The cost perspective of adopting Large Language Model-as-a-Service

Vasiliki Liagkou, Evangelia Filiopoulou, George Fragiadakis, Mara Nikolaidou *Member, IEEE*, Christos Michalakelis
Department of Informatics & Telematics
Harokopio University
Tavros, Greece
{vliagkou, evangelf, gfragi, mara, michalak}@hua.gr

2024 IEEE International Conference on Joint Cloud Computing (JCC) | 979-8-3503-8733-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/JCC62314.2024.00020

***Abstract*—Large Language Models (LLMs) are pivotal in generative AI applications. Consequently, major cloud providers, such as Amazon, Azure, and Google, introduce the offering of LLM-as-a-Service (LLMaaS) products to enable businesses to leverage NLP, data analysis, and predictive modeling in their cloud solutions. This paper explores the incorporation of LLM-as-a-Service solutions into business workflows with a focus on inference costs. We review various LLMaaS offerings and conduct a comparative analysis based on a real-world case study of an AI chatbot.**

***Index Terms*—Large Language Models, Inference Cost, Generative AI, Case Study, Cloud Platforms, LLM as a Service**

I. Introduction

The launch of GPT [1] in 2022 marked a significant milestone for Large Language Models (LLMs) and Generative AI (GenAI), drawing unprecedented attention from academia and industry. By leveraging natural language processing (NLP) to generate human-like text, LLMs have revolutionized human-machine interactions. Businesses are using LLMs to develop applications that automate tasks and enhance productivity [2]. With a projected CAGR of 33.2%, the global LLM market is expected to grow from USD 6.4 billion in 2024 to USD 36.1 billion by 2030 [3]. The LLM industry features many companies, with OpenAI leading at 59.1% market share [4], followed by Meta [5], Anthropic [6], and Google [7] as key players in the sector.

Due to computational needs, LLM deployment costs rise with size and complexity. Numerous businesses choose Models as a Service to reduce infrastructure costs. Leading cloud providers, including Amazon, Google, and Microsoft, offer LLM-as-a-Service with various deployment and price options. LLMaaS strategies differ in cost, making selection difficult. Concerns around data privacy, inference costs, and performance must be addressed before growing LLM consumption, despite lacking research on cost structures [8].

This study constitutes a first effort to provide a cost-based analysis of LLMaaS offerings and explore the cost perspective of using these new products. It is based on LLMaaS products offered by OpenAI and major cloud providers, Amazon, Google, and Azure. We investigate LLMaaS pricing models and examine how privacy and performance affect LLMaaS selection decisions. In addition, we use a real-world case study to explore the inference costs of several LLMaaS options.

II. LLM-as-a-Service

LLM-as-a-Service (LLMaaS) is a cloud-based delivery model that provides open-source and closed-source pre-trained LLMs as managed services to users, delivered either through proprietary APIs or hosting solutions.

Utilizing LLM-as-a-Service to develop an LLM application involves either engaging with an LLM provider directly or leveraging the services of a cloud provider.

LLM vendors provide consumers with API access solutions for utilizing closed-source models. Examples include the OpenAI API for GPT models or the Anthropic API for Claude models. These models offer a faster time to market with advanced features without requiring specialized technical expertise.

When leveraging a cloud provider for LLMaaS, users have the option to customize, deploy, and use pre-trained open-source or closed-source LLM models hosted in the cloud environment and managed by the providers. There are two approaches to leveraging a cloud provider:

a) Managed LLMs by cloud providers: Cloud providers manage pre-trained models, infrastructure, and tooling, offering both commercial and open-source LLMs via API interfaces. Users benefit from cloud flexibility, scalability, and robust security measures for data protection.

b) Hosted LLMs on Cloud Providers: Cloud providers enable deploying pre-trained open-source models, facilitating customization of both model and infrastructure to suit user needs and accommodate changing demand. This approach offers control but requires technical expertise and operational overhead.

