Leveraging RAFT for Accessible Mental Wellness Support

Irine Juliet Otieno
irine.juliet.otieno@yale.edu

Advisor: Prof. Arman Cohan
arman.cohan@yale.edu

1 Introduction

The global burden of mental health problems continues to grow, particularly affecting low-resource communities with limited access to care [1]. In response to the gap in mental health care access, individuals are increasingly turning to LLM-powered chatbots for support [2]. This project aims to explore the application of a new technique in domain-specific NLP, Retrieval-Augmented Fine-Tuning (RAFT), in developing a mental health support chatbot. While not intended to replace professional therapy or expert counseling, this chatbot has the potential to serve as an accessible companion in times of distress, offering immediate support and enhancing mental health literacy. Through this exploration, I hope to contribute to ongoing efforts in applying AI to enhance mental wellness, particularly for those who might otherwise lack access to support, providing a 24/7 accessible resource to alleviate the burden on traditional services.

2 Background

The evolution of chatbots in mental health support has progressed from early rule-based systems like ELIZA to sophisticated AI-driven approaches, showing promise in reducing symptoms of depression and anxiety [3,4]. Wysa, an AI-powered chatbot for therapy, exemplifies this advancement by using cognitive-behavioral techniques to provide mental health support [5].

General AI chatbots like ChatGPT often lack specialized knowledge and emotional attunement required for nuanced support.These limitations can be addressed by fine-tuning large language models on mental health domain-specific datasets, potentially improving the model's ability to generate empathetic responses. Retrieval-Augmented Generation (RAG) represents another significant development, aiming to provide more informed responses by sourcing external documents during conversations [6]. However, these approaches on their own still have challenges. Fine-tuning methods can lead to oversimplification or hallucination of inaccurate information due to limited training data. RAG systems, while improving response relevance, may introduce irrelevant information due to their reliance on semantic proximity for document retrieval.

The sensitive nature of mental health support demands a more robust approach that can combine specialized knowledge, contextual understanding, and accurate information retrieval while minimizing the risk of misinformation. This context sets the stage for exploring the recently introduced Retrieval-Augmented Fine-Tuning (RAFT), a method that combines RAG and fine-tuning to boost large language models' performance in specialized domains. By incorporating domain-specific documents during the fine-tuning process, RAFT enables the model to learn patterns specific to the target domain while also enhancing its ability to understand and utilize external context effectively [7]. This project will specifically explore its performance in the nuanced and sensitive field of mental health support.

3 Project Description

This project will utilize Retrieval-Augmented Fine-Tuning (RAFT), a cutting-edge technique recently introduced by Computer Science researchers at UC Berkley. The methodology involves fine-tuning a

base model, potentially LLaMA2, on a carefully curated dataset. This dataset will include mental health-related questions, "oracle" documents containing relevant answers, "distractor" documents without useful information, and chain-of-thought style answers detailing the reasoning process based on oracle documents.

By leveraging RAFT, the chatbot is expected to learn to effectively prioritize and integrate useful information while filtering out irrelevant content. This approach aims to enable the chatbot to provide potentially more accurate, contextually relevant, and personalized mental health support, drawing from a broad repository of psychological knowledge. The goal is to enhance the quality and reliability of AI-driven mental health support.

The potential success of this approach is supported by compelling evidence from the original RAFT paper, which demonstrated significant improvements over baseline models, domain-specific fine-tuned approaches, and standard RAG systems across different datasets [7]. These improvements suggest that RAFT could be particularly beneficial in the nuanced field of mental health support, where accurate information retrieval and contextually appropriate responses are crucial.

4 Devilerables and timeline

1. Project Proposal (Sept 5 - Sept 12)
    - Literature review on RAFT and mental health AI applications
    - Project scope definition
    - Preliminary project plan
    - Deliverable: Project Proposal Document (Due: Sept 12)
2. Data Collection and Preparation (Sept 13 - Oct 3)
    - Curated dataset of mental health-related questions and answers
    - Collection of "oracle" documents from reputable mental health sources
    - Collection of "distractor" documents
    - Development of chain-of-thought style answers
    - Deliverable: Prepared Dataset for RAFT Training
3. RAFT Model Development and fine-tuning (Oct 4 - Oct 24)
    - Implementation of RAFT architecture
    - Model finetuning on prepared dataset
    - Initial performance evaluation
    - Deliverable: Trained RAFT Model
4. Chatbot Interface Development (Oct 25 - Nov 7)
    - Design and implementation of a basic chatbot interface
    - Integration of RAFT model with the interface
    - Deliverable: Basic Working Chatbot Prototype
5. Testing and Refinement (Nov 8 - Nov 21)
    - System testing and bug fixes
    - Performance evaluation and comparison with baseline models
    - Final adjustments based on testing results
    - Deliverable: Testing Results and Performance Report
6. Project Wrap-up (Nov 22 - Nov 28)
    - Finalize project report
    - Prepare project presentation
    - Final prototype refinement
    - Deliverables: Final Project Report, Presentation, and Working Prototype

4.1 Key Deliverables

1. Project Proposal (Sept 12)
2. Prepared Mental Health Dataset for RAFT Training
3. Implemented and Trained RAFT Model
4. Basic Chatbot Interface with Integrated RAFT Model
5. Testing Results and Performance Comparison Report
6. Final Project Report and Presentation
7. Working Prototype of the Mental Health Support Chatbot

References

[1] World Health Organization. (2022). World mental health report: Transforming mental health for all. Geneva: World Health Organization. https://www.who.int/publications/i/item/9789240049338

[2] Song, I., et al. (2024). The Typing Cure: Experiences with Large Language Model Chatbots for Mental Health Support. arXiv preprint arXiv:2401.14362v2.

[3] Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine.

[4] Fitzpatrick, K. K., et al. (2017). Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial.

[5] https://www.wysa.com/

[6] Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.

[7] Zhang, T., et al. (2024). RAFT: Adapting Language Model to Domain Specific RAG. arXiv preprint arXiv:2403.10131v2.