MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS
Learning and Dynamical Systems Group (Dr. Michael Muehlebach)

Master Thesis Project

Risk-Aware Selective Mixture-of-Experts for Efficient Language Modeling

Recent advances in language modeling demonstrated the effectiveness of Mixture-of-Experts (MoE) architectures, which route inputs selectively to specialized sub-networks ("experts") to improve model capacity without a proportional increase in computation. However, existing MoE models often rely on heuristic routing mechanisms that lack guarantees for efficiency, reliability, and robustness under uncertain or adversarial inputs [1,2].

This MSc project builds upon the **risk-aware selective language modeling framework** [1], which introduces an online, hierarchical batch selection mechanism for selective pretraining, improving efficiency in large-scale language model training. The goal of this project is to extend this framework to the MoE architectures by designing a risk-aware selective routing mechanism that:

- Learns to route inputs dynamically to experts based on confidence and risk estimates.
- Improves training and inference efficiency by activating only a subset of experts while maintaining predictive performance.
- Provides robustness against distribution shifts and potential adversarial inputs, as in [1].

The student will explore both theoretical and practical aspects, including the integration of risk measures into expert gating, implementation in state-of-the-art MoE models, experimental evaluation on benchmark language modeling datasets, and comparisons with standard MoE routing strategies. This work will contribute to more efficient and reliable MoE architectures and deepen understanding of principled routing mechanisms in expert-based networks.

Learning and Dynamical Systems Group (https://lds.is.mpg.de/)
The Learning and Dynamical Systems Group is part of the Max Planck Institute for Intelligent Systems in Tübingen, Germany. Our research lies at the intersection of machine learning, dynamical systems, and mathematical optimization.
This project will be in **collaboration with Prof. Volkan Cevher at EPFL**-LIONS research group.

Prerequisites
Strong analytical skills and programming experience (Python, PyTorch, C/C++ or similar languages and tools). Background in machine learning, statistics, or mathematical optimization is a plus.

Contact
If you have any questions do not hesitate to contact us. When applying for a project, please include your CV, bachelor's and master's transcripts, and a one-page letter of motivation describing your research interests and educational background.

Melis Ilayda Bal, mbal@tuebingen.mpg.de

Dr. Michael Muehlebach, michael.muehlebach@tuebingen.mpg.de

References
[1] Melis Ilayda Bal, Volkan Cevher, and Michael Muehlebach. (2025). ESLM: Risk-Averse Selective Language Modeling for Efficient Pretraining. arXiv preprint arXiv:2505.19893.
[2] Melis Ilayda Bal, Volkan Cevher, and Michael Muehlebach. (2025). Adversarial Training for Defense Against Label Poisoning Attacks. In The Thirteenth International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=UlpkHciYQP