

Post-Training and Reinforcement Learning for Large Language Models

Supervisors: Fatima Rani, Juan Cabrera

Context:

Large Language Models (LLMs) such as GPT and Llama have transformed natural language understanding and generation. However, pretrained LLMs often display issues like hallucinations, inconsistent reasoning, and unsafe responses, making them unsuitable for direct deployment in production environments. Post-training including fine-tuning and reinforcement learning with human feedback (RLHF) has emerged as a crucial step to align model behavior with human values and specific real-world tasks. This thesis aims to systematically study and apply these post-training techniques to enhance reasoning, reliability, and safety in LLMs, bridging the gap between research prototypes and production-ready AI systems.



Tasks:

- **Adapt pretrained LLMs:** Fine-tune open-source models (e.g., Llama, Mistral) for specific real-world tasks using supervised fine-tuning (SFT).
- **Apply reinforcement learning:** Implement RLHF or related methods (PPO, GRPO) to align model behavior with human or synthetic feedback.
- **Evaluate systematically:** Develop benchmarks to assess reasoning, factual accuracy, instruction following, and safety.
- **Iterate and optimize:** Use evaluation insights to refine datasets, rewards, and training strategies for continuous improvement.
- **Ensure efficiency and deployment readiness:** Employ lightweight fine-tuning methods (e.g., LoRA) and analyze trade-offs in cost, latency, and inference performance.

Evaluation Metrics:

- **Task Performance:** Instruction-following accuracy, reasoning benchmarks.
- **Safety and Robustness:** Toxicity and hallucination rate reduction.
- **Efficiency:** Fine-tuning cost, parameter efficiency, and inference latency.
- **Energy-efficiency trade-offs:** By analyzing the Pareto front between model accuracy and energy consumption during training and inference.

Importance:

Fine-tuning and reinforcement learning are key to transforming generic LLMs into task-optimized, safe, and deployable systems.

This Master/ Diplomarbeit thesis contributes to the understanding of how alignment methods and evaluation strategies affect reasoning and safety, offering insights applicable to future industrial LLM pipelines and the broader field of responsible AI.

Required skills:

- Python programming skills.
- Experience with PyTorch, or similar frameworks (optional Hugging Face Transformers) .
- Understanding of deep learning and reinforcement learning fundamentals.
- Willing to learn LLM architectures and fine-tuning methods.
- (Optional) Experience with data annotation, prompt engineering, or model evaluation tools.

Key words: Large Language Models (LLMs), Reinforcement learning with human feedback (RLHF), Supervised fine-tuning (SFT), Ennergy-Efficiency.

Language: English

Corresponding email: fatima.rani@tu-dresden.de

References:

1. Sun, H., & van der Schaar, M. (2025). Inverse reinforcement learning meets large language model post-training: Basics, advances, and opportunities. arXiv preprint arXiv:2507.13158.
2. Liu, K., Yang, D., Qian, Z., Yin, W., Wang, Y., Li, H., ... & Zhang, L. (2025). Reinforcement learning meets large language models: A survey of advancements and applications across the llm lifecycle. arXiv preprint arXiv:2509.16679.
3. Wang, S., Zhang, S., Zhang, J., Hu, R., Li, X., Zhang, T., ... & Hovy, E. (2024). Reinforcement learning enhanced llms: A survey. arXiv preprint arXiv:2412.10400.
4. Lai, H., Liu, X., Gao, J., Cheng, J., Qi, Z., Xu, Y., ... & Tang, J. (2025, July). A survey of post-training scaling in large language models. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2771-2791).
5. Kumar, K., Ashraf, T., Thawakar, O., Anwer, R. M., Cholakkal, H., Shah, M., ... & Khan, S. (2025). Llm post-training: A deep dive into reasoning large language models. arXiv preprint arXiv:2502.21321.