Optimizing a Polish LLM for Psychological Support: A Comparative Study of Fine-Tuning Methods

Volha Tsekalo
Wydział Psychologii i Kognitywistyki, Uniwersytet im. Adama Mickiewicza w Poznaniu
voltse@st.amu.edu.pl

In recent years, Natural Language Processing has evolved significantly since the introduction of Transformers. Large Language Models (LLMs) are now frequently used in various fields, such as essay writing assistance, e-commerce support, and psychological aid. As psychological support is in high demand (World Health Organization, 2018, 2022), LLMs can serve as a supplementary tool when therapy with a human specialist is not an option. However, most existing solutions focus on English. Since there are very few therapeutic tools available for Polish, it is crucial to adapt local models to ensure the language sounds natural and fits the cultural context.

Cognitive Behavioral Therapy (CBT) relies on a clear, logical structure and specific step-by-step rules, which makes it much easier to adapt for Artificial Intelligence training compared to other forms of therapy (Grodniewicz & Hohol, 2023). Therefore, the project utilized the “Cactus” dataset, selected for its high-quality, structured training examples that cover a diverse range of mental health topics and specifically mimic the flow of professional CBT interventions (Lee et al., 2024).

However, to address the limitations of this dataset, which prioritized empathetic questioning over deep explanation, a synthetic data augmentation pipeline using Gemini 1.5 Flash was implemented. This process integrated psychoeducational components to align the responses with CBT principles.

The study involved training Bielik 7B – an open-source Polish LLM developed by Speakleash (Ociepa et al., 2024). I will present the results of four training rounds, comparing two methods: Parameter-Efficient Fine-Tuning (LoRA; Hu et al., 2021) and Partial Fine-Tuning. These methods were tested on two datasets: the original dataset translated into Polish, and the Gemini-generated hybrid dataset. Preliminary results confirm the limitations of the original dataset, showing that the model learned to mimic the questioning style, but struggled to provide logical and helpful therapeutic advice. Finally, I will compare these outcomes with the performance of the model trained on the hybrid dataset based on metrics such as Perplexity and BERTScore.

Bibliography

Grodniewicz, J. P., & Hohol, M. (2023).
Waiting for a digital therapist: Three challenges on the path to psychotherapy delivered by artificial intelligence.

Frontiers in Psychiatry, 14, 1190084.
https://doi.org/10.3389/fpsyt.2023.1190084

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021).
LoRA: Low-rank adaptation of large language models.
arXiv. https://doi.org/10.48550/arXiv.2106.09685

Lee, S., Kim, S., Kim, M., Kang, D., Yang, D., Kim, H., & Yeo, J. (2024).
Cactus: Towards psychological counseling conversations using cognitive behavioral theory.
arXiv. https://doi.org/10.48550/arXiv.2407.03103

Ociepa, K., Flis, Ł., Wróbel, K., Gwoździej, A., & Kinas, R. (2024).
Bielik 7B v0.1: A Polish language model—Development, insights, and evaluation.
arXiv. https://doi.org/10.48550/arXiv.2410.18565

World Health Organization. (2018).
The mental health workforce gap in low- and middle-income countries: A needs-based approach.
https://pmc.ncbi.nlm.nih.gov/articles/PMC3044251/pdf/BLT.10.082784.pdf

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