downstream tasks, providing a promising research direction to maintain cloud-based LLMs' service continuity in constrained environments.

Limitations

While our LlamaDuo pipeline presents a promising solution for migrating capabilities from service-oriented LLMs to smaller local models, as depicted in Table 1, several limitations must be acknowledged. First, the reliance on synthetic datasets generated by the service LLM may introduce biases and safety issues inherent in the original model, potentially affecting the fine-tuned model's performance on specific tasks or datasets (Liu et al., 2024). Additionally, the effectiveness of the pipeline in transferring knowledge is contingent upon the quality and diversity of the synthetic data generated. If the data does not adequately cover the necessary scope, the fine-tuned model may struggle with tasks outside of the provided examples (Razeghi et al., 2022; Kandpal et al., 2023). Furthermore, the iterative fine-tuning process, while beneficial for performance enhancement, can be computationally intensive and time-consuming, potentially offsetting some gains in model efficiency, cost, and affordability. Another limitation is the potential plateau in performance gains after several SFT iterations, which could necessitate alternative strategies for further improvement, *e.g.,* reinforcement learning (RL) (Ouyang et al., 2022; Rafailov et al., 2023). Lastly, the pipeline assumes access to the service LLM for data generation, which may not always be feasible due to proprietary restrictions or API access limitations.

Ethical Considerations

Our work introduces several ethical considerations that require careful examination. Primarily, the process of generating synthetic datasets raises questions about data privacy and security, especially if the data contains sensitive or proprietary information. There is a risk that such data, if not properly anonymized and secured, could lead to privacy violations or unauthorized data exposure (Liu et al., 2024; Das et al., 2025). Moreover, the transfer of biases from the service LLM to the smaller model could perpetuate or even exacerbate existing biases, leading to unfair or discriminatory outcomes in certain applications. It is crucial to implement robust bias detection and mitigation strategies within the pipeline to safeguard against these risks. Additionally, the use of proprietary models for generating synthetic data necessitates transparency regarding data handling practices and the potential limitations of the resultant models (Wang et al., 2023b).

Broader Impact

Beyond the immediate focus of this paper, we believe that the introduction of the LlamaDuo pipeline has the potential to significantly impact the landscape of LLMs deployment, particularly in environments with constrained resources or stringent privacy requirements. By enabling the migration of capabilities from large service-oriented LLMs to smaller, locally manageable models, the pipeline empowers organizations to maintain LLMs functionalities independently of external service providers, enhancing operational resilience and reducing dependency. This can lead to increased accessibility to advanced LLMs for smaller entities or those operating in regions with limited internet connectivity.

Acknowledgements

Jing Tang's work is partially supported by National Key R&D Program of China under Grant No. 2024YFA1012700 and No. 2023YFF0725100, by the National Natural Science Foundation of China (NSFC) under Grant No. 62402410 and No. U22B2060, by Guangdong Provincial Project (No. 2023QN10X025), by Guangdong Basic and Applied Basic Research Foundation under Grant No. 2023A1515110131, by Guangzhou Municipal Science and Technology Bureau under Grant No. 2024A04J4454, by Guangzhou Municipal Education Bureau (No. 2024312263), by Nansha District Project (No. 2023ZD022), and by Guangzhou Industrial Information and Intelligent Key Laboratory Project (No. 2024A03J0628) and Guangzhou Municipal Key Laboratory of Financial Technology Cutting-Edge Research (No. 2024A03J0630). This work is supported by IITP grant funded by the Korea government(MSIT)[RS-2023-00215959, Development of Access Agnostic wired and wireless integrated optical access technology].

References

Nazmiye Ceren Abay, Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. 2019. Privacy preserving synthetic data release using deep learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference,*