# Open-source Language Models: Identifying Contributing Factors and Future Trends

Allison Huang
August 2023

**Abstract**

While the computer science domain has long embraced open research, the potential open-sourcing of artificial intelligence models poses significant risks. Open-source models provide deeper insights into how models work, but they are also susceptible to misuse. Using a collection of case studies, interviews, and research, this report attempts to identify the factors that contribute to open-source development of artificial intelligence models and draw out trends that may accelerate or hinder development in the future. It finds that access to inputs determine what kind of project an actor can take on, access to know-how enables the actor to turn inputs into a model, and incentives for and against open-source dictate how the model is released. Generally speaking, actors with more incentives to open-source models have access to fewer resources. As a result, open-source development has consistently trailed behind closed-source models, aided by releases from a small number of well-resourced actors like Meta. However, increasing demand for compute has become a larger barrier for open-source actors over time, and it is unclear whether that will prevent a model powerful enough to be truly dangerous from being open-sourced. Many uncertainties remain — for instance, new training methods like model imitation may reduce the size of the compute barrier, but the merits of these methods are still up for debate. Additionally, it is unclear what regulation from government actors may look like and whether it will target open-source development.

# Executive Summary

Open research has long been the default in the computer science domain, but the open-sourcing of capable language models poses unique risks. While open-source models provide valuable insights into how language models work, they are susceptible to misuse. This research aims to better understand the open-source landscape such that the benefits of open-source may be wielded while also reducing the risks that come with releasing a model for anyone to use freely. The case studies included in the report are GPT-2 from OpenAI, OPT from Meta, BLOOM from the BigScience Workshop, and Vicuna from the large Modeling Systems Organization (LMSYS).

This report identifies and analyzes three key factors determining the quality of available open-source models — access to inputs, access to know-how, and incentives for and against open-sourcing models. Inputs include computational power, training data, and algorithmic insights; know-how can take the form of talent, publication or documentation, and tooling. Incentives for and against open-source vary widely, with some supporting open-source stances to gain credibility and increase safety and others opposing it on the grounds of maintaining competitive advantage and reducing the risks of misuse.

Based on the contributing factors, actors in the open-source landscape can be clustered into three groups: big tech/for-profit AI labs, nonprofit/volunteer AI labs, and academia/individuals. Big tech/for-profit labs have access to enough inputs and know-how to pre-train large language models. Nonprofit/volunteer AI labs are often dependent on sponsorships for compute as an input, though they struggle to retain the know-how to scale these systems because industry can offer compensation in ways that they cannot. Academia/individuals are limited in resources such that they can only fine-tune models that were pre-trained by other actors.

From these factors, this report draws out a number of emerging trends.

- **Much of big tech and for-profit AI labs are experiencing a paradigm shift from open- to close-source.** Safety has been cited as the predominant reason for keeping these models under wraps, and there is nothing to indicate this changing in the near future. Some actors are exploring business models that utilize open-source models, but it is unclear how this will work.
- **Meta's releases have a disproportionate impact on open-source progress.** Meta has far more resources than any actor in the open-source space, and by releasing both high-quality pre-trained models and detailed documentation about the making of these models (beyond the information included in a typical technical report), it is providing a lot for other actors to build off of.
- **Scaling increases the compute barrier, but open-source actors tend to do more with less.** Over time, the amount of compute required to train a comparable model to the state-of-the-art at any given time has rapidly increased. However, the open-source community has been able to create quite capable models with less compute than closed-source actors, it is unclear how much of a hindering factor access to compute will be.

- **Model imitation may yield competitive, open versions of closed-source state-of-the-art models.** While the merits of model imitation (fine-tuning a smaller model on data from a high-quality, teacher model) are still up for debate, recent research has found promising results. The imitation and teacher model appear to be somewhat correlated, raising questions about what the imitation models can reveal about the teacher models.
- **The EU AI Act suggests how future regulation may address open-source development.** As the first attempt at comprehensive legislation addressing AI, the drafts of the EU AI Act and their eventual reconciliation provide potential frameworks under which open-source development could be regulated.

# 1. Introduction

A rapid improvement in the capabilities of artificial intelligence (AI) paired with increased accessibility to the public in recent years has brought more attention to the future impacts of AI — the good it can do, the threats it may pose, and how its role may evolve. Currently, the most capable, state-of-the-art (SOTA) models are closed-source. The model code, weights, and training data are not accessible to the public. Instead, all interaction with the model takes place through chat interfaces or APIs, if public access is granted.

However, as with any other technology, the open-source artificial intelligence community has demonstrated the ability to experiment and iterate incredibly quickly. Individuals and student researchers have been replicating models as early as the release of GPT-2 in 2019, and the online library Huggingface contains hundreds of AI models and datasets — some of which are competitive with ChatGPT.

Open-source development has significant benefits and risks relative to closed-source development. Open-source development has long been the default in the computer science domain, as it means that academics can build on research, developers can quickly stress-test their products, and individuals can collaborate at large to build creative solutions. Specifically, in AI, open research practices allow researchers to better study how models work and are necessary for certain types of safety research.

However, the risks arise because of how powerful AI models are. While most SOTA close-sourced models have stringent safety precautions (which are by no means perfect), many open-source models either lack the same guardrails or, by definition, cannot prevent users from removing them. This creates a big opportunity for misuse, and this was already demonstrated when OpenAI's text-to-image model DALLE-2 was copied by StabilityAI. Their copycat, Stable Diffusion, was open-sourced and consequently used to generate deepfakes (Casper, 2022).

Artificial intelligence models are still quite young today, but it is plausible that in the future, language models could democratize access to complex biological systems such that an individual with little biological expertise could cause a pandemic or unleash a bioweapon (Urbina, 2022). This is just one potential scenario, but it suggests the danger of capable, fully open language models. These risks have prompted concerns about the need for regulatory action, sparking debates on how much government intervention is necessary and what form it should take.

This report does not delve into the pros and cons of open-source, nor does it seek to make policy recommendations. Instead, it aims to study development of open-source language models over time to determine:

1. The contributing factors that affect how quickly open-source development progresses
2. Any trends that may change the rate of open-source development in the future

This report only includes language models which “model the generative likelihood of word sequences, so as to predict the probabilities of future (or missing) tokens” (Zhao, 2023). It does not take into account other types of AI models like diffusion (text-to-image) or game-playing models, and it does not differentiate between language models fine-tuned for different downstream tasks.

