# Doctoral School of Information and Biomedical Technologies Polish Academy of Sciences (TIB PAN)

## SUBJECT:

Toward Safer and More Effective Large Language Models in Mental Health Applications

## SUPERVISOR:


- Prof. Szymon Łukasik (main supervisor);
  Naukowa i Akademicka Sieć Komputerowa – Państwowy Instytut Badawczy (NASK-PIB), ul. Kolska 12, Warszawa
  szymon.lukasik@nask.pl

- Wojciech Kusa, PhD (auxiliary supervisor).
  Naukowa i Akademicka Sieć Komputerowa – Państwowy Instytut Badawczy (NASK-PIB), ul. Kolska 12, Warszawa
  wojciech.kusa@nask.pl


## DESCRIPTION:

Large Language Models (LLMs), such as GPT-4, PaLM, and LLaMA, have shown impressive capabilities in understanding and generating natural language. Their potential to support mental health—through emotionally aware responses, psychoeducational dialogue, or early-stage support—is increasingly being explored. However, their deployment in sensitive areas like mental health introduces risks, including misinformation, lack of empathy, unintended harm, or ethical violations. Addressing these challenges requires more than general-purpose training; it demands models that are carefully adapted, aligned, and optimized for the unique demands of mental health interactions.

Recent research has begun addressing these concerns. Datasets like HamRaz provide language-specific and culturally tailored resources for counseling conversations, while CACTUS introduces structured therapeutic approaches, such as Cognitive Behavioral Therapy (CBT), into LLM training. The PsycoLLM benchmark evaluates models on psychological knowledge and ethical reasoning, and the study Can AI Relate? reveals disparities in empathy across demographic lines, underscoring the need for fairness and consistency in emotionally sensitive tasks.

This PhD research focuses on training and adapting LLMs specifically for use in mental health-related dialogue. The project will investigate effective training strategies, domain adaptation techniques, and safety-enhancing model alignment approaches. The goal is to build models that are not only linguistically competent but also contextually appropriate, emotionally aware, and aligned with ethical norms relevant to mental health. To measure progress, the project will use existing mental health evaluation benchmarks where available, and supplement them with human evaluation and task-specific metrics (e.g., response helpfulness, emotional appropriateness, risk-avoidance).

The expected outcomes of this research include a set of fine-tuned LLMs specifically optimized for mental health-related dialogue, an analysis of training and alignment techniques most effective in this domain, and a set of reproducible methods for improving LLM safety and relevance in high-stakes NLP applications. These contributions will support both academic advancement in domain-specific LLM adaptation and practical efforts to build responsible, AI-assisted tools for mental health.

## REQUIREMENTS:

- MSc degree in Computer Science, Computational Linguistics, Artificial Intelligence, Mathematics, or a related field
- High programming skills in Python
- Experience with NLP tools and model evaluation frameworks
- Knowledge of current research on LLM safety, alignment, and multilingual NLP
- Understanding of ethical considerations in AI applications in healthcare;
- Advanced level of English (spoken and written);

## BIBLIOGRAPHY:

Abbasi, M. A., Mirnezami, F. S., & Naderi, H. (2025). HamRaz: A Culture-Based Persian Conversation Dataset for Person-Centered Therapy Using LLM Agents. arXiv preprint arXiv:2502.05982.

Lee, S., Kim, S., Kim, M., Kang, D., Yang, D., Kim, H., ... & Yeo, J. (2024). CACTUS: Towards Psychological Counseling Conversations using Cognitive Behavioral Theory. Findings of EMNLP 2024.

Hu, J., Dong, T., Gang, L., Ma, H., Zou, P., Sun, X., ... & Wang, M. (2024). PsycoLLM: Enhancing LLM for Psychological Understanding and Evaluation. arXiv preprint arXiv:2407.05721.

Gabriel, S., Puri, I., Xu, X., Malgaroli, M., & Ghassemi, M. (2024). Can AI Relate: Testing Large Language Model Response for Mental Health Support. Findings of EMNLP 2024.

Zhang, M., Eack, S. M., & Chen, Z. Z. (2025). Preference Learning Unlocks LLMs' Psycho-Counseling Skills. arXiv preprint arXiv:2502.19731.