III. LLM-as-a-Service products overview

Eleven distinct products using GPT and Llama2 models were selected for analysis, considering their market prominence [4]. They are summarized in Table I. Selections included OpenAI's GPT products and offerings from major cloud providers, Amazon, Azure, and Google. The major characteristics of these products include:

A. Pricing policy

LLMaaS products are offered using two discrete pricing policies: token-based and time-based pricing.

979-8-3503-8733-9/24/$31.00 ©2024 IEEE
DOI 10.1109/JCC62314.2024.00020

TABLE I: Consolidated Overview of LLMaaS Products and Pricing

Provider	Service Type	Model Name	Pricing Model	License	Privacy	PTU
OpenAI	Managed	GPT 3.5 Turbo	Token-based	Closed-source		
	Managed	GPT 4 Turbo	Token-based	Closed-source		
Azure	Managed [*OpenAI Service*]	GPT 3.5 Turbo	Token-based	Closed-source	Only Region	
	Managed [*AI Studio*]	Llama2 70B	Token-based	Closed-source	Only Region	
	Hosted [*Model Catalog*]	Llama2 70B	Time-based	Open-source	✓	
AWS	Managed [*BedRock*]	Llama2 70B	Token-based	Closed-source	Only Region	
	Hosted [*SageMaker*]	Llama2 70B	Time-based	Open-source	✓	
	PTU [*BedRock*]	Llama2 70B	Time-based	Open-source	✓	✓
	Hosted [*SageMaker*]	Llama2 7B	Time-based	Open-source	✓	
Google	Hosted [*Model Garden*]	Llama2 70B	Time-based	Open-source	✓	
	Hosted [*Model Garden*]	Llama2 7B	Time-based	Open-source	✓	

1) Token-based pricing: In this pricing model, LLM and cloud providers charge based on token consumption per query, reflecting the input and response length. Costs vary with model size, as larger models incur higher expenses due to greater computational demands. Token-based pricing enables scalable usage without large upfront investments, allowing users to optimize for performance and cost efficiency [9].

2) Time-based pricing: Time-based pricing charges for LLM usage per hour rather than per token processed. Customers have two hourly pricing options:

- **Provisioned Throughput Unit (PTU)**: AWS and Azure offer PTU, a managed LLM service with guaranteed throughput, unavailable from Google. PTU ensures stable performance by allocating model capacity. Users can choose 1-month or 6-month terms for flexible commitments and costs.
- **Hosting Costs**: For open-source models, customers pay hourly based on computational resource usage, without extra inference and fine-tuning costs. Pricing varies by model size and region.

B. Privacy

Since LLM providers share user data during inference and fine-tuning, regulated businesses may avoid them. Cloud providers let consumers choose hosting regions and manage infrastructure, ensuring data integrity and confidentiality.

- Region: Cloud LLM APIs let users select hosting regions to meet varying privacy standards [10]. For instance, Azure's GPT API offers more region control compared to the OpenAI API.
- Infrastructure: Hosting an open-source model allows users full control over the model and data, which remain within their server environment.

C. Performance

Although LLM performance can be measured by various metrics, this study focuses on throughput, as it is the only metric provided within the LLMaaS pricing scheme [11].

- Throughput: LLM throughput, measured in tokens/s, affects speed. Managed LLMs may experience reduced throughput during peak usage. Users hosting models may require added resources for performance [12]. Azure and AWS offer PTU time-based pricing for guaranteed throughput.

IV. Inference Cost across LLMaaS

To compare LLMaaS inference costs, we are based on the development of a typical AI chatbot designed for the documentation of a worldwide telecommunications solutions provider. A typical conversation in which a user engages with the chatbot is presented in Table II.

TABLE II: Details of the conversation model

Aspect	Details
User-Model Interaction	5 queries/conversation and corresponding responses [13].
Average User Input	300 words/query: 1250 input tokens for 5 queries [14].
Average Model Output	150 words/response: 1000 output tokens/conversation [14].
Total Token Calculation	Each conversation uses a total of 2250 tokens (input + output).
Token Count Determination	OpenAI's tokenizer [15], all models generate equivalent number of tokens.

