Loquetier: A Virtualized Multi-LoRA Framework for Unified LLM Fine-tuning and Serving

Yuchen Zhang, Hanyue Du, Chun Cao, Jingwei Xu

State Key Laboratory for Novel Software Technology, Nanjing University, China

Github: https://github.com/NJUDeepEngine/Loquetier

- **Motivation**
- **Introduction**
- **Workflow Diagram & Framework Diagram**
- **Experiments**
- **Conclusion**

Motivation

Low-Rank Adaptation, LoRA

Critical challenges that existing approaches still face:

- **Weight-loading Inflexibility**
- **Task Imbalance**
- **Downtime During Switching**

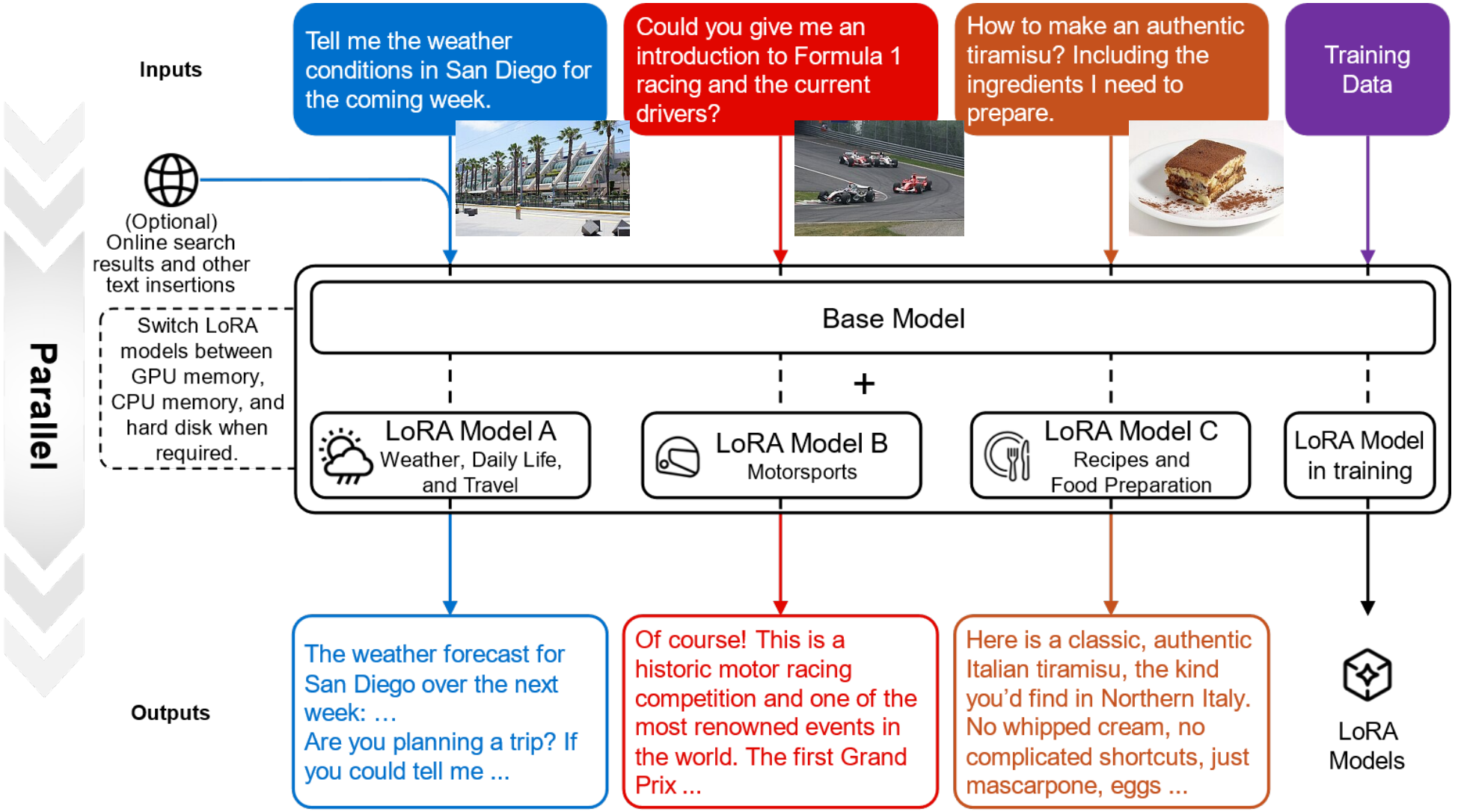
Introduction

We introduce **Loquetier**, a unified virtualization framework that integrates fine-tuning and serving of LLMs with LoRA-based PEFT.

