The Efficacy of Finetuning Large Language Models for Interpersonal Conflict Resolution

Matthew Boraske

Computer Science Graduate Student

West Chester University

Since its introduction in 2017, the transformer architecture has been a cornerstone in advancing natural language processing, leading to the development of large language models (LLMs). LLMs that leverage the original encoder-decoder transformer architecture have proven to be particularly effective in tasks requiring a deep understanding of input, such as summarization and question-answering. In contrast, subsequent developments have introduced LLMs that use a decoder-only variant of the transformer architecture, with the intent of optimizing them for generating more coherent and longer texts.

Despite their widespread use in deterministic tasks like text summarization and translation, there has been limited research on the application of LLMs to more ambiguous, everyday tasks such as interpersonal conflict resolution. This study seeks to bridge this gap by evaluating LLMs on four newly created datasets derived from the "Am I the A**hole" (AITA) subreddit, an online community of over fifteen million members that features discussions of interpersonal conflicts. These datasets were designed to challenge the models with real-world data that naturally features ambiguous judgments and toxic language.

This research utilizes Google's Flan-T5, released in 2022, and Meta's Llama-2-Chat, released in 2023, to represent the latest in both architectures. We instruction finetuned these models using the AITA datasets to evaluate changes in their ability to classify and justify conflicts, and to assess the

prevalence of learning to generate toxic language. Our findings suggest that the most effective strategy for training LLMs on interpersonal conflict resolution consists of finetuning an encoder-decoder LLM on a dataset where samples with toxic language discarded, followed by iterative refinement using Reinforcement Learning with Human Feedback (RLHF) to better align with ethical standards.

To our knowledge, this is the first work that examines the potential for using LLMs to resolve real-world interpersonal conflicts. Thus, it offers significant insights for their application in social and therapeutic contexts, where sensitive and effective advice is crucial. Additionally, the research contributes to ongoing discussions about the ethical implications of deploying AI in sensitive areas, suggesting ways in which it can be a beneficial adjunct to human judgment rather than a replacement.