Thesis Project Topic: Integrating Information Bottleneck for Enhanced Security Robustness in AI Alignment

We are seeking a motivated student to undertake a master thesis project focused on integrating the Information Bottleneck principle to improve the robustness of AI alignment. This project aims to explore how the Information Bottleneck can be applied to AI models to ensure they align better with human values and objectives, especially under conditions of uncertainty and adversarial environments.

Ideal Candidate:

- Strong interest in AI alignment, machine learning, and the theoretical underpinnings of AI safety.
- Basic knowledge of information theory, machine learning, and neural networks.
- Strong analytical and coding skills, with a passion for exploring innovative solutions to long-standing challenges in AI safety and robustness.

If you are eager to contribute to this cutting-edge research project and make a meaningful impact in the field of AI alignment, please contact derui.zhu@tum.de with your resume and transcript.