Performance Analysis of Llama 2 Among Other LLMs

Donghao HUANG[1,2], Zhenda HU[3], Zhaoxia WANG[1*]
School of Computing and Information Systems, Singapore Management University, Singapore[1]
Research and Development, Mastercard, Singapore[2]
School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai, China[3]
dh.huang.2023@smu.edu.sg; huzhenda2020@gmail.com; zxwang@smu.edu.sg*

***Abstract*—Llama 2, an open-source large language model developed by Meta, offers a versatile and high-performance solution for natural language processing, boasting a broad scale, competitive dialogue capabilities, and open accessibility for research and development, thus driving innovation in AI applications. Despite these advancements, there remains a limited understanding of the underlying principles and performance of Llama 2 compared with other LLMs. To address this gap, this paper presents a comprehensive evaluation of Llama 2, focusing on its application in in-context learning — an AI design pattern that harnesses pre-trained LLMs for processing confidential and sensitive data. Through a rigorous comparative analysis with other open-source LLMs and OpenAI models, this study sheds light on Llama 2's performance, quality, and potential use cases. Our findings indicate that Llama 2 holds significant promise for applications involving in-context learning, with notable strengths in both answer quality and inference speed. This research offers valuable insights for the fields of LLMs and serves as an effective reference for companies and individuals utilizing such large models. The source codes and datasets of this paper are accessible at https://github.com/inflaton/Llama-2-eval.**

***Index Terms*—large language model, in-context learning, generative pre-trained transformer, model evaluation**

I. Background and Introduction

On July 18, 2023, Meta announced the release of Llama 2, the next generation of their open-source large language model (LLM) [1]. Llama 2 is available for both research and commercial use, with Microsoft as its preferred partner. This open-source model serves as a foundational AI technology that empowers businesses and developers to create advanced applications such as chatbots and personal assistants, which require the generation of human-like text. It offers a solution for organizations that may lack the resources to develop such complex models from scratch.

According to the official Hugging Face organization for Llama 2 models from Meta [1], Llama 2 comprises pretrained and fine-tuned generative text models ranging from 7 billion to 70 billion parameters among these models, Llama-2-Chat has been specifically optimized for dialogue-based use cases [1]. Notably, Llama-2-Chat models excel in open-source chat model benchmarks and receive favorable evaluations in terms of helpfulness and safety, rivaling some well-known closed-source models like ChatGPT [2], [3] and PaLM [4].

Upon Meta's announcement, we promptly submitted a request for a commercial license, which was granted shortly. This license enables us to conduct comprehensive comparisons between Llama 2 models and other large language models (LLMs). This paper delves into the capabilities of Llama 2, with a particular emphasis on its role in in-context learning — an emerging AI design pattern that harnesses pretrained LLMs to process confidential and sensitive data. We have made the following contributions:

- Developed unique metrics and methodologies to assess the in-context learning of Llama 2 against various LLMs.
- Performed an in-depth comparison of Llama-2 models with other top-tier LLMs, emphasizing their in-context learning strengths.
- Beyond corroborating Meta's claims about Llama-2's capabilities, we've highlighted its commercial potential and offered insights for seamless enterprise adoption.

II. Related Work

In recent years, the rapid advancement of large language models (LLMs) has transformed the landscape of natural language processing (NLP) and artificial intelligence (AI). With the introduction of Llama 2 [1], there has been a surge of interest in evaluating and harnessing its capabilities. In this section, we review related work that contextualizes our comprehensive evaluation of Llama 2 and its implications for AI applications, with a particular focus on in-context learning.

A. Advances in Large Language Models

The development of LLMs has been a focal point of research and innovation in the field of NLP [5]–[7]. LLMs have a great deal of potential as capable AI assistants that can carry out complex reasoning tasks in a variety of disciplines. Due to the ease with which they may promote human engagement through user-friendly chat interfaces, they have been quickly and widely accepted by the general public. Notable models like GPT-3 and others have demonstrated the potential of pre-trained LLMs in various applications, ranging from text generation to translation and dialogue systems [2], [7].

*Corresponding Author: Zhaoxia WANG (e-mail: zxwang@smu.edu.sg)
[1]https://huggingface.co/meta-llama

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00108

Our work builds upon this body of research by assessing the unique qualities of Llama 2, specifically in the context of in-context learning, and comparing it with existing LLMs.

