Homework 1

Your Name

Due date: September 11 at 11:59 PM

Question 1.1

Project Gutenberg and C4

1.1.A

For Project Gutenberg, the ”knowledge cutoff” is Mar 2023, and Apr 2019 for C4.

1.1.B

The data of Project Gutenberg sourced from 75,000 free eBooks, and data source of C4 is English-language text sourced from the public Common Crawl web scrape.

1.1.C

T5 was trained primarily on C4. GPT-Neo was trained on The Pile, which includes Project Gutenberg.

1.1.D

For C4, it was released under the license ODC-BY(Open Data Commons License Attribution family). Not all documents have the same copyright protections, like NYT news article(copyrighted) and U.S. government website page(public domain) hve different protections.

For Project Gutenberg, it has its own Project Gutenberg license. Not all documents have the same protection. Like shakespeare’s works don’t have copyright while some eBook is marked as copyrighted in its header.

1.1.E

For C4, it might perform poorly in mathematical reasoning problems, cause it’s sourced from filtered web pages like news and blogs. So it lack high quality reasoning example in the dataset to train on.

For project Gutenberg, it might perform poorly at answering some latest factual question like things happened recently or what’s the latest iPhone model. The reasons behind is that the dataset doesn’t include the most recent information so that you are unable to answer such questions.

Question 1.2

1.2.A

1. Anthropic Claude 3 family.
2. DeepSeek-R1.
3. BigScience BLOOM.

1.2.B

Claude 3 family were trained on a mixture of publicly available internet text, data from labeling services and synthetic data generated internally. It's not that clearly how the data were processed.

DeepSeek-R1's model card emphasizes the use of reinforcement learning and synthetic reasoning traces, including domains like math, code and logic. But the sources and how it was trained on remain unclear.

BigScience BLOOM was trained on ROOTS corpus: A 1.6TB Composite Multilingual Dataset, which includes 46 natural languages and 13 programming languages. The data were tokenized by BPE algorithm with a simple pre-tokenization rule without normalization.

1.2.C

One use case could be that when you are faced with strict regulation or legal compliance that you might need to resort to models trained on publicly accessible data.

Question 2.1

2.1.A

6368 pages in total.

2.1.B

Table tags are dropped and only the cell text remains. Code inside tags becomes plain text.

2.1.C

Headers are converted to plain text with no indication of level or hierarchy, and everything inside $\langle img \rangle$ get removed.

2.1.D

CommonCrawl's WET extraction removes scripts/styles, normalize white spaces. And ours is a naive get_text() plus simple pii masking that preserves javascript/css. for llm training, CC's cleaner extraction is generally better.

Question 2.2

2.2.A

2709 passed out of 6368 records processed. That being said, 1.27 billion documents in the Common Crawl would be considered low-quality.

2.2.B

Two likely low-quality cases that slip through: (1) boilerplate navigation text like "Home — Contact — About" that has punctuation but little meaning, (2) spammy ad text padded with punctuation to evade filters.
Extra heuristics like dictionary coverage might work but training a classifier would be more effective because it can learn subtle signals beyond hard rules.

2.2.C

Non-English text often fails the heuristics, like punctuation differences, or bad-word mismatches. So many valid non-English documents may be dropped. This biases the final dataset toward English and languages with Latin-script punctuation.

2.2.D

A data cleaning stage you haven't implemented yet is near-duplicate document detection across the corpus.

Question 2.3

2.3.A

Yes, my deduplicate code would incorrectly merge files that are only partial similar. File storage system requires exact match.

2.3.B

Deduplication might cause some purposely repetitive datasets removed/simplified, and increase the perplexity of the model.(arXiv:2107.06499)

Question 2.4

2.4.A

138 seconds. it would take 720 days to process the entire CommonCrawl.

2.4.B

1. Using multi-thread, multi-processor or distributed framework the parallelize the process. 2. Replace with a faster library.

2.4.C

1. Tulu SFT data have chat or instruction format with role tags, like ”system” or ”user”. advantage: it aligns with chat models.
2. Yes both training and validation dataset.

Question 3.1

I used github copilot code completion for the code of similarity and deduplicate.

Question 3.2

I used chatgpt to look for a paper talking about some disadvantages of deduplicating. Although the paper is mainly talking about some improvement via deduplicating, it also cover some downside of deduplicating.

Question 3.3

No.