Table 2: Monthly operational cost comparison between Gemma 7B and GPT4o under different workloads. For GPT4o, input and output token counts are represented in the format input/output.

	Light Workload		Heavy Workload	
	Gemma 7B	**GPT4o**	**Gemma 7B**	**GPT4o**
Fine-tuning	Cloud $800	- -	Cloud $800	- -
Serving Specs.	1 x L4 $2,539	300M/30M $1,950	8 x L4 $20,312	1500M/150M $9,750
Serving Elec.	165 kWh $30	- -	1319 kWh $240	- -
2 Months	$3,369	$3,900	$21,592	$19,500
12 Months	$3,699	$23,400	$23,992	$117,000

ence on the performance of the Gemma 7B model, suggesting that larger local LLMs exhibit diminished sensitivity to the choice of service LLM as a judge. To qualitatively demonstrate the differences when using various types of service LLMs as evaluators, Figure 3 presents the results as KDE plots, characterized by the dataset volume. We observe that GPT4o maintains consistency in its evaluations across both similarity and precision metrics. In contrast, Claude 3 Sonnet is found to be more lenient in scoring, while Gemini 1.5 Flash assigns higher precision scores but significantly lower similarity scores. This underscores the importance of strategically aligning the selection of service LLMs with specific task requirements.

4.5 Cost of Long-term Deployment

We elucidate the cost-effectiveness of our proposed LlamaDuo pipeline, by conducting a long-term operational cost comparison between the fine-tuning of the small LLMs (Gemma 7B) and the token-based API usage of service LLMs (GPT4o). In the context of local LLM deployment, the QLoRA fine-tuning process of Gemma 7B, utilizing a dataset containing 256K samples, necessitates approximately one hour to complete a single experiment on $8 \times$ A100 GPUs. This process incurs an estimated cost of $50, based on the price provided by Google Cloud Platform. Accounting for multiple iterations of hyperparameter optimization, we estimate that the total fine-tuning cost remains below $800, which is deemed to be negligible. Deploying a single instance of the Gemma 7B model with support for a 1024 context length necessitates 24GB of GPU memory, making the L4 GPU an appropriate choice. Depending on the projected workload, the Gemma 7B model can be deployed either on a single server equipped with one L4 GPU ($2,539) or across eight servers, each with one L4 GPU, with each server hosting a replica of the model instance

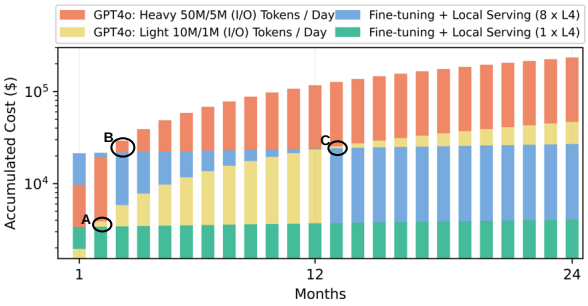

Figure 4: Long-term operational cost comparison between fine-tuning a local LLM and API-based token usage of GPT4o.

($20,312). In addition, the power consumption for each server is approximately $30 per month. For GPT4o, as of August 2024, the pricing is $5 and $15 per million tokens for input and output, respectively. We estimate that a light workload, utilizing 10 million input tokens and 1 million output tokens per day, incurs a daily cost of $65. Conversely, a heavy workload, consuming 50 million input tokens and 10 million output tokens per day, is estimated to cost $325 daily. The monthly operational cost comparison between Gemma 7B and GPT4o under different workloads is summarized in Table 2, demonstrating a significant advantage in fine-tuning and deploying a local LLM. Moreover, as depicted in Figure 4, after the first two months, the cost of using GPT4o under both light and heavy workloads exceeds that of setting up and running a local model deployed on $1 \times$ L4 GPU and $8 \times$ L4 GPU, respectively, as indicated by markers **A** and **B**. After one year, GPT4o's costs surpass those of deploying a local model in all scenarios, as denoted by marker **C**. These findings highlight the substantial economic benefits of investing in local LLM fine-tuning and deployment for long-term use. Avoiding recurring token-based charges and maintaining control over model customization further enhances the appeal of the LlamaDuo for cost-conscious users and scenarios.

5 Conclusion

In this study, we introduce LlamaDuo, the first automatic LLMOps pipeline designed to facilitate the seamless migration from service-oriented LLMs to smaller, locally manageable models. We conduct extensive experiments and analysis across a range of tasks with popular service and local LLMs to substantiate that LlamaDuo guarantees smaller local LLMs possess the potential to match or even exceed the performance of service LLMs in specific