where $\bigcup \psi(\cdot,\cdot,\cdot)$ represent a series of data post-processing operations, $\mathcal{D}_{synth}^{(t)}$ denotes synthetic data generated from service LLMs at t-th cyclicality, $\mathcal{S}_{\text{LLM}}$ and $\text{prompt}^{(synth)}$ are the service LLM and system prompt used for the data synthesis, respectively.

4 Experiments

In this section, we present a comprehensive evaluation of our LlamaDuo across a series of settings, demonstrating its robust performance and adaptability in real-world scenarios.

4.1 Experimental Settings

Tasks and coverage dataset. We select four categories of downstream tasks-summarization, classification, coding, and closed QA-based on their prevalent use and relevance to the operational scope of service LLMs. We utilize the open-source "No Robots" (Rajani et al., 2023) dataset as the coverage dataset. This coverage dataset consists of 10K high-quality prompt and response pairs across 10 categories, crafted by expert annotators. Specifically, we utilize four subsets of the coverage dataset, each corresponding to our targeted tasks. These subsets serve as seeds for generating synthetic data that can closely align with user expectations for LLM interactions.

Service and local LLMs. Considering the API cost effectiveness, rate limit, and model utility, we select popular service LLMs including GPT4o by OpenAI, Claude 3 Sonnet by Anthropic, and Gemini 1.5 Flash by Google to serve as synthetic data generators and judges. As for the small-scale local LLMs to be fine-tuned, we opt for the open-source Gemma 2B and 7B (Gemma Team, 2024), Mistral 7B (Jiang et al., 2023), and LLaMA3 8B (Meta, 2024) as the base models. This selection is motivated by our aim to rigorously evaluate the efficacy and adaptability of our proposed pipeline across diverse settings. The varying scales of base models facilitate a nuanced comparison, allowing us to assess the impact of model scale on performance improvements. However, as a model-agnostic LLMOps pipeline, our LlamaDuo can be generalized to various forms of service and local LLMs beyond the aforementioned models.

4.2 Implementation Details

We implement LlamaDuo using PyTorch and conduct experiments on $8 \times$ A100 (80GB) GPUs.

Synthetic dataset by service LLMs. We utilize the seeds selected from the train subset of the coverage dataset to prompt service LLMs to generate datasets, each comprising 300k samples. The specific prompt for data generation is presented in Figure 6 of Appendix A. Subsequently, we employ Locality-Sensitive Hashing (LSH) with MinHash and Rouge scoring mechanisms for data deduplication. Specifically, the LSH MinHash can efficiently identify and remove duplicate data samples, while the Rouge scoring mechanism ensures that the curated data exhibits high-quality and meaningful variations. After that, we acquire 256k samples for summarization tasks and 128k for other tasks.

Fine-tuning Local LLMs. We proceed to fine-tune the small local LLMs with $2^n\text{k}, n \in \{0, 1, \ldots, 8\}$ volumes of the synthetic dataset. To efficiently customize local LLM for a specific downstream task within constrained environments, we leverage QLoRA (Dettmers et al., 2024) for parameter-efficient fine-tuning with superior cost-effectiveness. The detailed configurations, which are tailored according to dataset sizes and tasks, can be found in Appendix B.

Batch inference. Each fine-tuned local model is prompted to generate $K = 4$ distinct responses, with each prompt sampled from the test subsets of the coverage dataset. To ensure fair comparisons, we maintain a consistent batch inference configuration across all fine-tuned models. The detailed configuration is depicted in Appendix B.

Service LLMs as judges. Following (Zheng et al., 2024), we employ pairwise comparison and single answer grading strategies to evaluate the response quality of the fine-tuned local LLMs. The corresponding prompts are given in Figure 5 of Appendix A. We utilize similarity and precision metrics. The similarity metric assesses the degree of correspondence between the generated responses and the ground truth, while the precision metric evaluates the accuracy of the match between the input prompts and their corresponding responses. To ensure reliability and mitigate inherent biases in the results, both metrics are quantified on a 0 to 100 scale, with each sample undergoing evaluation $M = 10$ times. The score of coverage percentage is set to $\zeta \in \{50, 70\}$.

4.3 Experimental Results

This section delves into the effectiveness and adaptability of the LlamaDuo pipeline, spanning different tasks with varying degrees of complexity, in-