REPORT ON PIRATED CONTENT USED IN THE TRAINING OF GENERATIVE AI

RIGHTS ALLIANCE
FOR THE CREATIVE INDUSTRIES ON THE INTERNET

March 2025

Thomas Heldrup
Head of Content Protection & Enforcement / Danish Rights Alliance

Table of contents

Introduction

It is now well understood that the major AI companies have collected and used copies of copyright protected content to train generative AI, such as LLMs, without permission from the relevant rightsholders. But what is becoming more apparent by the day is the prevalence of content sourced from pirate sites in AI training data.

This report focuses on how providers of generative AI models have used copies of copyright protected content sourced from "classic" pirate sources such as illegal filesharing sites and illegal movie streaming sites. The report also describes how generative AI developers obtain datasets containing pirated content.

Datasets from Common Crawl, a US based nonprofit that crawls and scrapes text from the internet, is included in the report despite it not being a pirate source in the

"classic" sense. But Common Crawl never obtained permission from rightsholders to copy, store and distribute the massive amounts of protected content it does including press publications, books and song lyrics. Common Crawl is also one of the most popular sources of training data for LLMs.

Summary of findings

Table of findings – provider, models and selected datasets used for training

Provider	Model	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
Apple	OpenELM	Books3	Opensubtitles	Common Crawl		
Anthropic	Claude	Books3	Opensubtitles	Common Crawl		
DeepSeek	DeepSeek V3	Books3	Opensubtitles	Common Crawl		
DeepSeek	DeepSeek -VL	Anna's Archive				
Meta	Llama	Books3	LibGen	Anna's Archive	Z-lib	Common Crawl
Microsoft	Phi	Books3	Common Crawl			
NVIDIA	Nemo Megatron	Books3	Opensubtitles	Common Crawl		
OpenAI	GPT	LibGen	Common Crawl	YouTube transcripts		
Runway AI	Gen3-alpha	Watchseries streaming sites	Kisscartoon.sh	YouTube		
Suno Inc	Suno	"tens of millions of recordings" which "includes essentially all music files of reasonable quality that are accessible on the open Internet"				

As described in detail below, generative AI model providers across the spectrum of LLMs, image, video and music generation have obtained infringing copies of copyright protected works that were sourced from what I call "classic pirate sources" such as illegal filesharing and streaming sites.

Table of findings – content type in selected datasets

Dataset	Content type
Anna's Archive	books
Books3	books
CommonCrawl	text from websites incl. press publications and song lyrics
Kisscartoon.sh	cartoons (video)
LibGen	books
OpenSubtitles	subtitles to movies and tv
Suno dataset	recorded music
Watchseries	movies and tv
YouTube	video
Z-lib	books

The datasets used by the AI model providers contain content spanning books, press publications, song lyrics, movie and tv subtitles, recorded music, movies and tv.

Table of findings – company behind selected dataset, sub-datasets and distribution

Dataset	Company	Subdataset	Subdataset2	Subdataset3	Distribution
Redpajama	Together.ai	Books3	Common Crawl		HuggingFace
Slimpajama	Cerebras	Books3	Common Crawl		HuggingFace
The Pile	EleutherAI	Books3	OpenSubtitles	Common Crawl	Torrent
LibGen	unknown				Torrent
Anna's Archive	unknown				Torrent
Common Crawl	The Common Crawl Foundation				Direct download, other datasets
YouTube	Runway AI				Internal dataset made with stream ripping software
Kisscartoon	Runway AI				Internal dataset made with stream ripping software
Watchseries	Runway AI				Internal dataset made with stream ripping software
YouTube transcripts	OpenAI				Internal dataset made with Whisper transcription model
Z-lib	unknown				Torrent

This report also shows how AI model providers have used publicly available datasets compiled by third parties that contain infringing copies of copyright protected works from illegal sources. These third-party datasets are distributed in many ways ranging from AI focused user-generated sharing platforms such as

HuggingFace and Kaggle, but also via Torrent filesharing or simply sharing directly by persons via online messaging and chat services such as Discord.

The wide range of dataset distribution methods reveals how AI model providers don't always engage in direct or commercial negotiations and agreements with the providers of AI training datasets. In many cases they obtain datasets from user-generate platforms such as HuggingFace or via downloading of torrent files without any interaction with the company compiling the dataset.

