



EleutherAI

Annual Report 2025

Open AI Research, Infrastructure, and Community

Contents

Contents	2
A Message from the Executive Director	3
Our Mission & Theory of Change	4
1 · Building the Open-Source Ecosystem	5
2 · Select High-Profile Research Projects	7
3 · Building the Scientific Ecosystem	10
4 · Policy	12
Organization: Finances, Independence & Governance	13

A Message from the Executive Director

Despite the upheaval in the research and AI communities, 2025 was an excellent year for EleutherAI. EleutherAI researchers and collaborators released a record 31 papers in 2025. This portfolio is headlined by the Common Pile v0.1 and Deep Ignorance, two projects that I believe we will eventually look back on as among the most influential research EleutherAI has ever produced. Across 20 submissions to the leading publication venues in our field (ICLR, ICML, COLM, and NeurIPS), papers bearing an EleutherAI affiliation achieved a 60% acceptance rate, over twice the average. Five junior researchers had their first first-author papers accepted at one of these venues.

The breadth of the community producing this work is at least as important as its volume. Seventy people published under an EleutherAI affiliation in 2025, only 12 of whom were members of our staff. Thirty-seven were publishing with EleutherAI for the first time. Twenty-two of those new collaborators came through the Summer of Open AI Research (SOAR), a program we piloted in 2025. SOAR brought people with little or no prior research experience together with mentors from across the EleutherAI community to conduct original research and write papers. Its success demonstrated something fundamental about our model: given access to mentorship, collaborators, and a serious research community, people who have historically lacked conventional pathways into AI research can make meaningful scientific contributions. We have already begun planning to run the program again in 2026.

Our community expanded institutionally as well. We continued to work closely with Hugging Face across research, community-building, and policy. Together with Hugging Face and the University of Edinburgh, we spent 2025 building the Evaluating Evaluations Coalition, a new community dedicated to improving how AI evaluations are developed and assessed. I have been delighted by its first year and am excited to see that community continue to grow. EvalEval joins the Data Provenance Initiative as community-driven organizations that EleutherAI is serving as the fiscal sponsor for.

The reach of EleutherAI's research also continued to grow. Our body of work received approximately 20,000 citations during the year, bringing our cumulative citation count above 43,000. But the part of this result that makes me proudest is not the total. For the first time, the majority of the citations EleutherAI received were to papers I did not co-author. I see that as a powerful measure of how far the organization has come. Our influence is increasingly distributed across a large and growing community rather than concentrated in our earliest projects or founders. We have become capable of supporting researchers, collaborations, and ideas that extend well beyond any one person. One of my goals for 2026 is for EleutherAI's work to have received more than twice as many citations as my own.

We enter 2026 with a stronger research program, a larger community, and clearer evidence that open, collaborative AI research can produce work of lasting scientific value. I am enormously proud of what this community accomplished in 2025, and even more excited about what we are becoming.

Stella Biderman

Executive Director, EleutherAI

Our Mission & Theory of Change

Artificial intelligence powered by large-scale generative models is transforming the world, and it will keep doing so. But an overwhelming share of the research and expertise in these systems sits inside the trillion-dollar companies that sell them commercially. That concentration is a problem. Public accountability for corporate actors, functioning democratic oversight, and equitable access to AI's benefits all depend on a robust, empowered scientific establishment that is independent of corporate funding and influence, and does work that is socially important but not commercially incentivized. EleutherAI's primary mission is to be an exemplar of that ideal. We also want to lower the ladder behind us, so that others face fewer barriers to doing this kind of work.

Our four areas of work

To that end, we concentrate on four areas, which organize the rest of this report:

- **Building the open-source ecosystem:** releasing datasets, libraries, and models used across academia, nonprofits, and industry.
- **Research:** studying the capabilities, biases, potential, and risks of large-scale AI.
- **Building the scientific ecosystem:** practicing “science in the open,” running research-mentorship programs, and supporting fledgling organizations and researchers around the world with guidance, funding, and fiscal sponsorship.
- **Policy:** informing regulators about the technology and the likely impact of proposed rules.