The report focuses on language models for two reasons. First, language models are arguably the most capable AI models as of now, making the risk of misuse the greatest. Second, language models have been developed at a range of sizes and by a variety of different actors, meaning that there are more data points from which to draw out trends.

Ultimately, this project hopes to set the stage for future research on the role of open-source development in artificial intelligence, such that the benefits of open research can be leveraged and the threats of misuse can be reduced.

## 1.1 Background

This section provides very brief definitions on relevant terms that are referenced throughout the report. Most of these terms are further elaborated upon in various sections.

The training of large language models is typically split into two parts:

1. **Generative pre-training** — Feeding a model on a massive corpus of text data such that given an input, it learns to predict probabilities of future tokens
2. **Supervised fine-tuning** — Used to hone a model for a specific downstream task by training it on labeled data

**Transformer** — An architecture for handling sequential data like text that is now the predominant architecture for text-to-text models; it is further discussed in Section 3.2

**Compute** — Computational power used to train models; it is further discussed in Section 3.1

# 2. Case studies

These case studies are chosen in an attempt to pick models that, taken together, are representative of LLM development. They vary in origin, age, and size (Table 1):

**Table 1:** Case studies

| | **Actor** | **Year** | **Size of largest** | **Training compute** | **Training hardware** |
|---|---|---|---|---|---|

| | | | version (parameters) | (FLOPs) | |
|---|---|---|---|---|---|
| **GPT-2** | OpenAI | 2018 | 1.5B | 1.49E+21* | No public information available |
| **OPT** | Meta | 2022 | 175B | 4.30E+23* | 992 A100s |
| **BLOOM** | BigScience | 2022 | 176B | 1.80E+23* | 384 A100s |
| **Vicuna** | LMSYS | 2023 | 13B | 4.10E+18~ | 8 A100s |

*Compute (FLOPs) from Epoch (2023).
~Calculating using Epoch's compute calculator (method 2), assuming a 30% utilization rate.

Informal interviews were conducted with people who worked on these models.

# 2.1 GPT-2 (2019)

## Development

GPT-2 was OpenAI's second generation of language model, the largest version containing 1.5B parameters. It was released in stages, a break from the open-source paradigm that had been considered default before. OpenAI feared misuse if the models were to be fully open-sourced, and the staged releases were a chance to see how a given model was used before releasing a more powerful version (Solaiman, 2019). The releases were as follows:

- February 2019 — Initial release and 124M parameter version (OpenAI, 2019a)
- May 2019 — 355M parameter version
- August 2019 — Six month follow up and 774M parameter version (OpenAI, 2019c)
- November 2019 — Full 1.5B version release (OpenAI, 2019b)

GPT-2 sought to build off its predecessor, GPT, by developing better generalization abilities. While past models were able to learn patterns from their training data, they had trouble generalizing to data that they had not encountered before (Radford, 2019).

While GPT-2 uses the same architecture as GPT with some slight modifications, it differs significantly from GPT in its size and training data. The largest version of GPT-2 was over ten times larger than GPT, and GPT-2 was trained on a new dataset, WebText. Past models had been trained only on one type of data — for instance, only fiction books. A large web scrape known as CommonCrawl had been used to solve this problem, but CommonCrawl has been known to include noisy and low-quality data (Zhao, 2023). To combat these shortcomings, WebText was created by scraping outbound links on Reddit with at least three karma. Karma was used as a heuristic that others found the content useful or interesting (Radford, 2019).

## Inspired models and replications

The implications of GPT-2 for open-source language model development arguably lie in the models it inspired that emerged between the initial release of GPT-2 in February 2019 and the full open-sourcing the 1.5B version in November 2019. These models essentially stress-tested OpenAI's staged release mechanism, finding that it was not particularly effective given how easy it was for small parties with little experience and few resources to come close to and eventually replicate GPT-2 (Shevlane, 2019).

The information contained in the rest of this section ("Inspired models and replications") is largely drawn from Toby Shevlane's thesis "The Artefacts of Intelligence: Governing scientists' contribution to AI proliferation." Shevlane covers the following three models in detail (Table 2).

**Table 2:** Inspired models and replications of GPT-2

| | **Date** | **Actor(s)** | **Objective** | **Data** | **Cost allocated** | **Compute source** |
|---|---|---|---|---|---|---|
| **Grover** | May 2019 | UW and Allen Institute for AI | fake news detection | RealNews | $35,000 | Google |
| **(Unnamed attempted replication)** | June 2019 | Connor Leahy | GPT-2 replica | attempted replica of WebText | $43,000 | Google |
| **OpenGPT-2** | August 2019 | Vanya Cohen and Aaron Gokaslan | GPT-2 replica | OpenWebTextCorpus (attempted replica of WebText) | $50,000 | Google |

**May 2019 — Researchers from University of Washington and AllenAI published a paper about their model, Grover.**

Grover was trained to be used for fake news detection, though it is included here because it uses the same architecture as GPT-2. Like OpenAI, they did not release the largest versions of the model, though researchers were granted access to the model on a case by case basis.

Grover was trained on RealNews, a subset of news articles from CommonCrawl created for the project.

Total cost of Grover was approximately $35,000 — compiling the dataset cost ~$10,000, and the remaining ~$25,000 was used for training, which occurred on Tensor Processing Units

(TPUs) provided by Google (Zellers, 2019). TPUs are Google's machine learning accelerators that are accessible via Google Cloud.

**June 2019 — Connor Leahy, an undergraduate in Germany, attempted to replicate but did not release a copy of GPT-2 1.5B.**

Connor had never used a transformer or a TPU before, but he attempted to create GPT-2 based on the information provided in the paper. Because the technical paper did not provide all the information necessary to perfectly replicate GPT-2, Connor's model differed from the original in his training setup, hyperparameters, and dataset (Leahy, 2019b).

As a result, his final model was significantly weaker in capabilities than GPT-2, as measured by perplexity on different datasets (Leahy, 2019b). Perplexity is a metric used to represent how well a model predicts the next token (Huggingface, n.d.). However, the first author of Grover thought that Connor's model, at its core, was mostly correct. Given more time and resources, he thought Connor would be able to get much better results (Shevlane, 2022).

Google again sponsored computing power for the project, and Connor estimated that he used $43,000 worth of compute (Leahy, 2019a).

