EleutherAI on AI, Innovation, and the Open Source Ecosystem

Senator Schumer, Members of the AI Insight Forum, and my fellow roundtable contributors, thank you for convening and bringing your insights to this important and timely discussion. My name is Stella Biderman and I am the Executive Director of EleutherAI. I am excited to speak with you today about the importance of increasing funding for open artificial intelligence research so that the United States can maintain its leadership role in innovation and AI safety, and the perils that we will face if the current trend of secrecy and privatization is allowed to continue.

EleutherAI is an open source AI research non-profit based in the U.S. specializing in large-scale generative AI technologies, especially text models. We were founded in 2020 by independent researchers who felt that technologies like GPT-3 were going to be the future of artificial intelligence, but felt that the non-release of the technology stifled innovation and specifically research on the capabilities, limitations, and biases of the technology. To solve this problem, we took it upon ourselves to train and publicly release the first open source GPT-3-style language models in the world. Over the next two years we released three models that each held the title "the largest open source language model in the world" at the time of their release [1, 2, 3], and subsequently have been used by hundreds of researchers around the world to study essential questions about hows and whys of large language models behavior [1, 2, 3, 4]. We also have worked with generative AI algorithms in other modalities, including the first publicly released text-to-image model and models for scientific applications such as protein folding, mathematics, and materials science research. Finally, in addition to training large language models we also do cutting-edge research on questions of high social importance such as measuring models' internal beliefs, understanding how models replicate or exacerbate existing biases and social inequities, and how model behavior can be constrained or controlled at deployment time.

Open Research is Essential to Innovation in AI

It is a common narrative that the scale at which today's large scale models operate makes open source researchers irrelevant, but nothing could be further from the truth. The core innovation engine in the world of artificial intelligence has always been open source researchers, whether it is the invention of the algorithms that underlie models like ChatGPT and Stable Diffusion or key innovations in improving training efficiency and architecture design. The key components of every large language model in the world today were created by researchers openly releasing their work, and much of it was done by staff at universities and non-profit organizations.

Many of the most pressing questions about artificial intelligence are not questions about making technology work; they're questions about making technology work right and work for everyone. Questions about reliability, controllability, and transparency are very easy to answer in the open source space. Because anyone can inspect the training data of models trained by EleutherAI, they can determine if their writing was in the training dataset, whether the model is inventing something new versus leveraging knowledge acquired through training, and explore the social biases encoded in the training data. These things are not possible for closed models, leaving these essential questions about how the technology will interact with society largely unanswerable. Even worse, companies are known to actively censor such research and regularly suspend accounts for researchers trying to test the limitations of their models. We cannot allow AI systems to be widely deployed and simultaneously inscrutable to anyone outside the tech companies developing the models or company-approved outside audits.

More Compute for Open Research is Needed

Overwhelmingly the funding and resources in the United States are going to a small number of very well resourced companies that have made the financial decision to stop sharing their research with the rest of the world. It is absolutely in the best interests of and within the capabilities of the United States government to counter this trend. EleutherAI is the only non-profit organization in the United States with expertise in training very large language models, but if we are going to continue to advance this transformative technology while minimizing its potential harms to society it is essential that we provide increased funding and opportunities for other researchers to contribute to this field.

I am proud that EleutherAI, along with the Canadian research institute Mila and the German non-profit LAION, currently holds the only major grant on the US government's largest supercomputer conducting research to advance AI. We have had the privilege of working closely with the staff at Oak Ridge National Lab to teach them about large language models (LLM) and bring the capability to train them to the Summit and Frontier supercomputers. However we are just one organization and very much the exception rather than the rule: further investment in public research is essential. **EleutherAI urges you to support increasing investment by the Department of**

Energy in sponsoring AI research and funding the National AI Research Resource (NAIRR).

