OPEN-SOURCE AND PROPRIETARY LLM MODELS

Pan Singh Dhoni, Independent Researcher

***Abstract*—In December 2022, the landscape of artificial intelligence was profoundly transformed with the advent of ChatGPT, a testament to the burgeoning capabilities of Large Language Models (LLMs). This breakthrough highlighted the power of advanced computational resources, particularly those furnished by leading chip manufacturers. The rapid proliferation of LLMs has been further evidenced by the increasing number of models being uploaded to platforms like Hugging Face [1] from developers around the globe. This research article delves into the comparative analysis of open-source and proprietary LLMs. It examines key metrics such as the volume of training tokens and the total parameters utilized in these models. By exploring these dimensions, the study aims to elucidate the benefits and trade-offs inherent to both categories, providing a comprehensive understanding of their respective impacts on the field of AI.**

***Index Terms*— open-source, proprietary, LLM**

I. INTRODUCTION

In today's technology landscape, organizations across various sectors are striving to harness the potential of Large Language Models (LLMs) to enhance their operations and services. Recognizing this demand, the research and development community is focused on creating advanced and highly sophisticated models. Hugging Face, a leading platform for model sharing and collaboration, reports that it has over 400,000 registered models, a number that continues to rise. This repository encompasses a diverse range of models, including LLMs and others. It is important to note, however, that while the total number of models exceeds 400,000, the subset of LLMs constitutes thousands, rather than hundreds of thousands, reflecting the specialized nature and complexity of these models.

This research article begins with an overview of Large Language Models (LLMs), highlighting their design and architecture as explored in the literature [2-5]. It examines the performance potential of open-source models, emphasizing that they can yield comparable or superior results. The study differentiates between open-source and proprietary LLMs, analyzing various open-source models with a focus on their training data, token usage, and model size parameters. A similar analysis is conducted for proprietary models. The article concludes by discussing the advantages and disadvantages of both types of models, providing a balanced view of their respective strengths and limitations.

II. LARGE LANGUAGE MODEL (LLM)

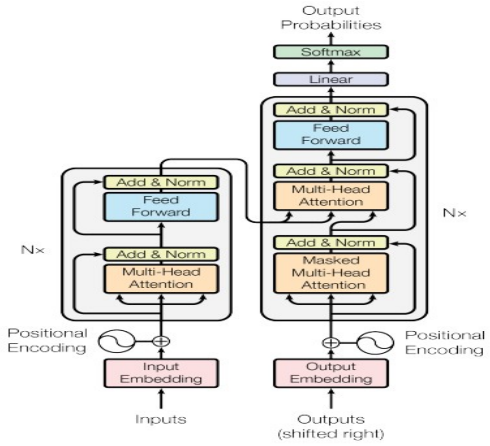
A large language model (LLM) is an advanced type of artificial intelligence algorithm that leverages neural network architectures, typically with billions of parameters, to analyze, interpret, and generate human language. These models utilize self-supervised learning, where they are trained on vast amounts of text data to perform various natural language processing (NLP) tasks. Applications of LLMs include text generation, machine translation, summarization, text-based image generation, automated coding, and conversational agents like chatbots.

III. METHODOLOGY

Design:

The Transformer model utilizes a structure of stacked self-attention layers and pointwise fully connected layers within both its encoder and decoder. Figure 1 illustrates the encoder on the left and the decoder on the right.

Figure 1: Image from a presentation by Ashish Vaswani, Aidan Gomez, one of eight co-authors of the 2017 paper that defined transformers (https://arxiv.org/abs/1706.03762).



.

IV. Open-source and Proprietary LLM's

Open source: Open-source LLM models are free to use and modify. However, for full-feature access, including support, organizations may need to purchase a license. This article will compare a selection of these models, focusing on key aspects such as the number of tokens they were trained on, their model size in terms of parameters, the organizations that developed them, and their release year.

Below tables have taken information for Llama 2 [6], BLOOM [7], Falcon 180B [8], XGen-7B [9], OPT-175B [10] and BERT [11-13].

Table 1. Open-Source Models

Open-Source LLM Models	**Trained on tokens**	**Model Size (Parameters)**	**Owned By**	**Year**
Llama 2	2 T	7B - 70 B	Meta	2023
BLOOM	366 B	176 B	Big Science open research	2022
Falcon 180B	3.5 T	180 B	Technology Innovation Institute	2023
XGen-7B	1.5 T	7 B	Salesforce	2023
OPT-175B	180 B	175 B	Meta	2022
BERT	3.3 B	340 M	Google	2018

Table 2. Open-Source Models Training on tokens.

Open-Source LLM Models	**Training on tokens (Billions)**
Llama 2	2000
BLOOM	366
Falcon 180B	3500
XGen-7B	1500
OPT-175B	180
BERT	3300

.