**August 2019 — Two masters students from Brown University created and released OpenGPT-2 with an accompanying blog post explaining how they did it.**

The two students had no prior experience in language modeling. They based their implementation on Grover, though it is not clearly stated in the OpenGPT-2 blog post whether this is because they had full access to Grover or if there was a significant difference between Grover and GPT-2 that made Grover better. However, the objective of the model was changed to match that of GPT-2.

The two students tried to replicate WebText based on the information in the paper, resulting in the creation of OpenWebTextCorpus. As with Connor Leahy's efforts, OpenWebTextCorpus is not a perfect replication of WebText because OpenAI did not include all the details of how WebText was created in the GPT-2 paper.

The final model had very similar perplexities to GPT-2, making it a successful replication of GPT-2.

Google sponsored compute for OpenGPT-2 as well, and the blog post estimates that $50,000 worth of compute were used to train the model (Cohen, 2019).

## Takeaways

- **Compute is a barrier to open-source replication.** All three models relied on compute sponsorships from Google to train their models.

- **Companies are incentivized to sponsor compute.** Google sponsored compute for these models because having the code optimized for TPUs incentivized future work to continue using TPUs (as opposed to switching to another company's hardware), and the publishing of models trained on TPUs was good marketing (Shevlane, 2022).
- **GPT-2 marked the beginning of a shift away from the open-source paradigm.** This was cemented with the release of GPT-3 which was only accessible via an API. Google Deepmind followed this move, keeping subsequent SOTA models closed-source.

## 2.2 OPT from Meta (2022)

### Development

OPT-175B was trained with the goal of providing a model with capabilities comparable to those of GPT-3 for open access to researchers (Zhang, 2022). Until that point, only models on the order of magnitude of tens of billions of parameters had been open-sourced. By providing access to a model with over 100 billion parameters, Meta provided researchers with insights into scaling language models that had previously been kept behind closed doors.

The model architecture and hyperparameters were based on that of GPT-3. The pre-training corpus was created by combining subsets of open-source datasets including RoBERTA, the Pile, and PushShift.io Reddit.

OPT was trained on 992 80GB A100 GPUs using optimization tools like Fully Sharded Data Parallel (FDSP) and Megatron-LM. The final model had a carbon footprint that was approximately 1/7th that of GPT-3.

OPT was open-sourced along with a 114-page logbook documenting the day-to-day training process of the model. The logbook provides a much more granular view of the training process, detailing hardware bugs and loss spikes, among other things.

### Takeaways

- **Detailed documentation is a significant vehicle for know-how.** While other releases often were limited to an academic paper with some implementation details, Meta's documentation detailed iterations of the model at different points in time, things they tried, as well as things that went wrong.

## 2.3 BLOOM (2022)

### Organization

BLOOM (BigScience Large Open-science Open-access Multilingual Language Model) was a model created during the BigScience Workshop, a "social experiment in value-grounded collaborative research" (Allen Institute, 2023). Over 1.5 years, over 1,200 researchers with

backgrounds in varying areas were organized into 30 working groups in an effort that ultimately produced BLOOM, the ROOTS corpus (the data that BLOOM was trained on), and several scientific papers (BigScience Workshop, 2022).

The project was made possible by a compute grant from the French government on a national supercomputer, worth ~$5M (Allen Institute, 2023).

BLOOM was entirely a volunteer effort, with a handful of people working full-time on the project overseeing academics and industry employees who made time to contribute when they could. A lead of a working group said in an interview that they found that this arrangement led to slow progress. Because of the sheer magnitude of people involved, they said that there was a "collective action problem where no one felt ownership for the project." Combined with a lack of enforcement mechanism, organizers carried a lot of the burden of the project (Allen Institute, 2023).

## Development

BLOOM was trained on the 1.6TB ROOTS Corpus, which includes 46 natural languages and 13 coding languages. Because traditional methods of filtering data often end up hurting already marginalized populations, there was an emphasis on getting data that was high quality, representative, and inclusive and governing it in a way that was fair to data providers.

Much of the corpus was curated by BigScience members through localized hackathons in a bottom-up approach. GitHub code was added as a source of training data for code. The corpus was completed with sources from OSCAR, which ultimately comprised 38% of the corpus. OSCAR (Open Super-large Crawled Aggregated coRpus) is a multilingual dataset created by filtering CommonCrawl.

To determine the model architecture and parameters, the training team referenced the GPT-3 paper and conducted their own experiments, producing the paper "What Language Model to Train if You Have One Million GPU Hours?". A member of the training team told me they effectively threw everything at the wall to see what stuck (and what scaled) because there had been such little public information on creating a pre-trained model of this scale.

That member also mentioned that the computational budget for these experiments was purely based on hardware constraints. The workshop received upgraded hardware partway through the project, effectively segmenting the compute they had access to since one cannot run different hardware in parallel.

The final training run took 3.5 months on 48 nodes of 8 A100s.

## Motivation & Openness

The values defining the workshop were inclusivity, diversity, openness, and responsibility. These values informed tangible decisions and attributes of the model, including its multilingual

capabilities and open-sourcing. Moreover, the workshop found a number of problems with the LLM landscape at the time — namely:

1. There was a lack of linguistic diversity with most models being trained primarily on either English or Chinese data
2. Industry labs tightly controlled almost everything relating to the models and their development and deployment
3. It was an echo chamber of machine learning scientists and engineers with few people of other backgrounds
4. A number of ethical, social, and environmental considerations were not taken into account, including the carbon footprint of training models and the importance of quality data in properly representing marginalized populations (Allen Institute, 2023; BigScience Workshop, 2022).

### Takeaways

- **Open-source projects often require more organizational overhead.** At this point in time, open-source projects are more likely to be volunteer work because there are more commercial incentives to keep projects closed-source. Volunteer work inherently lacks the structure and incentives of industry, as employees are part of a clear hierarchy, are compensated for their work, and face consequences if they don't complete their work. An organizer on the project explained in an interview that this places an extra burden on organizers, which scales as projects get larger.
- **There is motivation to work on LLMs that value openness and collaboration.** The participation of over 1,200 individuals on this volunteer project indicates some level of dissatisfaction with how the biggest actors in the LLM landscape are keeping their models closed-source.

## 2.4 Vicuna (2023)

### Base model

In February 2023, Meta released LLaMA, a range of foundation models with sizes ranging from 7B to 65B. It was an application of the scaling laws that emerged from Google DeepMind's paper on Chinchilla (Touvron, 2023). Chinchilla showed that language models trained according to previous scaling laws established by Kaplan, 2020 were undertrained — they could be trained on far more tokens of data and see improvements in capabilities before increasing their parameter count (Hoffman, 2022). As a result, LLaMA was much smaller than SOTA models at the time and previous Meta models like OPT.

