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

SUBJECT:

Toward Safer and More Effective LLMs in Multilingual Contexts: Evaluations Across European Languages

SUPERVISOR:

- Prof. Aldo Lipani (main supervisor);
  University College London
  aldo.lipani@acm.org

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

DESCRIPTION:

The rapid development of Large Language Models (LLMs), such as GPT, PaLM, LLaMA, and DeepSeek, has led to significant breakthroughs in natural language understanding, reasoning, and generation. These models are increasingly used in high-stakes applications such as education, healthcare, and legal advisory systems. However, the vast majority of evaluation frameworks and safety alignment efforts focus almost exclusively on English. This introduces a systemic bias, as a significant portion of the global population—including speakers of Polish and other European languages—remains underrepresented in both training data and performance benchmarks.

This PhD research will address the critical gap in evaluating the safety, effectiveness, trustworthiness and reasoning capabilities of LLMs in non-English European languages. It will focus specifically on Polish, German, French, and a broader selection of Slavic and Romance languages. The project will explore multilingual safety auditing by developing tools to detect harmful, biased, or misleading outputs that are more likely to go unnoticed in less frequently analyzed languages. A particular emphasis will be placed on constructing multilingual input-output safety guards, grounded in linguistic and cultural context, to evaluate whether models behave consistently and ethically across languages.

In addition to safety, the research will examine reasoning and task-related performance in multilingual contexts—especially in domains such as mathematics, logic, and programming. A key research question concerns how reasoning capabilities are preserved or distorted during alignment processes in multilingual settings. Emerging models like DeepSeekMath and DeepSeek-R1 show promising capabilities in logical reasoning and mathematical problem solving, but these results are almost entirely reported in English. By extending such evaluations to other languages, this research aims to shed light on cross-linguistic transfer of reasoning and general knowledge.

Ultimately, this work aspires to contribute to the development of safer, fairer, and more inclusive LLMs. It may also uncover new alignment strategies or training techniques that enhance multilingual reasoning fidelity, supporting the broader scientific effort to create globally robust AI systems. The expected outcomes will not only improve practical deployment in Europe but also push forward the theoretical understanding of multilingual alignment and reasoning in large-scale models.

REQUIREMENTS:

- MSc degree in computer science, computational linguistics, mathematics, AI, or 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*

Inan H. et all. (2023) Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations

Kanepajs A. et all. (2024) *Towards Safe Multilingual Frontier AI*

Kumar P. et all. (2025) *PolyGuard: A Multilingual Safety Moderation Tool for 17 Languages*

Sharma M. et all. (2025) Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming

Singh S. et all. (2024) *Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation*

Wang B. et all. (2024). DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models.