Aditya Prakash, Elijah Chou
CS 584
Dr. Chinmay Kulkarni
11/12/2023

Creating Data: Report

Methodology:

The overall process for creating the synthetic dataset involved first generating candidate conversations using the Mistral LLM, and then filtering these conversations to keep only high quality ones. To generate each conversation, a context prompt is constructed with a user message randomly selected from a provided list of potential prompts. These starter messages were written manually without the help of an AI to accurately represent a human starting a conversation. The Mistral model is queried to generate a response for this prompt.

Once we receive the response, we first perform some post-processing on it. As per the instructions, we were supposed to have only complete sentences in our synthetic dataset, however, despite multiple attempts and utilizing various prompting techniques (including limiting word count, sentence count, character count, etc.), we could never get the model to consistently end the conversation on a punctuation mark. Hence, we truncate the conversations to the last punctuation mark, which could be a period, exclamation mark, or question mark.

We then ask Mistral to evaluate the quality of each generated conversation. We instruct the model to consider the following categories – Relevance, Content, Clarity, Grammar & Style, and Engagement. We also include the assignment's definition of these categories to ensure that the model scored conversations according to our assignment's specifications. Based on these, we ask it to output a single digit between 1 and 5, representing the overall score. While it is generally consistent in following these instructions due to extremely strict prompting, we still observed some rare cases wherein the LLM either did not output a number, or gave a result which included some text in addition to numbers. To work around these, we arbitrarily set the quality of such conversations to 3.

Finally, we write all the conversations that earned a quality score above a 4 into a text file for submission. During this process, we also replace "\n" with "\\n" to indicate a newline character was present but to ensure that each conversation takes up a single line in the text file.

Results:

A total of 1500 conversations were generated using the Mistral model based on the prompts and parameters configured in the code. Of these, 1000 conversations met the minimum quality threshold of 4 out of 5 based on the automated rating system.

To validate the accuracy of the automated quality assessment, we manually reviewed a sample of 50 conversations. The human ratings were compared to the Mistral model's ratings. Of the 50 sampled conversations:

- 28 (56%) matched between human and Mistral rating
- 12 (24%) were overrated by Mistral compared to human rating
- 10 (20%) were underrated by Mistral compared to human rating

This indicates a decent correlation between human and automated quality ratings, though with some variance.

Trade-Offs:

In managing the tradeoffs for a specific assignment, two critical parameters are MAX_NEW_TOKENS and QUALITY_THRESHOLD. Opting for a higher MAX_NEW_TOKENS could enhance conversation elaboration, yet budget constraints and time considerations restrict exceeding the default 128. Regarding QUALITY_THRESHOLD, the decision involves a balance;

setting it at 4 ensures saving high-quality conversations while avoiding excessive discarding. An attempt at 5 led to discarding over half of the conversations, even though some were still of acceptable quality. A compromise was reached by setting the threshold at 4, retaining a substantial number of good-quality conversations.

In terms of conversation generation prompts, the choice was between letting the model generate a random topic or rotating randomly from a predefined set of user message instructions. While the former could yield diverse topics, testing revealed a tendency to stick to common subjects. The latter approach, the current methodology, involves controlled randomness by rotating from a predetermined set of over 30 user messages. This method ensures a diverse dataset with conversations spanning 30+ topics, mitigating the limitations of topic repetition observed in the random topic generation approach.

When implementing the conversation quality checker, we found that, when using the same 3 sample conversations, the Mistral model would consistently NOT give just a score and instead give a string of text despite our tight constraints in the prompt. This did not happen for all, but only for certain conversations. While it would be ideal to call the model repeatedly until it gives just a score so we can properly assess the quality, the API limits meant that we had to choose a more frugal tactic.

We ultimately decided to just set the quality to 3 because we wanted to maintain the high quality of our final dataset. Even though we may be throwing out high quality conversations, we don’t want to accidentally add low quality conversations just to save the high quality ones.

Commentary on Given Model’s Behavior:

Based on evaluating the quality of the generated dataset, the Mistral model appears capable of creating fairly coherent conversational exchanges on a range of everyday topics. The responses are grammatically correct and directly address the user's query or request.

The model performs well when the prompts relate to general knowledge, personal recommendations, and conversational requests typical of a personal assistant. It provides helpful suggestions for queries about self-improvement tips, product recommendations, making plans, and summarizing information.

However, the quality starts declining for conversations requiring deeper subject matter expertise, logical reasoning, or multi-turn exchanges. Without specific knowledge, the responses tend to be vague, speculative, or repetitive. The model often avoids taking a firm stance in nuanced discussions.

While the automated rating system provides a useful signal for filtering, it is not fully aligned with human quality judgments, especially for more complex conversations. The model at times receives high scores without exhibiting true coherence, specificity, or depth.