To compare inference costs, we calculate the cost per conversation across discrete products.

a) Cost/conversation for Token-Based products: Equation 1 calculates the cost per conversation for token-based pricing products.

$$\text{Cost per Conversion} = \left(\frac{InTC \times \text{Price per 1000 InT}}{1000}\right) + \left(\frac{OuTC \times \text{Price per 1000 OuT}}{1000}\right) \quad (1)$$

where $InTC$= Input Tokens per Conversation, $OuTC$=Output Tokens per Conversation, InT= Input Tokens and OuT= Output Tokens

TABLE III: Conversation Cost in Token-based Products

Token-based products	Cost/conversation($)
OpenAI GPT 3.5-Turbo	0.0021
OpenAI GPT 4-Turbo	0.0425
Azure OpenAI Service GPT3.5	0.0021
Azure AI Studio Llama2-70B	0.0050
AWS Bedrock Llama2-70B	0.0037

TABLE IV: AWS model configuration performance analysis

Model	Details
SageMaker Llama2-70B	1546 tokens/s on ml.p4d.24xlarge [16]
SageMaker Llama2-7B	1207 tokens/s on ml.g5.12xlarge [16]
Bedrock PTU Llama2-70B	Cost: $21.18/hr, Max Processing: 300,000 tokens/min [18].

b) Cost/conversation for Time-Based products: To compute the inference cost per conversation for each time-based product, we convert the hourly costs into the expense of generating output for each conversation, using the Equation 2 [16].

$$\text{Cost/Conversion} = \frac{OuTC}{Throughput/3600} \times \text{Instance Cost/Hour} \quad (2)$$

where $OuTC$ = Output Tokens per Conversation,

V. Results

A. Inference cost for token-based products

Table III breaks down the inference cost per conversation for managed token-based offerings. This is based on Equation 1 and publicly available pricing data from OpenAI, Azure, and AWS.

The cost comparison shows significant variations. OpenAI GPT-3.5 is the cheapest due to its competitive pricing strategy for market adoption and its lack of privacy and performance guarantees. Compared to other token-based systems, GPT-4 Turbo is the most expensive, but its greater capabilities justify its high cost despite lacking privacy and performance assurances [17]. Although Azure offers GPT models at OpenAI's costs, pricing may vary dependent on agreements with Microsoft [11]. Finally, AWS Bedrock is the most expensive token-based solution for Llama2-70B since its agents eliminate the need for clients to implement this functionality.

B. Inference cost for time-based products

Time-based products include hosted services for Llama2-70B and Llama2-7B from cloud providers and the managed AWS Bedrock PTU Llama2-70B. Table V shows results derived from Equation 2. Throughput for hosted Llama2-70B and Llama2-7B is based on AWS benchmarks in SageMaker [16]. Table IV summarizes the configuration details. Similar cost calculations apply to Azure and Google.

TABLE V: Conversation Cost in Time-based Products

Time-based product	Cost/Conversation($)
AWS SageMaker Llama2-70B	0.0068
AWS SageMaker Llama2-7B	0.0016
AWS Bedrock PTU Llama2-70B	0.0027
Azure Model Catalog Llama2-70B	0.0067
Google Model Garden Llama2-70B	0.0072
Google Model Garden Llama2-7B	0.0007

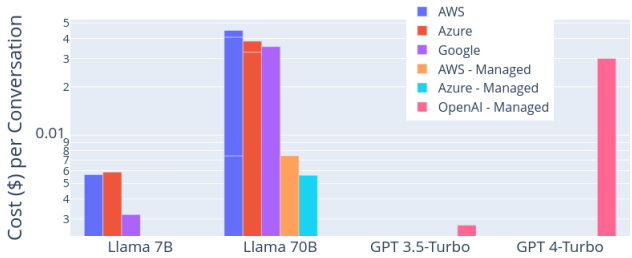

Fig. 1: Cost Breakdown by Provider and Model

Table V illustrates the strong correlation between LLMaaS costs, hardware configurations, and model throughput. Throughput varies per instance or model. Adjusting token generation rates or hourly instance fees may significantly affect inference costs. In addition, hosted LLMaaS for Llama2-70B costs approximately the same across three major cloud providers with similar hardware specifications, demonstrating a consistent cost structure.

