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

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

Availability of training data for LLMs

- Open Access: 46.0%
- Not reproducible: 27.0%
- Not released: 9.5%
- Gated: 17.5%
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
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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Files</th>
<th>Exact Duplicates</th>
<th>License Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Count</td>
<td>Percentage</td>
</tr>
<tr>
<td>The Stack v1</td>
<td>262,678,972</td>
<td>16,122,976</td>
<td>6.14%</td>
</tr>
<tr>
<td>RedPajama</td>
<td>28,793,312</td>
<td>1,579,521</td>
<td>5.49%</td>
</tr>
<tr>
<td>The Pile</td>
<td>18,044,000</td>
<td>4,113,263</td>
<td>22.80%</td>
</tr>
<tr>
<td>CodeParrot</td>
<td>18,695,559</td>
<td>2,681,590</td>
<td>14.34%</td>
</tr>
<tr>
<td>GitHub-Code</td>
<td>115,086,922</td>
<td>5,537,734</td>
<td>4.81%</td>
</tr>
<tr>
<td>CodeClippy</td>
<td>71,140,482</td>
<td>7,993,768</td>
<td>11.24%</td>
</tr>
</tbody>
</table>

Results - RQ2

- All datasets had comments referencing a strong copyleft license

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Files</th>
<th>Exact Duplicates</th>
<th>License Comments</th>
<th>License Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Count</td>
<td>Percentage</td>
<td>Count</td>
</tr>
<tr>
<td>The Stack v1</td>
<td>262,678,972</td>
<td>16,122,976</td>
<td>6.14%</td>
<td>2,067,830</td>
</tr>
<tr>
<td>RedPajama</td>
<td>28,793,312</td>
<td>1,579,521</td>
<td>5.49%</td>
<td>15,544</td>
</tr>
<tr>
<td>The Pile</td>
<td>18,044,000</td>
<td>4,113,263</td>
<td>22.80%</td>
<td>823,546</td>
</tr>
<tr>
<td>CodeParrot</td>
<td>18,695,559</td>
<td>2,681,590</td>
<td>14.34%</td>
<td>2,844,150</td>
</tr>
<tr>
<td>GitHub-Code</td>
<td>115,086,922</td>
<td>5,537,734</td>
<td>4.81%</td>
<td>7,548,615</td>
</tr>
<tr>
<td>CodeClippy</td>
<td>71,140,482</td>
<td>7,993,768</td>
<td>11.24%</td>
<td>2,823,923</td>
</tr>
</tbody>
</table>
Results - RQ3

Is other sensitive information included in public code datasets?
Results - RQ3

- There is more information than just licenses in code comments

![Figure 4: Restrictions on sharing and distributing code contained in a file, extracted from the RedPajama dataset](image)

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