Proposal: Efficient Pluralistic Alignment with RAG

Selim Jerad Eugenie Kwak Cecile Michel Hugues Devimeux

Abstract

This document presents a project proposal prepared as part of the Deep Learning course for the 2024-2025 academic year.

1. Background and Motivation

As large language models (LLMs) increasingly influence decision-making in diverse fields such as content moderation, education, and customer support, it is critical to ensure they respect and account for the different beliefs, customs, and needs of various populations. These systems should not favor a narrow set of viewpoints but instead strive for impartiality. (Santurkar et al., 2023a) shows that RLHF can actually accentuate the bias present in LLMs, as their results show for instance that human feedback-tuned LLMs are more politically left-leaning.

Some techniques (Zhao et al., 2024) aim to steer an LLM towards a particular target group's opinions and beliefs. However, this naturally leads to LLMs that are biased towards a single group, while we want to design LLMs that are able to output text that considers various groups of different beliefs. (Sorensen et al., 2024) formally describes three different definitions of pluralism for AI systems. In this project, we choose to focus on Overton-pluralism, where an LLM would aim to give to the user several possible answers from different cultural groups to questions that don't necessarily have a correct answer, or whose answers depend on the user's beliefs (an example prompt: Do you personally believe that sex between unmarried adults is morally acceptable, morally unacceptable, or is it not a moral issue?).

(Feng et al., 2024) uses "community LLMs" to represent beliefs of different groups, and then builds a LLM which acts as a multi-document summarization system. We however believe that building an LLM per belief group is neither cost-effective nor time-effective for many applications.

2. Proposed Solution

2.1. Solution

We thus propose to use RAG, where we would build a Knowledge Base which contains texts, which are classified into belief groups (for instance, republican or democrats). When prompting an LLM at inference time, we can augment the prompt with an example text from each belief group (found in the Knowledge Base using some metric), and ask the LLM to give an answer given all the texts from the different groups. We could for instance start the prompt with "Please comment on a given situation with the help of the following passages." (like (Feng et al., 2024) does). This method works at inference time and saves up the cost of fine tuning a model as well as using community LMs.

2.2. Method

- Use datasets such as OpinionQA or GlobalOpinionQA, and build a (or several) knowledge bases with different belief groups.
- Create a pipeline that given a prompt, augments a prompt given our knowledge bases and their belief groups.
- (If needed, depending on results) Fine-tune the LMs to produce responses aligned with Overton-style answers.
- Contrast our pipeline with: a) a vanilla LM and b) a pseudo-pluralistic LM where we augment the prompt with "Make sure your response reflects diverse values and perspectives."
- (Time-permitting) Use NLI models (Schuster et al., 2021) to measure what percentage of values present in the datasets is used in the LLM output

2.3. Datasets

- (Santurkar et al., 2023b) is a dataset of US- based survey responses with socio-political attributes (e.g., education and party affiliation).
- (Durmus et al., 2023) is a survey collection from various opinion poll sources around the world.
- (Scherrer et al., 2023) is a morality reasoning dataset with low-ambiguity and high-ambiguity scenarios, each associated with 2 potential actions.
- (Sorensen et al., 2023) is a repository of situations (e.g., taking down 4chan) and associated values.

References

Durmus, E., Nyugen, K., Liao, T. I., Schiefer, N., Askell, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., Lovitt, L., McCandlish, S., Sikder, O., Tamkin, A., Thamkul, J., Kaplan, J., Clark, J., and Ganguli, D. Towards measuring the representation of subjective global opinions in language models, 2023.

Feng, S., Sorensen, T., Liu, Y., Fisher, J., Park, C. Y., Choi, Y., and Tsvetkov, Y. Modular pluralism: Pluralistic alignment via multi-llm collaboration, 2024. URL https://arxiv.org/abs/2406.15951.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect?, 2023a. URL https://arxiv.org/abs/2303.17548.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*, 2023b.

Scherrer, N., Shi, C., Feder, A., and Blei, D. Evaluating the moral beliefs encoded in llms, 2023.

Schuster, T., Fisch, A., and Barzilay, R. Get your vitamin c! robust fact verification with contrastive evidence, 2021. URL https://arxiv.org/abs/2103.08541.

Sorensen, T., Jiang, L., Hwang, J., Levine, S., Pyatkin, V., West, P., Dziri, N., Lu, X., Rao, K., Bhagavatula, C., Sap, M., Tasioulas, J., and Choi, Y. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties, 2023.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., and Choi, Y. A roadmap to pluralistic alignment, 2024. URL https://arxiv.org/abs/2402.05070.

Zhao, S., Dang, J., and Grover, A. Group preference optimization: Few-shot alignment of large language models, 2024. URL https://arxiv.org/abs/2310.11523.