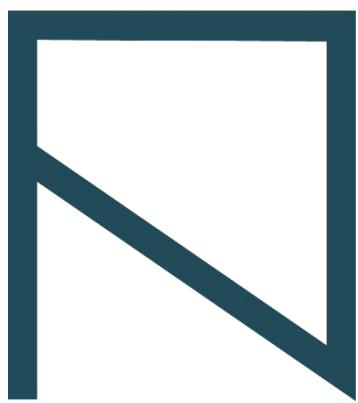


Report on
AI model providers' training data transparency
and
enforcement of copyrights



**RIGHTS
ALLIANCE**

FOR THE CREATIVE INDUSTRIES
ON THE INTERNET

September 5, 2024

Thomas Heldrup

Head of Content Protection & Enforcement / Danish Rights Alliance

Table of contents

<i>Introduction</i>	- 3 -
<i>Summary of findings</i>	- 6 -
<i>Text generating or multimodal AI models</i>	- 8 -
Case 1: Meta’s Llama models	- 8 -
Case 2: Mistral AI	- 14 -
Case 3: Google Gemini and Gemma	- 16 -
Case 4: OpenAI’s GPT	- 20 -
Case 5: Microsoft’s Phi	- 25 -
Case 6: Eleuther AI’s GPT-NeoX-20B	- 33 -
Case 7: Anthropic’s Claude	- 38 -
<i>Music generating AI models</i>	- 40 -
Case 8: Suno AI’s Suno	- 40 -
Case 9: Uncharted Labs’ Udio	- 42 -
<i>Video generating AI models</i>	- 43 -
Case 10: OpenAI’s Sora	- 43 -
Case 11: Runway AI’s Gen3-alpha	- 44 -
<i>Image generating AI models</i>	- 45 -
Case 12: Stability AI’s Stable Diffusion	- 45 -
Case 13: Black Forest Labs’ Flux.1	- 49 -

Introduction

Transparency is crucial

In this report we have collected our research on how providers of general purpose AI models are transparent about the content used for training their models. The aim of this report is to show the crucial nature of a transparency obligation on model providers and what degree of transparency is required for rightsholders to exercise and enforce their rights.

If a copyright holder wants to hold a provider of a general-purpose AI model accountable for the unauthorized copying that can take place during the collection, training (pre-training and finetuning) and deployment of such an AI model, the rightsholder must according to current European and Danish law document or render probable that content belonging to the rightsholder has been part of the model's training data. This can be achieved when:

The AI model provider is **transparent** about:

- **What** content was included in the training data (e.g. title and copyright information).
- **Where** the content was initially collected from (e.g. URL or name of platform/service).
- **When** the content was initially collected (i.e. date and time).

There are cases pending in the USA where rightsholders have been able to **generate output** with AI models that **closely resembles or is identical to content owned by the rightsholders** whereby they argue the output cannot have been generated without the AI model being trained on the content.¹

¹ The New York Times Company v. Microsoft Corporation (1:23-cv-11195)
Concord Music Group, Inc. v. Anthropic PBC (3:23-cv-01092)
UMG Recordings, Inc. v. Suno, Inc. (1:24-cv-11611)
UMG Recordings, Inc. v. Uncharted Labs, Inc. (1:24-cv-04777)

In most cases it is extremely cumbersome if not impossible to generate output that is identical or closely resembles copyright protected content in an AI model's training data. This may be a result of the AI model provider:

1. having implemented "**guardrails**" that limit what the model can generate, and/or
2. because the copyright protected **content is only represented a single or few times in the training data**. AI models will usually only generate exact copies ("memorization") or output that closely resembles copyright protected content in the training data if the copyright protected content was represented many times in the training data.

This leaves transparency as the crucial factor for rightsholders if they want to:

- (1) **discover** unauthorized copying of their content,
 - (2) **license** their content for AI training
- or
- (3) **initiate enforcement activities** based on unauthorized copying of their content in AI training.

Scope of the report

In this report we describe the transparency offered by the model providers, but we also include additional information about the training data that was provided e.g. in public court documents or internal documents leaked to the press. By combining what the providers disclose willingly and unwillingly we get a better understanding of the true scope of the training data used.

We look at text, music, video and image generating models. Note this is not an exhaustive list of models and providers.

Comment on transparency and the EU AI Act Art. 53 transparency obligation

We comment on the current level of transparency offered by the model providers and assess if this enables rightsholders to exercise and enforce their rights.

This assessment is closely tied to the transparency obligations in the EU AI Act Art. 53 (1)(d) whereby providers of general-purpose AI models have to provide a sufficiently detailed summary about content used as training data. As noted in the AI Act preamble 107 **the summary must facilitate rightsholders to exercise and enforce their rights.**

Article 53 1. Litra d:

Providers of general-purpose AI models shall: Draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office.

Preamble (107):

*In order to increase transparency on the data that is used in the pre-training and training of general-purpose AI models, including text and data protected by copyright law, it is adequate that providers of such models draw up and make publicly available a sufficiently detailed summary of the content used for training the general-purpose AI model. While taking into due account the need to protect trade secrets and confidential business information, **this summary should be generally comprehensive in its scope instead of technically detailed to facilitate parties with legitimate interests, including copyright holders, to exercise and enforce their rights under Union law, for example by listing the main data collections or sets that went into training the model, such as large private or public databases or data archives, and by providing a narrative explanation about other data sources used.** It is appropriate for the AI Office to provide a template for the summary, which should be simple, effective, and allow the provider to provide the required summary in narrative form. (our own highlight)*

Summary of findings

In order for rightsholders to exercise and enforce their rights against providers of general-purpose AI models the rightsholders must be able to determine whether or not their content was in the model in question's training data. For this to happen it is our experience that the following information needs to be made available by the model provider:

1. A **list of datasets** used, including:
 1. **Name of the dataset**
 - If the dataset is publicly available then the same name must be used as the original dataset.
2. A **narrative explanation of all datasets** used, including:
 1. **Content type(s)** present in the dataset (e.g. books, music, webpages),
 2. **Reference to publicly available research papers or third-party reports** describing the contents of the dataset in detail (it must be validated to be properly describing the data and the provider must keep a copy of papers/reports in the case it gets removed from the internet),
 3. **What period the dataset covers** (only required if data source has many datasets scraped over a period of time e.g. Common Crawl. Here it won't be sufficient to only name the dataset without also mentioning period e.g. Common Crawl August 2024 Crawl Archive (CC-MAIN-2024-33)),
 4. **When the datasets were collected** by the provider,
 5. If **synthetic datasets** were used:
 - **What model** was used to generate the dataset,
 - If the provider of the generating model has not been transparent on its training data then a **narrative explanation** similar to the above 2.1-2.4 of the datasets used to train the generating model must be provided.

Our report shows that all state-of-the-art model providers refrain from providing the necessary degree of transparency that would allow rightsholders to exercise and enforce their rights. Instead the model providers use phrases like “*publicly available content*” collected from the “*open web*” to describe their training data. Sometimes the model providers do not mention training data at all.

We show how some model providers have previously provided degrees of transparency that potentially could lead rightsholders to determine if their copyright protected content was used to train the models, but that the same providers have now reduced their level of transparency so rightsholders cannot exercise and enforce their rights anymore.

We also describe how some model providers who use terms like “open source” to describe their AI models are not in fact open regarding the content used to train their models.

Text generating or multimodal AI models

Case 1: Meta's Llama models

Transparency provided:

In February 2023 Meta published a research paper *LLaMA: Open and Efficient Foundation Language Models*² detailing the training data used to train a collection of language models they call "Llama". This is the first version of the general purpose Llama model with subsequent releases of Llama 2 and 3 described below.

Meta describes its Llama models as "open source AI".

In the **Llama 1** paper's section "2.1 Pre-training Data" (see the following screenshot) Meta lists the datasets that went into the training data and they provide a narrative explanation of each dataset. The list includes a *title* e.g. "Github", "Wikipedia", and "Books3".

The narrative explanation includes but not always the:

- *Content type* e.g. "books".
- *Reference to research papers* describing the individual datasets in more detail e.g. "Gao et al., 2020"³ describing the dataset "Books3".
- *What period the datasets cover* e.g. "[...] CommonCrawl dumps ranging from 2017 to 2020 [...]" or "[...] Wikipedia dumps from the June-August 2022 period [...]".

² <https://arxiv.org/abs/2302.13971>

³ <https://arxiv.org/abs/2101.00027>

2 Approach

Our training approach is similar to the methods described in previous work (Brown et al., 2020; Chowdhery et al., 2022), and is inspired by the Chinchilla scaling laws (Hoffmann et al., 2022). We train large transformers on a large quantity of textual data using a standard optimizer.

2.1 Pre-training Data

Our training dataset is a mixture of several sources, reported in Table 1, that cover a diverse set of domains. For the most part, we reuse data sources that have been leveraged to train other LLMs, with the restriction of only using data that is publicly available, and compatible with open sourcing. This leads to the following mixture of data and the percentage they represent in the training set:

English CommonCrawl [67%]. We preprocess five CommonCrawl dumps, ranging from 2017 to 2020, with the CCNet pipeline (Wenzek et al., 2020). This process deduplicates the data at the line level, performs language identification with a fastText linear classifier to remove non-English pages and filters low quality content with an n-gram language model. In addition, we trained a linear model to classify pages used as references in Wikipedia v.s. randomly sampled pages, and discarded pages not classified as references.

C4 [15%]. During exploratory experiments, we observed that using diverse pre-processed CommonCrawl datasets improves performance. We thus included the publicly available C4 dataset (Raffel et al., 2020) in our data. The preprocessing of C4 also contains deduplication and language identification steps: the main difference with CCNet is the quality filtering, which mostly relies on heuristics such as presence of punctuation marks or the number of words and sentences in a webpage.

Github [4.5%]. We use the public GitHub dataset available on Google BigQuery. We only kept projects that are distributed under the Apache, BSD and MIT licenses. Additionally, we filtered low quality files with heuristics based on the line length or proportion of alphanumeric characters, and removed boilerplate, such as headers, with regular expressions. Finally, we deduplicate the resulting dataset at the file level, with exact matches.

Wikipedia [4.5%]. We add Wikipedia dumps from the June-August 2022 period, covering 20

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

Table 1: **Pre-training data.** Data mixtures used for pre-training, for each subset we list the sampling proportion, number of epochs performed on the subset when training on 1.4T tokens, and disk size. The pre-training runs on 1T tokens have the same sampling proportion.

languages, which use either the Latin or Cyrillic scripts: bg, ca, cs, da, de, en, es, fr, hr, hu, it, nl, pl, pt, ro, ru, sl, sr, sv, uk. We process the data to remove hyperlinks, comments and other formatting boilerplate.

Gutenberg and Books3 [4.5%]. We include two book corpora in our training dataset: the Gutenberg Project, which contains books that are in the public domain, and the Books3 section of ThePile (Gao et al., 2020), a publicly available dataset for training large language models. We perform deduplication at the book level, removing books with more than 90% content overlap.

ArXiv [2.5%]. We process arXiv Latex files to add scientific data to our dataset. Following Lewkowycz et al. (2022), we removed everything before the first section, as well as the bibliography. We also removed the comments from the .tex files, and inline-expanded definitions and macros written by users to increase consistency across papers.

Stack Exchange [2%]. We include a dump of Stack Exchange, a website of high quality questions and answers that covers a diverse set of domains, ranging from computer science to chemistry. We kept the data from the 28 largest websites, removed the HTML tags from text and sorted the answers by score (from highest to lowest).

Tokenizer. We tokenize the data with the byte-pair encoding (BPE) algorithm (Sennrich et al., 2015), using the implementation from SentencePiece (Kudo and Richardson, 2018). Notably, we split all numbers into individual digits, and fallback to bytes to decompose unknown UTF-8 characters.

Screenshot 1 - LLaMA: Open and Efficient Foundation Language Models section 2.1

In July 2023 Meta published a research paper *Llama 2: Open Foundation and Fine-Tuned Chat Models*⁴. In the **Llama 2** paper's section "2.1 Pretraining Data" the training data is summarised as follows:

"Our training corpus includes a new mix of data from publicly available sources, which does not include data from Meta's products or services. We made an effort to remove data from certain sites known to contain a high volume of personal information about private individuals. We trained on 2 trillion tokens of data as this provides a good performance–cost trade-off, up-sampling the most factual sources in an effort to increase knowledge and dampen hallucinations."

2 Pretraining

To create the new family of LLAMA 2 models, we began with the pretraining approach described in Touvron et al. (2023), using an optimized auto-regressive transformer, but made several changes to improve performance. Specifically, we performed more robust data cleaning, updated our data mixes, trained on 40% more total tokens, doubled the context length, and used grouped-query attention (GQA) to improve inference scalability for our larger models. Table 1 compares the attributes of the new LLAMA 2 models with the LLAMA 1 models.

