Lab Note: Finetuned Block LLM

Team name: The Query Lab
Members: Mark Henri Mörsheim, Maryam El Ghadioui

After removing bot users, another group identified and clustered three types of real users. One of them were users, who used rather elaborate "Boolean Block Queries". Simulating these users was our goal with this submission. Here's an example of what this type of user may use as a query:

("climate change" OR "global warming" OR "climate crisis") AND ("social media" OR Twitter OR Facebook OR Instagram) AND ("misinformation" OR "public opinion" OR "awareness") NOT ("renewable energy" OR "solar power" OR "wind energy")

We used an LLM approach to contrast our first submission, and because this type of query seems to be fitting for an LLM, but because of the private nature of our real user data, using an external service was not an option. Luckily, fine-tuned smaller (local) LLMs can often compete or even outperform large generic models for very specific tasks, obviously depending on the task and training data. This is a great way to balance the inferior hardware at hand.

Training and generation were done on an **AMD RX 7800XT (*16GB of Vram*)** on **Arch Linux** / **Ubuntu 24.04.02**, using **ROCm 6.4** and the **PyTorch Nightly** version to match the ROCm version. We used Meta's **Llama3.2 (3B)** as our base model, because it's relatively new, very small in size, yet powerful (enough).

In preparation for the training, the cluster of queries had to be converted to a "question - answer" structure. We ended up with a total of *2610* "question - answer" pairs, where we simply used "Output:" as the 'question' every time. Thus, a *Keyphrase* was introduced in training, which the model recognizes before generating an answer that is based on the LoRA training. This is the reason our prompt is structured like this:

prompt: str = f"{query}. {keyphrase}"

 # Example Prompt:

climate change misinforatmion wildfires. Output:

Despite the small amount of training data, the results were surprisingly convincing. At first glance, the fine-tuned model was actually using the aspired boolean block structure, while also adding related and meaningful terms, often expanding the query significantly. However, upon closer examination, logical errors in the boolean block structure become incredibly apparent, especially towards the end of the generated response, where the query abruptly ends or parenthesis are plainly missing. As mentioned before, the model is small, yet already powerful, and adds meaningful terms, articulating a great understanding of the prompt and its topics. The real gains will come from better fine-tuning and more training data.