As independence from the financial and political priorities of corporations is a primary goal of EleutherAI, we are very cautious about who we accept funding from and under what terms. In particular, we have a blanket policy of rejecting any targeted funding for projects that we did not already desire to do.

2025 at a glance

- **31** peer-reviewed papers in 2025, published at ACL, NAACL, ICLR, ICML, COLM, NeurIPS, and TMLR.
- **~20,000** citations to EleutherAI's body of work during 2025. This measures uptake of our accumulated research, not only work released this year.
- **99 participants** in the first SOAR mentorship program: 12 projects, 500+ applicants
- **\$4.0M** in 2025 cash revenue, up from \$2.81M in 2024 and \$1.96M in 2023, plus about \$1.0M in donated compute. We ended the year with a \$1.54M operating surplus.
- **12 staff** a small core team, working alongside a much larger volunteer and collaborator community.

1 · Building the Open-Source Ecosystem

Reproducible AI research needs open ingredients: data, model suites, and infrastructure. Most of the field cannot otherwise access them, and markets have little incentive to provide them. Building these ingredients and releasing them freely is the foundation of everything else we do.

Flagship releases

- **The Common Pile v0.1:** an 8TB dataset of public-domain and openly licensed text from 30 sources, including about 300,000 public-domain books. We built it with partners including Hugging Face and the Allen Institute for AI. We used it to train the Comma v0.1 7B models, which perform comparably to models trained on unlicensed data.
- **PolyPythias:** fifty language-model pre-training runs studying stability and outliers, extending the widely-used Pythia suite.
- **RWKV-7 “Goose”:** RWKV was the first Recurrent Neural Network (RNN) to match the performance of transformers at the 10B+ scale. RWKV-7 continues that line of work. Its key innovation is a smarter way of updating the model’s internal memory as it reads. This lets it reliably keep track of information across long inputs. In a formal sense, that makes it more expressive than transformers: it can handle certain patterns that transformers provably cannot. It does this at constant memory and constant per-token cost, so running it does not get more expensive as inputs grow.

Training infrastructure: GPT-NeoX

GPT-NeoX is our open-source library for training large language models at scale. It is one of EleutherAI’s most widely adopted contributions, and it remains the backbone of our own model training. In 2025 models trained with GPT-NeoX included our own *Deep Ignorance*, the Hubble suite for studying LLM memorization, and Geodesic Research’s Alignment Pretraining models. We ported GPT-NeoX to the UK’s new Isambard-AI supercomputer, enabling users to train 10B+ parameter models efficiently across hundreds of GH200s for the first time. We have also continued to work supporting the U.S. Department of Energy’s FORGE benchmark, which uses the GPT-NeoX library to evaluate the throughput of its supercomputers.

This builds on GPT-NeoX’s strong record of adoption over its lifetime: the library has powered widely-used models such as Stability AI’s StableLM and Together AI’s RedPajama-INCITE. It has been ported onto leading supercomputers including Frontier, Summit, LUMI, and Leonardo, and is the library that the U.S. Department of Energy uses to benchmark its supercomputers via FORGE.

Evaluation infrastructure: the LM Evaluation Harness

The Language Model Evaluation Harness (lm-eval) is our open-source framework for standardized, reproducible LLM evaluation. It remains one of the most widely used tools of its kind, drawing roughly 1,200 citations in 2025. Growth in the use of this library has slowed as more evaluation work has moved to agentic and multi-turn settings that lm-eval does not yet support. Even so, it continues to see heavy academic and real-world use.

Releasing software open source makes usage data hard to track, but we can reasonably approximate it. A manual audit of 100 papers citing lm-eval in 2025 showed that 74 of them used lm-eval to run their evaluations, built on our library, or used leaderboards powered by lm-eval such as the Open LLM Leaderboard. Extrapolating across the 1,200 citations lm-eval received in 2025, that suggests over 850 papers written in 2025 use or build on the library. This number doesn’t count papers that build on libraries built on lm-eval such as AI2’s OLMES or the UK AI Security Institute’s Inspect-AI which uses several of our implementations.