2.1 Pretraining Data

Our training corpus includes a new mix of data from publicly available sources, which does not include data from Meta's products or services. We made an effort to remove data from certain sites known to contain a high volume of personal information about private individuals. We trained on 2 trillion tokens of data as this provides a good performance–cost trade-off, up-sampling the most factual sources in an effort to increase knowledge and dampen hallucinations.

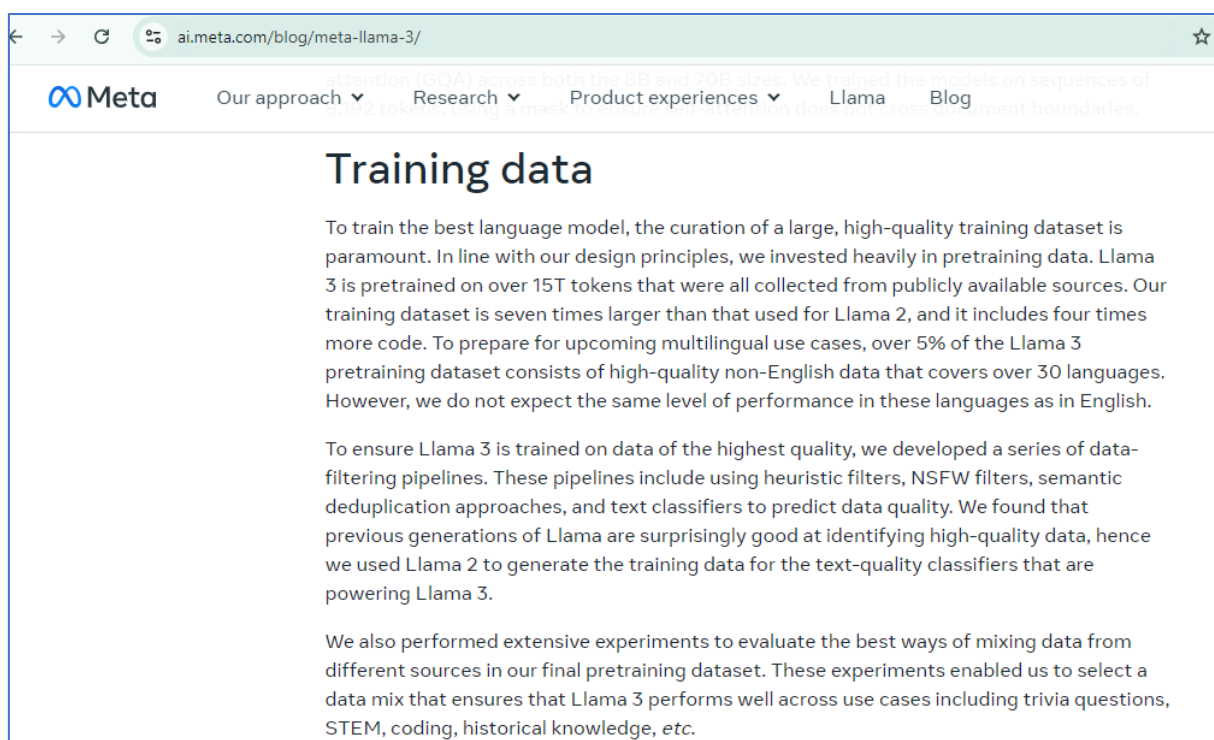
We performed a variety of pretraining data investigations so that users can better understand the potential capabilities and limitations of our models; results can be found in Section 4.1.

Screenshot 2 - Llama 2: Open Foundation and Fine-Tuned Chat Models section 2.1

⁴ <https://arxiv.org/abs/2307.09288>

In April 2024 Meta introduced **Llama 3** with a blogpost on their website⁵. Under the title “Training data” the training data is summarised as follows:

“To train the best language model, the curation of a large, high-quality training dataset is paramount. In line with our design principles, we invested heavily in pretraining data. Llama 3 is pretrained on over 15T tokens that were all collected from publicly available sources. Our training dataset is seven times larger than that used for Llama 2, and it includes four times more code. To prepare for upcoming multilingual use cases, over 5% of the Llama 3 pretraining dataset consists of high-quality non-English data that covers over 30 languages. However, we do not expect the same level of performance in these languages as in English.”



Screenshot 3 – Llama 3 blogpost “Training data” section on <https://ai.meta.com/blog/meta-llama-3/> (screenshot taken AUG 21, 2024)

⁵ <https://ai.meta.com/blog/meta-llama-3/>

Additional information about Meta's training data:

A class action lawsuit has been filed against Meta in the United States District Court of California by authors Richard Kadrey et.al.

In its answer to first consolidated amended complaint Meta admits in section 69: "Meta admits that portions of Books3, among many other materials, were used as training data for Llama 2 prior to its public release in July 2023."⁶

15	69. Meta admits that portions of Books3, among many other materials, were used as	
16	training data for Llama 2 prior to its public release in July 2023. Except as expressly admitted,	
17	Meta denies the allegations in paragraph 69.	

Screenshot 4 -

<https://storage.courtlistener.com/recap/gov.uscourts.cand.415175/gov.uscourts.cand.415175.72.0.pdf>

Comment on Meta's transparency:

Llama 1: While Meta does not provide a list of the specific content used to train its LLaMA model i.e. the *title and copyright information (e.g. publisher)* of the books in the "Books3" dataset, Meta's summary does list dataset *titles* and a narrative explanation of the datasets. Notably in the case of the "Books3" dataset there is also a *reference to another research paper* that allows for further research to be undertaken that may reveal the content used to train the model.

Indeed, in looking at the cited paper we were able to determine the source of the books in the "Books3" dataset as the "*Bibliotik private tracker made available by Shawn Presser (Presser, 2020)*"⁷. By following this citation we were able to locate a

⁶

<https://storage.courtlistener.com/recap/gov.uscourts.cand.415175/gov.uscourts.cand.415175.72.0.pdf>

⁷ <https://arxiv.org/abs/2101.00027>

copy of the dataset “Books3”⁸, thus giving us a complete list of titles and also a copy of the books in “Books3”. This in turn enabled us to scan and locate books belonging to Danish publishers and authors that we represent.

After having located illegal copies of books we represent that were present in “Books3” we were then able to initiate enforcement activities such as sending takedown notices to hosting providers of “Books3” namely to the sites The-Eye.eu and Huggingface.com. Meta never replied to our notice.

It should be noted that our successful takedown notice rendered other rightsholders unable to determine if their rights had been infringed as the dataset is no longer publicly available.

To conclude, the limited transparency provided by Meta enabled us to gather information from additional third-party sources and thereby determine that unauthorized copies of copyright protected content had indeed been used to train Meta’s AI model. The challenging task in relying on a list with titles and a narrative explanation containing reference to third-parties with the necessary information on content is that these datasets may be removed from the web in the future. This is the case with the Books3 dataset where we were able to determine copyright infringement, but other rightsholders cannot as the dataset is no longer available to download from the original host.

Llama 2 & 3: Meta makes no effort to be transparent about the content used to train its Llama 2 and Llama 3 models besides saying it originates from “*publicly available sources*”. This level of transparency cannot be used to determine whether copyright protected content has been used to train the AI models or exercise and enforce rights.

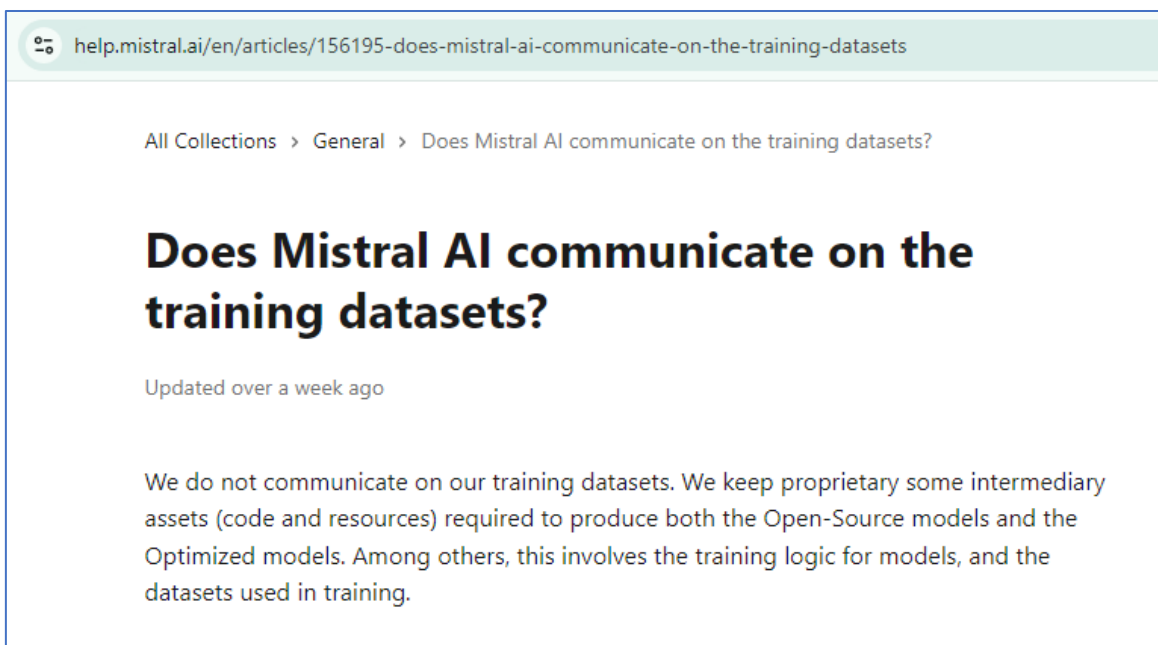
⁸ <https://x.com/theshawwn/status/1320282149329784833?lang=en>

Case 2: Mistral AI

Transparency provided:

Mistral AI currently provides 2 general purpose models: “Mistral Nemo” developed in collaboration with NVIDIA and “Mistral Large 2”. Mistral uses “open source” “open” and “open-weight” to describe its models.

Mistral AI has publicly stated that it does not intend to communicate on training datasets according to an article found in their “*Help Center and Frequently Asked Questions*”⁹.

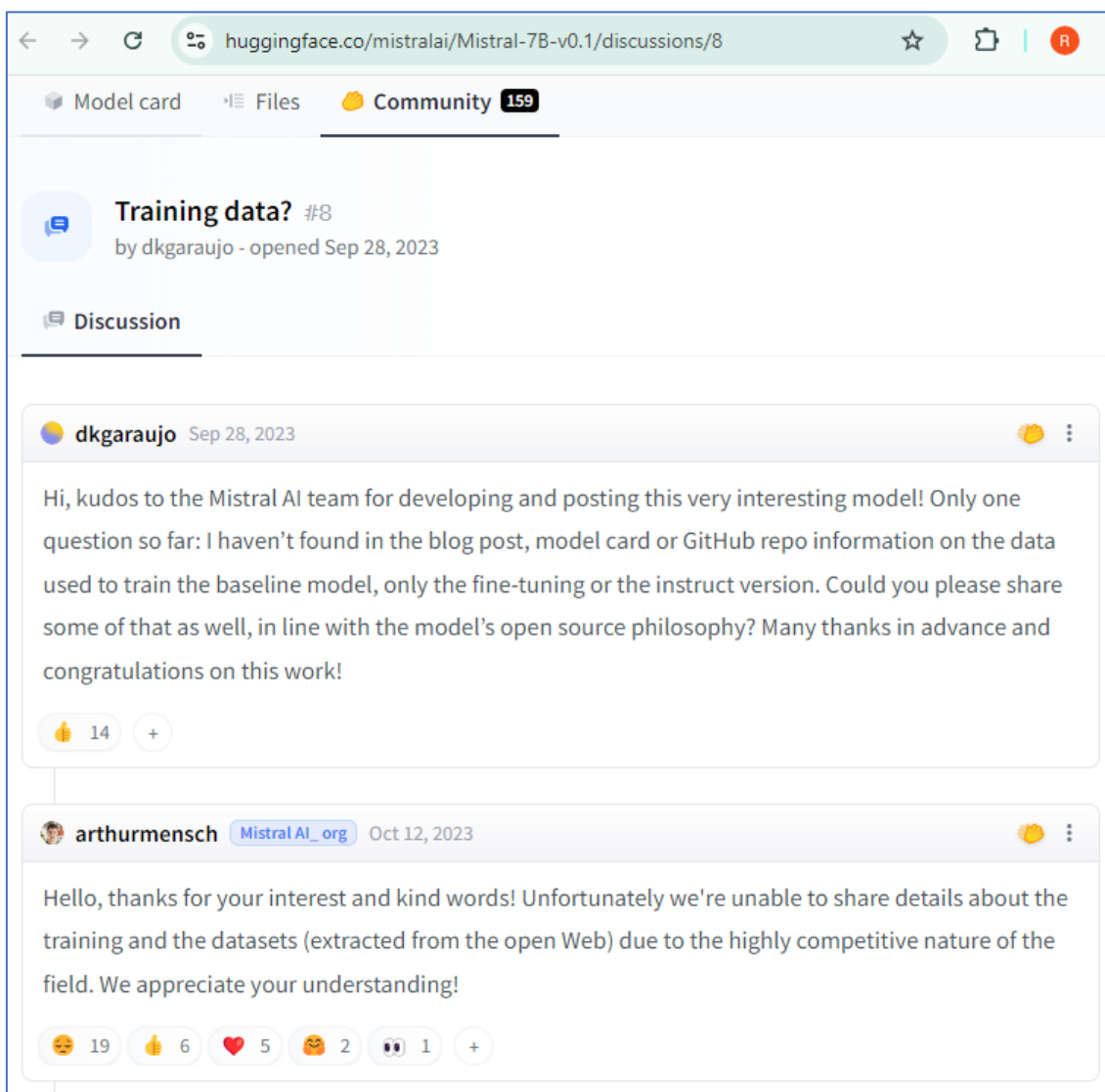


The screenshot shows a web page from help.mistral.ai. The breadcrumb trail is "All Collections > General > Does Mistral AI communicate on the training datasets?". The main heading is "Does Mistral AI communicate on the training datasets?". Below the heading, it says "Updated over a week ago". The main text reads: "We do not communicate on our training datasets. We keep proprietary some intermediary assets (code and resources) required to produce both the Open-Source models and the Optimized models. Among others, this involves the training logic for models, and the datasets used in training."

