Comparison of Large Language Models for Deployment Requirements

Alper Yaman†*, Jannik Schwab†, Christof Nitsche†, Abhirup Sinha† and Marco Huber†
†Department Cyber Cognitive Intelligence
Fraunhofer Institute for Manufacturing Engineering and Automation IPA, Stuttgart, Germany
Email: firstname.lastname@ipa.fraunhofer.de

***Abstract*—Large Language Models (LLMs), such as Generative Pre-trained Transformers (GPTs) are revolutionizing the generation of human-like text, producing contextually relevant and syntactically correct content. Despite challenges like biases and hallucinations, these Artificial Intelligence (AI) models excel in tasks, such as content creation, translation, and code generation. Fine-tuning and novel architectures, such as Mixture of Experts (MoE), address these issues. Over the past two years, numerous open-source foundational and fine-tuned models have been introduced, complicating the selection of the optimal LLM for researchers and companies regarding licensing and hardware requirements. To navigate the rapidly evolving LLM landscape and facilitate LLM selection, we present a comparative list of foundational and domain-specific models, focusing on features, such as release year, licensing, and hardware requirements. This list is published on GitLab and will be continuously updated.**

Keywords-generative AI; large language models; model comparison, HuggingFace.

I. Introduction

Large Language Models (LLMs) like Generative Pre-trained Transformer (GPT) are advanced Artificial Intelligence (AI) models designed to generate human-like text in response to the input they receive. These foundational models differ in underlying architecture, training procedures, and training data. They are trained on vast datasets containing a diverse range of internet text. They work by predicting the next word in a sequence, making them proficient at generating coherent sentences, and even writing poems or computer scripts.

The ability of LLMs to generate contextually relevant and syntactically correct text has revolutionized fields, such as content creation, customer service, and software development. LLMs are also integral in developing tools for language translation, summarization, and question-answering systems, enhancing accessibility and efficiency. Furthermore, they contribute significantly to research in natural language understanding and generation, pushing the boundaries of AI's capabilities in understanding complex language constructs.

However, LLMs can produce hallucinations, i.e., generating biased or incorrect information, which raises major concerns about their use in sensitive areas like law and healthcare. To address these drawbacks, pre-trained models are fine-tuned with domain-specific, task-specific corpora or instructions. Another method is Mixture-of-Experts (MoE) LLMs, where a set of LLMs (experts) attend to different parts of the input space. This concept is similar to ensemble methods in traditional machine learning, where the outputs from a set of models are voted to provide a single, more accurate outcome.

Despite these challenges, LLMs continue to be a pivotal area of research and development, resulting in a vast number of scientific articles. New jargon has rapidly emerged concerning the operation and evaluation of LLMs, including terms, such as prompt engineering, instruction-based fine-tuning [1], and Retrieval Augmented Generation (RAG) [2]. Additionally, the evaluation of the accuracy and performance of LLMs has been questioned, leading to the proposal of various metrics [3]. Multiple surveys have been published that provide comprehensive insights into recent advancements [4][5], discuss evaluation metrics from the perspective of explainability [6], and aim to align LLMs with human expectations [7].

In addition to closed-source cloud-based LLMs like ChatGPT, numerous models have been uploaded to HuggingFace for community use. However, these models vary in features, such as model size, embedding dimensions, and max token count, with details listed on platforms like HuggingFace and Github, and surveys [4][5]. This variability makes it challenging for companies and researchers to select an LLM that meets specific requirements, particularly when the model is intended for local deployment.

The aim of this study is to provide a comparative list of foundational and domain-specific models to support companies and researchers in selecting LLMs. In section II, we explain some of the existing LLMs lists, their content, and the parameters with which they are compared. In section III, we detail which models are selected and which features are compared. In section IV, basic statistics about the listed LLMs are provided, and a part of the comparison list is shown. In section V, further information is given about how the list will be maintained in the feature and the limitations of this study.

II. Related Work

As of May 2024 when this study was performed, HuggingFace had approximately 65 pre-trained LLMs for text generation tasks pertaining to the English language. Additionally, many fine-tuned models, based on the pre-trained models, have been uploaded to HuggingFace [8]. This platform has a couple of leaderboards that compare the fine-tuned models using a framework for few-shot language model evaluation [9]. The Open LLM Leaderboard compares models regarding their type, architecture, model precision, average accuracy, as well as accuracy values calculated separately using various datasets and benchmarks. Another leaderboard is Massive Text Embedding Benchmark (MTEB) Leaderboard illustrating the model size, memory usage, embedding dimensions, max tokens, average overall accuracy from 56 datasets, and average accuracies for classification, clustering, pair classification, reranking, retrieval, STS, and summarization from 12, 11, 3, 4,

15, 10, and 1 datasets, respectively [10]. A total of 281 models are compared with 159 datasets for 113 languages. LMSYS Chatbot Arena Leaderboard is a crowdsourced open platform to evaluate LLMs [11]. As of April 24, 2024, 91 models were evaluated using 800,000 human pairwise comparisons to rank them with the Bradley-Terry model [12]. Additionally, there are some Github repositories [13] and websites [14] that provide rough comparisons. Note that none of these leaderboards provides comprehensive details when companies and researchers encounter technical challenges when they deploy an LLM on their own hardware.

