LlamaDuo: LLMOps Pipeline for Seamless Migration from Service LLMs to Small-Scale Local LLMs

Chansung Park[*♢], **Juyong Jiang**[*♡], **Fan Wang**[*♡], **Sayak Paul**[♠], **Jing Tang**[†♡♣]
[♢]Electronics and Telecommunications Research Institute
[♡]The Hong Kong University of Science and Technology (Guangzhou)
[♣]The Hong Kong University of Science and Technology
[♠]Hugging Face
{deep.diver.csp,csjuyongjiang,csfanwang,spsayakpaul}@gmail.com
jingtang@ust.hk

Abstract

The widespread adoption of cloud-based proprietary large language models (LLMs) has introduced significant challenges, including operational dependencies, privacy concerns, and the necessity of continuous internet connectivity. In this work, we introduce an LLMOps pipeline, "LlamaDuo", for the seamless migration of knowledge and abilities from service-oriented LLMs to smaller, locally manageable models. This pipeline is crucial for ensuring service continuity in the presence of operational failures, strict privacy policies, or offline requirements. Our LlamaDuo involves fine-tuning a small language model against the service LLM using a synthetic dataset generated by the latter. If the performance of the fine-tuned model falls short of expectations, it is automatically improved through additional fine-tuning using extra similar data generated by the service LLM. This multi-turn process guarantees that the smaller model can eventually match or even surpass the service LLM's capabilities in specific downstream tasks, offering a practical and scalable solution for managing AI deployments in constrained environments. Extensive experiments with leading-edge LLMs are conducted to demonstrate the effectiveness, adaptability, and affordability of LlamaDuo across various downstream tasks. Our pipeline implementation is available at https://github.com/deep-diver/llamaduo.

1 Introduction

The emergence of LLMs has significantly transformed a myriad of tasks and domains (Chowdhery et al., 2023; Gemini Team, 2023; Achiam et al., 2023; Touvron et al., 2023; Zhao et al., 2023; Jiang et al., 2024a,b). In particular, cloud-based proprietary LLMs, referred to as service models, such as GPT-4 (Achiam et al., 2023), Gemini 1.5 (Gemini Team, 2023), and Claude 3 (Anthropic, 2024), have exhibited exceptional capabilities when compared to their smaller, open-source counterparts (Chang et al., 2024). A notable survey involving 70 AI industry leaders from diverse enterprises reveals that approximately 80% of the enterprise market share is dominated by closed-source platforms, with a significant portion of this share attributed to OpenAI (Wang and Xu, 2024).

However, the increasing reliance on cloud-based service models presents significant challenges in terms of operational dependencies (Achiam et al., 2023), privacy concerns (Wu et al., 2024), and accessibility challenges (Ray, 2023). These challenges manifest in various ways, including potential service disruptions, heightened risks to data privacy due to the transmission of sensitive information to external providers, mandatory internet connectivity for utilization, and inconsistencies stemming from updates to service providers' LLMs (Hadi et al., 2023; Zhao et al., 2023). Additionally, the transition from proof-of-concept (PoC) development utilizing service LLMs to deployment with local models frequently leads to diminished prompt effectiveness owing to differences between models, subsequently resulting in a suboptimal experience for end-users (Naveed et al., 2023; Lyu et al., 2024). To address these concerns and ensure consistent service delivery, it is imperative to develop smaller, locally manageable LLMs that can operate independently of cloud-based infrastructures.

Recent studies have demonstrated that the strategic fine-tuning of smaller and open-source LLMs with high-quality synthetic data (Wang et al., 2023b; Xu et al., 2023) generated by service LLMs can achieve performances that are on par with, or even surpass, those of proprietary LLMs in specific downstream tasks (Chiang et al., 2023; Taori et al., 2023a; Luo et al., 2023; Abdin et al., 2024; Zhou et al., 2024). Motivated by these findings, we introduce an LLMOps pipeline namely LlamaDuo

[*]Equal contributors: Chansung Park, Juyong Jiang, and Fan Wang. Listing order is random.
[†]Corresponding author: Jing Tang.