Screenshot 5 - <https://help.mistral.ai/en/articles/156195-does-mistral-ai-communicate-on-the-training-datasets> (screenshot taken AUG 21, 2024)

Meanwhile, Mistral AI CEO Arthur Mensch (HuggingFace.com username “*arthurmensch*”) replied to a question on Hugging Face about the training data for Mistral-7B that they had “*extracted [training datasets] from the open Web*”.

⁹ <https://help.mistral.ai/en/articles/156195-does-mistral-ai-communicate-on-the-training-datasets>



Screenshot 6 - <https://huggingface.co/mistralai/Mistral-7B-v0.1/discussions/8>

Comment on Mistral AI's transparency:

This level of transparency cannot be used to determine whether copyright protected content has been used to train the AI models or exercise and enforce rights.

Case 3: Google Gemini and Gemma

Transparency provided:

The general purpose AI model **Gemini 1.0** was released by Google in December 2023 together with a technical report *Gemini: A Family of Highly Capable Multimodal Models*¹⁰.

In the Gemini 1.0 report section 4. “Pre-Training Dataset” Google provides the following information about the training data:

“Gemini models are trained on a dataset that is both multimodal and multilingual. Our pre-training dataset uses data from web documents, books, and code, and includes image, audio, and video data.”

4. Pre-Training Dataset

Gemini models are trained on a dataset that is both multimodal and multilingual. Our pre-training dataset uses data from web documents, books, and code, and includes image, audio, and video data.

We use the SentencePiece tokenizer (Kudo and Richardson, 2018) and find that training the tokenizer on a large sample of the entire training corpus improves the inferred vocabulary and subsequently improves model performance. For example, we find Gemini models can efficiently

tokenize non-Latin scripts which can, in turn, benefit model quality as well as training and inference speed.

The number of tokens used to train the largest models were determined following the approach in Hoffmann et al. (2022). The smaller models are trained for significantly more tokens to improve performance for a given inference budget, similar to the approach advocated in Touvron et al. (2023a).

We apply quality filters to all datasets, using both heuristic rules and model-based classifiers. We also perform safety filtering to remove harmful content based on our policies. To maintain the integrity of evaluations, we search for and remove any evaluation data that may have been in our training corpus before using data for training. The final data mixtures and weights were determined through ablations on smaller models. We stage training to alter the mixture composition during training – increasing the weight of domain-relevant data towards the end of training. We find that data quality is an important factor for highly-performing models, and believe that many interesting questions remain around finding the optimal dataset distribution for pre-training.

Screenshot 7 - <https://arxiv.org/pdf/2312.11805>

¹⁰ <https://arxiv.org/abs/2312.11805>

Gemini 1.5 was released March 2024 together with a technical report *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*¹¹.

In the Gemini 1.5 report section 4. "Training Infrastructure and Dataset" Google provides the following information about the training data:

*"Like Gemini 1.0 series, Gemini 1.5 models are trained on multiple 4096-chip pods of Google's TPUv4 accelerators, distributed across multiple datacenters, and on a variety of multimodal and multilingual data. **Our pre-training dataset includes data sourced across many different domains, including web documents and code, and incorporates image, audio, and video content.** For the instruction-tuning phase we finetuned Gemini 1.5 models on a collection of multimodal data (containing paired instructions and appropriate responses), with further tuning based on human preference data. We refer readers to the Gemini 1.0 Technical Report (Gemini-Team et al., 2023) for further information."* (our highlight)

Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context

4. Training Infrastructure and Dataset

Like Gemini 1.0 series, Gemini 1.5 models are trained on multiple 4096-chip pods of Google's TPUv4 accelerators, distributed across multiple datacenters, and on a variety of multimodal and multilingual data. Our pre-training dataset includes data sourced across many different domains, including web documents and code, and incorporates image, audio, and video content. For the instruction-tuning phase we finetuned Gemini 1.5 models on a collection of multimodal data (containing paired instructions and appropriate responses), with further tuning based on human preference data. We refer readers to the Gemini 1.0 Technical Report (Gemini-Team et al., 2023) for further information.

Screenshot 8 - <https://arxiv.org/pdf/2403.05530>

Gemma

Google describes its Gemma models as "open models" because the models' weights are available for download. Gemma 1 was released in FEB, 2024 and Gemma 2 was released JUL 31, 2024.

¹¹ <https://arxiv.org/abs/2403.05530>

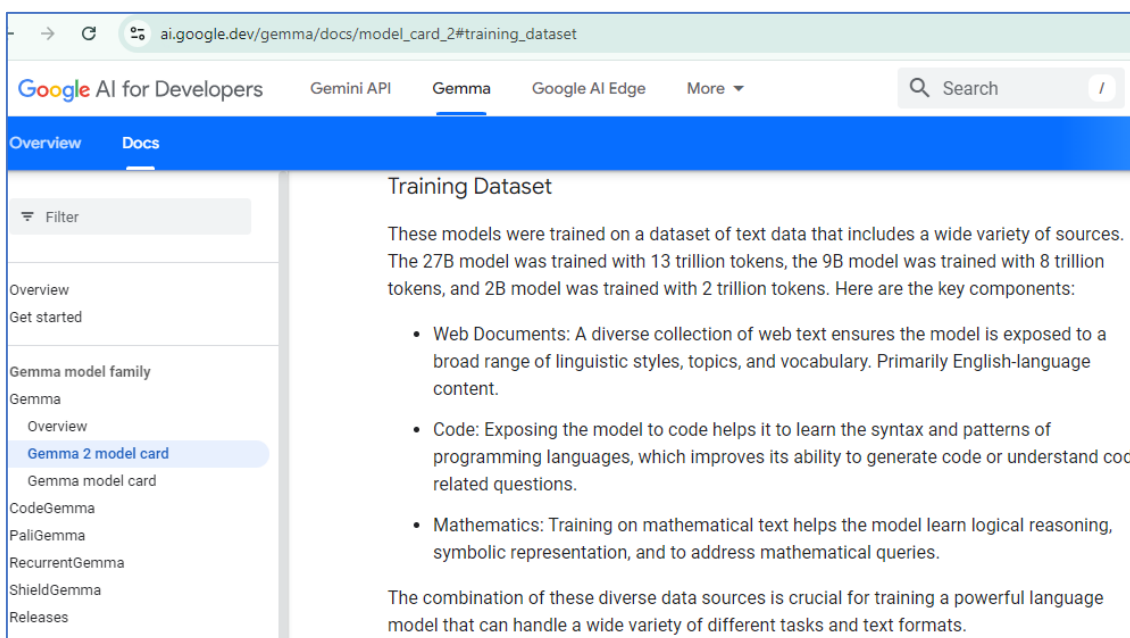
In the Gemma 2 model card¹² section “*Training Dataset*” Google provides the following information about the training data:

“These models were trained on a dataset of text data that includes a wide variety of sources. The 27B model was trained with 13 trillion tokens, the 9B model was trained with 8 trillion tokens, and 2B model was trained with 2 trillion tokens. Here are the key components:

- *Web Documents: A diverse collection of web text ensures the model is exposed to a broad range of linguistic styles, topics, and vocabulary. Primarily English-language content.*
- *Code: Exposing the model to code helps it to learn the syntax and patterns of programming languages, which improves its ability to generate code or understand code-related questions.*
- *Mathematics: Training on mathematical text helps the model learn logical reasoning, symbolic representation, and to address mathematical queries.*

The combination of these diverse data sources is crucial for training a powerful language model that can handle a wide variety of different tasks and text formats.”

¹² https://ai.google.dev/gemma/docs/model_card_2#training_dataset



Screenshot 9 - https://ai.google.dev/gemma/docs/model_card_2#training_dataset (screenshot AUG 22, 2024)

Comment on Google's transparency:

While Google provides a list of *types of content* e.g. web documents and books there is no list of titles to datasets or a narrative explanation including references to third-parties with the necessary information on content used.

This leaves rightsholders with no means to determine if their content has been used without permission to train Google's AI models and therefore no way to exercise and enforce their rights.

Case 4: OpenAI's GPT

Transparency provided:

OpenAI have released a series of AI models, called Generative Pretraining Transformers or "GPT".

The first version of the GPT-models, the **GPT-1**, had a research paper released on 11 June 2018 titled *Improving Language Understanding by Generative Pre-Training*¹³ in which OpenAI detail the first training data used for their model:

"We use the BooksCorpus dataset for training the language model. It contains over 7,000 unique unpublished books from a variety of genres including Adventure, Fantasy, and Romance."

Unsupervised pre-training We use the BooksCorpus dataset [71] for training the language model. It contains over 7,000 unique unpublished books from a variety of genres including Adventure, Fantasy, and Romance. Crucially, it contains long stretches of contiguous text, which allows the generative model to learn to condition on long-range information. An alternative dataset, the 1B Word Benchmark, which is used by a similar approach, ELMo [44], is approximately the same size

Screenshot 10 - https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

In 14 Feb 2019 OpenAI release their next version, the **GPT-2** and with it the research paper *Language Models are Unsupervised Multitask Learners*¹⁴. The research paper states the following in section 2.1. "Training Dataset":

"[W]e scraped all outbound links from Reddit, a social media platform, which received at least 3 karma. This can be thought of as a heuristic indicator for

¹³ https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

¹⁴ https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

whether other users found the link interesting, educational, or just funny. The resulting dataset, WebText, contains the text subset of these 45 million links.”¹⁵

QA task without a reward signal by using forward prediction of a teacher’s outputs. While dialog is an attractive approach, we worry it is overly restrictive. The internet contains a vast amount of information that is passively available without the need for interactive communication. Our speculation is that a language model with sufficient capacity will begin to learn to infer and perform the tasks demonstrated in natural language sequences in order to better predict them, regardless of their method of procurement. If a language model is able to do this it will be, in effect, performing unsupervised multitask learning. We test whether this is the case by analyzing the performance of language models in a zero-shot setting on a wide variety of tasks.

2.1. Training Dataset

Most prior work trained language models on a single domain of text, such as news articles (Jozefowicz et al., 2016), Wikipedia (Merity et al., 2016), or fiction books (Kiros et al., 2015). Our approach motivates building as large and diverse a dataset as possible in order to collect natural language demonstrations of tasks in as varied of domains and contexts as possible.

A promising source of diverse and nearly unlimited text is web scrapes such as Common Crawl. While these archives are many orders of magnitude larger than current language modeling datasets, they have significant data quality issues. Trinh & Le (2018) used Common Crawl in their work on commonsense reasoning but noted a large amount of documents “whose content are mostly unintelligible”. We observed similar data issues in our initial experiments with

Common Crawl. Trinh & Le (2018)’s best results were achieved using a small subsample of Common Crawl which included only documents most similar to their target dataset, the Winograd Schema Challenge. While this is a pragmatic approach to improve performance on a specific task, we want to avoid making assumptions about the tasks to be performed ahead of time.

Instead, we created a new web scrape which emphasizes document quality. To do this we only scraped web pages which have been curated/filtered by humans. Manually filtering a full web scrape would be exceptionally expensive so as a starting point, we scraped all outbound links from Reddit, a social media platform, which received at least 3 karma. This can be thought of as a heuristic indicator for whether other users found the link interesting, educational, or just funny.