When I go to AI conferences presently, the two countries that stand out the most to me are China and France. Both countries have made massive investments in bringing computing resources to the hands of their universities and it shows. I regularly meet French and Chinese PhD students who are doing work that their American peers would not dream of, and they always credit it to their respective governments' willingness to make the investment to make computing resources available to them. In the LLM space specifically, several countries have caught up with the US in the past few years: of the five largest publicly released language models in the world only the smallest was trained in the US, with the others being trained in the United Arab Emirates, France, China, and Russia. No academic institution[1] in the US has ever trained a LLM with 20 billion parameters or more, but five have done so in France, China, and the United Arab Emirates. **By studying how programs in these countries have succeeded in promoting AI research, we can learn about how the United States can return to the forefront of innovation.**

Non-Release Limits the Impact of AI

In the past few years there have been several examples of models developed that *could* have a substantial impact on the practice of science in the United States but did not because they were held privately by the organizations that developed them. The best example of this is Minerva, a ChatGPT-for-mathematicians, which was developed by Google in June 2022. Unfortunately it has seen almost no actual application because mathematicians are not allowed to use it. Similarly, Google collaborated with several academics to build a model called Baldur on top of Minerva which claimed to be able to supercharge work in formal reasoning. Again, the model was never made accessible to anyone outside of the company and the academics who created it lost access to it after the project was completed.

The mathematics community feels this loss strongly. I am an invited speaker at a series of workshops being held by the National Academies on "AI to Assist Mathematical Reasoning," and the inaccessibility of these models and subsequent loss of potential mathematical innovation was a major topic of discussion. At the most recent workshop I shared that EleutherAI – in collaboration with several other attendees – was in the

[1] Some US academics have participated in training models by collaborating on projects led by EleutherAI or others but no US academic institution has played a significant role in these projects.

middle of reproducing the Minerva model and a dozen academics came up to me afterwards to thank me for doing this important work and to talk about how excited they were about it. We just released our first set of models for mathematics and have had academics at over a dozen universities across the country reach out to tell us about how excited they were for this project and how it was going to help them do mathematics research.

Openness is a Tool for Good

There is no one-size-fits-all answer to what is "ethical" or what is "acceptable use." Determining these things requires context-specific examination of factors that cannot be known to model developers ahead of time. Thus, **transparency is a powerful tool for ethical deployment of AI because it enables downstream users to make informed decisions about what is or is not appropriate for them to use in their context.** Knowledge about the model's functionality, limitations, and biases lets the model deployer create an AI system that works for them and their intended user. In situations when the model developer is different from the model deployer, providing comprehensive information about the AI model is often the best way to promote ethical use.

Another benefit of increased transparency is that open source evaluation code lets us quantify model capabilities in a reliable and reproducible manner. In machine learning research, we unfortunately see instances of the same language model achieving different scores on the same benchmark between different papers. Our ability to evaluate the capabilities of models is being undermined by a lack of transparency, and the only way to fix this is for the evaluation code, training data, and experimental setup to be available for scrutiny: which means open source. **We need reliable evaluation in order to ensure sustainable progress and innovation in the generative AI space.**

Closed Models are not Safe: Their Dangers are Just Hidden

AI safety is a real and pressing issue, but it is not as simple as applying "safety filters" or hiding models behind commercial APIs. It is well-known in the machine learning world that the alleged safety filters deployed on popular commercial models are much better at making models look safe than actually be safe. The safety protocols that have been trumpeted as the solution by tech companies are very easily subverted and are frequently ineffective patches over deeper issues. Policymakers and society at large should be critical of such inadequate solutions. Many risks associated with generative

AI are complex sociotechnical problems, and thus require proportionately comprehensive solutions.

This is exacerbated by the fact that the market is full of “AI Safety” products that are ineffective at best and harmful at worst. While there are lots of organizations promoting “AI detectors” for text and images, the truth is that we do not know how to do this reliably. This is again something that has been widely known among experts, but our lack of ability to evaluate privately held tools makes it hard for researchers to effectively advocate on their own. **Regulators and researchers must work together to halt the spread of misinformation about “AI safety” tools through testing and creating standards.**

Reversing the Trend towards Closedness

We currently live in a world that is heavily trending towards non-disclosure of essential training details, including at companies that continue to release their final trained models openly. The reason for this is simple: all incentives point towards making less open systems. Openness is a liability, despite many parties benefiting greatly from greater access to model components, including training data. There are many research questions that require access to the data. Some of these questions are of great interest to policymakers and creators: what increases the likelihood of a language model reproducing long strings of words from its training data? Can a model derive scientific facts not stated explicitly in their training data?

Every month I speak with another organization that would love to be more open and more transparent about their work but who does not feel comfortable doing so due to a lack of legal clarity.

If there is one thing that this forum takes away from my statement, let it be that clarity is needed regarding copyright and AI training in the US. Any intervention in this legal matter must recognize that access to training data is important not only for companies developing and commercializing models, but also for researchers, auditors, and society at large.