Another finding in the report shows how some AI model providers will copy protected works from platforms such as YouTube, Netflix and other streaming platforms violating both the terms and conditions of those platforms, but more critical also the copyrights of the rightsholders whose content is available on the platforms. The AI providers use software called stream ripping that circumvents DRM measures applied to the platforms.

Datasets

One of the most popular datasets used by AI providers containing infringing copies of copyright protected works is "**The Pile**"[1]. This dataset was compiled by the US based nonprofit Eleuther AI and distributed on servers belonging to a group of data hoarders calling themselves The-Eye. Books3 was later distributed on HuggingFace, BitTorrent and servers belonging to individuals who wanted to continue spreading the dataset. The datasets Books3 and OpenSubtitles are sub-datasets in The Pile together with a host of other subdatasets.

Books3 contains 196.640 files containing the plain text of books. The books in Books3 originated from the illegal filesharing site Bibliotik.me, a notorious pirate site, also known as a shadow library, dedicated to sharing illegal copies of books.[2] The

[1] The Pile: An 800GB Dataset of Diverse Text for Language Modeling https://arxiv.org/pdf/2101.00027
[2] https://x.com/theshawwn/status/1320282149329784833?lang=en

dataset is now distributed via BitTorrent and by individuals on various online platforms and servers.

OpenSubtitles contains files with plain text subtitles to movies and tv-shows. The dataset compiled by J. Tiedemann sourced the subtitles from OpenSubtitles.org[3], which is a pirate site where users upload and share infringing copies of subtitles. The dataset contains almost all released movies and tv-shows up until 2016. The dataset is now distributed on HuggingFace and other AI sharing platforms and by the creator on his website.

The Pile also contains data from **Common Crawl**. Common Crawl (CC) is a US based nonprofit that crawls the internet creating copies of all text found on the websites it visits.[4] CC then stores the text on servers donated by Amazon Web Services and lets anyone download the datasets for free. CC does not ask permission of rightsholders to copy, store and distribute their copyright protected content. Instead, CC requires website owners to implement an opt-out of its copying etc. by expressing this in a robots.txt file on the website disallowing CC's crawler CCBot. The content in CC's datasets include press publications, books and song lyrics. Besides being distributed by CC itself, the Common Crawl datasets are also available in countless variants on HuggingFace including in the datasets Redpajama and Slimpajama.

Redpajama was compiled by a US based company Together.AI.[5] Books3 was a sub dataset in Redpajama including many parts of Common Crawl datasets. Redpajama also included C4 as a subdataset, which was collected by Google and contains many parts of Common Crawl's datasets. Redpajama is distributed on HuggingFace and Kaggle.

---

[3] https://aclanthology.org/L16-1559.pdf
[4] https://commoncrawl.org
[5] https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T later updated with a version 2 https://www.together.ai/blog/redpajama-data-v2

Slimpajama was collected by the US company Cerebras and contains a filtered version of the Redpajama dataset including Books3 and Common Crawl datasets.[6] Slimpajama is distributed on HuggingFace, Kaggle and GitHub.

LibGen (short for Library Genisis) is a classic pirate filesharing site dedicated to sharing illegal copies of books.[7] LibGen was "forked" and is now available in numerous versions run by different people. BitTorrent technology is the distribution method used to obtain the illegal book copies available on LibGen.

Z-lib (short for Z-library) is a classic pirate filesharing site dedicated to sharing illegal copies of books that has also been the target of the FBI resulting in multiple domains related to the service being seized in recent years.[8]

Anna's Archive is also a classic pirate filesharing site dedicated to sharing illegal copies of books. Anna's Archive works as an aggregator of shadow libraries such as LibGen and Z-Library that together contain millions of illegal copies of e-books. Anna's Archive has realized that their illegal book datasets are in high demand from developers of LLMs and they have a whole page on their website offering help to access the datasets for LLM training. BitTorrent technology is the distribution method used to obtain the illegal book copies available on Anna's Archive.