The resulting dataset, WebText, contains the text subset of these 45 million links. To extract the text from HTML responses we use a combination of the Dragnet (Peters & Lecocq, 2013) and Newspaper¹ content extractors. All results presented in this paper use a preliminary version of WebText which does not include links created after Dec 2017 and which after de-duplication and some heuristic based cleaning contains slightly over 8 million documents for a total of 40 GB of text. We removed all Wikipedia documents from WebText since it is a common data source for other datasets and could complicate analysis due to over-

¹<https://github.com/codelucas/newspaper>

Language Models are Unsupervised Multitask Learners

lapping training data with test evaluation tasks.

Screenshot 11 - https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

¹⁵ "Language Models Are Unsupervised Multitask Learners" Radford et. Al., 14 Feb 2019, page 3

In May, 2020 OpenAI released their model **GPT-3** with a paper titled *Language Models are Few-Shot Learners*¹⁶.

In the paper section 2.2. “Training Dataset” OpenAI explains how they used a filtered version of Common Crawl webscrapes, their own WebText dataset, Wikipedia and “two internet-based books corpora (Books1 and Books2)” to train their AI model. The specifics of Books1 and Books2 have never been shared publicly by OpenAI.

Datasets for language models have rapidly expanded, culminating in the Common Crawl dataset² [RSR⁺19] constituting nearly a trillion words. This size of dataset is sufficient to train our largest models without ever updating on the same sequence twice. However, we have found that unfiltered or lightly filtered versions of Common Crawl tend to have lower quality than more curated datasets. Therefore, we took 3 steps to improve the average quality of our datasets: (1) we downloaded and filtered a version of CommonCrawl based on similarity to a range of high-quality reference corpora, (2) we performed fuzzy deduplication at the document level, within and across datasets, to prevent redundancy and preserve the integrity of our held-out validation set as an accurate measure of overfitting, and (3) we also added known high-quality reference corpora to the training mix to augment CommonCrawl and increase its diversity.

Details of the first two points (processing of Common Crawl) are described in Appendix A. For the third, we added several curated high-quality datasets, including an expanded version of the WebText dataset [RWC⁺19], collected by scraping links over a longer period of time, and first described in [KMH⁺20], two internet-based books corpora (Books1 and Books2) and English-language Wikipedia.

Table 2.2 shows the final mixture of datasets that we used in training. The CommonCrawl data was downloaded from 41 shards of monthly CommonCrawl covering 2016 to 2019, constituting 45TB of compressed plaintext before filtering and 570GB after filtering, roughly equivalent to 400 billion byte-pair-encoded tokens. Note that during training, datasets are not sampled in proportion to their size, but rather datasets we view as higher-quality are sampled more frequently, such that CommonCrawl and Books2 datasets are sampled less than once during training, but the other datasets are sampled 2-3 times. This essentially accepts a small amount of overfitting in exchange for higher quality training data.

²<https://commoncrawl.org/the-data/>

Screenshot 12 - <https://arxiv.org/abs/2005.14165>

In March 2023 OpenAI then released their **GPT-4** model along with a technical report *GPT-4 Technical Report*¹⁷. In the report section 2 “Scope and Limitations of this Technical Report” OpenAI writes:

¹⁶ <https://arxiv.org/abs/2005.14165>

¹⁷ <https://arxiv.org/abs/2303.08774>

“Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.”

This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model [39] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [40]. Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.

Screenshot 13 - <https://arxiv.org/pdf/2303.08774>

But OpenAI goes on to write the following in the section 2.7 “Privacy”:

“GPT-4 has learned from a variety of licensed, created, and publicly available data sources, which may include publicly available personal information.”

2.7 Privacy

GPT-4 has learned from a variety of licensed, created, and publicly available data sources, which may include publicly available personal information. [59, 60] As a result, our models may have knowledge about people who have a significant presence on the public internet, such as celebrities and public figures. GPT-4 can also synthesize multiple, distinct information types and perform multiple steps of reasoning within a given completion. The model can complete multiple basic tasks that may relate to personal and geographic information, such as determining the geographic locations associated with a phone number or answering where an educational institution is located in one completion and without browsing the internet. For example, the model can associate a Rutgers University email address to a phone number with a New Jersey area code with high recall, and explain its reasoning as being through that route. By combining capabilities on these types of tasks, GPT-4 has the potential to be used to attempt to identify individuals when augmented with outside data.

Screenshot 14 - <https://arxiv.org/pdf/2303.08774>

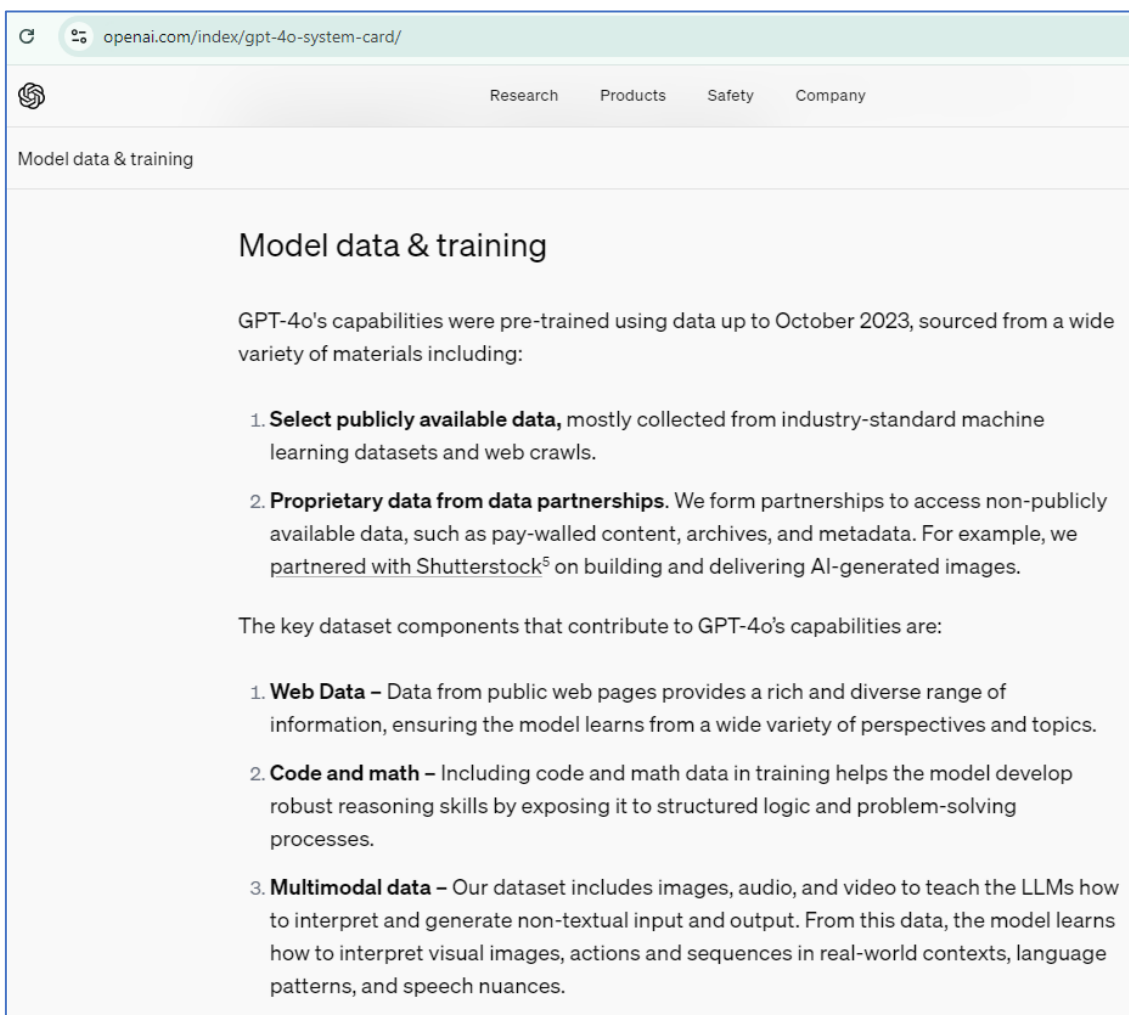
OpenAI released **GPT-4o** in MAY 2024 and published a system card in AUG 2024 *GPT-4o System Card*¹⁸.

In the section “Model data & training” OpenAI explains how they've used:

¹⁸ <https://cdn.openai.com/gpt-4o-system-card.pdf>

“data up to up to October 2023, sourced from a wide variety of materials including:

- (1) Select publicly available data, mostly collected from industry-standard machine learning datasets and web crawls.
- (2) Proprietary data from data partnerships. We form partnerships to access non-publicly available data, such as pay-walled content, archives, and metadata. For example, we partnered with Shutterstock on building and delivering AI-generated images.”



The screenshot shows the OpenAI website page for the GPT-4o system card. The browser address bar displays 'openai.com/index/gpt-4o-system-card/'. The page header includes the OpenAI logo and navigation links for Research, Products, Safety, and Company. The main content area is titled 'Model data & training' and contains the following text:

GPT-4o's capabilities were pre-trained using data up to October 2023, sourced from a wide variety of materials including:

1. **Select publicly available data**, mostly collected from industry-standard machine learning datasets and web crawls.
2. **Proprietary data from data partnerships**. We form partnerships to access non-publicly available data, such as pay-walled content, archives, and metadata. For example, we partnered with Shutterstock⁵ on building and delivering AI-generated images.

The key dataset components that contribute to GPT-4o's capabilities are:

1. **Web Data** – Data from public web pages provides a rich and diverse range of information, ensuring the model learns from a wide variety of perspectives and topics.
2. **Code and math** – Including code and math data in training helps the model develop robust reasoning skills by exposing it to structured logic and problem-solving processes.
3. **Multimodal data** – Our dataset includes images, audio, and video to teach the LLMs how to interpret and generate non-textual input and output. From this data, the model learns how to interpret visual images, actions and sequences in real-world contexts, language patterns, and speech nuances.

Screenshot 15 - <https://openai.com/index/gpt-4o-system-card/> (taken AUG 26, 2024)

Comment on OpenAI's transparency:

With its latest models OpenAI is providing a level of transparency that rightsholders cannot use to determine whether or not their content has been used to train OpenAI's models and thereby exercise and enforce their rights.

Looking at how OpenAI has previously been transparent about using Common Crawl to train its models "*Select publicly available data, mostly collected from industry-standard machine learning datasets and web crawls*" most likely covers Common Crawl. It can also cover their "WebText" dataset with Reddit comments.

The information that data was collected up to October 2023 does not provide rightsholders with sufficient knowledge to exercise and enforce their rights.

Case 5: Microsoft's Phi

There are interesting features to be observed in Microsoft's language models "Phi" as they are trained extensively on synthetically generated training data from OpenAI's models.

Microsoft released **Phi-1** in JUN 2023 together with a paper *Textbooks Are All You Need*.¹⁹ This language model was trained specifically for coding. In the paper section 2 "*Training details and the importance of high-quality data*" Microsoft lists datasets collected by others by providing *titles of the dataset* ("*The Stack*" and "*StackOverflow*"). Microsoft also provides *references to papers* by third-parties describing the datasets in more detail ("*Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, et al. The stack: 3 tb of permissively licensed source code. arXiv preprint arXiv:2211.15533, 2022.*").

¹⁹ <https://arxiv.org/abs/2306.11644>

Microsoft also describes in section 2.2 “*Creation of synthetic textbook-quality datasets*” how it used OpenAI’s GPT-3.5 model to generate datasets with synthetic textbook-like content.

In this work, we address this challenge directly and show that by intentionally selecting and generating high-quality data, we can achieve state-of-the-art results on code-generation tasks with a much smaller model and less compute than existing approaches. Our training relies on three main datasets:

- A *filtered code-language* dataset, which is a subset of The Stack and StackOverflow, obtained by using a language model-based classifier (consisting of about 6B tokens).
- A *synthetic textbook* dataset consisting of <1B tokens of GPT-3.5 generated Python textbooks.
- A small *synthetic exercises* dataset consisting of ~180M tokens of Python exercises and solutions.

Screenshot 16 – phi-1 technical report <https://arxiv.org/pdf/2306.11644>

Microsoft released **Phi-1.5** in SEP 2023 together with a paper *Textbooks Are All You Need II: phi-1.5 technical report*.²⁰ This version of Phi is a general purpose model that can generate poems, emails, summarize, code etc.

In the paper section 2.2 “*Training data*” Microsoft describes how it has created a new “*synthetic, “textbook-like” data (roughly 20B tokens) for the purpose of teaching common sense reasoning and general knowledge of the world (science, daily activities, theory of mind, etc.). We carefully selected 20K topics to seed the generation of this new synthetic data. In our generation prompts, we use samples from web datasets for diversity.*”

²⁰ <https://arxiv.org/abs/2309.05463>

2.2 Training data

Our training data for **phi-1.5** is a combination of **phi-1**'s training data (7B tokens) and newly created synthetic, "textbook-like" data (roughly 20B tokens) for the purpose of teaching common sense reasoning and general knowledge of the world (science, daily activities, theory of mind, etc.). We carefully selected 20K topics to seed the generation of this new synthetic data. In our generation prompts, we use samples from web datasets for diversity. We point out that the only non-synthetic part in our training data for **phi-1.5** consists of the 6B tokens of filtered code dataset used in **phi-1**'s training (see [GZA⁺23]).

