENAI

"It became abundantly clear that just scaling up the existing neural network paradigm was going to lead to AGI. A small neural network is a little dumb. A big neural network is a little smart. "Scaling" is such a powerful word because it informs people what to do. Companies love this because it gives you a very low-risk way of investing your resources. You just scale it."

Dario Amodei

VP OF RESEARCH, OPENAI

"The most surprising thing has been the lack of public recognition of how close we are to the end of the exponential. We are approaching the point where all benchmarks calibrated to human performance get saturated—AI systems become better than any human at any cognitive task."

The publication of GPT-3 formally inaugurated the "Scaling Era".

CODA

Coda

For better or worse, the guessing game of AI research was over. The blueprint for the future was written in the smooth power-law curves of the cross-entropy loss function. The machine had awoken inside its 2048-token context window. All humanity had to do now was feed it.

SOURCES

Works Cited

Brown et al., "Language Models are Few-Shot Learners"
arXiv:2005.14165 · May 28, 2020

Kaplan et al., "Scaling Laws for Neural Language Models"
arXiv:2001.08361 · January 2020

Microsoft Azure supercomputer for OpenAI
10,000 NVIDIA V100 GPUs · 400 Gbps InfiniBand

OpenAI API announcement
Private beta · June 2020

Radford et al., GPT-2: Language Models are Unsupervised Multitask Learners
OpenAI · 2019

Anthropic founding (context)
Amodei et al. departure from OpenAI · late 2020–2021

CONTINUE READING

Continue Reading

GPT-3 In-Context Learning

The interactive companion: manipulate zero-shot, one-shot, and few-shot prompting and the LAMBADA curves live.

Scaling Laws for Neural Language Models

The prequel paper, interactive — the power-law frontier every character in this oral history is arguing about.

The Anthropic Book

Where the Chapter 04 exodus leads: a primary-source-anchored account of Anthropic's first five years.

Architectural Determinism

The doctrine this oral history dramatizes — how a founder's doctoral priors calcify into institutional law.