Our main contributions:

- **Virtual Module**
- An optimized **computation flow** with an **SMLM kernel design**
- Extensive experiments show that Loquetier outperforms existing systems across diverse scenarios.

Workflow Diagram



*Images from Wikipedia via Wikimedia Commons

Framework Diagram

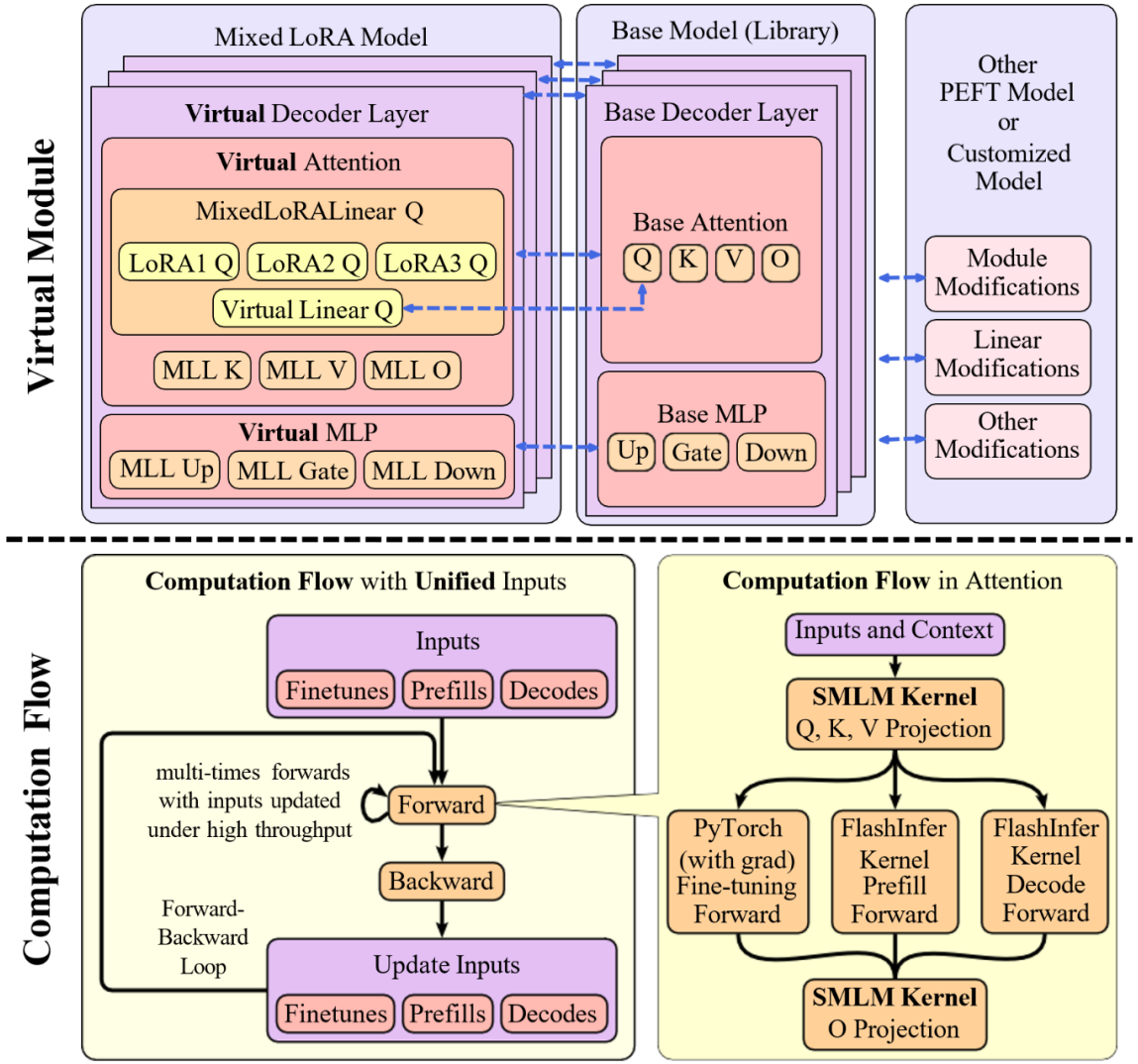
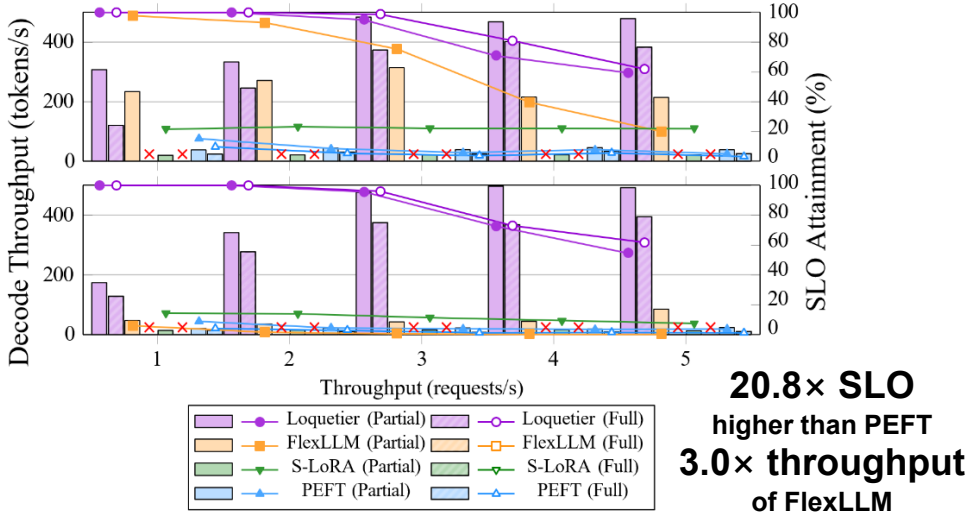