To keep lm-eval central as evaluation evolves, we are undertaking a large-scale refactor. It will make the library easier to contribute new tasks to, easier for AI agents to use, and simpler to maintain. We are also in discussions with the EvalEval Coalition about adopting it and sharing the maintenance load

2 · Select High-Profile Research Projects

Interpretability

Modern neural networks can be evaluated from the outside, but their internal computation remains difficult to understand. Mechanistic interpretability attempts to identify the representations and computations that produce model behavior. Better methods could help researchers explain capabilities, diagnose failures, and assess whether models are behaving for the reasons we expect. The field is still far from those goals, however, and progress requires evaluating interpretability methods as critically as the models they are intended to explain. Much of our work in interpretability this year has focused on studying sparse autoencoders (SAEs), a widely used methodology to decompose a model's activations into components that appear more interpretable to humans that we introduced in our 2023 paper *Sparse Autoencoders Find Highly Interpretable Features in Language Models*.

Sparse autoencoders do not recover a unique set of “true” features. A natural hope is that an SAE discovers a stable set of features already present in the model. *Sparse Autoencoders Trained on the Same Data Learn Different Features* tested that assumption by training SAEs on the same models and data while varying only their random initialization. The resulting decompositions differed substantially: in one experiment involving 131,000 learned features in Llama 3 8B, only about 30% were shared across random seeds. The result held across layers, datasets, model families, and SAE architectures. This does not make SAEs useless, but it does change how their outputs should be interpreted: as potentially useful decompositions of activation space, not exhaustive or uniquely correct inventories of the features a model “really” uses.

Transcoders provide a stronger decomposition for some interpretability tasks. In a related comparison, *Transcoders Beat Sparse Autoencoders for Interpretability* evaluated SAEs against transcoders, which learn to predict the output of a neural-network component from its input. Across the models and evaluation methods studied, transcoder features produced more interpretable and more faithful accounts of model computation. The paper also introduced skip transcoders, which reduced reconstruction error without sacrificing measured interpretability. These results recommend a practical change for researchers studying feed-forward networks: where both approaches are applicable, transcoders may be a better default than residual-stream SAEs. They do not resolve the broader validation problems identified by the field, but they improve the tools available for investigating them.

Evaluation

AI evaluation determines which capabilities become visible and which systems appear to be making progress. Many widely used benchmarks are narrow, translated from English, close to saturation, or poorly matched to the claims made about them. In 2025, EleutherAI-affiliated research expanded evaluation across languages, cultures, forms of generation, and scientific reasoning, and found important limitations in current systems.

Global PIQA: culturally grounded evaluation at unprecedented linguistic breadth. Most multilingual benchmarks either cover relatively few languages or translate the same source material across them. Translation makes comparison easier, but it can erase the local knowledge that language models are supposed to understand. Global PIQA was built through a participatory process involving 335 researchers from 65 countries. Its 116 language varieties span 14 language families and 23 writing systems; in its non-parallel portion, more than half of the examples refer to local foods, customs, traditions, or other culturally specific knowledge.

State-of-the-art models performed well in aggregate but unevenly across languages. Accuracy gaps reached 37 percentage points between languages, despite a 50% random baseline, and open models generally trailed proprietary ones. The result complicates broad claims about multilingual competence: a model can achieve a

strong average while remaining substantially less useful in lower-resource languages. Global PIQA is both an evaluation artifact and an experiment in how benchmarks can be built. Community members contributed knowledge in their own languages and were credited as co-authors rather than treated only as anonymous annotators. The benchmark emerged from the MRL 2025 shared task described in the scientific-ecosystem section of this report.

KMMLU: measuring Korean knowledge without translating an English test. KMMLU contains 35,030 expert-level questions across 45 subjects, drawn from original Korean examinations rather than translations of English-language benchmarks. One fifth of the questions require knowledge specific to Korean culture. When the researchers evaluated leading public and proprietary models, none exceeded 60%, well below the 80% passing threshold of the source examinations. The finding showed that strong performance on translated benchmarks can obscure meaningful gaps in Korean linguistic, cultural, and professional knowledge. The benchmark and evaluation code were released publicly and KMMLU was integrated into the LM Evaluation Harness.

