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

SUBJECT:

Training Dynamics and Knowledge Control in Foundation Models

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:

As foundation models become increasingly central to contemporary AI systems, understanding and controlling how they acquire, retain, and forget information is essential for their safe and effective deployment. This dissertation investigates the training dynamics and knowledge control mechanisms in large-scale pretrained models, focusing on how iterative optimization procedures affect the stability and structure of acquired knowledge.

Special attention is given to training strategies that alternate between supervised learning and alignment steps, such as reinforcement learning from human feedback (RLHF) or preference-based optimization. These approaches are studied both for their ability to improve alignment and for their impact on catastrophic forgetting—the unintended loss of previously acquired capabilities.

Another core focus is machine unlearning: the targeted removal of specific information or behaviors without retraining from scratch. Case studies include the removal of toxic content, sensitive personal data, and copyright-protected material. The work also investigates whether some domains of knowledge are more "forgettable" than others, and whether models can be trained to resist unintended knowledge loss.

Together, these investigations aim to build a deeper understanding of how foundation models acquire, adjust, and relinquish knowledge—laying the groundwork for safer, more controllable AI systems.

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
- Familiarity with European languages and linguistic typology
- Advanced level of English (spoken and written);

BIBLIOGRAPHY:

DeepSeek, (2025) *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*

Machine Unlearning: A survey https://arxiv.org/pdf/2306.03558

Knowledge Unlearning for LLMs: Tasks, Methods, and Challenges https://arxiv.org/pdf/2311.15766

How Do Large Language Models Acquire Factual Knowledge During Pretraining? https://arxiv.org/pdf/2406.11813

Exploring Forgetting in Large Language Model Pre-Training https://arxiv.org/pdf/2410.17018

Preventing Catastrophic Forgetting in Continual Learning of NewNatural Language Tasks https://arxiv.org/abs/2302.11074