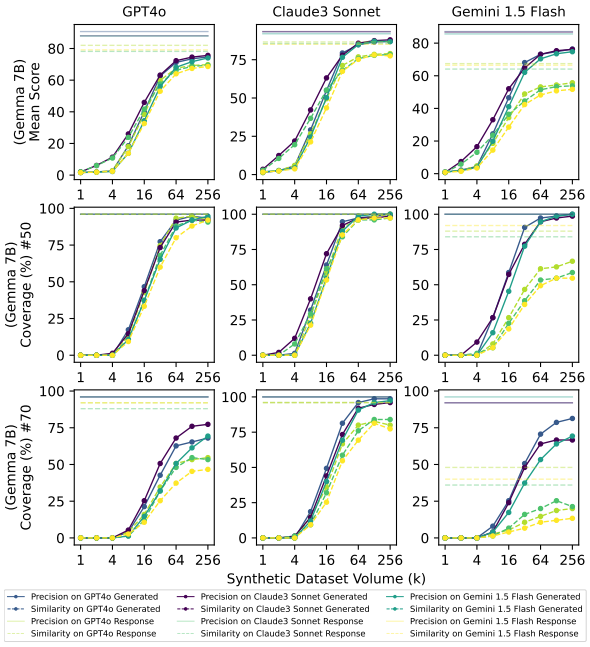

Figure 2: Performance of Gemma 7B fine-tuned on varied volumes of synthetic dataset produced by various service LLMs including GPT4o, Claude 3 Sonnet, and Gemini 1.5 Flash. The first to third columns represent the performance of the model evaluated by GPT4o, Claude 3 Sonnet, and Gemini 1.5 Flash as judges, respectively. The first row show mean scores, while the second and third rows shows the coverage percentage with 50 and 70 scores, respectively.

4.4 In-depth LLMOps Pipeline Analysis

In this section, we conduct an in-depth analysis of LlamaDuo through summarization task. Notably, the experimental findings exhibit a consistent pattern across various tasks, underscoring the generalizability of LlamaDuo.

Impact of synthetic dataset volume. We explore how the volume of synthetic dataset influences the performances of fine-tuned local LLMs, aiming to elucidate a scaling law (Kaplan et al., 2020; Hoffmann et al., 2022) on how the performance of fine-tuned models changes as the number of synthetic dataset samples increases. Overall, the Gemma 7B model exhibits consistent performance improvements and comes closer to the performance of service LLMs with increasing volumes of synthetic data, as assessed through precision and similarity metrics by diverse evaluators, as depicted in Figure 2. This suggests that fine-tuning local LLMs with synthetic data, which mimics the characteristics and patterns of real-world data, can bring the same effect as actual data. Thus, it paves a new way to eliminate the challenges of data scarcity, privacy concerns, and high costs associated with crafting

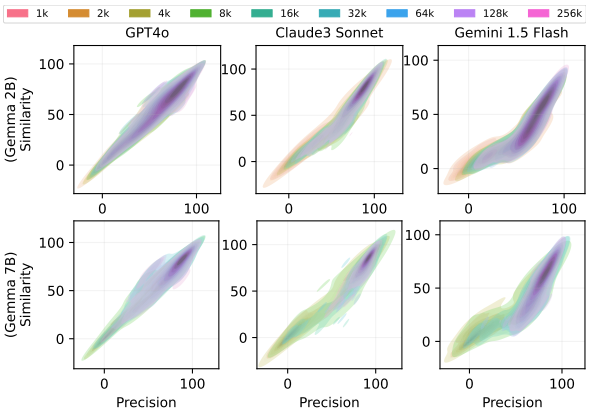

Figure 3: The KDE Plots of Precision v.s. Similarity by varied synthetic dataset volumes with 2^nk, $n \in \{0, 1, \ldots, 8\}$ and various evaluators with GPT4o, Claude 3 Sonnet, Gemini 1.5 Flash as judges from first to third columns, while the first and second rows represent the results of Gemma 2B (first row) and Gemma 7B (second row), respectively.

data (Liu et al., 2024). Notably, we observe that the synthetic data generated by Claude 3 Sonnet results in the highest-performing models, outperforming those fine-tuned with data produced by GPT4o and Gemini 1.5 Flash, in descending order. Moreover, when the synthetic dataset volume ranges from 64k to 256k, the Gemma 7B model reaches the performance saturation point and achieves performance that is much closer to, or equal to, that of service LLMs. This demonstrates the efficacy of our LlamaDuo in enabling smaller models to replicate or even surpass the performance of service LLMs in specific downstream tasks.

Impact of service LLMs as data generator and judge. As shown in Figure 2, we observe that the choice of service LLM for data generation does not significantly impact the performance of the fine-tuned models. Specifically, (1) a consistent trend of performance enhancement is observed with the increased volume of synthetic data, irrespective of the service LLM employed for data generation; (2) the local LLMs fine-tuned on synthetic data generated by GPT4o and Claude3 Sonnet typically lead to slightly better performance than those by Gemini 1.5 Flash. On the other hand, employing different service LLMs as judges manifests a more pronounced impact on the performance of the fine-tuned local LLMs. Overall, GPT4o and Gemini 1.5 Flash emerge as more rigorous judges compared to Claude 3 Sonnet, with Gemini 1.5 Flash assigning notably lower similarity scores. Moreover, we observe that in data sparsity scenarios (1k to 4k), the type of evaluators has minimal influ-