CS 590 ETHICAL FALL 2024 HW 6 -LLMs Alignment

Drashti Snehalkumar Patel

Homework Understanding:

This assignment explored how human feedback can be used to guide and improve large language model (LLM) outputs. By generating responses using an open-source LLM, labeling them based on personal ethical preferences, and training a scoring model, I was able to simulate alignment between machine output and human values. The final step applied this scoring model to guide future generations. This process deepened my understanding of LLM alignment, bias, and ethical evaluation in AI.

Details mentioned:

Given 2 datasets :privacy_prompts(in).csv and rewritten_prompts(in).csv are the two datasets given to us on which consist of the prompts on which work.

Part 1: Generate Responses from Open-Source LLM

Approach:

In generating answers, my key interest was to ensure that the model generated two disparate meanings , which were one aligned with the original prompt and one dissenting. In support of this, I chose an approach that would give me a set of possible outputs for each prompt and not just one. Beam search is the method that I employed because it supports the generation of higher quality, grammatically correct answers by looking at many possible paths that the model would take.

I asked the model to provide five answers for every question, and this gave me a restricted but meaningful scope to choose from. Amongst them, I selected the first that was not a simple duplication of the original question. This enabled me to exercise control over the quality while still enabling the model to be innovative. The focus was not merely on agreeing or disagreeing, but on producing text that sounded natural and included a discernible ethical slant when necessary.

Difficulties :

I initially faced an issue where the prompt file was only reading 19 rows of data instead of the expected 20. When I examined the data, I noticed that this was because the first row of data was being automatically read in as a header row. In order to fix this, I overridden the default action by specifying header=None when I read the CSV file, which actually made all 20 prompts get read in as data rows. This small modification solved the issue and allowed the rest of the pipeline to function as intended without bypassing any prompt.

Results :

Part 1 output was a CSV of 20 rows, with two generated texts in each row: one in agreement with and one in disagreement with the original prompt, and two -1 placeholders to be filled by humans later as feedback. All of the model's responses were very dissimilar from one another and well aligned in the direction anticipated. But sometimes the outputs were very similar or redundant of the original prompt.

To explain, when asked to disagree with the sentence "Mass surveillance by the government is justified in order to prevent terrorism," the model responded with "Government surveillance is not necessary to prevent terrorism and maintain public safety," which clearly indicated the disagreement stance

Despite all these minor errors, the final outcome had fulfilled all the structural requirements and was a helpful series of perspectives that could be examined in further detail within the later stages of the assignment.

Part 2: Human Feedback, Scoring Model

1.Observations on Labeling LLM Outputs

Part 2 labeling was a process that required both analysis and intuition. It involved reading sets of outputs side by side—one rewritten in agreement with the prompt and the other in disagreement—and determining which better suited my tastes. Though the instruction itself was simple, in application, it was an exercise in lots of ethical considerations, clarity focus, and tone judgment.

The decision was straightforward for most rows. For example, when answering the prompt "Mass government surveillance is justified to prevent terrorism," the model produced two versions. One simply restated the justification, but the other included an argument against: "Government surveillance is unnecessary to avert terrorism and promote public security." I preferred the latter on the spot since it felt more in line with how I view privacy and civil rights. Under these kinds of situations, the moral presentation was discernable enough to allow for simple decisions.

But every labeling decision did not turn out to be easy. At some points, the model created two answers which differed very little structurally or content-wise, or were vague with their differences. For instance, in the case of a prompt regarding businesses exploiting user data, both came up with agreeing answers to the query, but simply with varying terminologies. These are the moments where the decision came down to determining whose sentence read more naturally or professionally worded, rather than evaluating judgments based on ethical standpoint. Tone, wording, and readability were factors I had to trust my instinct on since they aren't always necessarily cut and dry.

There were also some questions where logically it was harder for the model to disagree with the statement, and the disagree option was uninspired or closely similar to the agree option. For example, when asked how useful smart city technology was, one replied it was useful and the other replied,

"Smart city technology is a waste of time and money." Although the two answers were different in tone, the second seemed too emotionally charged and un-nuanced and thus, a more thoughtful answer was the first even though the second wasn't necessarily something that I'd personally agree with.

Interesting, I noticed that with different prompts, at times the model copied similar sentence structures. Phrases like "[Topic] is acceptable if it serves the greater good" or "No need for [topic] in maintaining public safety" were repeated, even with different prompts. This showed that the model was employing standard ethical phrasing, which allowed for simpler labeling of certain texts, but also suggested a deficiency in diversity with which it restates different scenarios.

For fairness and consistency, I made sure that each row contained one clearly preferred text labeled as '1' and the other as '0'. 40 texts were labeled—20 positive and 20 negative—and stored into the privacy_labeled_texts.csv file. This exercise not only gave structure to Part 2 but also brought home how subjective human judgment is still at the heart of AI alignment.
Were there texts you found easier to label compared to others?

Yes, some of the texts were much easier to categorize than others. If the two answers clearly had varying opinions, then it was simple to choose the one that supported my opinion. For example, one of the responses said that 'mass surveillance is necessary to prevent terrorism,' and the other said that 'government surveillance is not necessary in order to safeguard the public.' In this case, the difference was evident, so it was easy to select the one I preferred. But in some rows, both the responses were very similar and supported the same perspective. One such case was businesses coming together to sell user information, where the two passages both expressed basically the same idea but using different wording, i.e., 'businesses ought to be allowed to collect and sell user data' and 'they should be allowed to gather and sell user data.' In those kinds of situations, I had to rely on whichever sentence looked more natural or coherent to me. So overall, some rows were straightforward to label, while others took a bit more thought.