Figure 2: Training on tokens vs LLM models

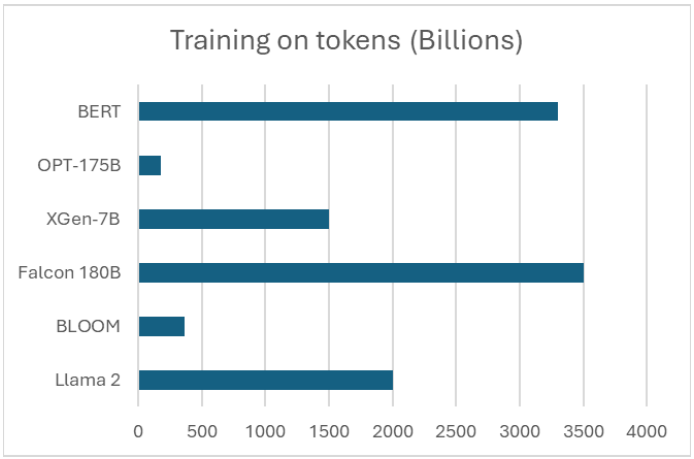
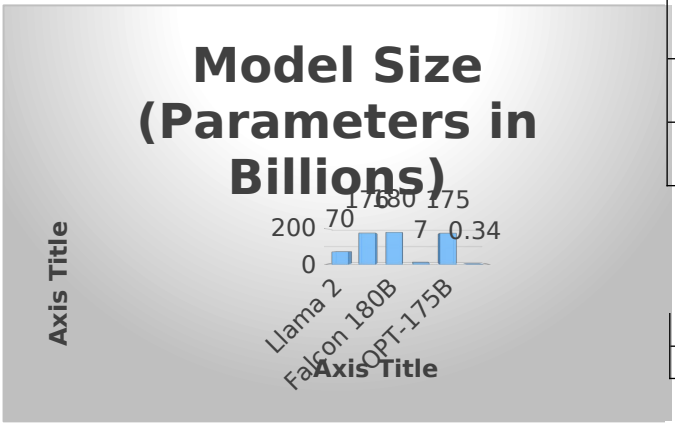


Table 3. Open-Source Model Size (Parameters in Billions).

Open-Source LLM Models	Model Size (Parameters in Billions)
Llama 2	70
BLOOM	176
Falcon 180B	180
XGen-7B	7
OPT-175B	175
BERT	0.34

Figure 3: Model Size in Billions



LLM Models	tokens. (Billions)	Parameters (Billions)		
gpt4	300	175	OpenAI	2023
Claude 3	0.001	10	Anthropic	2023
gemini	200	130	Google	2023

Open-Source Benefits & Risks:

Table 5: Benefits & Risk

Fine-Tunning	Bias
Community	Security

Open-Source vs proprietary LLM's:

Table 6: Difference between Open-Source & Proprietary LLM

Feature	Open-Source LLMs	Proprietary LLMs
Cost	Free to use, with some degree	Requires a paid license
Customization	High, with full access to source code	Limited by licensing terms
Support	Community-based support	Professional support from the company
Transparency	Fully transparent	Often opaque, limited insight into internals
Updates and Maintenance	Community-driven	Regular, company-provided updates

Proprietary: Proprietary models are owned by companies that have full control over their usage. Access to these models typically requires a license, which imposes restrictions and costs to use.

Below tables have taken information for gpt4 [14-16], Claude 3 [17-18], gemini [19].

Table 4. Proprietary Model Size (Parameters in Billions)

Open-Source	Trained on	Model Size	Owned By	Year

Table 7: Extension of Table 6

Open-Source Models	Proprietary (Paid) Models
• Use as off-the-shelf or fine tune • provides flexibility for customizations • can be smaller in size to save cost • commercial / non-commercial use	• Usually offered as LLMs-as-a-service • Some can be fine-tuned • Restrictive licenses for usage and modification
Examples: LLaMA/ mosai etc.	Examples: OpenAI, ChatGPT
Pros: Development is slow	Pros: Development is fast, quality will be better
Cons: Cost is low, no vendor lock-in	Cons: Cost is high, Vendor lock-in, Data privacy/ Security

V. Conclusions

This paper provides a comprehensive comparison between open-source and proprietary large language models (LLMs), focusing on key parameters such as cost, customization, support, and transparency. It highlights the distinct benefits and risks associated with open-source models, such as cost-effectiveness and flexibility, against the professional support and advanced features of proprietary models. By delineating these differences, this paper aims to inform stakeholders of the critical considerations involved in selecting the most appropriate LLM for their needs, balancing the trade-offs between openness and controlled access.

.

References

1. https://huggingface.co/
2. Reto Gubelmann, et. al: https://link.springer.com/chapter/10.1007/978-3-031-64312-5_21
3. Gautam B, Anupam Purwar, et. al: https://arxiv.org/abs/2406.11424
4. Sanjay Kukreja, et. al: https://dl.acm.org/doi/abs/10.1145/3647782.3647803
5. Zilin Ma, et.al: https://dl.acm.org/doi/abs/10.1145/3613905.3651100
6. https://llama.meta.com/llama2/
7. Teven Le Scao, et ai: https://ar5iv.labs.arxiv.org/html/2211.05100
8. Ebtesam Almazrouei, et. al: https://arxiv.org/abs/2311.16867
9. https://blog.salesforceairesearch.com/xgen/
10. https://ai.meta.com/blog/democratizing-access-to-large-scale-language-models-with-opt-175b/
11. https://www.datacamp.com/blog/what-is-bert-an-intro-to-bert-models
12. https://huggingface.co/docs/transformers/model_doc/bert
13. https://huggingface.co/blog/bert-101
14. https://openai.com/index/gpt-4/
15. https://www.technologyreview.com/2023/03/14/1069823/gpt-4-is-bigger-and-better-chatgpt-openai/
16. https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/
17. https://claudeaihub.com/claude-3-models-compared/
18. https://paperswithcode.com/paper/model-card-and-evaluations-for-claude-models
19. https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#gemini-15