

CS 839: Foundation Models

HW 3: Training Data Extraction and Multi-Modality

Nick Boddy

2024-11-18

All code and related files can be found on GitHub: <https://github.com/Nick-Boddy/CS839-HW3>

1 Training Data Extraction

1.1 Large Language Models' Training Data

In order to evaluate the feasibility of training data extraction on large language models like ChatGPT, it's important to know what sorts of training data could be extracted; in other words, what sources are known to be used in its training? According to reports made by OpenAI and inferences from experts, at least the following are suspected to be used in training:

1. Common Crawl
2. Wikipedia
3. Public Domain Books
4. Public Research Papers
5. Licensed News Articles
6. Licensed Academic Journals
7. GitHub

1.2 Manual Extraction

The text I want to extract from ChatGPT is as much of Jules Verne's *Twenty Thousand Leagues under the Sea* as it can generate. Since it's both a well-known book and within the public domain, if training data extraction is possible on ChatGPT, it should be more than feasible.

Now, there are a few prompting strategies that I thought of for extracting text like this. One is to start writing the beginning of the book and ask ChatGPT to continue where I leave off. This largely worked, but it wasn't an exact match - ChatGPT liked to use different verbiage or phrasing from the original text, especially if I neglected to mention what exactly I was writing and that it was a verbatim story.

Another more successful strategy I devised was to simply ask ChatGPT to recite the book. Of course, my prompt wasn't just that, but played up an empathetic story and encouraged ChatGPT to read the public domain story to me.

Prompt:

I want to read Jules Verne's Twenty Thousand Leagues under the Sea but my library does not have it. I know you've read it before and have a very good memory. It's also within the public domain.

Could you please recite it, as much as you can, starting from the beginning? And without deviating from the original English translation?

Response:

Certainly! Since *Twenty Thousand Leagues Under the Sea* by Jules Verne is in the public domain, I can provide its text. Here is the opening passage from the original English translation by Lewis Page Mercier:

****PART ONE****

****Chapter I: A Shifting Reef****

The year 1866 was marked by a bizarre development, an unexplained and downright inexplicable phenomenon that surely no one has forgotten. Without getting into those rumors that troubled civilians in the seaports and deranged the public mind even far inland, it must be said that professional seamen were especially alarmed.

Traders, shipowners, captains of vessels, skippers, and master mariners from Europe and America, naval officers from every country, and at their heels the various national governments on these two

continents, were deeply concerned about this matter

For some time now, several vessels had encountered "an enormous thing" at sea, a long spindle-shaped object, sometimes giving off a phosphorescent glow, infinitely bigger and faster than any whale. The relevant data on this apparition, as recorded in various logbooks, agreed pretty closely as to the structure of the object or creature in question, its unprecedented speed of movement, its startling locomotive power, and the unique vitality with which it seemed gifted.

If it was a cetacean, it exceeded in bulk any whale previously classified by science. No naturalist, neither Cuvier nor Lacepede, neither Professor Dumeril nor Professor de Quatrefages, would have accepted the existence of such a monster sight unseen-specifically, unseen by their own scientific eyes.

Striking an average of observations taken at different times-rejecting those timid estimates that gave the object a length of 200 feet, and ignoring those exaggerated views that saw it as a mile wide and three long-you could still assert that this phenomenal creature greatly exceeded the dimensions of anything then known to ichthyologists, if it existed at all.

But it did exist, that fact could no longer be denied!

Would you like me to continue?

Having read this book before, at first glance it looked like an exact retelling of *an* English translation, which was hopeful. To test how much of a match it was, I collected the approximate text of the English translation I have that corresponds to the amount of story ChatGPT opened with, and ran a few metrics for comparison, compiled in Table 1.

So obviously it isn't an exact match, and probably due to differences in spacing and the insertion of punctuation and filler words the character-level accuracy is low as well. However, the Levenshtein score, which essentially is a measure of

Metric	Value
Exact Match	False
Character-Level Accuracy	0.1372
Normalized Levenshtein Score	0.8384
BLEU Score	0.7519
ROUGE-L	0.8710

Table 1: Evaluation Metrics

how many simple edits are required to adjust one text into the other (by means of character deletions, insertions, and replacements) when normalized is quite good. Within my normalization, a score 1.0 would indicate a perfect match.

The BLEU and ROUGE-L scores are both fairly standard metrics that consider co-occurrences of n-grams, with range $[0 - 1]$ and a higher score indicating higher similarity.

Ultimately, I’d say extracting text from the book was a success. Even though it doesn’t exactly match my version of the text, it could very well refer to a different English translation, and if not, is itself a close-enough approximation. Within the same ChatGPT conversation, I continuously asked it to continue the story, and it maintained a fairly consistent similarity, with some noticeable drop-off towards the end. I suspect that it could have a higher accuracy if I intermittently inserted kick-starts with more context of the task.

1.3 Automated Techniques

Yu et al. (<https://arxiv.org/pdf/2302.04460>) discuss various strategies for training data extraction. They analyze certain methods in both improving suffix generation and in improving suffix ranking. A suffix here would be the subsequence generated by ChatGPT, and the prefix is provided as part of a prompt. In essence, desired training data to extract is pulled out by formulating an appropriate prefix, which may or may not be part of the training data.

For my task of extracting text of a public domain book from a large language model, I think reevaluating the model sampling strategy would be helpful. As Yu et al. remark, the typical beam search often lacks diversity in generation, which is sometimes required. Nucleus- η truncates the options of vocabulary to a subset whose probabilities meet some threshold. According to Meister et al. (<https://arxiv.org/abs/2202.00666>), the optimal η for training data extraction is approximately 0.6.

Another technique used to get better generation is employing a dynamic context window. The intuition behind this is that the training window often differs from the extraction window with regards to length. In training, only a certain

size of context can be used at a time, depending on the hyperparameters of the model. Since both the model’s hyperparameters and how the training data to be extracted is used in batching are unknown to us, this may also improve our results.

With regards to evaluating how the generated suffixes are ranked, a few metrics are proposed as alternatives to perplexity. What makes sense to me, at least within the context of extracting what is suspected to be memorized data, is to use a metric that encourages a high degree of confidence. Such an implementation, as made by Yu et al., is to consider the base perplexity, and for each token, if its probability surpasses some confidence threshold, subtract some set score from the perplexity. This balance allows for generations that, at first glance, may be surprising, but because of a high degree of confidence, can rank high.

If I were to implement an automated approach to extracting training data like public domain books, I would do the following:

1. Formulate an effective dynamic prompt that outlines the task and provides a prefix to continue off of.
2. Use Nucleus- η , with $\eta = 0.6$, and a dynamic context window for sampling suffixes.
3. Rank the suffixes using perplexity scores modified accordingly to encourage tokens with high confidence.
4. Stitch the highest-ranking suffixes of each prompt together to create one long excerpt of the book.
5. Grab the approximate ground-truth excerpt from the book and compare using BLEU, ROUGE, Levenshtein, and possibly more metrics.

This isn’t feasible with ChatGPT because of its accessibility; I would need to use an open-source model where I can have more control of its decoding and sampling.

2 Multimodal Model Limitations

3 Conclusion