The model was available via a request form that was intended to give researchers access. The blog post announcing the release states:

> Access to the model will be granted on a case-by-case basis to academic researchers; those affiliated with organizations in government, civil society, and academia; and industry research laboratories around the world. (Meta, 2023b)

The model was leaked online a week later, and it has since been widely downloaded (Vincent, 2023). A number of fine-tuned variants have since been created and open-sourced, including Stanford's Alpaca and LMSYS's Vicuna. Alpaca was released just two weeks after LLaMA and Vicuna just a week after Alpaca.

Given that these researchers very likely would have received access to LLaMA upon request, these models would have likely been created regardless of whether or not LLaMA was leaked. However, the leak is still significant in that it demonstrates unintentional, unrestricted access to a model as a result of attempted, controlled access.

LLaMA was particularly significant because its small size meant that it required less computing power to run and fine-tune it (so researchers could build upon it quickly) but it still performed quite well compared to other models of similar size.

## Development

Vicuna was fine-tuned using 70k ChatGPT conversations collected through ShareGPT, a website allowing users to publish their conversations with ChatGPT. The training process was similar to that of Alpaca but "enhanced the training scripts [...] to better handle multi-round conversations and long sequences" (LMSYS, 2023).

Vicuna was trained using PyTorch FSDP on 8 A100 GPUs in one day.

## Takeaways

- **Individuals/academic researchers are currently dependent on high-quality pre-trained models.** They do not have access to enough compute to even consider pre-training a model; they can only finetune them. It is worth noting that this wasn’t always the case — the models inspired by GPT-2 showed that it was possible for individuals to obtain access to enough compute to train a near-SOTA model.
- **SOTA models are a source of data for training future models.** A member of LMSYS explained in an interview that they thought the ShareGPT data was what made Vicuna better than other fine-tuned LLaMA variants. Because conversations on ShareGPT have been deliberately shared by the users who generated them, the resulting dataset is already somewhat curated to include interesting data and exclude low quality data.

# 3. Contributing factors

This section covers the contributing factors for open-source development. Analysis of how these factors interact with each other is in Section 4 — this section just seeks to provide an explanation of the factor and how it contributes to open-source development. The factors have been divided into three categories:

1. **Access to inputs** — This determines what type of project is possible for the actor involved.
2. **Access to know-how** — This is required to turn the inputs into a model.
3. **Incentives to open-source** — This determines to what degree the final output is published.

## 3.1 Access to inputs

The three inputs to AI are compute, data, and algorithmic insights (Amodei, 2018).

### Compute

Compute refers to computational power used for training and inference. Compute can be accessed on a user's local device or server, and large amounts of compute are often accessible to users via the cloud.

The cost of the compute required to train modern LLMs is incredibly high. For instance, the training costs of GPT-3 175B and LLaMA 65B were over $1M (2020 USD), PaLM is estimated to have cost over $3M (2020 USD) (Epoch, 2022), and Sam Altman has said that GPT-4 cost over $100M (2023 USD) to train (Knight, 2023).

Because of this, access to compute is typically a product of funding or sponsorship. Big tech is able to funnel profits from other products into buying compute, and top AI labs raise money to buy compute. Nonprofit or volunteer efforts to train LLMs are often reliant on sponsorship of compute, as seen in the inspired models and replications of GPT-2 and the training of BLOOM.

Actors with access to significant amounts of compute may sponsor projects for two reasons. First, compute-intensive projects offer them a chance to stress test their infrastructure, as in the cases of BLOOM (Allen Institute, 2023). Second, for industry actors with proprietary hardware, the software that is written will be optimized for their hardware. This means that future projects are incentivized to continue using the same hardware, which increases usage of that line of hardware in addition to increasing publicity around it (Shevlane, 2022).

### Data

The data that a model is trained on greatly impacts the performance of the final model. Traditionally, larger datasets and data of higher quality have shown to produce favorable results.

The process of building datasets requires a lot of resources and typically consists of multiple steps:

1. **Collection** — When finding data, there are typically two types of corpora to pull from. General corpora consist of data used to build language modeling and generalization capabilities (e.g. Web text, books, etc.). Specialized corpora include data used to improve performance on downstream tasks (e.g. multilingual data, scientific text, code, etc.)
2. **Processing** — The data is cleaned in several steps. First, low quality data is filtered out, typically based on some sort of indicator (though this varies for different datasets). The data is then deduplicated at the sentence, document, and dataset level because repetition in data has been found to lead to lower performance in models. Lastly, personally identifiable information is removed (Zhao, 2023).

Open-source datasets have also been created and released, such as The Pile from EleutherAI and the ROOTs corpus from BigScience.

### Algorithmic insights

Algorithms influence how the other inputs are used to produce the final model. The development of large language models has seen a number of algorithmic insights:

- The introduction of the transformer architecture and self-attention offered a number of benefits over the previous practice of using recurrent neural networks to process text, including that transformers could be parallelized and that they could learn relationships between words (Google Cloud Tech, 2022).
- Experimentation using encoder-only, decoder-only, or encoder-decoder transformer architectures have yielded varying results, though all SOTA models over 100B parameters are decoder-only (BigScience Workshop, 2022).
- Scaling laws have established relationships between the amount of compute and data used to train a model of any given size to maximize performance while using compute efficiently. Initial scaling laws introduced by OpenAI recommended scaling the amount of compute used to train a model, training data, and parameters proportionally (Kaplan, 2020). Google DeepMind's paper on Chinchilla found that this resulted in undertrained models, and that training small models (in terms of parameter size) on more data resulted in more efficient inference (Hoffman, 2022).
- Dense transformers cannot scale forever because inference becomes incredibly expensive — GPT-4 is rumored to have a mixture of experts (MoE) architecture to combat this (Patel, 2023).

## 3.2 Access to know-how

### Talent

Language modeling at scale requires specific skill sets beyond just designing the model. There are two primary challenges to large language modeling. The first is a memory issue — it can be hard to fit parameters in the memory of the GPU. The second is that one can run into very long training times (because of the number of operations that must be carried out) if the algorithms, software, and hardware stack are not optimized properly (Smith, 2022).