We remark that the experience gained in the process of creating the training data for both **phi-1** and **phi-1.5** leads us to the conclusion that the creation of a robust and comprehensive dataset demands more than raw computational power: It requires intricate iterations, strategic topic selection, and a deep understanding of knowledge gaps to ensure quality and diversity of the data. We speculate that the creation of synthetic datasets will become, in the near future, an important technical skill and a central topic of research in AI.

Screenshot 17 - phi-1.5technical report <https://arxiv.org/pdf/2309.05463>

Microsoft released **Phi-2** in DEC 2023 together with a blogpost

*Phi-2: The surprising power of small language models.*²¹ In the blogpost section "Key Insights Behind Phi-2" Microsoft provides the following details about training data:

"Firstly, training data quality plays a critical role in model performance. This has been known for decades, but we take this insight to its extreme by focusing on "textbook-quality" data, following upon our prior work "Textbooks Are All You Need." Our training data mixture contains synthetic datasets specifically created to teach the model common sense reasoning and general knowledge, including science, daily activities, and theory of mind, among others. We further augment our training corpus with carefully selected web data that is filtered based on educational value and content quality."

²¹ <https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/>

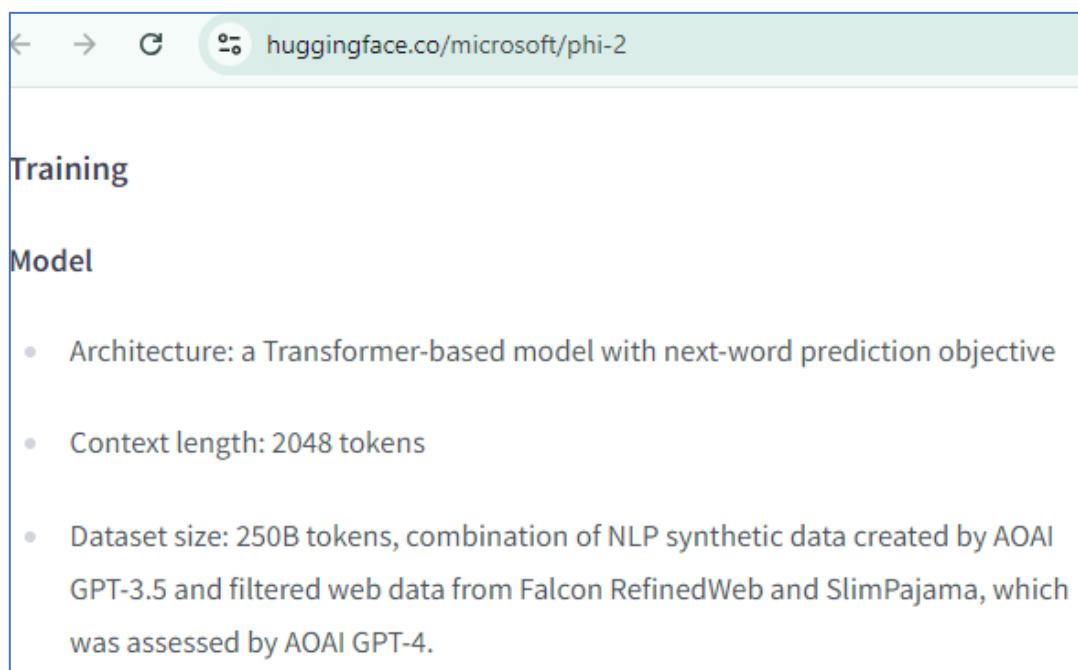
Firstly, training data quality plays a critical role in model performance. This has been known for decades, but we take this insight to its extreme by focusing on “textbook-quality” data, following upon our prior work “[Textbooks Are All You Need](#).” Our training data mixture contains synthetic datasets specifically created to teach the model common sense reasoning and general knowledge, including science, daily activities, and theory of mind, among others. We further augment our training corpus with carefully selected web data that is filtered based on educational value and content quality. Secondly, we use innovative techniques to scale up, starting from our 1.3 billion parameter model, Phi-1.5, and embedding its knowledge within the 2.7 billion parameter Phi-2. This scaled knowledge transfer not only accelerates training convergence but shows clear boost in Phi-2 benchmark scores.

Screenshot 18 - Phi-2 blogpost <https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/>

Microsoft has more detail on the “*carefully selected web data*” on its Hugging Face page for Phi-2²² mentioning the datasets “*Falcon RefinedWeb*” and “*SlimPajama*”:

“*Dataset size: 250B tokens, combination of NLP synthetic data created by AOAI GPT-3.5 and filtered web data from Falcon RefinedWeb and SlimPajama, which was assessed by AOAI GPT-4.*” (AOAI: Azure OpenAI)

²² <https://huggingface.co/microsoft/phi-2>



Screenshot 19 - <https://huggingface.co/microsoft/phi-2> (screenshot AUG 22, 2024)

Microsoft released **Phi-3-mini** in APR 2024 with a technical report *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone*.²³ Here Microsoft writes:

"Our training data of consists of heavily filtered publicly available web data (according to the "educational level") from various open internet sources, as well as synthetic LLM-generated data. Pre-training is performed in two disjoint and sequential phases; phase-1 comprises mostly of web sources aimed at teaching the model general knowledge and language understanding. Phase-2 merges even more heavily filtered webdata (a subset used in Phase-1) with some synthetic data that teach the model logical reasoning and various niche skills."

²³ <https://arxiv.org/pdf/2404.14219>

Training Methodology. We follow the sequence of works initiated in “Textbooks Are All You Need” [GZA⁺23], which utilize high quality training data to improve the performance of small language models and deviate from the standard *scaling-laws*. In this work we show that such method allows to reach the level of highly capable models such as GPT-3.5 or Mixtral with only 3.8B total parameters (while Mixtral has 45B total parameters for example). Our training data consists of heavily filtered publicly available web data (according to the “educational level”) from various open internet sources, as well as synthetic LLM-generated data. Pre-training is performed in two disjoint and sequential phases; phase-1 comprises mostly of web sources aimed at teaching the model general knowledge and language understanding. Phase-2 merges even more heavily filtered webdata (a subset used in Phase-1) with some synthetic data that teach the model logical reasoning and various niche skills.

Screenshot 20 - Phi-3 Technical Report <https://arxiv.org/pdf/2404.14219>

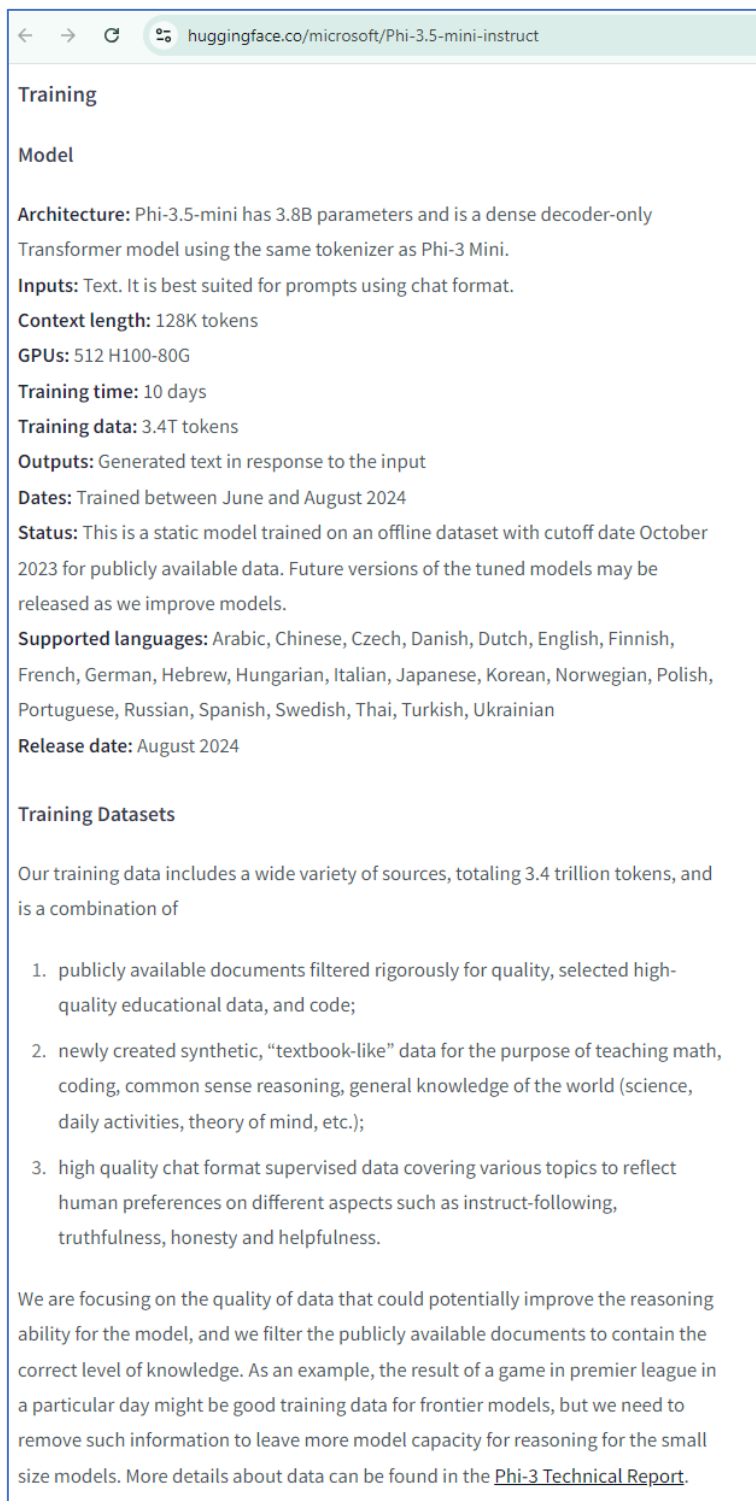
Microsoft released **Phi-3.5-mini** in AUG 2024 on Hugging Face.²⁴ Here Microsoft writes:

“Our training data includes a wide variety of sources, totaling 3.4 trillion tokens, and is a combination of

1. publicly available documents filtered rigorously for quality, selected high-quality educational data, and code;
2. newly created synthetic, “textbook-like” data for the purpose of teaching math, coding, common sense reasoning, general knowledge of the world (science, daily activities, theory of mind, etc.);
3. high quality chat format supervised data covering various topics to reflect human preferences on different aspects such as instruct-following, truthfulness, honesty and helpfulness.

We are focusing on the quality of data that could potentially improve the reasoning ability for the model, and we filter the publicly available documents to contain the correct level of knowledge. As an example, the result of a game in premier league in a particular day might be good training data for frontier models, but we need to remove such information to leave more model capacity for reasoning for the small size models. More details about data can be found in the [Phi-3 Technical Report](#).”

²⁴ <https://huggingface.co/microsoft/Phi-3.5-mini-instruct>



← → ↻ 🌐 huggingface.co/microsoft/Phi-3.5-mini-instruct

Training

Model

Architecture: Phi-3.5-mini has 3.8B parameters and is a dense decoder-only Transformer model using the same tokenizer as Phi-3 Mini.

Inputs: Text. It is best suited for prompts using chat format.

Context length: 128K tokens

GPUs: 512 H100-80G

Training time: 10 days

Training data: 3.4T tokens

Outputs: Generated text in response to the input

Dates: Trained between June and August 2024

Status: This is a static model trained on an offline dataset with cutoff date October 2023 for publicly available data. Future versions of the tuned models may be released as we improve models.

Supported languages: Arabic, Chinese, Czech, Danish, Dutch, English, Finnish, French, German, Hebrew, Hungarian, Italian, Japanese, Korean, Norwegian, Polish, Portuguese, Russian, Spanish, Swedish, Thai, Turkish, Ukrainian

Release date: August 2024

Training Datasets

Our training data includes a wide variety of sources, totaling 3.4 trillion tokens, and is a combination of

1. publicly available documents filtered rigorously for quality, selected high-quality educational data, and code;
2. newly created synthetic, “textbook-like” data for the purpose of teaching math, coding, common sense reasoning, general knowledge of the world (science, daily activities, theory of mind, etc.);
3. high quality chat format supervised data covering various topics to reflect human preferences on different aspects such as instruct-following, truthfulness, honesty and helpfulness.

We are focusing on the quality of data that could potentially improve the reasoning ability for the model, and we filter the publicly available documents to contain the correct level of knowledge. As an example, the result of a game in premier league in a particular day might be good training data for frontier models, but we need to remove such information to leave more model capacity for reasoning for the small size models. More details about data can be found in the [Phi-3 Technical Report](#).