Figure 1: The framework diagram of Loquetier.

Experiments

Models:	Llama3-8B
Hardware:	NVIDIA A6000 48G GPUs (infer-only); NVIDIAH800 80G GPUs (other tasks).
Tasks:	Inference-only;
	Fine-tuning-only;
	Unified fine-tuning and inference;
	Mutable capacity allocation simulation;
	Simulated real-world workload (based on BurstGPT [2] datasets)

Experiments



20.8× SLO
higher than PEFT
3.0× throughput
of FlexLLM

Figure 2: Comparison of the performance of Loquetier and 3 other baselines in inference tasks.

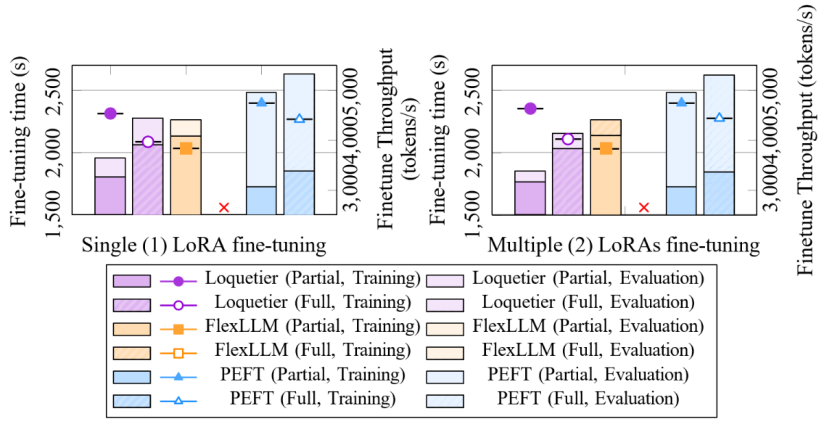
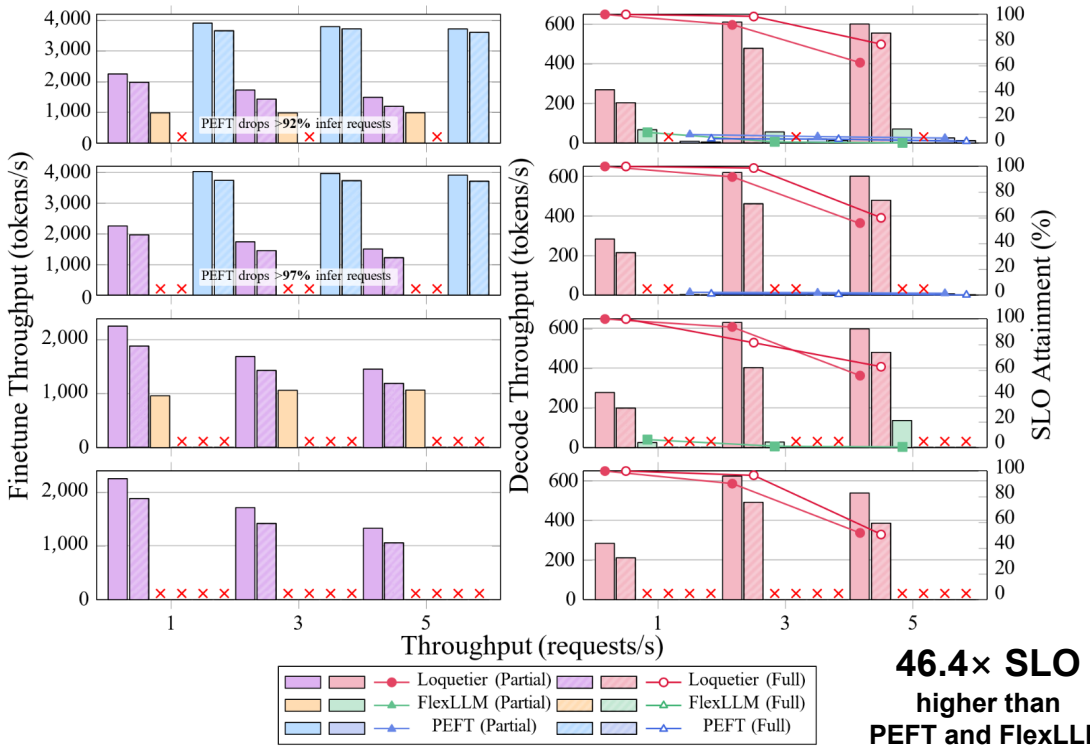


Figure 3: Comparison of the performance of Loquetier, FlexLLM and PEFT in fine-tuning tasks.



46.4× SLO
higher than
PEFT and FlexLLM

Figure 4: Comparison of the performance of Loquetier and PEFT in unified tasks. The 4 subplots correspond respectively to single-finetune & single-infer, single-finetune & multi-infer, multi-finetune & single-infer, and multi-finetune & multi-infer.