As a result, engineers with proper expertise in distributed training, the practice of splitting up a training workload among several nodes of computers, are needed in addition to scientists in order to train a large model. Indeed, in a survey of over 400 AI researchers, 90% of researchers indicated that “specialized knowledge, talent, or skills,” was the most important factor in their most significant project (Musser, 2023).

These skills likely go beyond knowledge of the latest algorithmic innovations because many decisions often have to be made throughout the training process, so building intuition around large language models also seems extremely valuable.

## Publication and documentation

The publication and documentation of models can provide varying levels of insight into how a model was trained.

To borrow some terminology from Ben Cottier’s “Understanding the diffusion of large language models,” an artifact is defined as “a product of AI research or development.” This may include model code, model weights, training data, or documentation of the training process. Diffusion is defined as “when an actor acquires an existing artifact.” Cottier also provides a taxonomy of diffusion mechanisms. Of those, only two are discussed here because the impact of the release of open- and closed-sources fall within them, and providing further distinctions doesn't appear to add more utility.

1. **Open publication** — Before 2020, this was the primary diffusion mechanism in the LLM development landscape. With the release of GPT-3, most of the biggest actors stopped participating with a few exceptions (most notably, Meta). Cottier makes a distinction between this, "theft," and "leaks." While the extent to which information from an actor is unintentionally spread often impacts future release practices, the distinction is unhelpful in the context of open-source development because the implications of open publication, theft, and leaks are all the same.
2. **Incremental research** — This is the result of closed-source published research. Information can be acquired in any of the following ways: First, models are accompanied by technical papers that often disclose implementation details such as the architecture of hyperparameters of a model. Second, the release of SOTA models often provide proof of concept for an idea or strategy that otherwise may have been risky to pursue. Third, closed-source models may still be accessible to the public in some limited and potentially monitored way. For instance, GPT-3 was made available via a public API that was

monitored by OpenAI. While the model was kept under wraps, it was still possible to extract model outputs that were later used as training data for other open-source models.

It is worth noting that there is a difference between the documentation that accompanies open-source releases and the technical papers that accompany most releases but are the primary source of information about a closed-source model. A member of the BLOOM training team explained in an interview that Meta's documentation is particularly useful because it details things that go wrong, attempts to fix them, and in general are much more conducive to replication or further research. Technical papers often include much less granular detail about the process of the model, leaving researchers to try to fill gaps themselves.

### Tooling

Tooling has been extensively used in the training of large language models, particularly to optimize distributed learning. The Megatron-LM transformer framework from NVIDIA and DeepSpeed (deep learning optimization library) from Microsoft were used in BLOOM, GPT-Neo-X, and OPT-175B.

## 3.3 Incentives for and against open-source

Any given actor may have a number of reasons for and against open-sourcing their work and must weigh both sides when determining how to release their work.

Reasons for open-sourcing work:

1. Open-sourcing work establishes **credibility**. Publishing work in full allows the authors to take credit for their contributions, and it allows other researchers to build on top of it, which in turn boosts prestige. This also helps organizations that produce valuable research attract talent and money. This paradigm was inherited from broader research in computer science, where it has been the default release practice.
2. When a base model is open-sourced, the open-source community typically iterates on it extremely quickly. This benefits the creator of the base model, as they have essentially received **free labor** built on their architecture (Google, 2023). This is particularly helpful for actors who may feel like they are behind competitors.
3. Open-sourcing a model may be part of a **business model** — for instance, one in which access to the model is free but the company charges for fine-tuning services. This provides an opportunity for the creator of a model to capture the value of the free labor they've received on it.
4. Some parties believe that open-sourcing models will help **make AI safer** overall. There are a couple of reasons for this.
    a. Interpretability research is reliant on open-sourced models because researchers must have access to the full internals of the model.

b. When open-sourced, models are tested much more rapidly and intensely which leads to quicker discovery and fixing of issues or vulnerabilities.
c. Publishing datasets allows others to check them for issues such as bias.

Reasons against open-sourcing work:

1. Keeping models closed-source and having all usage flow through a central mechanism such as an API **increases safety** (Shevlane, 2022). Open-source models are far more susceptible to being misused because use of the models cannot be monitored.
2. Keeping models closed-source allows companies to maintain a **competitive edge**. By showing off what they have created without allowing others to access the internals or replicate it, companies can stay ahead of competitors while still earning a bit of credibility by publishing a technical report. This is important if models can be monetized, better performance will attract more customers.

# 4. Trends

## 4.1 Clusters of actors

Based on the identified contributing factors, certain clusters of actors emerge:

- Big tech/for-profit AI labs — ex. Anthropic, Google DeepMind, Inflection, OpenAI, StabilityAI
- Nonprofit/volunteer AI labs — ex. Allen Institute for AI, EleutherAI
- Academia/individuals — ex. Large Modeling Systems Organization

The gaps between clusters can roughly be explained by the three categories of factors (access to inputs, access to know-how, and incentives to open-source ). Note that these are generalizations of actors — there are exceptions in each cluster.

### Access to inputs

**Big tech/for-profit AI labs have access to enough inputs to pre-train large models.** These actors can purchase compute and build datasets using profits from other revenue streams or from venture capital.

**Nonprofit/volunteer AI labs have access to inputs to pre-train models.** These actors can often gain access to inputs but are more dependent on external actors. Compute is often sponsored, as in the case of BigScience's BLOOM (sponsored by French supercomputer). In the past, most of the pre-trained models from these actors have been <100B parameters, with larger projects requiring collaboration between actors and thus, more organizational overhead.

**Academia/individuals have limited access to inputs such that they can only fine-tune models.** These actors typically have the least access to inputs. Pre-training a language model

is simply out of the question for most university groups, so many resort to fine-tuning existing models. This means that they do not need to collect as much data since less data is required for fine-tuning a model than pre-training one.

## Access to know-how

**Big tech/for-profit AI labs have talent with expertise to pre-train large models.** These actors can attract top talent with high salaries and benefits. They are likely also in the best position to build talent in-house, as they have the easiest access to inputs and the most senior experience.

**Nonprofit/volunteer AI labs struggle to retain engineering talent to pre-train large models.** These actors often lose talent to industry players because industry players have more resources to compensate employees. For instance, EleutherAI began as a volunteer effort to train and release LLMs. It has produced a number of successful models, datasets, and frameworks, but found that people would use these projects as a chance to upskill and eventually leave to industry positions (Weights and Biases, 2023). The organization has since worked to form a nonprofit so it can hire employees, but a current member explained in an interview that it currently finds itself somewhat limited by salary budgets as it's incredibly hard for nonprofits to compete with the sky-high salaries that industry can offer. A member of the training team on BLOOM in an interview also cited lack of experience as the biggest hindering factor for training the model.

