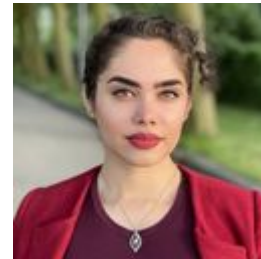


An Exploratory Investigation into Code License Infringements in Large Language Model Training Datasets

Jonathan Katzy, Razvan-Mihai Popescu, Maliheh Izadi, Arie van Deursen



Who are we?

AI Enabled Software Engineering Lab
AISE

Software Engineering Research Group
SERG

Jonathan Katzy, Razvan-Mihai Popescu, Maliheh Izadi, Arie van Deursen



Leading question

Are permissively licensed datasets
permissively licensed?

Context

- Lawsuits
 - The Pile (Books3)
 - Getty Images vs. Stable Diffusion
 - The New York Times vs. OpenAI

Context

- Lawsuits
- Issues
 - For profit use of copyrighted data
 - Outputs that can harm data holders
 - Memorization of data

Context

- Lawsuits
- Issues
- Claims
 - Damages and lost revenue
 - Deletion of datasets and models

Research questions

- Is there interest in permissively licensed code datasets?
- Are there traces of strong copyleft licenses in publicly available datasets?
- Is other sensitive information included in public code datasets?

Approach

- Gather literature surveys

Approach

- Gather literature surveys
- Extract models

Approach

- Gather literature surveys
- Extract models
- **Extract and collect datasets**

Approach

- Gather literature surveys
- Extract models
- Extract and collect datasets

RQ1: Is there interest in permissively licensed code datasets?

Approach

- Gather literature surveys
- Extract models
- Extract and collect datasets
- Collect strong copyleft licensed code

GPL 2.0, GPL 3.0, AGPL

Approach

- Gather literature surveys
- Extract models
- Extract and collect datasets
- Collect strong copyleft licensed code
- Compare overlap between licensed code and dataset

Calculate SHA-256 hash of all files, from our collected dataset as well as publicly available datasets.

Approach

- Gather literature surveys
- Extract models
- Extract and collect datasets
- Collect strong copyleft licensed code
- Compare overlap between licensed code and dataset

RQ2.1: Are there traces of strong copyleft licenses in publicly available datasets?

Approach

- Compare overlap between licensed code and dataset
- Extract first comment

Regex search for any comment block, or multiline comment that starts in the first 20 characters

Approach

- Compare overlap between licensed code and dataset
- Extract first comment
- Search for licenses

Regex search for language referring to GPL 2.0, GPL 3.0, and AGPL licenses

Approach

- Compare overlap between licensed code and dataset
- Extract first comment
- Search for licenses

RQ2.2: Are there traces of strong copyleft licenses in publicly available datasets?

Approach

- Compare overlap between licensed code and dataset
- Extract first comment
- Search for licenses
- Search for distribution intent

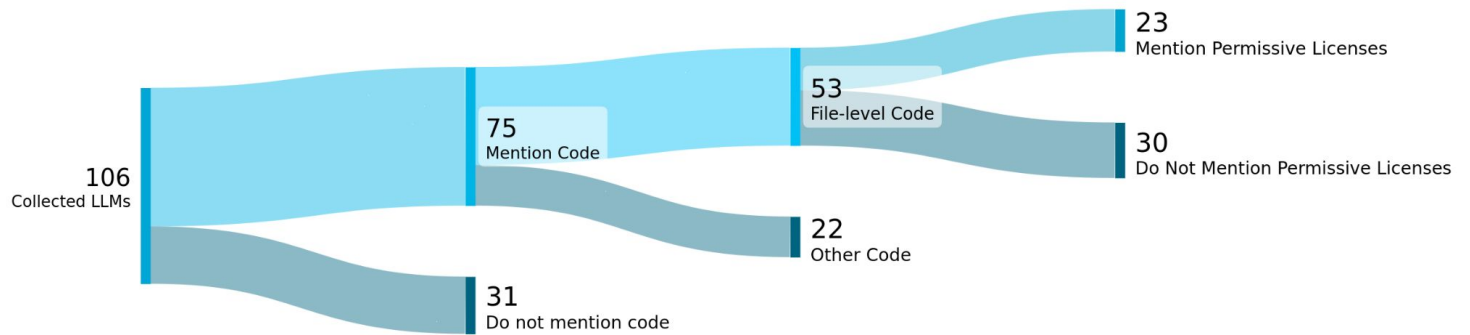
Regex search for terms such as “confidential”, “do not share”, etc...