Apple

Apple has developed and provides access to a range of LLMs called OpenELM.[9]

For pre-training of the model, Apple used *The Pile* (contains Books3, OpenSubtitles and Common Crawl) and *Redpajama* (contains Books3, Common Crawl).

---

[6] https://cerebras.ai/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama
[7] https://en.wikipedia.org/wiki/Library_Genesis
[8] https://techhq.com/2023/11/how-is-z-library-down-again-alternatives-ebooks/
[9] OpenELM: An Efficient Language Model Family with Open Training and Inference Framework https://arxiv.org/pdf/2404.14619

Source	Subset	Tokens
RefinedWeb		665 B
RedPajama	Github	59 B
	Books	26 B
	ArXiv	28 B
	Wikipedia	24 B
	StackExchange	20 B
	C4	175 B
PILE		207 B
Dolma	The Stack	411 B
	Reddit	89 B
	PeS2o	70 B
	Project Gutenberg	6 B
	Wikipedia + Wikibooks	4.3 B

Table 2. Dataset used for pre-training OpenELM.

Screenshot 1 - Table from OpenELM: An Efficient Language Model Family with Open Training and Inference Framework https://arxiv.org/pdf/2404.14619

Anthropic

Anthropic has developed and provides access to a range of LLMs called Claude. In a research paper by Anthropic titled *A General Language Assistant as a Laboratory for Alignment* they describe how their language models are trained on The Pile dataset (contains Books3, OpenSubtitles and Common Crawl).[10]

> 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 2 – text from A General Language Assistant as a Laboratory for Alignment

Anthropic's use of The Pile as training data for Claude was confirmed in a WIRED article:

"The Pile includes a very small subset of YouTube subtitles," Jennifer Martinez, a spokesperson for Anthropic, said in a statement confirming use of the Pile in Anthropic's generative AI assistant Claude. "YouTube's terms cover direct use of its platform, which is distinct from use of the Pile dataset. On the point about potential violations of YouTube's terms of service, we'd have to refer you to the Pile authors."[11]

[10] *A General Language Assistant as a Laboratory for Alignment* https://arxiv.org/pdf/2112.00861
[11] https://www.wired.com/story/youtube-training-data-apple-nvidia-anthropic/

DeepSeek

DeepSeek has developed and provides access to a range of LLMs called DeepSeek.

The DeepSeek models are trained on The Pile dataset (contains Books3, OpenSubtitles and Common Crawl) and Redpajama (contains Books3 and Common Crawl).[12]

DeepSeek has developed and provides access to a Vision-Language called DeepSeek-VL.[13] This model is trained on more than 1 million illegal copies of e-books sourced from Anna's Archive.

[12] DeepSeek LLM Scaling Open-Source Language Models with Longtermism https://arxiv.org/pdf/2401.02954.
[13] DeepSeek-VL: Towards Real-World Vision-Language Understanding https://arxiv.org/pdf/2403.05525

Meta

Meta has developed and provides access to a range of LLMs called Llama.

Meta has published a research paper *LLaMA: Open and Efficient Foundation Language Models*[14] describing how they used the Books3 dataset to train their model.

In the class action lawsuit filed against Meta in the United States District Court of California by authors Richard Kadrey et.al.[15] Meta has admitted to using Books3 to train its Lama models 1-3.

> Plaintiffs' demands for just the alleged "Shadow Datasets" would be incredibly burdensome exercise that would provide minimal if any relevant information for the issues in this case, as Plaintiffs' works, to the extent they are included in the training data, are a miniscule portion of the overall training data set, and Meta has already admitted that text from each of the Plaintiffs' books was included in the Books3 dataset used to train Meta's Llama models. (*See* amended responses to RFAs 3–6, ECF 352, Ex. A.) Meta has also already produced documentation to Plaintiffs that identifies the datasets that are being used for ongoing Llama 4 training. Llama 5 remains at early stages of planning and it is not yet known what datasets will be used for training at this time.