**Academia/individuals lack the experience to pre-train models.** Due to a lack of inputs, these actors typically cannot build up the skill sets required for pre-training large language models. A professor at Brown University said in an interview that they think their students are smart enough to learn to pre-train models, but they simply have never had the chance. However, this assessment refers specifically to the skill sets for training language models at scale — the results of fine-tuned models from academia have been quite successful.

## Incentives for and against open-source

**Big tech/for-profit AI labs have few incentives to open-source their work.** With a couple of exceptions (Meta, StabilityAI), most of these actors choose not to open-source their work. Fears of misuse are most commonly cited as rationale for not publishing openly (OpenAI, Anthropic), though the incentive to maintain competitive advantage is likely at play as well. A primary reason for open-sourcing is that allowing others to build onto models and build application layers over them can be a useful way to even the playing field for actors who are behind (Abboud, 2023).

**Nonprofit/volunteer AI labs have many incentives to open-source their work.** These actors typically open-source their work to build credibility and in turn attract talent and money. Some actors in this category also open-source their work to increase safety.

**Academia/individuals have many incentives to open-source their work.** Publishing research openly is the default in academia, and many of these actors follow that paradigm for purposes of maintaining credibility and allowing others to build on what they have done.

## 4.2 Emerging trends

### Most big tech actors and for-profit AI labs have experienced a paradigm shift from open- to closed-source.

The default paradigm inherited from broader computer science research was that of publishing everything. Because of that, much of the initial research in large language modeling was open-sourced (ex. BERT, GPT).

However, as language models became more capable, a number of reasons to stop open-sourcing models emerged. First, there were concerns that models would be intentionally misused to do harm (this was the primary reason cited for limiting access to models). Second, it became apparent that there is an enormous opportunity for profit, and close-sourcing models helps actors maintain a competitive edge in the field.

This has led to several experiments on how to deploy models. GPT-2 attempted a staged release, and GPT-3 only allowed API access. The latest SOTA models have been similarly restricted — either monitoring use through APIs or not allowing any public access to the models. This attitude towards deployment has also affected publication practices. The GPT-3 paper was delayed and purposefully intended to be dull (Shevlane, 2022), and the GPT-4 paper lacked fewer implementation details relative to previous releases. This keeps in line with OpenAI's charter, which stated in 2018 that it "expect[s] that safety and security concerns will reduce [its] traditional publishing in the future."

Whether open-source development might prove competitive with the closed-source state of the art was hotly debated when a leaked document from Google argued that open-source was rapidly catching up to companies developing closed-source models like Google and OpenAI and that these companies have no advantage (and some clear disadvantages) in the long term. However, as of now, neither Google nor OpenAI seem to have signaled any changes in their stance towards open-source development.

### Meta's releases have a disproportionate impact on open-source progress.

Meta is releasing two types of artifacts that are useful for various actors participating in open-source development:

- High-quality pre-trained models of various sizes that are used by individuals and academia
- Detailed documentation about training said models (especially larger models) that is a source of knowhow for nonprofit/volunteer labs

If Meta were to stop open-sourcing their work, there are other actors who may be able to partially fill the gaps but it would likely slow down the rate of open-source development.

Pre-trained models that have comparable performance to LLaMA have been developed by smaller actors, such as Falcon from UAE's Technology Innovation Institute, CerebrasGPT from Cerebras, or MPT from MosaicML, but often on longer timelines. The lack of documentation for large models may be something that is harder to replace, as it is significantly harder for actors who are not in the "big tech/for-profit AI labs" cluster to train models at that scale in the first place.

This demonstrates the impact of just one actor from "big tech/for-profit AI labs" cluster open-sourcing their work. The sheer amount of resources that this cluster of actors has access to is so much greater than that of any other type of actor, that the impacts of their releases are outsized, should they choose to release them openly. It is worth noting that Meta has even more money and talent than most of the actors in this cluster, though.

It appears likely that there will be actors from this cluster that continue to open-source their work, though they are definitely the minority of the cluster. Meta does not plan to change this open release strategy any time soon, as demonstrated with the release of LLaMA 2 which was released with fewer restrictions than its predecessor (Meta, 2023a). Smaller actors in the cluster like StabilityAI, Cerebras, and MosaicML are also open-sourcing models, and new companies like Mistral AI have said they will join in soon (Lunden, 2023).

## Scaling increases the compute barrier, but open-source actors tend to do more with less.

As models have gotten increasingly bigger, compute has become a prohibitive barrier for certain actors. The inspired models and replications of GPT-2 demonstrated that individuals could secure enough compute (albeit, via sponsorships) to train a model comparable to the SOTA at that time. This is simply not true as of now because it would take far more compute and hands on deck to train models on par with the current SOTA. How scaling plays out in the future will affect what actors with limited resources can do.

However, there are a couple of counterfactors:

- The cost of compute goes down over time. Notably, the cost of compute is going down slower than the demand is increasing (Hobbhahn & Besiroglu, 2022; Sevilla, 2022).

- Open-source is able to achieve competitive results to closed-source with fewer resources (compute). Currently, pre-training models in the 10–100B parameter range has allowed the open-source community to come close to achieving parity with GPT-3.
- If the ultimate concern is misuse of models, the danger arises once the capabilities of models pass a certain threshold, regardless of whether they are SOTA or not (e.g. this bar could be lower than the SOTA at a given point in time). Depending on where that threshold is, compute may not end up being a barrier to developing a dangerous open-source model.

## Model imitation may yield competitive, open versions of closed-source SOTA models.

Model imitation has emerged as a strategy to improve small LLMs by finetuning them on data from larger LLMs. The release of LLaMA spurred a string of these models. Vicuna was the strongest of the LLaMA iterations, boasting performance at 92% of ChatGPT quality, as evaluated by GPT-4. However, it was soon critiqued by Gudibande et al. (2023) which argued that model imitation is only able to mimic the style but not the reasoning abilities of larger models.

Microsoft trained Orca in an attempt to use explanation tuning to teach the model reasoning skills. It surpassed Vicuna's performance in Vicuna's original evaluation setting, in addition to outperforming it on benchmarks like Big Bench Hard. While it still lags significantly behind GPT-4, Orca indicates that explanation tuning may be a promising avenue of research (Mukherjee, 2023).