Screenshot 21 - Phi-3.5-mini <https://huggingface.co/microsoft/Phi-3.5-mini-instruct> (AUG 22, 2024)

Comment on Microsoft's transparency:

Microsoft lists *titles of datasets* and *references to papers* by third-parties with more details for Phi-1 and Phi-2 such as the dataset "SlimPajama" used in training Phi-2. Looking closer at SlimPajama²⁵ we can see that this dataset contains content from e.g. Common Crawl, Google's C4 (also sourced from Common Crawl), Github, Books (Books3 is found in RedPajama²⁶ dataset used in part in SlimPajama). By downloading the SlimPajama dataset it would be possible to search for specific content, but this is a very cumbersome process.

Microsoft does not list datasets used or give narrative explanations of content used to train Phi-3-mini and later models. Here Microsoft uses descriptions such as "*publicly available web data*" from "*various open internet sources*".

Microsoft does explain how it used Open AI's GPT-3.5 to generate textbook-like content to train its models. While this provides transparency into *how* Microsoft generated the training data and with *what tools* (GPT-3.5) this does not provide any transparency about the real training data since OpenAI is not itself transparent about the training data for its AI models.

Our assessment is that rightsholders have a very limited way of determining if their content has been used to train Microsoft's early Phi models. But Microsoft has ceased from listing dataset titles from Phi-3-mini and forward. The lack of transparency into what OpenAI's GPT-3.5 model is trained on makes the knowledge about Microsoft using this model to generate synthetic training data useless when it comes to determining what specific content was used to train the Phi-models and if there is a basis to exercise and enforce copyrights.

²⁵ <https://huggingface.co/datasets/cerebras/SlimPajama-627B>

²⁶ <https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T>

Case 6: Eleuther AI's GPT-NeoX-20B

Transparency provided:

The US based open-source organization EleutherAI released their GPT-NeoX-20B model in APR 2022 together with a technical paper *GPT-NeoX-20B: An Open-Source Autoregressive Language Model*.²⁷ In the paper section 3.1 "Training Data" the training data is described by providing a list of dataset *titles* and *references* to *self-published and third party papers* with more details on the training content.

<p>et al., 2020) to reduce memory consumption by distributing optimizer states across ranks. Since the weights and optimizer states of a model at this scale do not fit on a single GPU, we use the tensor parallelism scheme introduced in Shoeybi et al. (2020) in combination with pipeline parallelism (Harlap et al., 2018) to distribute the model across GPUs. To train GPT-NeoX-20B, we found that the most efficient way to distribute the model given our hardware setup was to set a tensor parallel size of 2, and a pipeline parallel size of 4. This allows for the most communication intensive processes, tensor and pipeline parallelism, to occur within a node, and data parallel communication to occur across node boundaries. In this fashion, we were able to achieve and maintain an efficiency of 117 teraFLOPS per GPU.</p> <p>3.1 Training Data</p> <p>GPT-NeoX-20B was trained on the Pile (Gao et al., 2020), a massive curated dataset designed specifically for training large language models. It consists of data from 22 data sources, coarsely broken down into 5 categories:</p> <ul style="list-style-type: none"> • Academic Writing: Pubmed Abstracts and PubMed Central, arXiv, FreeLaw,⁵ USPTO Backgrounds,⁶ PhilPapers,⁷ NIH Exporter⁸ • Web-scrapes and Internet Resources: <p>⁵https://www.courtlistener.com/ ⁶https://bulkdata.uspto.gov/ ⁷https://philpapers.org/ ⁸https://exporter.nih.gov/</p>	<p>CommonCrawl, OpenWebText2, StackExchange,⁹ Wikipedia (English)</p> <ul style="list-style-type: none"> • Prose: BookCorpus2, Bibliotik, Project Gutenberg (PG-19; Rae et al., 2019) • Dialogue: Youtube subtitles, Ubuntu IRC,¹⁰ OpenSubtitles (Lison and Tiedemann, 2016), Hacker News,¹¹ EuroParl (Koehn, 2005) • Miscellaneous: GitHub, the DeepMind Mathematics dataset (Saxton et al., 2019), Enron Emails (Klimt and Yang, 2004) <p>In aggregate, the Pile consists of over 825 GiB of raw text data. The diversity of data sources reflects our desire for a general-purpose language model. Certain components are up-sampled to obtain a more balanced data distribution. In contrast, GPT-3's training data consists of web-scrapes, books datasets, and Wikipedia. When comparing results in this work to GPT-3, the training data is almost certainly the biggest known unknown factor. Full details of the Pile can be found in the technical report (Gao et al., 2020) and the associated datasheet (Biderman et al., 2022).</p> <p>It is particularly notable that the Pile contains a scrape of StackExchange preprocessed into a Q/A form. There is a significant and growing body of work on the influence of the syntactic structure of finetuning data on downstream performance (Zhong et al., 2021; Tan et al., 2021;</p> <p>⁹https://archive.org/details/stackexchange ¹⁰https://irclogs.ubuntu.com/ ¹¹https://news.ycombinator.com/</p>
---	---

Screenshot 22 - <https://arxiv.org/pdf/2204.06745>

²⁷ <https://arxiv.org/pdf/2204.06745>

EleutherAI is also responsible for “the Pile” dataset mentioned as training data for GPT-NeoX-20B. “The Pile” is a massive dataset for training general-purpose language models. The Pile was made available for free online together with the technical paper from DEC 2020 *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*.²⁸

In the paper's section 2 “The Pile Datasets” the 22 sub-datasets are listed with dataset *titles* and a narrative explanation is provided with:

- *Content type* e.g. “Books3 is a dataset of books [...]”.
- *Source of the content* e.g. “[...] the Bibliotik private tracker made available by Shawn Presser (Presser, 2020).”.
- *References to third-party papers* describing the dataset in more detail.
- *Reason for why the sub-dataset is included*.

²⁸ <https://arxiv.org/pdf/2101.00027>

all processing performed on each dataset (and the Pile as a whole) in as much detail as possible. For further details about the processing of each dataset, see Section 2 and Appendix C.

²<https://github.com/EleutherAI/the-pile>

2 The Pile Datasets

The Pile is composed of 22 constituent sub-datasets, as shown in Table 1. Following Brown et al. (2020), we increase the weights of higher quality components, with certain high-quality datasets such as Wikipedia being seen up to 3 times (“epochs”) for

Component	Raw Size	Weight	Epochs	Effective Size	Mean Document Size
Pile-CC	227.12 GiB	18.11%	1.0	227.12 GiB	4.33 KiB
PubMed Central	90.27 GiB	14.40%	2.0	180.55 GiB	30.55 KiB
Books3 [†]	100.96 GiB	12.07%	1.5	151.44 GiB	538.36 KiB
OpenWebText2	62.77 GiB	10.01%	2.0	125.54 GiB	3.85 KiB
ArXiv	56.21 GiB	8.96%	2.0	112.42 GiB	46.61 KiB
Github	95.16 GiB	7.59%	1.0	95.16 GiB	5.25 KiB
FreeLaw	51.15 GiB	6.12%	1.5	76.73 GiB	15.06 KiB
Stack Exchange	32.20 GiB	5.13%	2.0	64.39 GiB	2.16 KiB
USPTO Backgrounds	22.90 GiB	3.65%	2.0	45.81 GiB	4.08 KiB
PubMed Abstracts	19.26 GiB	3.07%	2.0	38.53 GiB	1.30 KiB
Gutenberg (PG-19) [†]	10.88 GiB	2.17%	2.5	27.19 GiB	398.73 KiB
OpenSubtitles [†]	12.98 GiB	1.55%	1.5	19.47 GiB	30.48 KiB
Wikipedia (en) [†]	6.38 GiB	1.53%	3.0	19.13 GiB	1.11 KiB
DM Mathematics [†]	7.75 GiB	1.24%	2.0	15.49 GiB	8.00 KiB
Ubuntu IRC	5.52 GiB	0.88%	2.0	11.03 GiB	545.48 KiB
BookCorpus2	6.30 GiB	0.75%	1.5	9.45 GiB	369.87 KiB
EuroParl [†]	4.59 GiB	0.73%	2.0	9.17 GiB	68.87 KiB
HackerNews	3.90 GiB	0.62%	2.0	7.80 GiB	4.92 KiB
YoutubeSubtitles	3.73 GiB	0.60%	2.0	7.47 GiB	22.55 KiB
PhilPapers	2.38 GiB	0.38%	2.0	4.76 GiB	73.37 KiB
NIH ExPorter	1.89 GiB	0.30%	2.0	3.79 GiB	2.11 KiB
Enron Emails [†]	0.88 GiB	0.14%	2.0	1.76 GiB	1.78 KiB
The Pile	825.18 GiB			1254.20 GiB	5.91 KiB

Table 1: Overview of datasets in the Pile before creating the held out sets. Raw Size is the size before any up- or down-sampling. Weight is the percentage of bytes in the final dataset occupied by each dataset. Epochs is the number of passes over each constituent dataset during a full epoch over the Pile. Effective Size is the approximate number of bytes in the Pile occupied by each dataset. Datasets marked with a † are used with minimal preprocessing from prior work.

each full epoch over the Pile. Detailed information about the construction of each dataset is available in Appendix C.

2.1 Pile-CC

Common Crawl is a collection of website crawls from 2008 onwards, including raw web pages, metadata and text extractions. Due to the raw nature of the dataset, Common Crawl has the advantage of including text from diverse domains, but at the cost of varying quality data. Due to this, use of Common Crawl typically necessitates well-designed extraction and filtering. Our Common Crawl-based dataset, Pile-CC, uses just-Text (Endrédy and Novák, 2013) on Web Archive files (raw HTTP responses including page HTML) for extraction, which yields higher quality output than directly using the WET files (extracted plain-text).

2.2 PubMed Central

PubMed Central (PMC) is a subset of the PubMed online repository for biomedical articles run by the United States of America’s National Center for Biotechnology Information (NCBI), providing open, full-text access to nearly five million publications. Most publications indexed by PMC are recent, and their inclusion is mandated for all NIH funded research starting from 2008 by the NIH Public Access Policy. We included PMC in the hopes that it will benefit potential downstream applications to the medical domain.

2.3 Books3

Books3 is a dataset of books derived from a copy of the contents of the Bibliotik private tracker made available by Shawn Presser (Presser, 2020). Bibliotik consists of a mix of fiction and nonfiction books and is almost an order of magnitude

Screenshot 23 - <https://arxiv.org/pdf/2101.00027>

larger than our next largest book dataset (BookCorpus2). We included Bibliotik because books are invaluable for long-range context modeling research and coherent storytelling.

2.4 OpenWebText2

OpenWebText2 (OWT2) is a generalized web scrape dataset inspired by WebText (Radford et al., 2019) and OpenWebTextCorpus (Gokaslan and Cohen, 2019). Similar to the original WebText, we use net upvotes on Reddit submissions as a proxy for outgoing link quality. OpenWebText2 includes more recent content from Reddit submissions up until 2020, content from multiple languages, document metadata, multiple dataset versions, and open source replication code. We included OWT2 as a high quality general purpose dataset.

2.5 ArXiv

ArXiv is a preprint server for research papers that has operated since 1991. As shown in fig. 10, arXiv papers are predominantly in the fields of Math, Computer Science, and Physics. We included arXiv in the hopes that it will be a source of high quality text and math knowledge, and benefit potential downstream applications to research in these areas. ArXiv papers are written in LaTeX, a common typesetting language for mathematics, computer science, physics, and some adjacent fields. Training a language model to be able to generate papers written in LaTeX could be a huge boon to the research community.

2.6 GitHub

GitHub is a large corpus of open-source code repositories. Motivated by the ability of GPT-3 (Brown et al., 2020) to generate plausible code completions despite its training data not containing any explicitly gathered code datasets, we included GitHub in the hopes that it would enable better downstream performance on code-related tasks.

2.7 FreeLaw

The Free Law Project is a US-registered non-profit that provides access to and analytical tools for academic studies in the legal realm. CourtListener,³ part of the Free Law Project, provides bulk downloads for millions of legal opinions from federal and state courts. While the full dataset provides multiple modalities of legal proceedings, including dockets, bibliographic information on judges,

³<https://www.courtlistener.com/>

and other metadata, we focused specifically on court opinions due to an abundance of full-text entries. This data is entirely within the public domain.

2.8 Stack Exchange

The Stack Exchange Data Dump⁴ contains an anonymized set of all user-contributed content on the Stack Exchange network, a popular collection of websites centered around user-contributed questions and answers. It is one of the largest publicly available repositories of question-answer pairs, and covers a wide range of subjects—from programming, to gardening, to Buddhism. We included Stack Exchange in the hopes that it will improve the question answering capabilities of downstream models on diverse domains.

2.9 USPTO Backgrounds

USPTO Backgrounds is a dataset of background sections from patents granted by the United States Patent and Trademark Office, derived from its published bulk archives⁵. A typical patent background lays out the general context of the invention, gives an overview of the technical field, and sets up the framing of the problem space. We included USPTO Backgrounds because it contains a large volume of technical writing on applied subjects, aimed at a non-technical audience.