These tables compare the success scores of the LLMs along with their basic information (e.g., type and architecture) but omit the requirements for deployment. Including these requirements is essential to streamline the feasibility analysis process when selecting the most suitable LLM. Our comparison list addresses these needs by providing information on hardware and licensing requirements.

III. Proposed Work

In this study, we created an extensive comparison list of LLMs for researchers and companies to simplify LLM selection. Since there are numereous fine-tuned models, we primarily focused on covering base foundational LLMs, as much as possible. Nevertheless, some existing domain-specific (e.g., in the medical domain) fine-tuned models were included. We then defined the model features that help users to select the correct LLM. To easily distinguish between different LLMs, we provided both LLM names and families together with the model features, such as release year, license types, and hardware requirements.

The outcome of this study, in the form of a comparison table, is published on a GitLab page for community use. Since new LLMs and their derivatives are continually being developed, this is an ongoing effort, and the GitLab page will be updated regularly[15].

A. Model Selection

We selected 108 LLMs based on the criteria of being open-source and having been published in or after 2023. Approximately 20 of them are foundational LLMs, such as Mistral, LLaMA-2, LLaMA-3, Code LLaMA, Gemma, RecurrentGemma, Falcon, Dolly, etc. Some fine-tuned LLMs, such as BioMistral, Meditron, and Medicine-LLM, as well as several MoE LLMs (e.g., Mixtral, Grok-1, and DBRX) were included.

B. Model Features

We included information on LLM families and the versions existing within the LLM families. The sizes (i.e., number of parameters) and release dates were listed to track the gradual development in this field.

Furthermore, we also investigated the commercial aspects of the listed open-source LLMs and listed the license information. Since understanding the licenses can be difficult for readers, in another column, we clarified if the licenses allow for commercial usage of the model (with or without any restrictions) or not.

In addition, we included information on minimum memory requirements (RAM and vRAM) and required disk space for complete fine-tuning and inference. Note that these requirements are applicable for loading the 5-bit quantized versions of the models. Loading models with full-precision floating point numbers usually requires twice or four times more memory relative to their parameters.

IV. Results

A small subset of our resulting table is shown in Table I [15]. The information on LLMs, along with their families, license, and memory requirements is listed to provide a quick overview of the LLMs for the specific needs and use cases of researchers and companies.

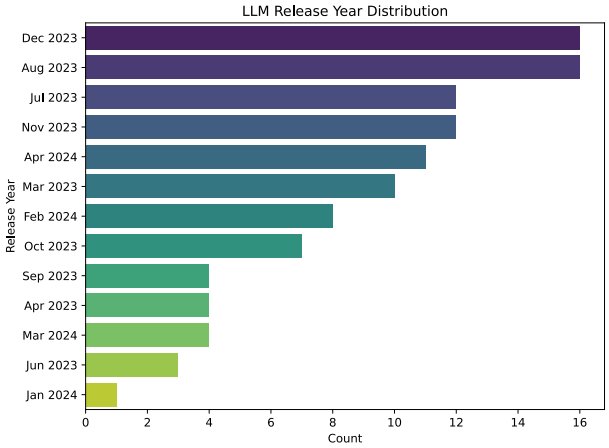

Figure 1. Release Year Distribution of Listed LLMs

Figure 1 shows the distribution of release date, indicating that; most of the LLMs we listed were released in 2023. Note that the most recent LLMs on our list were released in April 2024.

Figure 2 shows the distribution of model size, indicating that; most of our listed LLMs have 7 billion parameters. The size of the rest of the models ranges from 13 billion to 314 billion parameters). The lower number of parameters can allow an LLM to be deployed on edge devices, e.g., NVIDIA Jetson while the larger ones require more hardware resources.

Table II shows the distribution of license categories among our listed LLM models. Regarding commercial usage of the listed LLMs, around 51% of models have permissive licenses (Apache 2.0, MIT, Gemma) that allow for commercial usage without permission from model authors. Additionally, approximately 32% of listed LLMs have limited commercial usage licenses (LLaMA-2, LLaMA-3, DataBricks Open Model License). Models with such licenses require permission from model authors if commercial usage exceeds 700M monthly active users. In Table I, such models are denoted as “Partial” commercial usage.

