KonKIS 2024 – Session 3. Large Language Models

Title: "Enhancing Emotional Support Chatbots for Depression Using Reinforcement Learning with Expert Knowledge Integration"

Lingxiao Kong and Zeyd Boukhers

Fraunhofer Institute for Applied Information Technology (FIT), Sankt Augustin, Germany

Abstract:

Mental health has become a paramount concern in contemporary society, as an increasing number of individuals are grappling with depression (Depressive disorder (depression), 2024). People increasingly rely on partners, friends, or even chatbots as effective means to prevent and alleviate symptoms of depression (Rauws, 2022). When observing interpersonal interactions, it's evident that conversations are frequently driven by emotional responses rather than logical reasoning. For example, in reaction to a statement like, "*My teacher is so annoying. He insists that I attend his course precisely on time*", a supportive friend might empathize by saying, "*That sounds tough. Everyone has urgent matters sometimes. Being 10 minutes late shouldn't matter*". In contrast, a chatbot, designed to respond logically, might suggest; "*It seems important to adhere to the schedule as required by your course guidelines*".

To enhance the effectiveness of an emotional support chatbot specifically designed for supporting individuals with depression, integrating proven therapeutic approaches like Solution-Focused Brief Therapy (SFBT) and Cognitive Behavioral Therapy (CBT) can be beneficial. A practical approach is to involve developing datasets from expert therapeutic sessions to train the chatbot to ensure it learns to emulate effective therapeutic communication. Additionally, integrating therapeutic principles into the advanced layers of transformer models can allow the chatbot to generate responses that are not only contextually appropriate but also therapeutically beneficial (Qi Ge, 2023). This method enhances the chatbot's capability to provide meaningful support, making it a valuable tool in mental health care. However, these techniques pose the risk of catastrophic forgetting, where the model may lose earlier knowledge as it acquires new information.

We aim to examine the application of Reinforcement Learning (RL) to embed expert therapeutic knowledge in chatbots, focusing on Solution-Focused Brief Therapy (SFBT) and Cognitive Behavioral Therapy (CBT). By employing Inverse Reinforcement Learning (IRL), we deduce the reward functions implicit in annotated expert behaviors. This allows a pre-trained Large Language Model (LLM) to be fine-tuned through RL. The goal is to enable the chatbot to conduct multi-turn emotional support conversations that more effectively support and enhance the patient's emotional well-being, while mitigating the issue of catastrophic forgetting.

References

Depressive disorder (depression). 2024, June 24). Retrieved from WHO Official Website: https://www.who.int/news-room/fact-sheets/detail/depression

Qi Ge, L. L. (2023). Designing Philobot: A Chatbot for Mental Health Support with CBT Techniques. *Proceedings of 2023 Chinese Intelligent Automation Conference* .

Rauws, M. (2022). The Rise of the Mental Health Chatbot. *Artificial Intelligence in Medicine*, 1609–1618.