2.10 Wikipedia (English)

Wikipedia is a standard source of high-quality text for language modeling. In addition to being a source of high quality, clean English text, it is also valuable as it is written in expository prose, and spans many domains.

2.11 PubMed Abstracts

PubMed Abstracts consists of the abstracts from 30 million publications in PubMed, the online repository for biomedical articles run by the National Library of Medicine. While the PMC (see Section 2.2) provides full-text access, the subset of coverage is significantly limited and biased towards recent publications. PubMed also incorporates MEDLINE, which expands the coverage of biomedical abstracts from 1946 to present day.

⁴<https://archive.org/details/stackexchange>

⁵<https://bulkdata.uspto.gov/>

Screenshot 24 - <https://arxiv.org/pdf/2101.00027>

2.12 Project Gutenberg

Project Gutenberg is a dataset of classic Western literature. The specific Project Gutenberg derived dataset we used, PG-19, consists of Project Gutenberg books from before 1919 (Rae et al., 2019), which represent distinct styles from the more modern Books3 and BookCorpus. Additionally, the PG-19 dataset is already being used for long-distance context modeling.

2.13 OpenSubtitles

The OpenSubtitles dataset is an English language dataset of subtitles from movies and television shows gathered by Tiedemann (2016). Subtitles provide an important source of natural dialog, as well as an understanding of fictional formats other than prose, which may prove useful for creative writing generation tasks such as screenwriting, speechwriting, and interactive storytelling.

2.14 DeepMind Mathematics

The DeepMind Mathematics dataset consists of a collection of mathematical problems from topics such as algebra, arithmetic, calculus, number theory, and probability, formatted as natural language prompts (Saxton et al., 2019). One major weakness of large language models has been performance on mathematical tasks (Brown et al., 2020), which may be due in part to a lack of math problems in the training set. By explicitly including a dataset of mathematical problems, we hope to improve the mathematical ability of language models trained on the Pile.

2.15 BookCorpus2

BookCorpus2 is an expanded version of the original BookCorpus (Zhu et al., 2015), a widely used language modeling corpus consisting of books written by “as of yet unpublished authors.” BookCorpus is therefore unlikely to have significant overlap with Project Gutenberg and Books3, which consist of published books. BookCorpus is also commonly used as dataset for training language models (Radford et al., 2018; Devlin et al., 2019; Liu et al., 2019).

2.16 Ubuntu IRC

The Ubuntu IRC dataset is derived from the publicly available chatlogs⁶ of all Ubuntu-related channels on the Freenode IRC chat server. Chatlog data

⁶<https://irclogs.ubuntu.com/>

provides an opportunity to model real-time human interactions, which feature a level of spontaneity not typically found in other modes of social media.

2.17 EuroParl

EuroParl (Koehn, 2005) is a multilingual parallel corpus originally introduced for machine translation but which has also seen use in several other fields of NLP (Groves and Way, 2006; Van Halteren, 2008; Ciobanu et al., 2017). We use the most current version at time of writing, which consists of the proceedings of the European Parliament in 21 European languages from 1996 until 2012.

2.18 YouTube Subtitles

The YouTube Subtitles dataset is a parallel corpus of text gathered from human generated closed-captions on YouTube. In addition to providing multilingual data, Youtube Subtitles is also a source of educational content, popular culture, and natural dialog.

2.19 PhilPapers

The PhilPapers⁷ dataset consists of open-access philosophy publications from an international database maintained by the Center for Digital Philosophy at the University of Western Ontario. We included PhilPapers because it spans a wide body of abstract, conceptual discourse, and its articles contain high quality academic writing.

2.20 NIH Grant Abstracts: ExPORTER

The NIH Grant abstracts provides a bulk-data repository for awarded applications through the ExPORTER⁸ service covering the fiscal years 1985-present. We included the dataset because it contains examples of high-quality scientific writing.

2.21 Hacker News

Hacker News⁹ is a link aggregator operated by Y Combinator, a startup incubator and investment fund. Users submit articles defined as “anything that gratifies one’s intellectual curiosity,” but submitted articles tend to focus on topics in computer science and entrepreneurship. Users can comment on submitted stories, resulting in comment trees discussing and critiquing submitted stories. We

⁷<https://philpapers.org/>

⁸<https://exporter.nih.gov/>

⁹<https://news.ycombinator.com>

Screenshot 25 - <https://arxiv.org/pdf/2101.00027>

scrape, parse, and include these comment trees since we believe they provide high quality dialogue and debate on niche topics.

2.22 Enron Emails

The Enron Emails dataset (Klimt and Yang, 2004) is a valuable corpus commonly used for research about the usage patterns of email. We included Enron Emails to aid in understanding the modality of email communications, which is typically not found in any of our other datasets.

2019) and GPT-3 (Brown et al., 2020), shown in Figure 2. We use all available versions of GPT-2, and all four versions of GPT-3 available via the OpenAI API. Because of the cost associated with using the OpenAI API, we evaluate on one-tenth of the respective test sets for most of the constituent datasets. We report the perplexity converted to bits per UTF-8 encoded byte (BPB). Importantly, we compute perplexity by evaluating each document independently within each dataset, as opposed to concatenating all documents as is common practice

Screenshot 26 - <https://arxiv.org/pdf/2101.00027>

Comment on EleutherAI's transparency:

EleutherAI provides *direct* access to a copy of its training data while also providing a list of *dataset titles* and a narrative explanation covering *content type*, *source of content* and references to third-party papers, thus enabling rightsholders to determine if their content has been used to train GPT-NeoX-20B. In turn this allows rightsholders to determine if they want to enforce their rights against the model provider.

EleutherAI did not respond to our notification that they had infringed upon our members' copyrights.

Case 7: Anthropic's Claude

Transparency provided:

Anthropic has released research papers exploring the training of language models. One such paper, released on 9 December, 2021, titled *A General Language Assistant as a Laboratory for Alignment* describes how Anthropic trained an early version of their language model:

"The natural language dataset was composed of 55% heavily filtered common crawl data (220B tokens), 32% internet books (128B tokens), and some smaller

distributions including OpenWebText, Wikipedia, Stack Exchange, Arxiv, Legal and Patent documents, Ubuntu-IRC discussion, and movie scripts, most of which we sourced from The Pile [GBB+20]."²⁹

The training dataset is composed of 90% natural language and 10% python code. All components of the NL and code datasets were globally fuzzily deduplicated [BMR⁺20], and we train for one epoch on all sub-components (i.e. we do not repeat any data). The natural language dataset was composed of 55% heavily filtered common crawl data (220B tokens), 32% internet books (128B tokens), and some smaller distributions including OpenWebText, Wikipedia, Stack Exchange, Arxiv, Legal and Patent documents, Ubuntu-IRC discussion, and movie scripts, most of which we sourced from The Pile [GBB⁺20].

Screenshot 27 - <https://arxiv.org/pdf/2112.00861>

Anthropic released their latest version of their Claude model: Claude 3.5 Sonnet in June, 2024 on Anthropic's website.³⁰ While the newer versions of Claude do not specify anything about the training data, Anthropic provide the following in their FAQ³¹:

"While it is not our intention to "train" our models on personal data specifically, training data for our large language models, like others, can include web-based data that may contain publicly available personal data. We train our models using data from three sources:

- 1. Publicly available information via the Internet*
- 2. Datasets that we license from third party businesses*
- 3. Data that our users or crowd workers provide"*

²⁹ "A General Language Assistant as a Laboratory for Alignment" 9 December 2021, Anthropic, p. 27 <https://arxiv.org/pdf/2112.00861>

³⁰ <https://www.anthropic.com/news/claude-3-5-sonnet>

³¹ <https://support.anthropic.com/en/articles/7996885-how-do-you-use-personal-data-in-model-training>

How do you use personal data in model training?

Updated over a week ago

About model training

Large language models such as Claude need to be 'trained' on text so that they can learn the patterns and connections between words. This training is important so that the model performs effectively and safely.

While it is not our intention to "train" our models on personal data specifically, training data for our large language models, like others, can include web-based data that may contain publicly available personal data. We train our models using data from three sources:

1. Publicly available information via the Internet
2. Datasets that we license from third party businesses
3. Data that our users or crowd workers provide

Screenshot 28 - <https://support.anthropic.com/en/articles/7996885-how-do-you-use-personal-data-in-model-training> (taken 5.9.24)

Comment on transparency:

Anthropic has in the past provided a list of some datasets used to train their models, but the transparency was lacking in many ways such as only mentioning the type of content used e.g. "internet books" or not providing a narrative explanation of the datasets. The current models' training data are described as originating from "Publicly available information via the Internet", which leaves rightsholders with no way to determine if their content has been used to train Anthropic models.

Music generating AI models

Case 8: Suno AI's Suno

Transparency provided by Suno AI:

Suno AI launched a "beta" version of its AI music generation service in JUL 2023 first through a channel on the social media website Discord and later via a web

interface. Suno launched a new version of its service dubbed “v3” in MAR 2024 and a version named “v3.5” in MAY 2024.

Initially Suno did not provide any public information on the content used to train its models running the Suno AI music generation service.

Additional information:

Suno AI was sued in the USA by several record companies alleging that Suno has violated their copyrights by reproducing their copyrighted recordings without permission and subsequently training Suno's AI models. In Suno AI's first answer to these allegations Suno AI admits that their AI model was trained on “*tens of millions of recordings*” which “*includes essentially all music files of reasonable quality that are accessible on the open Internet*”.³²

Comment on transparency:

The original level of transparency provided by Suno left rightsholders with no way of determining whether or not Suno had used their content to train its models. Suno only disclosed information regarding their training data after the record companies were able to generate output that was almost identical or so similar to recordings owned by the labels that it couldn't have been generated without Suno training its models on the record companies' content.

This said it is still uncertain whether the statements from Suno in the US case is sufficient to enable other rightsholders to exercise their rights. It may likely be the case that other rightsholders will also have to generate output indicating that the Suno model has been trained on their content.

32

<https://fingfx.thomsonreuters.com/gfx/legaldocs/zjvqymadevx/USA%20COURT%20MUSIC%20COPYRIGHTS%20sunoanswer.pdf>

Case 9: Uncharted Labs' Udio

Transparency provided:

Uncharted Labs launched Udio an AI music generator in APR 2024 on their website. Uncharted Labs has since released their V1.5 model with increased capabilities in JUL 2024.

Uncharted Labs has not willingly provided any transparency into what their models were trained.

Additional information:

Uncharted Labs was sued in the USA by several record companies alleging that Uncharted Labs has violated their copyrights by reproducing their copyrighted recordings without permission and subsequently training the Udio models. In Uncharted Labs' first answer to these allegations it admits that Udio was trained by showing it "*many instances of different kinds of recordings gathered from publicly available sources*" and that plaintiff's recordings were probably in the training data for Udio.³³

Comment on transparency:

The original level of transparency provided by Uncharted Labs left rightsholders with no way of determining whether or not the Udio model was trained on their content. Uncharted Labs only disclosed information regarding their training data after the record companies were able to generate output that was almost identical or so similar to recordings owned by the labels that it couldn't have been generated without Uncharted Labs training its models on the record companies' content.

³³ <https://www.courtlistener.com/docket/68878697/26/umg-recordings-inc-v-uncharted-labs-inc/>

Even with the statements from Uncharted Labs in the US case it is highly doubtful that this is sufficient to enable other rightsholders to exercise their rights. Other rightsholders would almost certainly also have to generate output indicating that the Udio model has been trained on their content.

Video generating AI models

Case 10: OpenAI's Sora

Transparency provided:

OpenAI's Sora AI video generator model has yet to be released to the public. The company has however released technical reports on the product in which they describe how it was developed.³⁴ Here OpenAI writes:

"We take inspiration from large language models which acquire generalist capabilities by training on internet-scale data."

Additional information:

In an interview on 13 March, 2024 with the CTO of OpenAI, Mira Murati, published by the Wall Street Journal, it is revealed that to train Sora OpenAI *"used publicly available data and licensed data."* When questioned further about whether they used videos from YouTube and Facebook, Murati says she is unsure about it and that she will not go into further detail about the data.³⁵

A New York Times article from April 2024 details how several of the bigger AI companies including OpenAI have scraped content (video transcripts) from YouTube to train their AI models. This not only violates YouTube's terms and

³⁴ <https://openai.com/index/video-generation-models-as-world-simulators/>

³⁵ <https://www.youtube.com/watch?v=mAUpxN-ElgU> (From time: 04:29)

conditions, but it was also done without the permission from rightsholders to the videos.³⁶ Whether OpenAI also took videos to train Sora has not been confirmed.³⁷

Comment on OpenAI's transparency:

The degree of transparency OpenAI provides i.e. that they train on publicly available data is insufficient for rightsholders to determine whether their content has been used to train Sora. There is no list of datasets or a narrative explanation that enables rightsholders to exercise their rights.