This further supports the premise that creating open-source models that have competitive performance will be possible with little compute and limited access to a SOTA or near-SOTA model.

Notably, these imitation models typically inherit the weaknesses of the teacher model (Mukherjee, 2023). Having an open-source version of a strong model may provide opportunities for further research, both good and bad. For instance, a transferable jailbreak using an adversarial suffix was generated using Vicuna, and it transferred better to GPT-based models (relative to other models, such as Bard or Claude v2) because Vicuna learned from ChatGPT (Zou, 2023).

## The EU AI Act suggests how future regulation may address open-source development.

The details of the European Union's AI Act are still being hammered out, and its approach to regulating open-source systems has been hotly debated. As is typical in the EU's legislative

process, there are currently three drafts of the act (each from different parties) and the final draft will be decided through a three-way negotiation (Lynch, 2023).

The latest draft from Parliament spares providers of open-source systems, instead imposing requirements on those using the open-source system for commercial purposes. This aims to be more precise than the previous two drafts from the European Commission and the Council of the European Union, which were more vague about the status of open-source and the limits of non-commercial use. However, Parliament does not include foundation models (a category that would include LLMs) in their open-source exemption, and the Council's draft includes open-source models in their definition of "General Purpose AI," (also intended to include LLMs) subjecting them to the requirements (AI Act, 2022).

While the Act will not be finalized until the end of 2023, this discussion reveals potential frameworks under which open-source may be regulated and the stakes at hand for any given actor. For instance, GitHub and HuggingFace have pushed for a number of changes that would better support the open-source landscape in Europe (David, 2023), and France has funded open-source projects in hopes that open research will help boost progress in establishing its own national AI industry (Chatterjee, 2023).

Notably, the EU AI Act is the first attempt at binding regulation over AI, so the tone it sets will likely influence future legislation from other regulatory bodies.

# 5. Conclusion

Examining a number of case studies of open-source large language model development provides a lens through which to analyze the factors that contribute to progress in the field and the trends that may shape the future trajectory of the field.

A primary concern of open-source models is their susceptibility to misuse. The threat of misuse posed by any given model arises from its capabilities in an absolute sense, which must be beyond a given threshold to equip a malicious actor to do harm (e.g. whether the model is considered SOTA is irrelevant). While a lack of compute limits much of open-source work and has become a bigger barrier over time, it is unclear whether it will prevent models from surpassing this threshold.

The case studies reveal that actors with the most resources — big tech and for-profit AI labs — are shifting away from open-source, citing concerns of misuse. Consequently, most actors in the open-source space tend to be strapped for inputs like compute and know-how like talent, and they benefit from wealthier actors providing artifacts like documentation or access to SOTA models. This suggests that it would be quite hard for open-source capabilities to consistently match or surpass closed-source capabilities.

Instead, open-source efforts trail behind closed-source models, continuing to improve steadily. It appears that there will continue to be some well-resourced actors open-sourcing their work, as Meta shows no sign of stopping and incentives exist for other actors to take on this role. These actors significantly empower the rest of the open-source ecosystem, indicating that this dynamic between open- and closed-source development seems likely to continue. As closed-source models develop potentially dangerous capabilities, open-source models may be close behind.

However, many aspects remain uncertain. As compute demands increase, it is unclear whether open-source efforts can keep up, especially given that business models that attempt to capture the value of open-source models have not been extensively tested. Further research on the robustness of these business models and how they appeal to companies may provide insight into their viability. The impact of regulation from government actors is also unclear, though it appears the EU AI Act may not hit open-source too hard. Moreover, whether shortcuts like model imitation prove to be successful will also influence how open-source models attempt to compete with closed-source models. Given that the links between imitation and teacher models may pose security hazards, further research is necessary to anticipate how this method in particular will be used.

# 6. Acknowledgements

I am immensely grateful to my mentor, Hadrien Pouget, for his invaluable feedback and guidance throughout the writing of this report. I'm also grateful to Zachary Rudolph for organizing the research fellowship through which this report was written.

# 7. References

Abboud, L., Murgia, M., Murphy, H., & Criddle, C. (2023, July 13). *Meta to release commercial AI model in effort to catch rivals*. @FinancialTimes; Financial Times. https://www.ft.com/content/01fd640e-0c6b-4542-b82b-20afb203f271

AI Act. (n.d.). Documents. The Artificial Intelligence Act. https://artificialintelligenceact.eu/documents/

Allen Institute. (2023). The BigScience Workshop [YouTube Video]. In *YouTube*. https://www.youtube.com/watch?v=XFSJ30ZaMW4

Amodei, D., & Hernandez, D. (2018, May 16). *AI and compute*. Openai.com. https://openai.com/research/ai-and-compute

BigScience Workshop. (2022). *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*. https://arxiv.org/pdf/2211.05100.pdf

Casper, S., Christoffersen, P., & Yew, R.-J. (2022, November 5). *The Slippery Slope from DALLE-2 to Deepfake Anarchy*. Effectivealtruism.org. https://forum.effectivealtruism.org/posts/Bnp9YDqErNXHmTvvE/

Chatterjee, M., & Volpicelli, G. (2023, August 4). *France bets big on open-source AI*. POLITICO; POLITICO. https://www.politico.eu/article/open-source-artificial-intelligence-france-bets-big/

Cohen, V. (2019, August 22). *OpenGPT-2: We Replicated GPT-2 Because You Can Too - Vanya Cohen - Medium*. Medium; Medium. https://medium.com/@vanya_cohen/opengpt-2-we-replicated-gpt-2-because-you-can-too-45e34e6d36dc

Cottier, B. (2022, December 21). Understanding the diffusion of large language models: summary. Effectivealtruism.org. https://forum.effectivealtruism.org/posts/nc3JFZbqnzWWAPkmz/understanding-the-diffusion-of-large-language-models-summary-1

David, E. (2023, July 26). *GitHub and others call for more open-source support in EU AI law*. The Verge; The Verge. https://www.theverge.com/2023/7/26/23807218/github-ai-open-source-creative-commons-hugging-face-eu-regulations

Epoch. (2022). *Parameter, Compute and Data Trends in Machine Learning*. Epochai.org. https://epochai.org/data/pcd

Google Cloud Tech. (2021). Transformers, explained: Understand the model behind GPT, BERT, and T5 [YouTube Video]. In *YouTube*. https://www.youtube.com/watch?v=SZorAJ4I-sA