Screenshot 3 – Meta's statement page 6 of 9
https://storage.courtlistener.com/recap/gov.uscourts.cand.415175/gov.uscourts.cand.415175.443.0.pdf

[14] *LLaMA: Open and Efficient Foundation Language Models*
https://arxiv.org/abs/2302.13971
[15] Kadrey v. Meta Platforms, Inc. (3:23-cv-03417)
https://www.courtlistener.com/docket/67569326/kadrey-v-meta-platforms-inc/

Court documents also show how Meta used LibGen, Z-Lib and Anna's Archive to create training datasets for an upcoming 4th version of the Llama model.[16]

Case 3:23-cv-03417-VC Document 377 Filed 01/08/25 Page 2 of 13

With just two hours left before the fact discovery cut-off on Friday, December 13, 2024, Meta produced some of the most incriminating internal documents it has produced to date relevant to Meta's copyright infringement claim and fair use defense, as well as Plaintiffs' proposed new claims. These documents concern Meta's torrenting and processing of pirated copyrighted works, including that: Meta's CEO, Mark Zuckerberg, approved Meta's use of the LibGen dataset notwithstanding concerns within Meta's AI executive team (and others at Meta) that LibGen is "a dataset we know to be pirated," Stein Reply Decl. ("Reply Ex."), Ex. A at 211699, 211702; top Meta engineers discussed accessing and reviewing LibGen data but hesitated to get started because "torrenting from a [Meta-owned] corporate laptop doesn't feel right 😃," Reply Ex. B at 204224; one of those engineers "filtered . . . copyright lines" and other data out of LibGen to prepare a CMI-stripped version of it to train Llama, Reply Ex. C at 204220-21; and, by January 2024, Meta had already torrented (both downloaded and distributed) data from LibGen, Reply Ex. D.[1] And just yesterday, when asked about the type of piracy described in the TACC, Mr. Zuckerberg testified that such activity would raise "lots of red flags" and "seems like a bad thing." Reply Ex. E (Zuckerberg Dep. Tr.) at 102:10–14; 98:24–99:2.

Screenshot 4 - https://storage.courtlistener.com/recap/gov.uscourts.cand.415175/gov.uscourts.cand.415175.423.0_1.pdf

[16] https://chatgptiseatingtheworld.com/wp-content/uploads/2025/01/Unredacted-Reply-of-Plaintiffs-1.pdf and https://storage.courtlistener.com/recap/gov.uscourts.cand.415175/gov.uscourts.cand.415175.423.0_1.pdf and https://storage.courtlistener.com/recap/gov.uscourts.cand.415175/gov.uscourts.cand.415175.443.0.pdf