Case 11: Runway AI's Gen3-alpha

Transparency provided:

Runway AI's video generator Gen3-alpha was released in JUN 2024. Runway AI writes in a blogpost on their website how Gen3-alpha was trained on both images and videos, but they don't go into further detail.³⁸

Additional information:

It was revealed by 404 Media that Runway AI had scraped thousands of YouTube videos to train their models on, without consent from the rightsholders.³⁹ The information was leaked to 404 Media including an internal Runway AI spreadsheet containing links to several thousand YouTube channels and videos. A former Runway employee told 404 Media, that it was a company-wide effort to compile the YouTube videos into a spreadsheet, which then would be used for training the Gen3-alpha model. While 404 Media were not able to confirm that every single one of the videos was used in training, it clearly shows that Runway were not holding back on scraping copyright protected content to train their models.

³⁶ <https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html?smid=nytcore-ios-share&sgroup=c-cb>

³⁷ <https://www.theverge.com/2024/4/6/24122915/openai-youtube-transcripts-gpt-4-training-data-google>

³⁸ <https://runwayml.com/research/introducing-gen-3-alpha>

³⁹ <https://www.404media.co/runway-ai-image-generator-training-data-youtube/>

The spreadsheet also contained a sheet called “Non-YouTube source” wherein one could locate links to e.g. Kisscartoon.sh known for pirating animated content, a Studio Ghibli Archive and several other websites containing pirated content such as the Watchseries brand. 404 Media were able to recreate videos with Gen3-alpha that very closely resembled videos in the leaked internal spreadsheet, rendering it probable that the videos were indeed used to train the model.

Comment on transparency:

Runway AI provides no official information on the content they have used to train their Gen-3 Alpha model, thus making it impossible for rightsholders to determine if their content was used in the training process.

Image generating AI models

Case 12: Stability AI’s Stable Diffusion

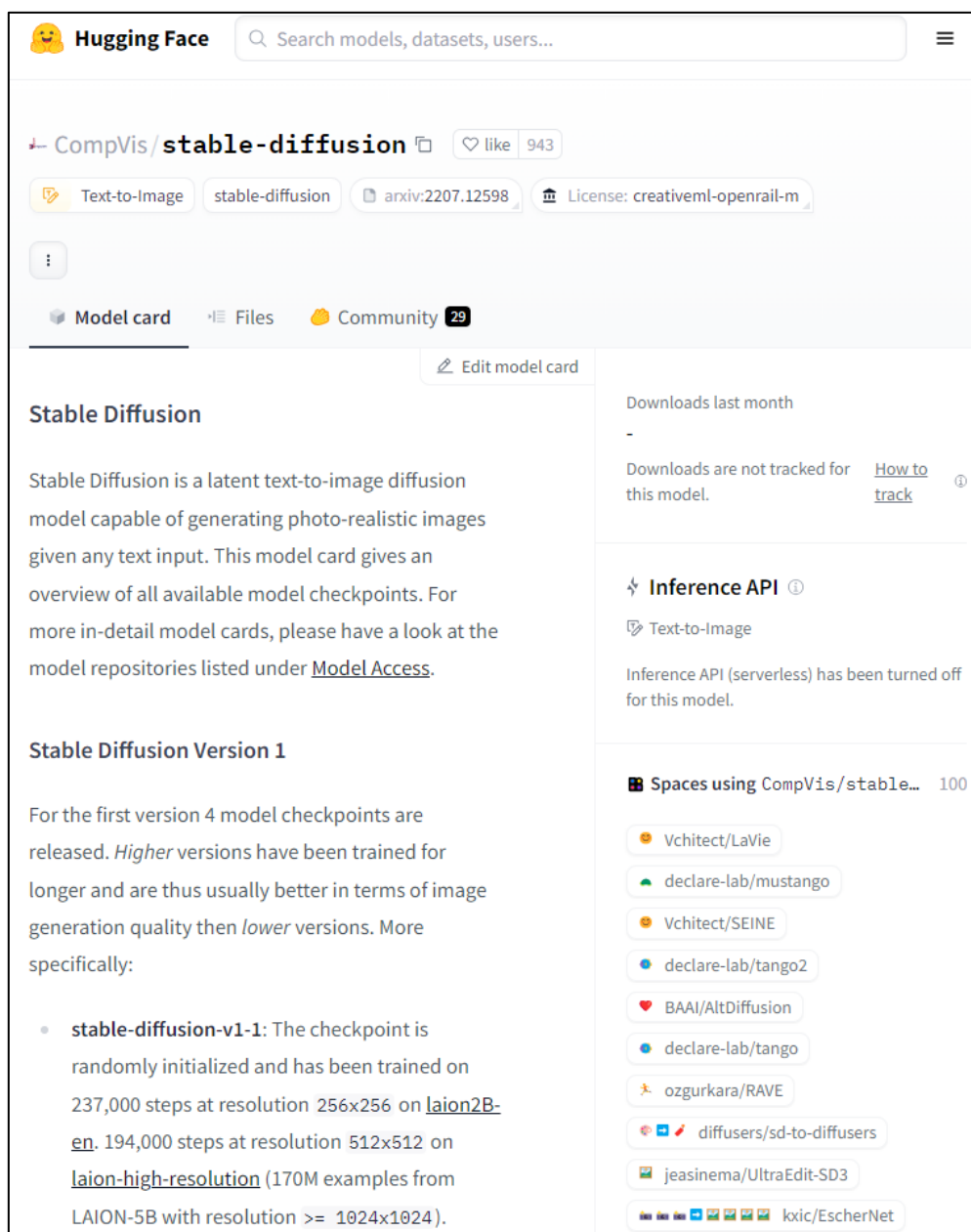
Transparency provided:

Stability AI has used “open source AI” and “open access” to describe its models.

When the Stable Diffusion V1 model was released in 2022⁴⁰ the training data for several iterations of the V1 model was detailed on the model’s model card⁴¹ as seen below:

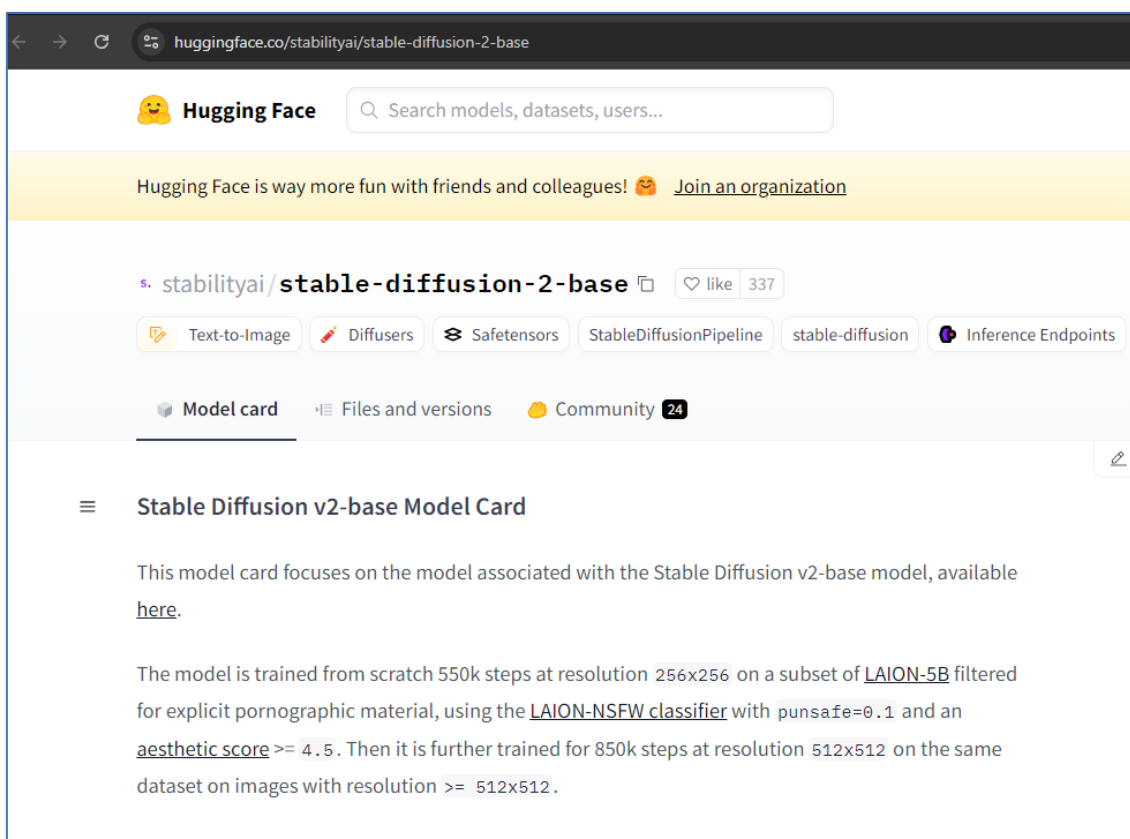
⁴⁰ <https://github.com/CompVis/stable-diffusion>

⁴¹ <https://huggingface.co/CompVis/stable-diffusion>



Screenshot 29 - <https://huggingface.co/CompVis/stable-diffusion>

The Stable Diffusion 2 model, which was released on 24 November 2023, was trained on a subset version of the LAION-5B dataset.



Screenshot 30 - <https://huggingface.co/stabilityai/stable-diffusion-2-base>

LAION is a non-profit organization providing large-scale machine learning models, datasets and related code to the public for free⁴². They have been funded by Stability AI in the development of their datasets⁴³. As shown above Stability AI have used several iterations of LAION's datasets in their previous models. The LAION datasets are publicly available and the LAION-400M and LAION-5B are both curated datasets, which have filtered image-text pairings using filtered Common Crawl data.

⁴² <https://laion.ai/about/>

⁴³ <https://laion.ai/blog/laion-5b/>

In this work, we address this challenge and make multimodal training more accessible by assembling a public dataset that is suitable for training large image-text models. Specifically, we introduce LAION-5B, the largest public image-text dataset containing over 5.8 billion examples (see Table 1 for a comparison). By starting from Common Crawl [1] and filtering this data source with an existing CLIP model, we derive a dataset consisting of three parts: 2.32 billion English image-text examples, 2.26 billion multilingual examples, and 1.27 billion examples that are not specific to a particular language (e.g., places, products, etc.). Beyond assembling the dataset, we also explore its ethical implications and flaws that emerge with large-scale data collection. By releasing LAION-5B publicly, we offer the first opportunity for the community to audit and refine a dataset of this magnitude.

Screenshot 31 - <https://arxiv.org/abs/2210.08402>

Stability AI currently provides two AI Image Generators; Stable Diffusion 3 Medium and Stable Diffusion XL, the former being the newest in their line of image generators.

Stable Diffusion 3 Medium was released on 12 June, 2024. Stability has provided the following about their training dataset:

*"We used synthetic data and filtered publicly available data to train our models. The model was pre-trained on 1 billion images. The fine-tuning data includes 30M high-quality aesthetic images focused on specific visual content and style, as well as 3M preference data images."*⁴⁴

Comment on transparency:

While Stability AI were initially transparent about their training data, they have now chosen to be opaque about what content has been used to train their newer models.

The transparency provided in the V1 model enables rightsholders to determine, whether their works were used in training or not, as LAION makes the datasets

⁴⁴ <https://huggingface.co/stabilityai/stable-diffusion-3-medium>

available to the public for further investigation. In the V2 model Stability AI only provides, that it is a subset variation of LAION-5B, which could make it harder to determine, what specific content was used in the model.

The transparency in the newest models, Stable Diffusion 3 medium and Stable Diffusion XL, does not provide rightsholders with any insight into what was used to train the models and therefore makes it impossible to determine if specific protected content was used in the training process.

Case 13: Black Forest Labs' Flux.1

Transparency provided:

Black Forest Labs released three text-to-image generators on 1 August, 2024.⁴⁵ The team behind the models have previously worked on the Stable Diffusion models and they have received funding from leading US VCs including Andreessen Horowitz. The Flux models are state-of-the-art models.

The three models are:

- FLUX.1 [pro], a paid image generator. No information is disclosed on the training data.
- FLUX.1 [dev], an open-weights model for non-commercial use. In this case being an open-weights model means that the model is free to download on HuggingFace, while not providing any further information on what training was based on or how they got the weights for the model.
- FLUX.1 [schnell], also an open-weights model as described above, tailored for local deployment and personal use. No further information on training data can be located in this model.

⁴⁵ <https://blackforestlabs.ai/announcements/>



X (Twitter) has partnered with Black Forrest Labs to provide X Premium and Premium+ users access to FLUX.1.⁴⁶

Comment on transparency:

Black Forest Labs provides no transparency into training data making it impossible to determine whether specific copyright protected content has been used to train the FLUX.1 models.

⁴⁶ <https://x.ai/blog/grok-2>