Exploring Empathy Using Direct Preference Optimization (DPO) in Large Language Models for Mental Health: A Study with GEMMA, LLAMA and BioBERT

Matheus Muniz Damasco[1]

[1]Departamento de Modelagem Computacional – Universidade Federal de Juiz de Fora
Juiz de Fora – MG – Brazil

matheus.damasco@estudante.ufjf.br

Abstract. *This study investigates the ability of large language models (LLMs), such as GEMMA, LLAMA, and specialized models like BioBERT, to simulate the role of psychologists with a focus on empathy. By applying the Direct Preference Optimization (DPO) technique, we aim to fine-tune these models using psychological counseling datasets to enhance their ability to generate emotionally appropriate and context-sensitive responses. Our approach compares the performance of generalist and specialized models.*

1. Introduction

In society, there has been a growing prevalence of mental health disorders affecting a significant number of individuals, from children to adults. As a result, there is an urgent need for innovative and accessible interventions. According to the World Health Organization, approximately 970 million people are affected by mental health conditions. This staggering number represents a major challenge for healthcare systems, impacting economic productivity and overall quality of life. Mental health disorders are not only associated with significant healthcare costs but also result in substantial productivity losses, further increasing the societal burden. Moreover, individuals with mental health issues experience a reduced life expectancy of about 10 to 20 years. Given the severe consequences of mental health disorders, finding solutions to alleviate the pressures faced by individuals has become an urgent priority.

In parallel with the rapid rise of mental health challenges, artificial intelligence (AI) has gained significant momentum, with Generative AI (GenAI) becoming increasingly prevalent. This simultaneous growth of mental health needs and GenAI presents an opportunity to leverage large language models (LLMs) for mental health counseling, potentially offering a transformative solution to address these challenges. By harnessing the power of LLMs, projects like this aim to develop prototypes capable of providing empathetic support to individuals facing mental health issues.

2. Problem

Although language models such as GEMMA, LLAMA, and BioBERT demonstrate the ability to understand and generate natural language responses, their alignment to exhibit empathetic behavior in psychological contexts remains a significant challenge. Empathy involves not only recognizing emotional cues but also responding with sensitivity and contextual appropriateness, which are subjective and complex requirements.

Empathy is a multi-dimensional construct encompassing two broad aspects: emotion and cognition [Sharma et al. 2020]. The emotional aspect pertains to the stimulation and resonance elicited in reaction to a user's experiences and feelings. In contrast, the cognitive aspect refers to the deliberate process of understanding and interpreting the user's experiences and emotions and effectively communicating that understanding back to them.

3. Hypothesis

By applying the Direct Preference Optimization (DPO) technique to language models such as GEMMA, LLAMA, and BioBERT, it is possible to enhance their responses to meet the empathetic standards expected in psychological interactions, surpassing traditional training methods.

4. Objective

The objective of this study is to develop a framework to fine-tune large language models (LLMs), such as GEMMA, LLAMA, and BioBERT, with a focus on enhancing their empathetic capabilities using the Direct Preference Optimization (DPO) technique. To achieve this, we aim to:

- Train both generalist and specialized models using datasets containing psychological interactions.
- Evaluate the impact of DPO on the ability of these models to generate empathetic and contextually appropriate responses.
- Compare DPO-optimized models with standard approaches, identifying their advantages and limitations.
- Analyze the performance differences between generalist models (GEMMA/LLAMA) and the specialized BioBERT model within the mental health domain.

5. Datasets

This study leverages three distinct datasets to train and evaluate large language models (LLMs) for empathetic interactions in psychological contexts. The selected datasets offer diverse annotations and scenarios relevant to therapeutic conversations.

5.1. AnnoMI: A Dataset of Expert-Annotated Counselling Dialogues

The AnnoMI dataset [Wu et al. 2022] is the first publicly accessible collection of professionally transcribed motivational interviewing (MI) dialogues, annotated by domain experts. It includes high- and low-quality MI interactions, making it a valuable resource for analyzing therapist and client behaviors. Key features are:

- **Metadata:** Includes transcript details, MI quality (high or low), topics, and URLs for original videos.
- **Annotations:** Each utterance is labeled with therapist behaviors (e.g., reflection, questioning) or client talk types (e.g., change, sustain).

5.2. Motivational Interviewing Dataset

This dataset [Welivita and Pu 2022] contains approximately 2,000 dialogues annotated with labels derived from the Motivational Interviewing Treatment Integrity (MITI) code. It focuses on listener utterances and provides insights into motivational interviewing techniques in peer support forums. Highlights include:

- **Scale:** A large-scale collection with diverse conversations.
- **Annotations:** Listener utterances are labeled based on the MITI framework to assess therapeutic quality.
- **License:** Distributed under the CC BY-NC-SA 3.0 license, emphasizing ethical use.

5.3. UFJF Addiction Counseling Dataset

This dataset contains 15 sessions between psychologists and patients, involving 13 unique patients seeking help for addictions such as smoking and alcoholism. It provides qualitative insights into therapeutic interventions for addiction. Key aspects include:

- **Focus:** Captures behaviors related to tobacco and alcohol use.
- **Expert Interaction:** Sessions involve trained psychologists engaging with patients.
- **Scope:** A compact dataset offering detailed interactions for addiction therapy.

6. Contributions

This study aims to deliver the following key contributions:

- Training Empathetic Models: Develop language models capable of generating empathetic and contextually appropriate responses in psychological interaction scenarios.
- Validation of DPO in Therapeutic Contexts: Establish a replicable methodology for leveraging Direct Preference Optimization (DPO) to align large language models with psychological and therapeutic standards.
- Provide a comprehensive analysis contrasting the performance of generalist models (GEMMA, LLAMA) with that of specialized models (BioBERT), highlighting their strengths and limitations in therapeutic applications.

References

Sharma, A., Miner, A. S., Atkins, D. C., and Althoff, T. (2020). A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Welivita, A. and Pu, P. (2022). Curating a large-scale motivational interviewing dataset using peer support forums. In *Proceedings of the 29th International Conference on Computational Linguistics*.

Wu, Z., Balloccu, S., Kumar, V., Helaoui, R., Reiter, E., Reforgiato Recupero, D., and Riboni, D. (2022). Anno-mi: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.