Experiments

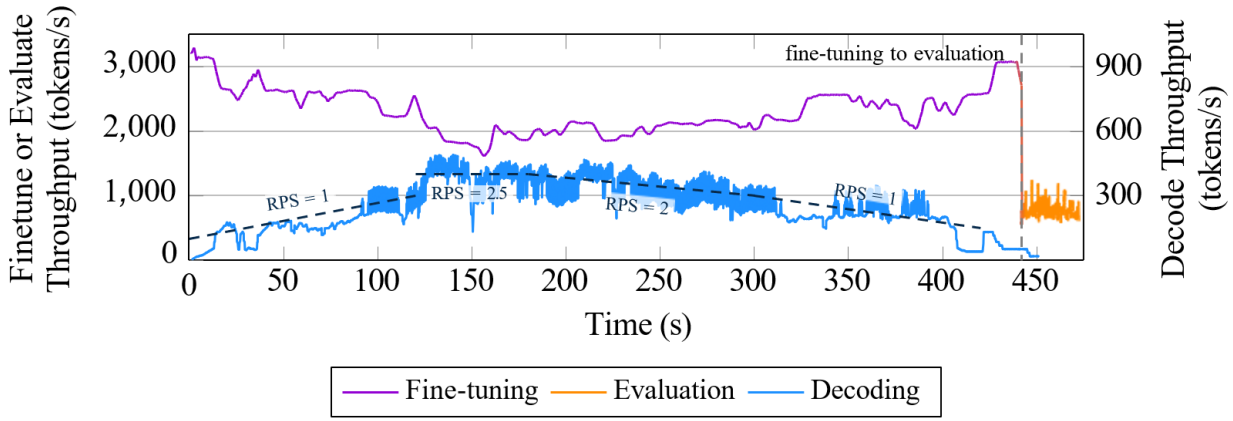


Figure 5: Performance of Loquetier under dynamic load in unified task.

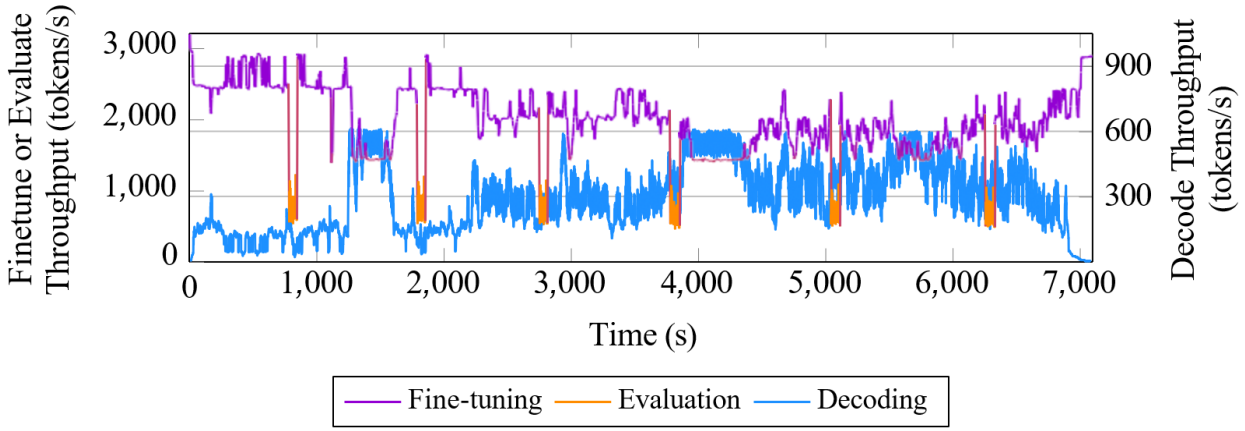


Figure 6: Performance of Loquetier under simulated real-world load in unified task.

Conclusion

We present **Loquetier**, a virtualized multi-LoRA framework that runs fine-tuning and inference tasks uniformly.

Loquetier performs well in the inference task, maintaining comparable efficiency in the fine-tuning task.

In the unified task and 2 simulations, inference efficiency is maintained as much as possible, well balancing the performance of fine-tuning and inference tasks.

Reference:

[1] Edward J Hu, et al. Lora: Low-rank adaptation of large language models. ICLR, 1(2):3,2022

[2] Yuxin Wang, et al. BurstGPT:A Real-World Workload Dataset to Optimize LLM Serving Systems. In Proceedings of the31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25),Toronto, ON, Canada, 2025. ACM. doi: https://doi.org/10.1145/3711896.3737413. URL https://doi.org/10.1145/3711896.3737413.

Thank you for your attention.