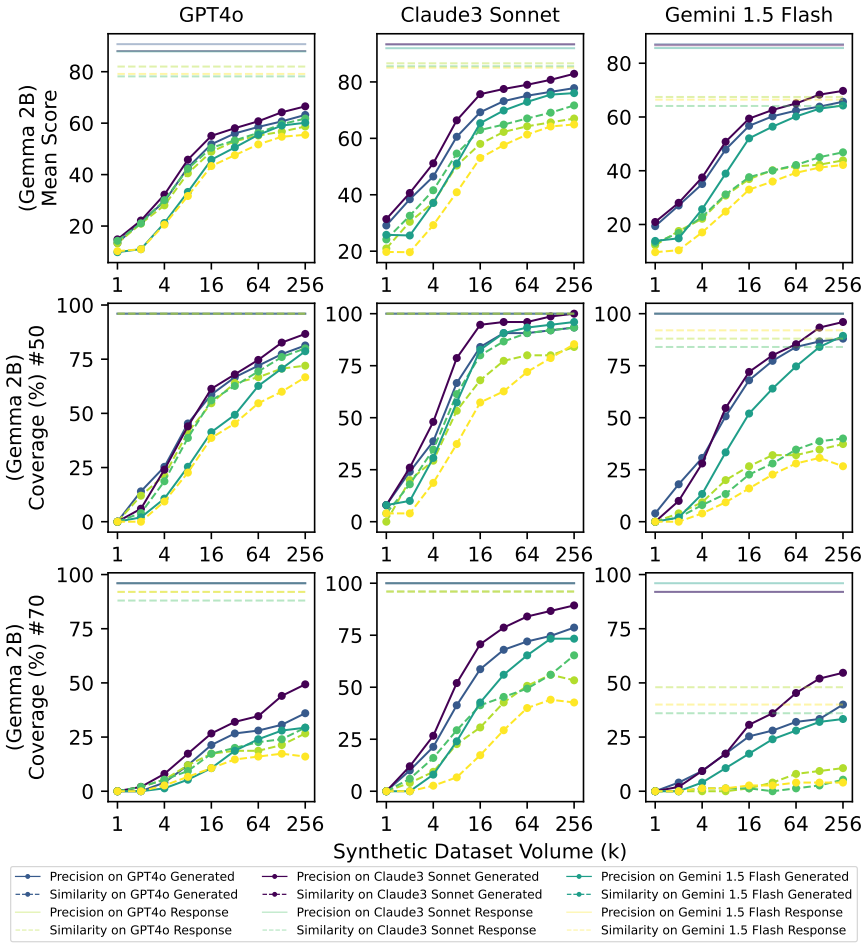

Figure 8: Performance of Gemma 2B fine-tuned on varied volumes of synthetic dataset producted by various service LLMs including GPT4o, Claude 3 Sonnet, and Gemini 1.5 Flash. The first to third columns represent the performance of the model evaluated by GPT4o, Claude 3 Sonnet, and Gemini 1.5 Flash as judges, respectively. The first row show mean scores, while the second and third rows show the coverage percentage with 50 and 70 scores, respectively.

cated computational budget.

C More Experimental Results

The performance of Gemma 2B fine-tuned on varied volumes of synthetic dataset produced by various service LLMs including GPT4o, Claude 3 Sonnet, and Gemini 1.5 Flash is shown in Figure 8.

D Case Study

This section delves into detailed case studies showcasing the enhanced capabilities of the aligned small-scale local LLMs. We use Gemma 2B and 7B models as examples to illustrate.

The cases (Figure 9-17) illustrate the performances of the aligned models across summarization, classification, coding, and closed QA tasks. Specifically, these models are tuned on distinct 128K datasets generated by GPT4o for each corresponding task. Each case provides evaluations by GPT4o, Claude 3 Sonnet, and Gemini 1.5 Flash, offering a comprehensive assessment of the precision and similarity of the models' responses.

To expand the scope of our analysis, we include two additional cases (Figure 11 and 12) to explore the summarization capabilities of the Gemma 2B and 7B models tuned with 256K synthetic datasets. These datasets are generated by GPT4o, Claude 3 Sonnet, and Gemini 1.5 Flash respectively, providing valuable insights into the models' adaptability to different training data sources.

The cases presented above demonstrate the capability of the aligned Gemma 2B and 7B models to produce high-quality responses. Additionally, the cases offer insight into how different service LLMs evaluate text. Through this comparative lens, we reveal discernible variances in judgment and assessment criteria, enriching our understanding of the models' operational dynamics.