Exhibit 44

eli Oct 27, 2022 10:53 AM

https://anthropic.slack.com/archives/C01HP5FJGM7/p1666893113444629

dario Replied on Oct 27, 2022 10:59 AM

Seems pretty interesting

Zac Replied on Oct 27, 2022 11:05 AM

If the recent copyright drama is legally well-founded, this doesn't actually help - even most of these permissive licenses require attribution!

Zac Replied on Oct 27, 2022 11:05 AM

Though IMO if we had to, we could probably bootstrap entirely off the CC0, MIT-0, Unlicense subset and a large dose of self-play; it'd just be a bit more expensive.

eli Replied on Oct 27, 2022 11:12 AM

imo this definitely helps, in that it sets a norm for containing provenance information and filtering based on license. it fixes the issue of trying to argue the distribution of the dataset itself is legal, because it's actually following the license terms and doesn't rely on the fair use argument. you're right that it doesn't help on the actual training side, because most models don't have a way to give attribution for the snippets they generate, but it does pave the way for more work towards the chain of custody suggestion

this is different and imo better/more defensible copyright-wise than e.g. laion which does not distribute or copy the image data, but which only provides an index of links. this dataset is an actual release of data rather than indexes (which can bitrot) and the distribution abides by copyright licenses

+1 1 1

dario Replied on Oct 27, 2022 11:18 AM

Also, if the issue is outputting copyrighted data verbatim (rather than training on it), there are various measures that can be taken against that (like filtering on the output)

+1 1

dawn Replied on Oct 27, 2022 11:17 AM

I wonder how they define "near deduplication" (which for them only filters out half the data)

filtering from 28TB down to 3TB based on licenses seems sad

eli Replied on Oct 27, 2022 11:21 AM

Near-deduplication was implemented in the pre-processing pipeline on top of exact deduplication. To find near-duplicates, MinHash with 256 permutations of all documents was computed in linear time. Locality Sensitive Hashing was used to find the clusters of duplicates. Jaccard Similarities were computed inside these clusters to remove any false positives and with a similarity threshold of 0.85. Roughly 40% of permissively licensed files were (near-)duplicates. See section 3 of the paper for further details.
The following are not stored:
Files that cannot contribute to training code: binary, empty, could not be decoded
Files larger than 1MB
The excluded file extensions are listed in Appendix B of the paper.
https://huggingface.co/datasets/bigcode/the-stack
not sure where "the paper" is

tristan Replied on Oct 27, 2022 11:23 AM

Note that code without a license is unusually likely to be bad code, so it may not be as sad as you'd think

100 1

Zac Replied on Oct 27 2022 11:29 AM

Yeah, strong agree there.

Zac Oct 27 2022 11:02 AM

replied to a thread:

https://anthropic.slack.com/archives/C01HP5FJGM7/p1666893113444629

If the recent copyright drama is legally well-founded, this doesn't actually help - even most of these permissive licenses require attribution!