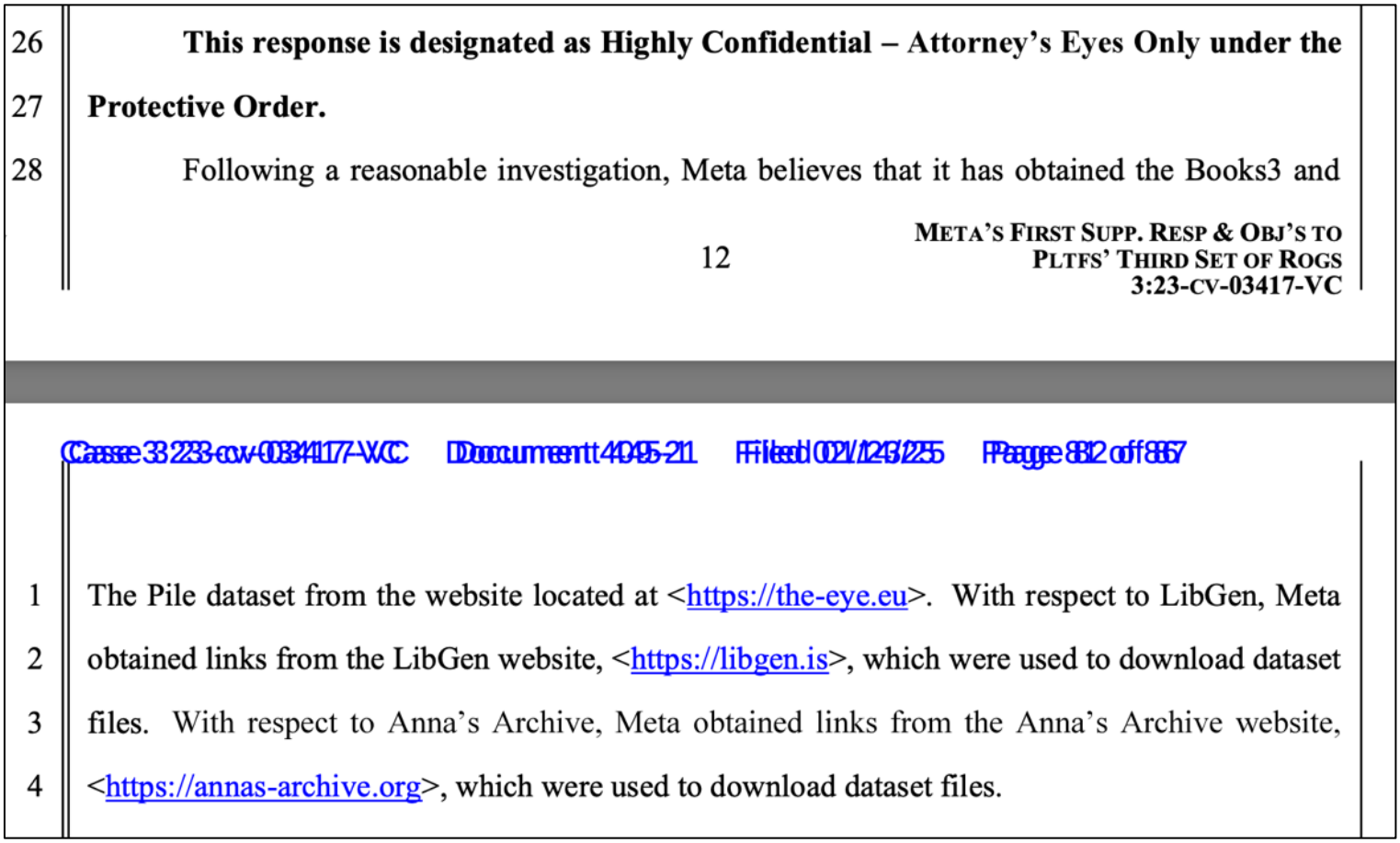
BSF

weeks ago before the Court adopted Meta's proposed amended case schedule. This discovery is needed—while Meta continued to argue late into fact discovery (and well after the deadline to serve additional discovery requests) that "[t]here is nothing about Meta's efforts with respect to Llama 4 that make any fact in dispute more or less likely," Dkt. 267 at 12, it later turned out that Meta trained that very model by torrenting massive quantities of copyrighted works. Dkt. 417-11 (April 2024 Meta WorkChat showing a minimum of *81.7 TB* torrented from Anna's Archive for Llama 4).

Screenshot 5 -
https://storage.courtlistener.com/recap/gov.uscourts.cand.415175/gov.uscourts.cand.415175.423.0_1.pdf

The problem is that Meta's witnesses have testified that additional "Shadow Datasets" were used as Llama 4 training data, and those datasets are not reflected in Llamas 1-3. For example, the evidence makes clear that over the last several months, Meta has increased its reliance on Shadow Datasets, including the notorious "Z-Lib" Shadow Dataset, which had numerous domains seized by the FBI in recent years.[1] Moreover, Meta has begun to source new copyrighted works for its LibGen dataset through the website Anna's Archive. Yet, Meta has not produced any of this data, and Plaintiffs still do not possess the full Llama 4 (or Llama 5) training datasets despite their relevance.

Screenshot 6 -
https://storage.courtlistener.com/recap/gov.uscourts.cand.415175/gov.uscourts.cand.415175.443.0.pdf



26 **This response is designated as Highly Confidential – Attorney's Eyes Only under the**
27 **Protective Order.**
28 Following a reasonable investigation, Meta believes that it has obtained the Books3 and

1 The Pile dataset from the website located at <https://the-eye.eu>. With respect to LibGen, Meta
2 obtained links from the LibGen website, <https://libgen.is>, which were used to download dataset
3 files. With respect to Anna's Archive, Meta obtained links from the Anna's Archive website,
4 <https://annas-archive.org>, which were used to download dataset files.