BiGGen Bench: evaluating generated answers with criteria specific to the task. Open-ended generation is difficult to score. Broad labels such as “helpful” or “harmless” can conceal which aspects of an answer succeeded or failed, while benchmarks concentrated on instruction-following provide limited coverage of other capabilities. BiGGen Bench evaluates nine capabilities across 77 tasks using criteria written for the particular instance being judged. Its authors used five evaluator models to assess 103 language models, allowing them to examine not only model performance but also the reliability and biases of model-based evaluation. This work received the NAACL 2025 Best Paper Award.

SPOT: current models cannot reliably verify scientific manuscripts. Proposals for “AI co-scientists” often emphasize generating hypotheses, code, or papers. SPOT asks whether models can perform the complementary quality-control task: finding consequential errors in scientific work. The benchmark contains 83 published papers and 91 errors serious enough to lead to errata or retraction, cross-validated with authors and human annotators. On this task, no evaluated model exceeded 21.1% recall or 6.1% precision. Even the strongest model rarely rediscovered the same errors across repeated runs, and domain experts found that its mistakes often resembled student-level misunderstandings. These results identify a substantial gap between current model performance and reliable automated scientific verification. They are a reason for caution, not a claim that AI cannot eventually contribute to scientific review.

Open Weight Safety

Open weight models allow researchers and users to inspect, modify, and deploy systems without depending on the original developer. Those properties enable independent research and diffuse technical capacity, but also make conventional safeguards fragile: once weights are public, a downstream actor can fine-tune the model to remove refusal behavior, and there is no practical way to recall every copy. Almost no work has been done on developing safeguards that make sense for open weight models, and we’ve focused a significant amount of effort on doing field-building work in this space.

Deep Ignorance: preventing dangerous knowledge during pretraining. Most existing safeguards are applied after a model has already learned potentially dangerous information. *Deep Ignorance: Filtering Pretraining Data Builds Tamper-Resistant Safeguards into Open-Weight LLMs* studied a different intervention: filtering dual-use biological material out of the pretraining data so that the model never internalizes the targeted knowledge in the first place. In collaboration with the UK AI Security Institute and researchers at the University of Oxford, the team trained several 6.9-billion-parameter models from scratch on filtered and unfiltered data.

The filtered models performed near chance on the study's biothreat-proxy evaluations while showing no observed degradation on unrelated general-capability benchmarks. They also resisted adversarial fine-tuning for as many as 10,000 steps and 300 million tokens of high-quality biological text, more than an order of magnitude beyond the post-training safeguards used as baselines. The filtering pipeline added less than 1% to training compute, making it highly efficient to apply at frontier scale. This paper was the runner-up for the Best Paper award at the NeurIPS Workshop on Biosecurity Safeguards for Generative AI and is currently under review at ICLR.

Video Deepfake Abuse: we know how this technology will play out. In *Video Deepfake Abuse: How Company Choices Predictably Shape Misuse Patterns* we examined how choices by model developers and distribution platforms shape the production of non-consensual intimate imagery and abusive synthetic media. We found that a small number of openly released video-generation models had become dominant tools for photorealistic abuse, echoing patterns previously observed in image generation. The paper argues that these harms are not an unavoidable consequence of model capability alone: training-data curation, post-training safeguards, release decisions, and platform moderation materially affect how easily models can be misused. No intervention eliminates the risk, but organizations retain consequential choices about how much harm they enable or mitigate.