Approach

- Compare overlap between licensed code and dataset
- Extract first comment
- Search for licenses
- Search for distribution intent

RQ3: Is other sensitive information included in public code datasets?

Results - Study collection



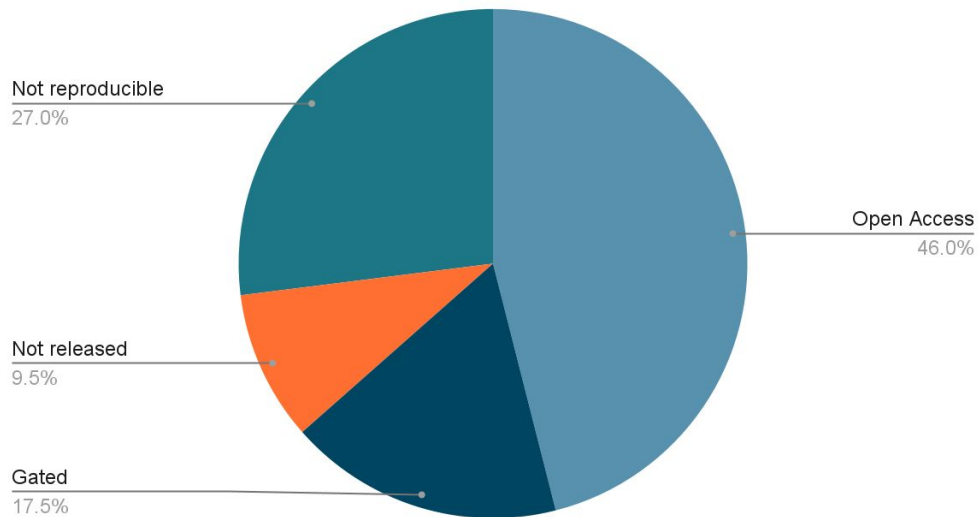
Results - Study collection

Table 3: File-Level Code datasets used for training foundational models

Ref	Dataset	Available	Count
1	Big Query	Pay-wall	10
2	The Pile	DMCA-takedown	12
3	The Stack v1	Open	8
4	RedPajama	Open	3
5	CodeParrot	Open	2
6	PaLM Dataset	Not Released	3
7	Roots	Not Open to All	1
8	SkyPile	Not Released	1
9	BigPython	Not Released	2
10	MassiveText	Not Released	1
11	GitHub-Code Dataset	Open	3
12	CodeClippy Dataset	Open	1
13	ExtraPythonData	Not Released	1
14	Code LLaMa Dataset	Not Released	1
15	Custom Dataset	Not Released	17

Results - Study collection

Availability of training data for LLMs



Results - RQ1

Is there interest in permissively licensed code datasets?

Results - RQ1

- Code is used more frequently in training

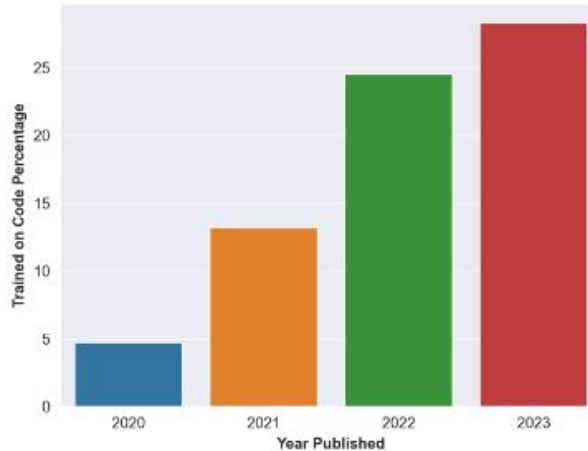


Figure 2: Percentage of LLMs trained on code per year over the total number of LLMs