B. In-Context Learning and Pre-trained LLMs

In-context learning, as an AI design pattern, has gained prominence for its ability to process confidential and sensitive data using pre-trained LLMs [8]. Prior studies have explored the use of LLMs for handling contextually relevant information and generating context-aware responses [9]. We aim to contribute to this area of research by providing a comprehensive evaluation of Llama 2's performance and suitability for in-context learning, shedding light on its strengths and potential use cases.

C. Comparative Analyses of LLMs

Comparative analyses of LLMs have been instrumental in assessing their capabilities and identifying their relative advantages. Existing research has compared different LLMs on various benchmarks, considering factors such as answer quality [10] [11], inference speed [11] [12], and safety [13]. Our study aligns with this research theme, as we conduct a rigorous comparative analysis between Llama 2 and other open-source LLMs and OpenAI models, aiming to provide insights into its competitive edge and unique contributions.

III. Methodology

In this section, we present our approach and methodology for the comprehensive evaluation of Llama 2, with a specific focus on its application in in-context learning, as well as its comparison with existing large language models (LLMs). Our methodology encompasses data collection and preprocessing, model selection, in-context learning workflow and evaluation metrics.

By adopting this systematic approach, we harness in-context learning's potential, ensuring efficient and secure interactions with Large Language Models.

A. Data Collection and Preprocessing

To rigorously evaluate Llama 2's capabilities, we employed a subset of the MS MARCO dataset, a comprehensive reading comprehension and question answering dataset curated by Microsoft [2]. We carefully handpicked 100 queries with well-formed answers across five categories (LOCATION, NUMERIC, PERSON, DESCRIPTION, ENTITY), resulting in a test dataset comprising 500 queries. Table I provides an overview of this evaluation dataset.

B. Model Selection

We purposefully chose our models to ensure a thorough comparative study. Llama 2 was our primary focus given its recent emergence and the notable features discussed in the related work section. Alongside it, we evaluated established LLMs like GPT-4, GPT-3.5, and several leading open-source models. However, due to our GPU cluster's memory constraints (48GB), we restricted our evaluation to open-source LLMs with up to 13B parameters. This selection framework was designed to benchmark Llama 2 against recognized standards, helping us discern its distinctive value in the domain.

C. Experiment Setup

To assess the in-context learning proficiency of chosen LLMs, we crafted a Python application leveraging LangChain [3], a renowned open-source platform tailored for seamless integration with LLMs. For each query in our evaluation set, the program creates a prompt that combines the query with its associated 10 passages for context. This prompt is forwarded to the LLM to produce a response. The generated answer is then compared with the established ground truth to assess its accuracy. Table II provides a visual representation of this process using sample data from a specific query.

Upon processing all 500 queries, the application computes various evaluation metrics, details of which are elucidated in the ensuing section.

D. Evaluation Metrics

When appraising the performance of Llama 2 and its counterparts, we adopted a suite of metrics specifically designed for in-context learning evaluations:

- Inference Speed: This metric gauges the rapidity of the models in processing queries and generating outputs, a pivotal attribute for applications necessitating real-time responses. It's quantified in terms of tokens generated per second.
- Answer Quality: A measure of the precision and pertinence of the answers produced by the models in the realm of in-context learning. Aligning with the official benchmarks for Question Answering and Natural Language Generation tasks [4], we opted for Rouge-L [14], [15] and Bleu-1 [16] to assess answer quality.

IV. Results and Discussions

In this section, we showcase the results of our in-depth comparison between Llama 2 and other prominent LLMs, especially those from OpenAI. Our analysis covers aspects such as answer quality and inference speed, and we explore their relevance in the context of in-context learning applications.

A. Inference Speed

We evaluated inference speeds by running our test program on Nvidia A40 and L40 GPUs for Llama-2 and various open-source models. To optimize costs, we restricted testing of OpenAI models to the A40. Figure 1 presents the results, highlighting that the newer L40 GPU enhances inference speed by 24% to 49%, contingent on the LLM in use.

[2]https://github.com/microsoft/MSMARCO-Question-Answering\#qa

[3]https://github.com/langchain-ai/langchain

[4]https://microsoft.github.io/msmarco/

TABLE I
OVERVIEW OF EVALUATION DATASET

ID	Answers	Passages	Query	Query ID	Query Type	Well Formed Answers
0	[2,662]	{'is_selected': [0, 0, 0, 1, 0, 0, 0, 0], 'pas...	albany mn population	15177	NUMERIC	[The population of Albany, Minnesota is 2,662.]
...	...	...	...	...	...	...
499	[African-Nguni]	{'is_selected': [0, 0, 1, 0, 0, 0, 0, 0], 'pas...	what ethnicity is the surname sabol	658265	PERSON	[The ethnicity of the surname Sabol is African...