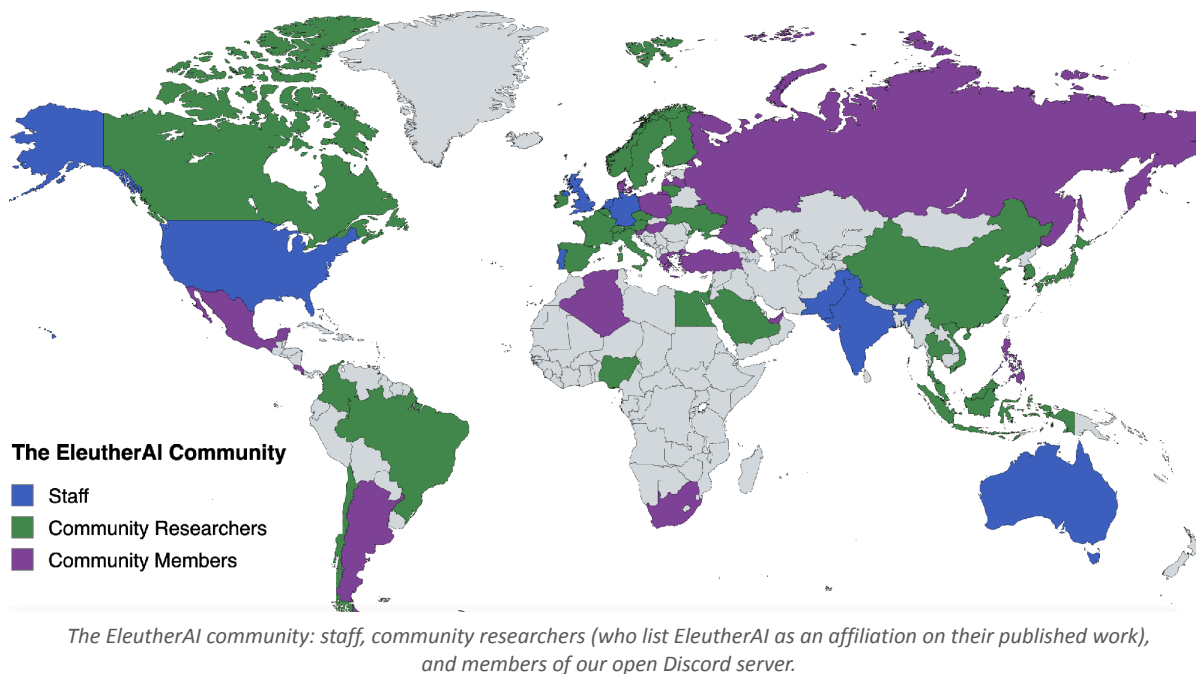
Building the field. Our work has excited several research labs, including the Apertus group, [FAR.AI](#), Geodesic Research, and Thinky Machines. We are exploring avenues for future collaboration with these groups and are applying for a grant with [FAR.AI](#) to run an open weight safety accelerator program in 2026.

3 · Building the Scientific Ecosystem

An independent scientific establishment is only as strong as the people in it. Beyond our own research, we work to widen the field. We practice science in the open, train new researchers, and support others doing this work.

A global community

We have always been more than our staff. Our work is carried out by a distributed, international community. That includes a core team, community researchers around the world who list EleutherAI as an affiliation on their published work, and the many members of our open, public Discord server. The community spans six continents. Widening it is central to both our mission and our impact.



Mentorship: SOAR

We ran the first **Summer of Open AI Research (SOAR)**, a mentorship program that pairs newcomers with mentors from EleutherAI and partner organizations (including Cohere and Apart) to do real, publishable open research. It drew more than 500 applications in two weeks, far beyond the roughly 100 we expected. From those, 142 people were accepted onto 12 projects, spanning mechanistic interpretability to AI safety. 99 participants remained active through to the end, a high retention rate for a fully online unpaid program.

Three SOAR teams released academic papers by the end of 2025, two of which have been accepted for publication already. Three more teams made substantial progress on papers that we anticipate being released in 2026 and one team's work was combined with a larger, on-going project at EleutherAI. Many teams had non-paper outputs, such as developing a map of [community-driven research orgs](#), using AI to improve moderation in our Discord server, and contributing to the back-end of the [share-lm project](#).

Testimonials from participants:

"Just after SOAR ended, I went on to be a MATS fellow in a very theory-heavy research stream. I think having participated in SOAR made the difference between me looking like an 'aspiring AI/ML researcher with lots of math background' instead of a 'physicist bandwagoning to AI.'"