Figure 1 provides a comparative view of the overall cost per conversation across all products. It demonstrates fluctuations in conversation unit prices among the products. Hence, the key factor in adopting an LLMaaS strategy is evaluating costs across different usage patterns on a daily basis to determine the most cost-effective service.

VI. Discussion

Figure 2 depicts the costs associated with daily conversation volumes when comparing token-based managed products to time-based hosted Llama2-70B. With AWS SageMaker Llama2-70B, there is an initial investment for the instance, then constant costs regardless of conversation volume until reaching the maximum supported conversations based on throughput. Beyond this threshold, scaling to a new instance becomes necessary for maintaining user experience and performance, as evident in the graph, particularly at 134,000 conversations. In contrast, token-based pricing models show linear cost increases with conversation volume, with GPT4 expenses rising rapidly, ultimately becoming prohibitively expensive.

Considering these observations, we can discern a pricing trend between token-based Llama2-70B and time-based hosted Llama2-70B. The latter consistently accrues higher costs with increased usage. Nevertheless, for customers with stringent privacy needs seeking to maintain authority over both the model and data, opting exclusively for hosted LLMs becomes imperative despite the associated higher expenses. Another

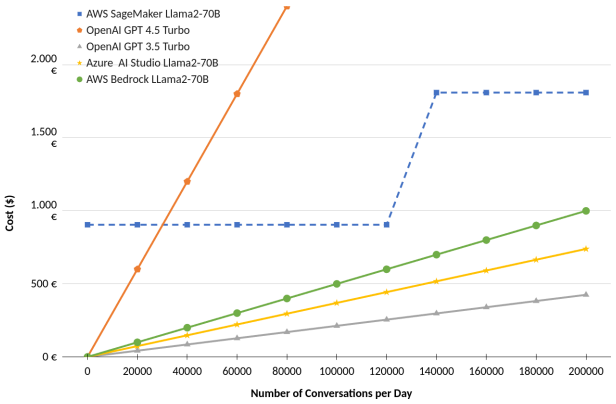

Fig. 2: Daily usage costs between token-based products and time-based (dotted line) AWS SageMaker Llama2-70B

way to lower the cost of Llama2 inference could be to use the smaller model Llama2-7B and find ways to improve its performance so that it is about the same as Llama2-70B [19].

The next step involves comparing the time-based hosted AWS SageMaker Llama2-7B, a smaller-scale alternative to Llama2-70B, against token-based products. Figure 3 illustrates that hosted Llama2-7B exhibits significantly lower costs compared to managed token-based Llama2-70B offerings. It also offers pricing similar to GPT3.5 while enabling complete data control. However, there's a usage threshold below which hosted Llama2-7B becomes more costly than AWS Bedrock Llama2-70B, Azure AI Studio Llama2-70B, and OpenAI GPT 3.5. These thresholds correspond to 30,000, 50,000, and 80,000 conversations per day for each respective service.

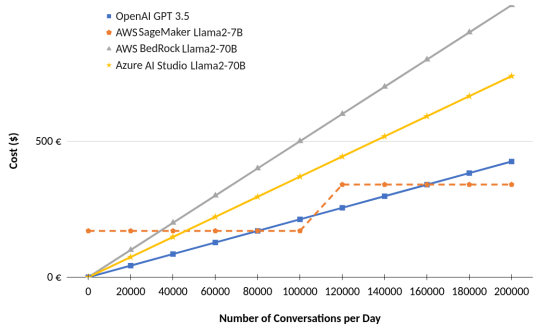

Fig. 3: Daily usage costs between token-based products and time-based (dotted line) AWS SageMaker Llama2-7B

Finally, AWS offers customers the chance to obtain Llama2-70B at a competitive price point comparable to token-based models, with performance guarantees. Figure 4 compares Llama2-70B in a token-based approach to Llama2-70B PTU in AWS. However, the break-even point for daily usage volumes, where the token-based solution becomes more cost-effective, is below 100,000 conversations.

