Research Interests

Large-scale pre-trained generative models, such as large language models (LLMs), have recently shown exceptional versatility and effectiveness across diverse applications, driven by their training on extensive web corpora. However, much of the training process relies on memorizing existing data and lacks a systematic stage for developing advanced problem-solving skills through exploration and practice, as humans do. This limitation hinders performance on complex reasoning tasks and often leads to the generation of unreliable content. Additionally, the growing scale of pre-trained models reduces their feasibility for deployment on resource-constrained devices.

Motivated by these challenges, I aim to focus my PhD research on the following topics: (i) Developing strategies for training generative models to enhance **problem-solving skills and reliability through self-exploration**, such as applying reinforcement learning. (ii) Improving the **computational efficiency** of large generative models in both training and inference. The following sections provide an overview of my research experiences, key insights, and future aspirations.

1 Advancing Generative Models via Self-exploration

Training generative models through self-exploration provides two major benefits. First, it reduces constraints on the model's expressiveness caused by the teacher-forcing paradigm. For example, a mathematical problem can be solved using various approaches or calculation orders, but fine-tuning models on a single path limits the diversity. Second, it enables the model to fully leverage the knowledge embedded in its parameters. Similarly, humans cannot master a subject by simply memorizing textbooks; practice and exercises are essential for deep understanding and effective application of the learned knowledge. As the saying goes, '***Practice makes perfect.***' These advantages have inspired my conviction in advancing generative models by developing self-exploration training strategies.

Knowledge is the Foundation of Problem-Solving and Exploration. Neither humans nor models can effectively develop problem-solving skills through exploration when faced with unseen domain-specific tasks. For instance, a student with no understanding of calculus would struggle to solve problems, let alone improve their skills to answer similar questions through mere trial and error, highlighting the importance of **a solid foundational knowledge base to support the self-exploration learning paradigm**. My previous research [1] showed that acquiring relevant knowledge before problem-solving training is crucial for effectively training LLMs on domain-specific tasks. Specifically, we curated two corpora for training the model to solve domain-specific tasks: unlabeled in-domain data to provide domain knowledge and task-oriented passages for downstream problem-solving. Through a two-stage training that mixed these corpora in different ratios, our experiments revealed that the optimal results were achieved when knowledge acquisition preceded problem-solving training. I believe this finding also holds true for exploration-based training, where prior knowledge is essential for exploring new skills.

Enhancing Reliability through Knowledge-driven Self-exploration. Although pre-trained LLMs have learned extensive world knowledge, they sometimes fail to apply it properly in answering questions, revealing a gap between what they know and what they say. To address this issue, in our work [2], we explored a self-consistency-based decoding strategy to enhance LLM reliability by generating the final answer conditioned on parallel aggregation of multiple initial outputs. Additionally, in our study [3], we tried to balance hallucination and descriptiveness in multimodal LLMs using contrastive decoding, integrating descriptive intermediate states with final-layer outputs to refine responses. Both studies reveal that although LLMs possess the knowledge to answer questions, they require further refinement to apply it effectively. This raises an important question for me: **Can LLMs improve their expressiveness with internal knowledge by learning from their own generations?** To address this, I would dive deeper into the root causes of unreliable generations and investigate how training via exploration can help overcome these challenges.

Training Strategies for Self-exploration. The core of the self-exploration training for generative models lies in their interaction with the external world: generating content, receiving feedback, and refining outputs based on the feedback. Specifically, external feedback enables the model to refine its generation through iterative search and optimize it using reinforcement learning or preference optimization. In [4], I initially explored training LLMs to generate summaries that better aligned with relevant queries by utilizing feedback from downstream retrieval tasks. By analyzing the experiments, I found that **the richness and diversity of the generations are vital for exploration**, as richer outputs expand search spaces, allowing models to explore more flexibly and align better with external feedback. This finding aligns with the breakthrough reasoning model O1 [1], where long-form thoughts optimized through reinforcement learning improve performance on complex reasoning tasks.

Future Work for PhD. I wish to develop effective strategies for training models through self-exploration from the following perspectives:

- Advancing algorithms for training generative models to **improve problem-solving skills and reliability via exploration**, including search, reinforcement learning, and preference optimization. This effort may also involve defining specific intermediate actions in generations and designing feedback strategies for generative models, such as fine-grained process feedback.
- Identifying how to balance acquiring new skills through exploration with preserving existing knowledge for models to achieve overall improvement.

