Towards Aligned and Efficient Large Language Models

Yu Meng

Department of Computer Science, University of Virginia
yumeng5@virginia.edu

Abstract

Large language models (LLMs) have rapidly transformed the landscape of AI, demonstrating remarkable capabilities across reasoning, communication, and problem-solving. Yet, realizing their full potential requires addressing two critical challenges. **First**, their behavior must be steered and refined after training to ensure reliability, safety, and alignment with human values and intentions. **Second**, their large scale comes with substantial costs in training and deployment, necessitating research into more efficient methods.

My research centers on advancing both of these fronts—making LLMs both aligned and efficient. On one side, I investigate post-training techniques that allow models to better reflect human preferences, demonstrate strong reasoning capabilities, and mitigate hallucination. On the other side, I study methods for improving data efficiency in training and inference efficiency in deployment. Together, these thrusts highlight a broader vision of enabling LLMs that are not only powerful, but also trustworthy and accessible at scale.

Post-Training: Aligning and Enhancing LLMs

While pretraining endows LLMs with broad linguistic and factual knowledge, their raw outputs are not inherently aligned with human goals. My work in post-training develops methods to refine and enhance these models so they can better align with human intentions, incorporate external knowledge, and reason more effectively.

First, to align models with subjective human values, I develop novel preference optimization techniques exemplified by SimPO (Meng, Xia, and Chen 2024). These methods enable models to learn nuanced, context-aware behaviors that are helpful and safe. **Second**, to combat the critical issue of hallucination, my work on retrieval-augmented generation (RAG), InstructRAG (Wei, Chen, and Meng 2025) improves LLMs' utilization of external knowledge through self-synthesized rationales, ensuring their outputs are factually accurate. **Third**, to elicit LLMs' reasoning capabilities for complex tasks, I investigate how reinforcement learning shapes model behavior and propose new training paradigms for complex reasoning (Zhu et al. 2025).

Collectively, these efforts create LLMs aligned with human preferences, grounded in facts, and transparent in logic.

Efficiency: Reducing Data and Inference Costs

Despite their extraordinary capabilities, the enormous scale of LLMs poses practical challenges, both in acquiring sufficient training data and in deploying them efficiently at inference time. My research tackles these bottlenecks by developing methods that reduce data annotation requirements and accelerate decoding, thereby broadening the accessibility and scalability of LLMs.

First, to alleviate the heavy reliance on large amounts of labeled data, I develop methods to train LLMs to produce high-quality, diverse datasets for specialized tasks (Meng et al. 2023). This approach dramatically lowers the barrier for building new, customized models by automating a traditionally manual and expensive process. **Second**, to address the computational expense of autoregressive decoding, I investigate methods that leverage predictions from early layers of an LLM (Wei et al. 2025). This approach reduces the latency of forward passes required per token, substantially improving inference-time efficiency while maintaining output quality. Such techniques expand the practical reach of LLMs to resource-constrained environments and reduce deployment barriers.

These contributions aim to develop LLMs that are powerful and scalable, lowering both the data and compute cost.

References

Meng, Y.; Michalski, M.; Huang, J.; Zhang, Y.; Abdelzaher, T.; and Han, J. 2023. Tuning language models as training data generators for augmentation-enhanced few-shot learning. In *ICML*.

Meng, Y.; Xia, M.; and Chen, D. 2024. SimPO: Simple preference optimization with a reference-free reward. In *NeurIPS*.

Wei, Z.; Chen, W.-L.; and Meng, Y. 2025. InstructRAG: Instructing Retrieval-Augmented Generation via Self-Synthesized Rationales. In *ICLR*.

Wei, Z.; Chen, W.-L.; Zhu, X.; and Meng, Y. 2025. AdaDecode: Accelerating LLM Decoding with Adaptive Layer Parallelism. In *ICML*.

Zhu, X.; Xia, M.; Wei, Z.; Chen, W.-L.; Chen, D.; and Meng, Y. 2025. The surprising effectiveness of negative reinforcement in LLM reasoning. In *NeurIPS*.