designed to automatically facilitate the seamless migration from service-oriented LLMs to smaller, locally manageable models without the need for human intervention. Our pipeline begins with utilizing a task-specific initial dataset, referred to as the coverage dataset, to fine-tune a smaller open-source LLM. The performance of fine-tuned local LLMs is evaluated using a service LLMs-as-a-Judge strategy (Zheng et al., 2024). If the performance of the fine-tuned model falls short of expectations, we improve it by iteratively fine-tuning on additional synthetic data generated by the service LLM. LlamaDuo ensures that the smaller model is capable of eventually matching or even surpassing the service LLM's performance in specific downstream tasks, offering superior long-term economic advantages. Therefore, it presents a practical and scalable solution for managing AI deployments in environments where resources are limited. We conduct extensive experiments and analyses across a range of typical tasks, using popular service LLMs such as GPT4o, Claude 3 Sonnet, and Gemini 1.5 Flash, as well as local LLMs, including Gemma 2B and 7B, Mistral 7B, and LLaMA3 8B, to demonstrate that our LlamaDuo guarantees the smaller local LLMs possess the potential to eventually match or even exceed the performance of service LLMs in specific downstream tasks. To summarize, our contributions are as follows:

- We introduce LlamaDuo, an efficient and affordable LLMOps pipeline designed to facilitate seamless migration from service-oriented LLMs to smaller, locally manageable models without human intervention, ensuring service continuity in constrained environments.

- We employ a multi-turn approach using task-specific synthetic data generated by service LLMs to ensure that LlamaDuo empowers the smaller model to eventually match or even exceed the performance of the service LLM in specific downstream tasks.

- We substantiate the pipeline's robust performance and adaptability in real-world context through comprehensive experiments across a range of typical tasks, employing popular service LLMs as synthetic data generators and judges for well-known small local LLMs.

- We emphasize the significant economic advantages of LlamaDuo for investing in smaller, locally manageable LLMs and their deployment for sustained use, as opposed to the transient benefits derived from the token-based API usage of service LLMs.

2 Related Work

2.1 Alignment with Instruction Tuning

LLMs pretrained on massive corpora demonstrate remarkable capabilities across a wide range of tasks (Zhao et al., 2023; Cai et al., 2024; Yoo et al., 2024; Wang et al., 2024a). Despite their capabilities, a notable challenge with LLMs is their misalignment with user instructions, which limits their practical applications in real-world scenarios (Xu et al., 2023; Wang et al., 2023b). The misalignment stems from the initial pretraining objective of LLMs, which focuses on minimizing generation errors rather than adhering to human instructions (Ouyang et al., 2022; Chung et al., 2024). To solve the mismatch, instruction tuning is proposed, which enables LLMs to complete diverse tasks from instructions without significant computational resources or alterations to the model's architecture(Longpre et al., 2023; Muennighoff et al., 2023; Taori et al., 2023b). Specifically, instruction tuning involves supplementary training of pretrained LLMs with datasets structured as instruction-output pairs (Zhang et al., 2023). The efficacy of instruction tuning is largely contingent upon the quality and diversity of the instruction datasets employed (Wang et al., 2024b). However, the process of curating high-quality, diversified data is fraught with challenges, including the extensive time required for creation, privacy concerns, high costs, and the need for substantial human labor (Xu et al., 2023). In response to these challenges, recent studies have explored innovative methods for constructing instruction datasets, notably the utilization of LLMs for data synthesis (Liu et al., 2024).

2.2 LLM-synthetic Instruction Data

LLMs have demonstrated an unprecedented ability to comprehend and execute natural language instructions (Ouyang et al., 2022; Chung et al., 2024; Touvron et al., 2023). This ability is attributed to the process of training LLMs using substantial instruction datasets (Wang et al., 2023b). However, acquiring massive instruction datasets is challenging due to data scarcity, privacy issues, low data quality, and prohibitive costs associated with manual data curation (Abay et al., 2019; Xu et al., 2023;