2 Efficient Language Modeling

The emergence of LLMs has attracted the attention of many researchers and internet users. However, their capabilities come with significant resource demands. Specifically, scaling up these models demands exponentially more GPU hours and data for training, while the computational cost of inference lowers their throughput. These challenges highlight the strong need to develop effective techniques for improving their efficiency in both training and inference.

Efficient Long Sequence Modeling. One of the most significant barriers to improve LLM efficiency is the $O(n^2)$ complexity of the attention mechanism in standard Transformers, which makes sequence length a persistent bottleneck in initial models. Beyond the computational cost, the KV cache for previous tokens grows linearly with the sequence length and sometimes consumes larger memory than the model itself. To shorten the memorized token sequence, attention scores offer an intuitive and effective way to assess token importance. Tokens with lower attention scores can be identified as less important and either discarded or offloaded to the CPU to save memory.

From another angle, in my work [5], my colleagues and I proposed to assess the importance of each token based on its contribution to task performance and applied reinforcement learning to train a token selector, where the reward function was designed as a combination of task performance and a length penalty. This initiative showed a promising way to **measure token importance beyond attention scores**, offering the possibility of determining token importance before attention computation.

Efficient Model Architecture. Although standard LLMs like LLaMA have demonstrated impressive capabilities, their parameter allocation can be further optimized. Simply reducing the number of key-value heads is an effective way to lower model parameters, resulting in only a slight decline in performance. Given that the Key vector in Transformer is only involved in computing the attention matrix, I made an initial attempt to reallocate part of the parameters from the Key projection to the Query and Value projections [6], where the model demonstrated superior performance compared to the standard version with the same training data. Additionally, the standard attention map during decoding in pre-trained LLMs has been demonstrated to be sparse, making its computations unnecessarily redundant. Beyond Transformer architecture, memory-efficient models like State Space Models (SSM), which use efficient hidden states to store previous information, along with their hybrid variants, have shown great potential in both efficiency and performance.

Future Work for PhD. Leveraging the sparsity of causal attention in standard Transformers and optimizing the inefficiency of sequence memorization are both critical to improving speed and memory efficiency in standard LLMs. In my PhD research, I wish to explore:

- How can the **attention sparsity** be leveraged to enhance LLM efficiency? How to **approximate full attention** with less computation while maintaining expressivity?
- How to effectively **measure the token importance**? How should less important tokens be processed for potential reuse? (e.g., offloading their KV cache)
- Why do **hybrid versions of sequence models**, such as SSM and attention mechanisms, complement each other, and how can they be integrated to achieve both optimal efficiency and proficiency?

[1] https://openai.com/index/learning-to-reason-with-llms/

References

[1] **Xiao Liang***, Xinyu Hu*, Simiao Zuo, Yeyun Gong, Qiang Lou, Yi Liu, Shao-Lun Huang, and Jian Jiao. Task oriented in-domain data augmentation. *EMNLP*, 2024.

[2] Yi Cheng, **Xiao Liang**, Yeyun Gong, Wen Xiao, Song Wang, Yuji Zhang, Wenjun Hou, Kaishuai Xu, Wenge Liu, Wenjie Li, et al. Integrative decoding: Improve factuality via implicit self-consistency. *arXiv preprint arXiv:2410.01556*, 2024.

[3] Yaoyuan Liang, Zhuojun Cai, Jian Xu, Guanbo Huang, Yiran Wang, **Xiao Liang**, Jiahao Liu, Ziran Li, Jingang Wang, and Huang Shao-Lun. Unleashing region understanding in intermediate layers for mllm-based referring expression generation. *Advances in Neural Information Processing Systems*, 2024.

[4] **Xiao Liang**, Xinyu Hu, Simiao Zuo, Jimi He, Yu Wang, Victor Ye Dong, Yeyun Gong, Kushal S Dave, Yi Liu, Qiang Lou, et al. What you see is what you get: Entity-aware summarization for reliable sponsored search. In *Neurips Safe Generative AI Workshop 2024*.

[5] Jiawen Xie, Pengyu Cheng, **Xiao Liang**, Yong Dai, and Nan Du. Chunk, align, select: A simple long-sequence processing method for transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 13500–13519, 2024.

[6] **Xiao Liang** and Yeyun Gong, et al. Differential rescaling of query, key and value for efficient language models. ***To Be Present***.