Charles R. W.

“SOAR gave me the opportunity to get my foot in the door and actually contribute to a publication-worthy AI research project. It connected me with talented mentors who helped me build the foundations for my research journey.”

Ayesha I.

“SOAR was my first experience in an AI research program. Working remotely with experienced mentors pushed me to think like a researcher, from reading literature critically to designing experiments systematically to knowing when to ask the right questions. I’d recommend SOAR to any early-stage researcher who wants to bridge the gap between coursework and real research.”

Eryawan P. Y.

Workshops & Shared Tasks

We co-organized two workshops with community shared tasks, bringing together hundreds of researchers from around the world to improve multilingual representation in natural language processing:

- **WMDQS:** the 1st Workshop on Multilingual Data Quality Signals, held at COLM 2025 with Common Crawl, MLCommons, and Johns Hopkins. It ran a language-identification shared task on Common Crawl web text.
- **The MRL 2025 shared task,** a distributed collaboration that produced Global PIQA across 100+ languages, with community contributors credited as co-authors.

Fiscal sponsorship & support for other organizations

We extend guidance, financial support, and fiscal sponsorship to fledgling organizations doing important open work. In 2025 we served as fiscal sponsor for:

- [the EvalEval Coalition](#): a research community developing scientifically grounded methods and infrastructure for broader-impact evaluations of AI.
- [the Data Provenance Initiative](#): a group building public measurement infrastructure for AI data and model ecosystems, which traces the provenance and licensing of training data.

We were also working to establish a funding and fiscal-sponsorship relationship with **UniverseTBD**, a collaboration applying large language models to astronomy and the creators of AstroLLaMA. We expected to formalize that partnership going forward.

4 • Policy

AI policy forms a steady part of EleutherAI's work. Though smaller in scale than our research program, it meets a consistent need for an informed voice on open weight models. The conversation surrounding AI regulation has matured substantially since we started policy work in 2023, but expertise at the intersection of AI research, open source, and policy is still rare. Bridging this gap is important to our mission.

Research inspired by policy, advocacy shaped by research

Our independence means we set our own agenda, and sometimes we choose research topics specifically informed by current policy questions. 2025 saw the culmination of two such projects:

The **Common Pile** is the largest LLM training dataset composed exclusively of public domain and openly licensed text. While we believe the use of works for LLM training is likely fair use under US law, we still see the Common Pile as vital to the research ecosystem. First, open science requires sharing and fully documenting our training data. Second, debates about copyright and AI were happening in the dark, since large AI developers have no incentive to share any useful details about their training data. By building a corpus of only openly licensed and public domain content, we could demonstrate the benefits and challenges of this approach. Third, we raised the bar for the quality of provenance data in datasets, which is crucial for many technical policy interventions in data privacy and attribution.

Since most AI safety practices are designed for models that at most provide API access to their users, very little research exists on safeguards tailored to open weight models. Consequently, policy discussions about the risks of open weight models lack empirical grounding. **Deep Ignorance**, completed in 2025, is another project designed to address a research gap directly relevant to an area of AI policy. Together with UK AISI, we produced the first quantitative comparison of two fundamental safety interventions: pretraining data filtering versus later fine-tuning.

Taken together, these projects illustrate how we use our expertise in AI to make progress on difficult policy questions. We keep our research and policy work closely linked, and we believe it is important for our advocacy to stay grounded in empirical findings.

Policy Engagement

2025 was an uncertain year for AI public policy, especially in the United States. Nevertheless, we continued advocating for policies informed by our research and supporting our allies in the open source community. Our primary focus in the US remains educating policymakers about open source and providing expert opinion rooted in our unique position as independent open model developers. Drawing on our AI evaluation experience, we advised the Department of Commerce on best practices for testing LLMs. At the state level, we worked to ensure emerging AI regulation does not inadvertently harm the open source ecosystem — many AI policies are designed with only the largest companies and proprietary models in mind, and as a result can have disproportionate impact on open models.