TABLE I. A SNAPSHOT OF THE TABLE OF CURRENT OPEN-SOURCE LLMS

						Fine-tuning	Inference		
Family	**Name**	**Release Year**	**Size (B Parameters)**	**License type**	**Commercial Usage**	**Min. GB GPU**	**Min. GB RAM**	**Min. GB GPU**	**Min. GB Disk Space**
Code	Code-13B	Dec 23	13	CC-BY-NC-ND 4.0	No	26	11.73	5.4	9.23
	Code-33B	Dec 23	33	CC-BY-NC-ND 4.0	No	66	25.55	13.5	23.05
CodeLLaMA	7B	Aug 23	7	LLaMA-2	Partial	14	7.28	2.8	4.78
	7B-Instruct	Aug 23	7	LLaMA-2	Partial	14	7.28	2.8	4.78
	7B-Python	Aug 23	7	LLaMA-2	Partial	14	7.28	2.8	4.78
	13B	Aug 23	13	LLaMA-2	Partial	26	11.73	5.4	9.23
	13B-Instruct	Aug 23	13	LLaMA-2	Partial	26	11.73	5.4	9.23
	13B-Python	Aug 23	13	LLaMA-2	Partial	26	11.73	5.4	9.23
	34B	Aug 23	34	LLaMA-2	Partial	68	26.84	14.2	23.84
	34B-Instruct	Aug 23	34	LLaMA-2	Partial	68	26.84	14.2	23.84
	34B-Python	Aug 23	34	LLaMA-2	Partial	68	26.84	14.2	23.84
LLaMA-2	7B	Jul 23	7	LLaMA-2	Partial	14	7.28	2.8	4.78
	7B-Chat	Jul 23	7	LLaMA-2	Partial	14	7.28	2.8	4.78
	7B-Coder	Dec 23	7	LLaMA-2	Partial	14	7.28	2.8	4.78
	13B	Jul 23	13	LLaMA-2	Partial	26	11.73	5.4	9.23
	13B-Chat	Jul 23	13	LLaMA-2	Partial	26	11.73	5.4	9.23
	70B	Jul 23	13	LLaMA-2	Partial	140	51.25	29.3	48.75
	70B-Chat	Jul 23	70	LLaMA-2	Partial	140	51.25	29.3	48.75
Med42	70B	Nov 23	70	Med42	No	140	51.25	29.3	48.75
Starling LM	7B-Alpha	Nov 23	7	CC-BY-NC 4.0	No	14	7.63	2.7	5.13
	Alpha 8X7B MoE	Dec 23	47	CC-BY-NC 4.0	No	94	34.73	17.3	32.23
WizardLM	7B-v1.0	Apr 23	7	Non-commercial	No	14	7.28	2.8	4.78
	13B-v1.2	Jul 23	13	LLaMA-2	Partial	26	11.73	5.4	9.23
	30B-v1.0	Jun 23	30	Non-commercial	No	60	25.55	13.5	23.05
	70B-v1.0	Aug 23	70	Non-commercial	No	140	51.25	29.3	48.75
Zephyr	3B	Nov 23	3	StabilityAI Non-Commercial Research Community License	No	6	4.49	1.2	1.99
	7B-Alpha	Oct 23	7	MIT	Yes	14	7.63	2.7	5.13
	7B-Beta	Oct 23	7	MIT	Yes	14	7.63	2.7	5.13
BioMistral	7B	Feb 24	7	Apache 2.0	Yes	14	7.63	2.7	5.13
	7B-DARE	Feb 24	7	Apache 2.0	Yes	14	7.63	2.7	5.13
	7B-TIES	Feb 24	7	Apache 2.0	Yes	14	7.63	2.7	5.13
	7B-SLERP	Feb 24	7	Apache 2.0	Yes	14	7.63	2.7	5.13
TinyLLaMA	1.1B-Chat-v1.0	Jan 2024	1.1	Apache 2.0	Yes	2.2	3.28	0.5	0.78