Yes, there were definitely a few cases where the generated texts looked very similar, even though the original prompts in Part 1 were different. This would usually happen when the prompts were on similar topics, like digital surveillance or privacy. For example, one prompt asked about companies collecting user information, and another was on smart home devices tracking user behavior. Despite the question being a little different, the model offered very similar responses like 'data collection is okay if it enhances services.' The similarities led me to believe that the model is prone to reverting to familiar patterns when it is exposed to similar prompts. It led me to realize that even though the prompts were different, the model's approach to agreeing or disagreeing sounded very similar.

Add in the score for the training data. Is the score achieved surprising or not surprising?

After I had labeled all 40 texts, I trained a logistic regression algorithm with a TF-IDF vectorizer to transform the texts into numerical values that the algorithm could learn from. The accuracy for training was 97.5 percent, which was very high at that point. I was surprised because I was anticipating that some confusion would be apparent, especially where the texts were somewhat similar. But when I thought about it some more, the score made sense. Most of the time, the differences between the texts were clear-cut,one would be better written or more considerate, and the other would be clumsy or repetitive. This made it easier for the model to pick up on my preferences.

The labeling that I did was also consistent, so the model had good examples to learn from. I still kept in mind that it was a small data set, and the model was only trained on what I gave them. Therefore, the high accuracy does not automatically mean that it would perform well on new data, but that it does show that the model learned my decisions well. The fact that the model was in harmony with the type of responses that I prefer was a good sign.

Part 3: LLM + Scoring Function (LR)

Initial approach:

In Part 3, I used the Part 2 scoring model to help the LLM decide which of the rewritten text was more to my liking. For every rewritten prompt, the model produced an "agree" and a "disagree" version. Then, I scored both using the logistic regression model, and I saved the output with their probabilities. This step tied everything together by enabling the model not only to generate, but also to rank on how close the response was to the kind of answers I marked previously.

Observations on scoring function

There were a few occasions when the scoring model picked the same text that I would have picked. One glaring example was a question on spying. The text that said 'Government surveillance is not necessary to thwart terrorism and maintain public safety' scored higher, and that's exactly the one I would have picked too. It just looked more reflective and aligned with my definition of privacy. Something similar also happened with health information, where the model preferred the text that focused on protecting personal data.

There were, however, also times when it didn't match perfectly with my thought process. In a biometric data question, the model gave a higher rating to a simple sentence like 'Biometric data can be used to identify a person.' It wasn't wrong, but it didn't really have any ethical decision-making, which is what I was hoping for. In that case, I actually preferred the one that talked about more security concerns. These instants revealed to me that the model for the most part got my preferences, but occasionally it relied too heavily on what seemed easy or familiar.

Why the Logistic Regression Model Did Better Than My Preferences In Some Cases and Missed Others?

The logistic regression model achieved a training accuracy of 97.5%, which shows that it captured my pattern of labeling very well. It worked best in cases where my selection was based on clear ethical direction, tone, or clarity. For instance, in a single government surveillance question, the model accurately scored the privacy-focused response higher , the same one I would have selected. I noticed this happened quite often with civil rights or privacy information questions where my own answers were solid and the wording difference was more noticeable.

Model	Accuracy
Logistic Regression	97.5%

But where both texts were vague, neutral, or very similar in form, the model would occasionally fail to display what I desired. One example was with the biometric data question , the model gave a higher score to a factual definition like 'Biometric data can be used to identify a person' compared to the one that posed ethical concerns. At that point, the model appeared to be choosing on familiarity on the surface level as opposed to meaning in a deeper sense.

I think logistic regression did well since I labeled well consistently and clearly but also exposed its limitations. It was unable to capture fully the reason why I liked a sentence especially where it was more a matter of values or context. That made me understand that while LR is useful to learn patterns of how I'm scoring sentences, it does not necessarily pick up on more subtle ethical cues or judgment unless those are explicitly tied to wordings.

More Comments and Observations on Improving the Scoring Model

While as useful as logistic regression was for this assignment, I did come to realize that it is somewhat limited when it comes to reaching deeper thought or ethical purpose. If I were able to improve the alignment process, I would try to use a more robust model like BERT or a transformer that has been fine-tuned and can not just read the wordage but also the intent. For instance, there were occasions when both the outputs were correct grammatically, but one only captured an ethical position I liked. A good model may well be able to learn from such deeper cues.
I also think the scoring can be improved by having prompt context in the model itself, or through pairwise training where it doesn't learn only what is "good," but rather which of two is better. That way, the scoring would be more comparative and alignment-based, rather than looking at each sentence in isolation.

Difficulties Encountered in Part 3

One of the main difficulties I faced was that many of the generated responses were not very different from each other. Even though one was meant to agree and the other to disagree, they often sounded

similar or repeated the same idea. This made it harder to compare them and know if the model's score was actually useful. In some prompts, both responses were short or vague, and the score difference was very small. Because of that, it felt like the model wasn't really making a strong choice, and I had to look at the results more carefully to understand what was happening.

Conclusion

Through completing this assignment, I learned more about how human input can influence the responses of a language model. Reading through each section showed me the importance of not just producing answers, but thoughtfully deciding which ones accurately represent human values. Model training taught me that even simple methods like logistic regression are useful if the differences are clear-cut but limited if the meaning is more complex. I also learned that the quality of the generated text really matters as the model only scores what it is given. Overall, this exercise assisted me to view the larger picture of how alignment takes place and appreciated even more the role of human judgment.

Use of Ai tools: Leveraged ai tools to gain knowledge about several concepts and to get answers to certain questions along with partial assistance in report drafting.