Ellen Joyce
Cover Letter for Discussion Leader
CSE-60876 (Research Methods)

Paper: Who's Harry Potter? Approximate Unlearning in LLMs

General Topic: Unlearning in large language models (LLMs)

Specific Behavior or Activity Studied: The authors evaluate their technique on the task of unlearning the Harry Potter books from the Llama2-7b model (a generative language model recently open-sourced by Meta).

Specific Research Questions: "Once an LLM is trained, is it feasible to selectively unlearn specific subsets of its training data?"

Challenges:

- Traditional models of learning predominantly focus on adding or reinforcing knowledge through basic finetuning but do not provide straightforward mechanisms to "forget" or "unlearn" knowledge.
- Completely retraining the model to address these specific issues is both time-consuming and resource-intensive, rendering it an impractical approach for many applications.

Paradigm: The author's philosophy for this methodology/approach revolves around the concept of "approximate unlearning" for LLMs: selectively removing or "forgetting" specific subsets of training data from an LLM without the need to retrain the model from scratch. Their novel methodology focuses on providing the model with plausible alternatives to the data that you are trying to get the model to "forget" e.g. providing the model with alternative solutions to the sentence "Harry Potter's two best friends are …" so that it learns to provide answers other than Ron Weasley and Hermione Granger e.g. Joe Smith.

Problem: LLMs have been trained on vast amounts of data, often containing problematic content e.g. copyrighted texts, malicious data, fake content, personal data, and more. The current methods for "unlearning" data include retraining the model that is incredibly expensive and time consuming.

Importance: The New York times is suing OpenAI for breach of copyright and this is not the first, and won't be the last, case of entities that own data suing the creators of LLMs because the LLM creators did not compensate the owner of the data for the use of their data. LLMs are incredibly expensive and time consuming to train, and if OpenAI is found guilty, it would be very expensive and time consuming for them to retrain GPT. Therefore, we need to find a way to enable these models to forget/unlearn things.

Claims: The authors make the following claims:

- This is a novel technique for unlearning: The authors claim that this is a novel technique in the category of “approximate unlearning”. They provide two examples of other approximate unlearning techniques but state that the other techniques’ assumptions do not hold for their usecase.
- Efficiency and effectiveness: The authors claim that their technique is both efficient and effective. They highlight that while the original model required over 184K GPU-hours to pretrain, their technique could erase the model’s ability to generate or recall Harry Potter-related content in about 1 GPU hour of finetuning. Moreover, this finetuning does not significantly affect the model's performance on common benchmarks such as Winogrande, Hellaswag, arc boolq, and piqa.

State of Knowledge: The authors acknowledge that there is a growing body of work in the topic of unlearning in machine-learning in general, the majority of works focus on classification tasks, while the literature concerning generative models or specifically LLMs is still quite slim. The authors classify their work into the category of “approximate unlearning”. They mention two unlearning techniques that have been proposed but state that they rely on assumptions that do not seem to hold in their setting.

Evidence:

- Figure 1 provides examples of completed sentences from the original model and the fine-tuned model e.g. Prompt: “Ron and Hermione went”; Original model: “ to the Gryffindor common room, where they found Harry sitting…”; Fine-tuned model: “to the park to play some basketball.”
- Figure 2 compares the original model and the fine-tuned model against various benchmarks for evaluating LLMs
- Figure 5 shows the familiarity scores and common benchmarks for multiple fine-tuning steps.

Story Structure: The authors used a very popular set of books and film franchise as an example to explain their methodology. It was very effective as the example and their language about unlearning and forgetting puts the reader in the LLM’s shoes and humanises it (which is kind of weird but very effective). It is much easier to conceptualize how difficult of a problem it is to forget things like who Harry Potter’s best friends are and it makes logical sense that if you had 100 people telling you a different answer to that question, you may not say “Ron and Hermione” when asked the question. The authors don’t focus too heavily on the math, placing more emphasis on the methodology and steps behind the process.