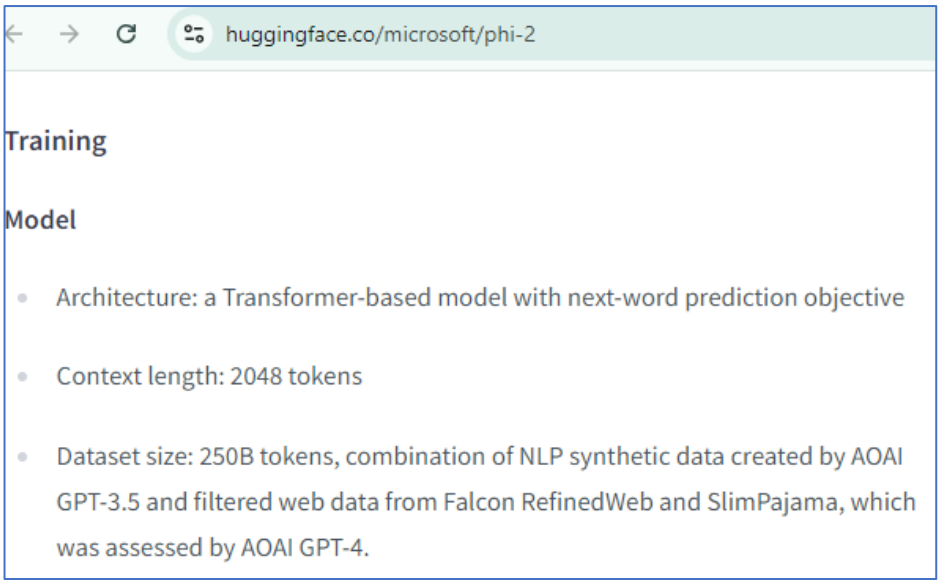
Screenshot 7 -
https://storage.courtlistener.com/recap/gov.uscourts.cand.415175/gov.uscourts.cand.415175.445.1.pdf

Microsoft

Microsoft has developed and provides access to a range of LLMs called Phi.

Looking at Phi version 2 Microsoft writes that they used the dataset Slimpajama (contains Books 3 and Common Crawl) to train their model.[17]

[17] https://huggingface.co/microsoft/phi-2



Screenshot 8 - https://huggingface.co/microsoft/phi-2

NVIDIA

Nvidia has developed and provides access to a range of LLMs, VLM and video models called Nvidia NeMo.

On the system card for NeMo Megatron-GPT 1.3B Nvidia writes that it has used The Pile (contains Books3, OpenSubtitles, Common Crawl) to train the model.[18]

OpenAI

OpenAI has developed and provided access to a range of LLMs called GPT.

In the GPT-3 paper OpenAI writes how it 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.[19]

[18] https://huggingface.co/nvidia/nemo-megatron-gpt-1.3B
[19] *Language Models are Few-Shot Learners* https://arxiv.org/abs/2005.14165

In the class action lawsuit filed against OpenAI in the United States District Court, N.D. California by authors Tremblay et.al[20] it has been revealed how OpenAI used datasets sourced from LibGen to train its GPT models.[21]

A New York Times article describes how OpenAI created a speech recognition tool called Whisper to transcribe audio from YouTube videos that was then used to train the GPT-4 model.[22]

Runway AI

Runway AI has developed and provided access to a video generating AI model called Gen3-alpha.

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.[23] 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.

---

[20] Tremblay v. OpenAI, Inc. (3:23-cv-03223) https://www.courtlistener.com/docket/67538258/tremblay-v-openai-inc/?filed_after=&filed_before=&entry_gte=&entry_lte=&order_by=desc

[21] https://storage.courtlistener.com/recap/gov.uscourts.cand.414822/gov.uscourts.cand.414822.254.0.pdf

[22] https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html

[23] 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, making it probable that the videos were indeed used to train the model.

Suno Inc

Suno has developed and provides access to a music generating AI model called Suno.

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*".[24]

There are only two likely options for Suno to obtain the "tens of millions of recordings" that are in their training dataset. The first would be to use classic pirate sites to download the recordings directly from cyberlockers or via BitTorrent technology. The second way would be to use stream ripper technology to circumvent the DRM protection mechanisms on streaming platforms making it possible to copy the recordings directly from the streaming platforms.

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