Google (2023, May 4). *Google "We Have No Moat, And Neither Does OpenAI."* Semianalysis.com; SemiAnalysis. https://www.semianalysis.com/p/google-we-have-no-moat-and-neither

Gudibande, A., Wallace, E., Snell, C., Geng, X., Liu, H., Abbeel, P., Levine, S., & Song, D. (2023). *The False Promise of Imitating Proprietary LLMs*. https://arxiv.org/pdf/2305.15717.pdf

Hobbhahn, M., & Besiroglu, T. (2022, June 27). Trends in GPU price-performance. Epoch. https://epochai.org/blog/trends-in-gpu-price-performance?r

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., De, D., Casas, L., Hendricks, L., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Van Den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., & Elsen, E. (2022). *Training Compute-Optimal Large Language Models*. https://arxiv.org/pdf/2203.15556.pdf

Hugging Face. (2023). *Perplexity of fixed-length models*. Huggingface.co. https://huggingface.co/docs/transformers/perplexity

Kaplan, J., Mccandlish, S., Openai, Openai, T., Openai, B., Openai, R., Openai, S., Openai, A., Openai, J., & Openai, D. (2020). *Scaling Laws for Neural Language Models*. https://arxiv.org/pdf/2001.08361.pdf

Knight, W. (2023, April 17). *OpenAI's CEO Says the Age of Giant AI Models Is Already Over*. WIRED; WIRED. https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/

Leahy, C. (2019a, June 6). *Replicating GPT2–1.5B - Connor Leahy - Medium*. Medium; Medium. https://medium.com/@NPCollapse/replicating-gpt2-1-5b-86454a7f26af

Leahy, C. (2019b, June 13). *Addendum: Evaluation of My Model - Connor Leahy - Medium*. Medium; Medium. https://medium.com/@NPCollapse/addendum-evaluation-of-my-model-e6734b51a830

LMSYS. (2023). *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality | LMSYS Org*. Lmsys.org. https://lmsys.org/blog/2023-03-30-vicuna/

Lunden, I. (2023, June 13). France's Mistral AI blows in with a $113M seed round at a $260M valuation to take on OpenAI | TechCrunch. TechCrunch. https://techcrunch.com/2023/06/13/frances-mistral-ai-blows-in-with-a-113m-seed-round-at-a-260m-valuation-to-take-on-openai/

Lynch, S. (2023). *Analyzing the European Union AI Act: What Works, What Needs Improvement*. Stanford HAI; Stanford University. https://hai.stanford.edu/news/analyzing-european-union-ai-act-what-works-what-needs-improvement

Meta. (2023a). Llama 2 - Meta AI. Meta.com. https://ai.meta.com/llama/

Meta. (2023b, February 24). *Introducing LLaMA: A foundational, 65-billion-parameter language model*. Meta.com. https://ai.meta.com/blog/large-language-model-llama-meta-ai/

Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H., & Awadallah, A. (2023). *Orca: Progressive Learning from Complex Explanation Traces of GPT-4*. https://arxiv.org/pdf/2306.02707.pdf

Musser, M., Gelles, R., Kinoshita, R., Aiken, C., & Lohn, A. (2023, June 5). *"The Main Resource is the Human" - Center for Security and Emerging Technology*. Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/the-main-resource-is-the-human/

OpenAI. (2019a). *Better language models and their implications*. Openai.com. https://openai.com/research/better-language-models#update

OpenAI. (2019b). *GPT-2: 1.5B release*. Openai.com. https://openai.com/research/gpt-2-1-5b-release

OpenAI. (2019c). *GPT-2: 6-month follow-up*. Openai.com. https://openai.com/research/gpt-2-6-month-follow-up

Patel, D., & Wong, G. (2023, July 10). *GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE*. Semianalysis.com; SemiAnalysis. https://www.semianalysis.com/p/gpt-4-architecture-infrastructure

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Sevilla, J. (2022). *Estimating Training Compute of Deep Learning Models*. Epochai.org. https://epochai.org/blog/estimating-training-compute

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022). COMPUTE TRENDS ACROSS THREE ERAS OF MACHINE LEARNING. https://arxiv.org/pdf/2202.05924.pdf

Shevlane, T. (2019). *The Artefacts of Intelligence: Governing Scientists' Contribution to AI Proliferation | GovAI*. Governance.ai. https://www.governance.ai/research-paper/the-artefacts-of-intelligence-governing-scientists-contribution-to-ai-proliferation

Smith, S., Patwary, M., Norick, B., Legresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., Zhang, E., Child, R., Aminabadi, R., Bernauer, J., Song, X., Shoeybi, M., He, Y., Houston, M., Tiwary, S., & Catanzaro, B.

(2022). *Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model*. https://arxiv.org/pdf/2201.11990.pdf

Solaiman, I., Brundage, M., Jack, O., Openai, C., Openai, A., Herbert-Voss, A., Openai, J., Openai, A., Openai, G., Wook, J., Openai, K., Kreps, S., Politiwatch, M., Newhouse, A., Blazakis, J., Mcguffie, K., & Openai, J. (2019). *OpenAI Report Release Strategies and the Social Impacts of Language Models*. https://arxiv.org/pdf/1908.09203.pdf

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. ArXiv.org. https://arxiv.org/abs/2302.13971

Urbina, F., Filippa Lentzos, Cédric Invernizzi, & Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, *4*(3), 189–191. https://doi.org/10.1038/s42256-022-00465-9

Vincent, J. (2023, March 8). *Meta's powerful AI language model has leaked online — what happens now?* The Verge; The Verge. https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse

Weights & Biases. (2023). How EleutherAI Trains and Releases LLMs: Interview with Stella Biderman [YouTube Video]. In *YouTube*. https://www.youtube.com/watch?v=aH1IRef9qAY

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., Choi, Y., & Allen, P. (2019). *Defending Against Neural Fake News*. https://arxiv.org/pdf/1905.12616.pdf

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, S., Sridhar, A., Wang, T., Zettlemoyer, L., & Ai, M. (2022). *OPT: Open Pre-trained Transformer Language Models*. https://arxiv.org/pdf/2205.01068.pdf

Zhao, W., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., & Liu, P. (2023). *A Survey of Large Language Models*. https://arxiv.org/pdf/2303.18223.pdf

Zou, A., Wang, Z., Kolter, J., & Fredrikson, M. (2023). *Universal and Transferable Adversarial Attacks on Aligned Language Models*. https://arxiv.org/pdf/2307.15043.pdf