Results - RQ1

- Interest in code dataset licensing growing fast

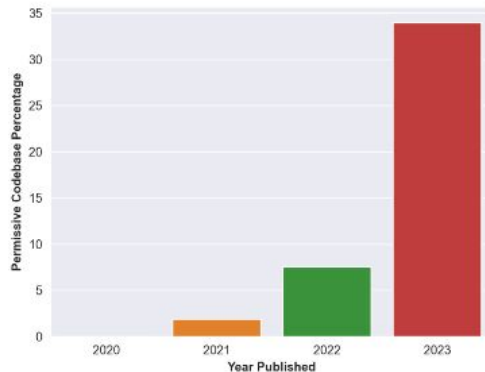


Figure 3: Percentage of LLMs trained on permissive file code per year over the total number of LLMs trained on file level code

Results - RQ2

Are there traces of strong copyleft licenses in publicly available datasets?

Results - RQ2

- All datasets had exact duplicate of code associated with a strong copyleft license

Table 5: Amount of code files found to be associated with a strong copyleft license

Dataset	Files	Exact Duplicates		License Comments	
		Count	Percentage	Count	Percentage
The Stack v1	262,678,972	16,122,976	6.14%	2,067,830	0.78%
RedPajama	28,793,312	1,579,521	5.49%	15,544	0.05%
The Pile	18,044,000	4,113,263	22.80%	823,546	4.56%
CodeParrot	18,695,559	2,681,590	14.34%	2,844,150	15.21%
GitHub-Code	115,086,922	5,537,734	4.81%	7,548,615	6.56%
CodeClippy	71,140,482	7,993,768	11.24%	2,823,923	3.97%

Results - RQ2

- All datasets had comments referencing a strong copyleft license

Table 5: Amount of code files found to be associated with a strong copyleft license

Dataset	Files	Exact Duplicates		License Comments	
		Count	Percentage	Count	Percentage
The Stack v1	262,678,972	16,122,976	6.14%	2,067,830	0.78%
RedPajama	28,793,312	1,579,521	5.49%	15,544	0.05%
The Pile	18,044,000	4,113,263	22.80%	823,546	4.56%
CodeParrot	18,695,559	2,681,590	14.34%	2,844,150	15.21%
GitHub-Code	115,086,922	5,537,734	4.81%	7,548,615	6.56%
CodeClippy	71,140,482	7,993,768	11.24%	2,823,923	3.97%

Results - RQ3

Is other sensitive information included in public code datasets?

Results - RQ3

- There is more information than just licenses in code comments

```
1 <Company> all rights reserved.  
2 this software contains proprietary and confidential  
3 information of <Company> and its contributors.  
4 use, disclosure and reproduction is prohibited without  
5 prior consent.
```

Figure 4: Restrictions on sharing and distributing code contained in a file, extracted from the *RedPajama* dataset

Table 6: Amount of code files found to be associated with some form of ownership/copyright disclaimer

Dataset	Copyright		First Comments
	Count	Percentage	
The Stack v1	5,073,823	6.54%	77,595,559
RedPajama	30,500	1.34%	2,281,378
ThePile	501,877	7.39%	6,794,995
CodeParrot	773,062	5.38%	14,372,397
GitHub-Code	2,669,845	5.89%	45,301,797
CodeClippy	1,695,556	6.72%	25,223,157

Conclusion

- Checking repo licenses is not enough
- More work needed
- Build on existing works

Implications

- No datasets is free of code licenses inconsistencies
- All models could output licensed code

Implications

- No datasets is free of code licenses inconsistencies
- All models could output licensed code

Who is responsible for what part of the LLM training pipeline?

Questions?



JKatzy.nl



J.B.Katzy@TUDelft.nl



@katzy_jonathan



jkatzy