VII. Conclusions

This study reviews 11 LLMaaS products and provides a cost-based evaluation. The findings indicate that token-based

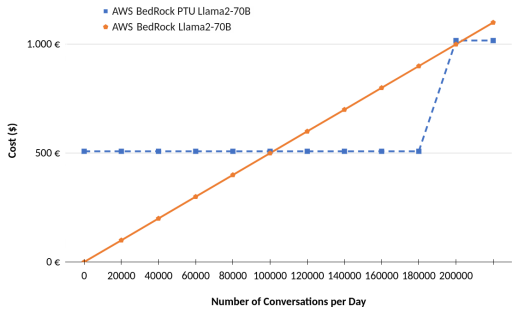

Fig. 4: Daily usage costs for token-based AWS Bedrock Llama2-70B vs time-based (dotted line) PTU Llama2-70B.

pricing presents a more appealing option for small-scale production applications. Cloud providers address the balance between cost and deployment simplicity through token-based pricing models. However, for larger-scale production scenarios where inference demands increase, time-based pricing models can prove advantageous. The optimal choice depends entirely on the specific use case, necessitating careful analysis by enterprises to recommend a deployment approach that aligns with business privacy and operational requirements.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
[2] [Online]. Available: https://www.grandviewresearch.com/industry-analysis/large-language-model-llm-market-report
[3] [Online]. Available: https://www.marketsandmarkets.com/Market-Reports/large-language-model-llm-market-102137956.html
[4] Arize. [Online]. Available: https://arize.com/blog/llm-survey/
[5] Meta. [Online]. Available: https://llama.meta.com/
[6] Anthropic. [Online]. Available: https://www.anthropic.com/
[7] Google. [Online]. Available: https://cloud.google.com/model-garden
[8] Mlops. [Online]. Available: https://mlops.community/wp-content/uploads/2023/07/survey-report-MLOPS-v16-FINAL.pdf
[9] S. Shekhar, T. Dubey, K. Mukherjee, A. Saxena, A. Tyagi, and N. Kotla, "Towards optimizing the costs of llm usage," *arXiv preprint arXiv:2402.01742*, 2024.
[10] U. P. Liyanage and N. D. Ranaweera, "Ethical considerations and potential risks in the deployment of large language models in diverse societal contexts," *Journal of Computational Social Dynamics*, vol. 8, no. 11, pp. 15–25, 2023.
[11] Azure. [Online]. Available: https://ai.azure.com/
[12] A. Ouyang, "Understanding the performance of transformer inference," Ph.D. dissertation, Massachusetts Institute of Technology, 2023.
[13] Outgrow. [Online]. Available: https://outgrow.co/blog/vital-chatbot-statistics
[14] L. Zheng, W.-L. Chiang, Y. Sheng, T. Li, S. Zhuang, Z. Wu, Y. Zhuang, Z. Li, Z. Lin, E. Xing *et al.*, "Lmsys-chat-1m: A large-scale real-world llm conversation dataset," *arXiv preprint arXiv:2309.11998*, 2023.
[15] OpenAI. [Online]. Available: https://openai.com
[16] AWS. [Online]. Available: https://aws.amazon.com/blogs/machine-learning/benchmark-and-optimize-endpoint-deployment-in-amazon-sagemaker-jumpstart/
[17] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
[18] Amazon. [Online]. Available: https://aws.amazon.com/
[19] Y. Chen, S. Qian, H. Tang, X. Lai, Z. Liu, S. Han, and J. Jia, "Longlora: Efficient fine-tuning of long-context large language models," *arXiv preprint arXiv:2309.12307*, 2023.