We also maintained our engagement in the European Union, building on prior work related to the AI Act. We weighed in on the Act's Code of Practice in order to ensure the Act's implementation remains feasible for open weight models. We attended the **AI Action Summit** in Paris, which was an excellent opportunity to highlight how open source models fit into broader national AI strategies. Our conversations with senior policymakers at the Summit focused on sharing our insights into the AI supply chain and the cost of training and operating language models — yet another area of AI practice where commercial incentives impair information sharing with regulators. We [supported our allies](#) in further open source advocacy in the EU.

Organization: Finances, Independence & Governance

Financial overview

Revenue

	2025	2024	2023
Total revenue (cash)	\$4,010,710	\$2,808,142	\$1,960,567

Expenses

Expense	2025	2024	2023
Personnel	\$1,585,889	\$1,952,770	\$1,179,028
Compute & cloud infrastructure	\$476,247	\$0	\$0
Travel & conferences	\$256,215	\$50,162	\$18,685
Contract & professional services	\$69,904	\$194,899	\$120,371
Software & subscriptions	\$36,000	\$69,450	\$25,139
Grants & scholarships	\$0	\$287,822	\$0
Office & other program costs	\$42,000	\$19,740	\$5,679
Total expenses (cash)	\$2,466,255	\$2,574,843	\$1,348,902

Net

	2025	2024	2023
Net assets, beginning of year	\$844,964	\$611,665	\$0
Change in net assets	\$1,544,454	\$233,299	\$611,665
Net assets, end of year	\$2,389,418	\$844,964	\$611,665

Note: numbers may not sum exactly due to rounding to the nearest dollar

Cash revenue grew from \$1.96M in 2023 to \$2.81M in 2024 to \$4.01M in 2025. We ended 2025 with an operating surplus of \$1,544,454 and net assets of about \$2.39M. We built this reserve deliberately, as several major grants were scheduled to end or decline in 2026 and we wanted a cushion to protect our research program while we secured new funding. About 85% of our spending goes to research programs each year.

We have additionally received \$1,000,000 in in-kind contributions each year that are not included above.

Funder Breakdown

EleutherAI's 2025 work was supported by a diversified base of philanthropic, public, and mission-aligned funders, totaling about \$4.0 million in cash and \$1.0 million in compute. Our 2025 supporters included:

- **Hugging Face:** \$1,000,000
- **#startsmall:** \$1,000,000 (via Fidelity Charitable), the 2025 installment of a three-year, \$3,000,000 commitment.
- **Good Ventures / Open Philanthropy:** \$880,758, supporting our interpretability research.
- **The UK AI Security Institute (AISI):** \$809,452 (£630,700), via DSIT and Capgemini, supporting our open-weight safety work, including *Deep Ignorance*.
- **The Mozilla Foundation:** \$200,000, supporting our openly-licensed data work.
- **Renaissance Philanthropy (AI for Math):** \$120,500, supporting AI4Math research.
- **CoreWeave:** roughly \$1,000,000 in donated cloud compute each year, including 2025 (in-kind).

Looking ahead: funding and sustainability

We enter 2026 anticipating a sharp decline in several major grants. Our three year grant from Hugging Face comes to an end after Q1, a loss of \$750,000 in funding. We also expect our support from the UK AI Security Institute to fall from about \$809,000 to \$200,000, as the large cost of Deep Ignorance pretraining was a major driver of the 2025 funding figure. In total, we expect roughly \$1.6 million less revenue in 2026 than in 2025. We planned for this well in advance, forming a relationship with #startsmall, whose multi-year commitment replaces part of the lost funding and provides bridge funding to cover what would otherwise be a shortfall. As a result, our board-approved 2026 budget stays balanced, with a modest projected surplus, even as total funding falls. We are also in active conversations with several of our long-term funders and expect them to increase their support beginning in FY 2027.

People & governance

Stella Biderman, Executive Director.

Board of Directors: Emad Mostaque, Colin Raffel

Legal: EleutherAI is a 501(c)(3) incorporated in Delaware in 2023. EIN: 92-2215190