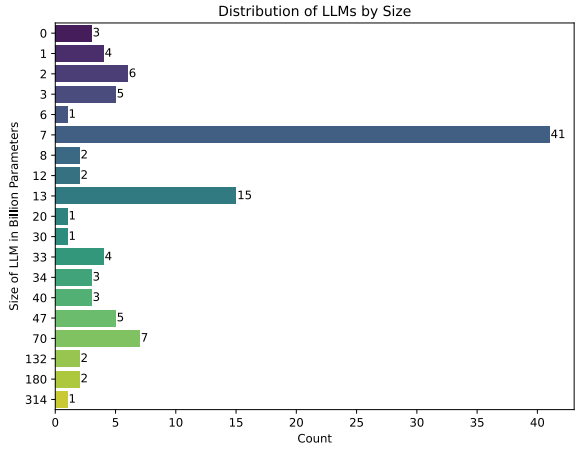

Figure 2. Distribution of LLM Size in Billion Parameters

TABLE II. LICENSE DISTRIBUTION OF OPEN-SOURCE MODELS IN OUR LIST

License Type	**Count**	**Percentage (%)**
Apache 2.0	36	33.33
LLaMA-2	29	26.85
Gemma	12	11.11
MIT	7	6.48
CC-BY-NC 4.0	5	4.63
CC-BY-NC-ND 4.0	4	3.70
LLaMA-3	4	3.70
Non-commercial	3	2.78
Microsoft Research License	2	1.85
Databricks Open Model License	2	1.85
Falcon-180B TII license	2	1.85
Med42 (derivative of LLaMA-2)	1	0.93
StabilityAI Non-Commercial Research Community License	1	0.93
Total	108	—

Our comparison table includes LLMs that have been specifically fine-tuned for the medical domain. Reducing hallucinations is particularly crucial in the medical field, as the generated responses may be used for diagnosis and treatment. Consequently, medical LLMs like BioMistral, Medicine-LLM, and Meditron have been fine-tuned by their developers using textual data from PubMed Central Open Access, internationally recognized medical guidelines, and a meticulously curated medical corpus.

V. CONCLUSION

In this paper, we proposed a comprehensive list of LLMs. This list is aimed at supporting researchers and companies in selecting LLM that is suitable for their use case, needs, and hardware requirements. This list is an ongoing effort and will be updated as new pre-trained or fine-tuned LLMs arrive.

Fine-tuning capability of LLMs has lad to many derivations of them for specific use cases. Since listing every fine-tuned LLM may not help researchers and companies and on the opposite; may confuse them more, this list does not cover all the fine-tuned versions of foundational LLMs. Another limitation is that the proposed list may not include the latest LLMs since the update frequency of the table may not align with the publication of new ones.

In future work, we will include more domain-specific models to list the LLM options for different applications. Furthermore, we will assess user feedback and highlight the advantages and disadvantages of the recommended deployments. Note that, in this study, the LLMs listed were not tested. The requirements provided by HuggingFace and the developers of LLMs will be verified as part of the future work.

ACKNOWLEDGMENT

We thank Nehal Darwish (University of Stuttgart, Institute of Industrial Manufacturing and Management (IFF)) for preparing the first draft of the comparison list.

REFERENCES

[1] S. Zhang *et al.*, "Instruction tuning for large language models: A survey," *arXiv preprint arXiv:2308.10792*, 2024.

[2] Y. Gao *et al.*, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2024.

[3] Y. Chang *et al.*, "A survey on evaluation of large language models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, pp. 1–45, Mar. 2024, ISSN: 2157-6904. DOI: 10.1145/3641289.

[4] W. X. Zhao *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[5] J. Huang and K. C.-C. Chang, "Towards reasoning in large language models: A survey," *arXiv preprint arXiv:2212.10403*, 2023.

[6] H. Zhao *et al.*, "Explainability for large language models: A survey," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 2, pp. 1–38, Feb. 2024, ISSN: 2157-6904. DOI: 10.1145/3639372.

[7] Y. Wang *et al.*, "Aligning large language models with human: A survey," *arXiv preprint arXiv:2307.12966*, 2023.

[8] E. Beeching *et al.*, "Open llm leaderboard," Accessed: 2024-05-28, 2023, [Online]. Available: https://huggingface.co/open-llm-leaderboard.

[9] L. Gao *et al.*, *A framework for few-shot language model evaluation*, version v0.0.1, Sep. 2021. DOI: 10.5281/zenodo.5371628.

[10] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "Mteb: Massive text embedding benchmark," *arXiv preprint arXiv:2210.07316*, 2022. DOI: 10.48550/arxiv.2210.07316.

[11] W.-L. Chiang *et al.*, "Chatbot arena: An open platform for evaluating llms by human preference," *arXiv preprint arXiv:2403.04132*, 2024.

[12] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952, ISSN: 00063444.

[13] E. Yan, "Open llms," Accessed: 2024-05-28, [Online]. Available: https://github.com/eugeneyan/open-llms.

[14] "The llm index," Accessed: 2024-05-28, [Online]. Available: https://sapling.ai/llm/index.

[15] A. Yaman, J. Schwab, C. Nitsche, A. Sinha, and M. Huber, "Gen-ai model overview table," Accessed: 2024-06-13, 2024, [Online]. Available: https://technology-project-aimv-projects-generative-ai-54af1e2b8cbbab0a.pages.fraunhofer.de (visited on 2024).