TABLE II
STEPS OF PROCESSING A QUERY

1) Retrieve Query	_____ is considered the father of modern medicine.
2) Create Prompt and Send it to LLM	System: Use the following pieces of context to answer the users question. If you don't know the answer, just say that you don't know, don't try to make up an answer. ---------- (... 10 passages inserted here as context ...) Human: _____ is considered the father of modern medicine.
3) Compare LLM Generated Answer against Ground Truth	***Answer***: Hippocrates is considered the father of modern medicine. ***Ground truth***: Hippocrates is considered the father of modern medicine.

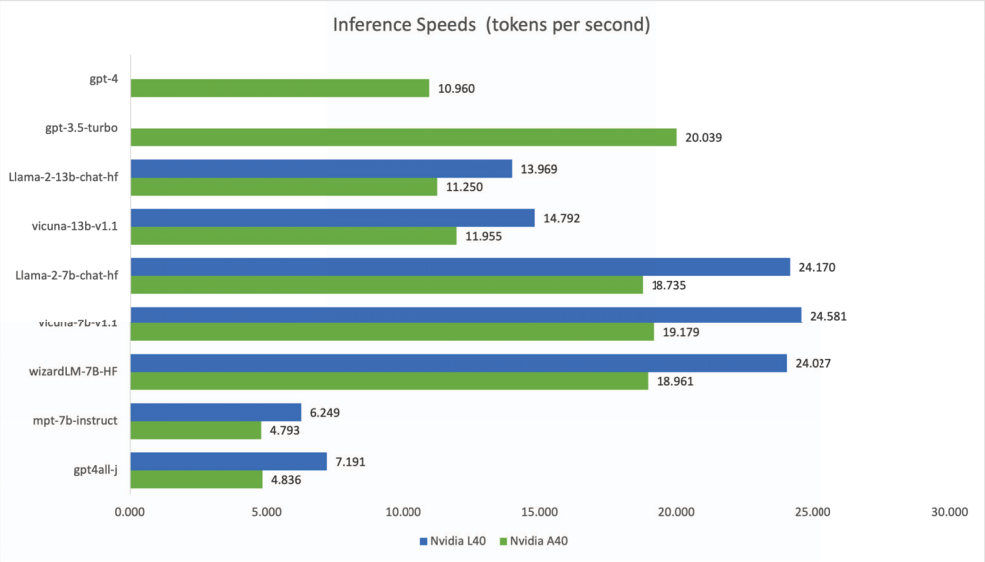

Fig. 1. Inference Speeds (running on Nvidia A40 and L40 GPUs)

B. Answer Quality

Figure 2 displays the answer quality metrics, specifically Rouge-L [14], [15] and Bleu-1 [16] scores, both overall and for each of the five categories: LOCATION, NUMERIC, PERSON, DESCRIPTION, and ENTITY. A notable observation is that while there are performance variations between different LLMs, all models consistently exhibit superior performance for queries in the LOCATION category compared to the others.

V. RESEARCH DISCOVERY

Within the scope of in-context learning, Llama-2 models perform on a level comparable to OpenAI's offerings.

- The Llama-2-13b-chat-hf model slightly edges out GPT-3.5-turbo in answer quality, but doesn't quite match up to GPT-4.
- When benchmarked on Nvidia A40 and L40 GPUs, Llama-2-13b-chat-hf shows faster inference speeds than GPT-4 but is slower than GPT-3.5-turbo. Using cutting-edge GPUs like Nvidia A100 or V100 might further optimize its performance.

(a) OVERALL

(b) LOCATION

(c) NUMERIC

(d) PERSON

(e) DESCRIPTION

(f) ENTITY

Fig. 2. Answer Qualities (running on Nvidia A40 GPUs)

Comparatively, Llama-2 models stand out amongst other open-source LLMs.

- With almost equivalent inference speeds, Llama-2-13b-chat-hf delivers superior answer quality compared to vicuna-13b-v1.1, despite the latter's claims of achieving over 90% of the quality seen in OpenAI ChatGPT and Google Bard [5].
- Llama-2-7b-chat-hf performs comparably to vicuna-7b-v1.1 and wizardLM-7b-HF in terms of speed and answer quality, and outperforms other 6B/7B LLMs.

In essence, our findings resonate with the claims made by Meta during the Llama-2 launch:

- Llama-2-Chat models consistently outperform other open-source alternatives across diverse benchmarks. In terms of user evaluations focused on utility and safety, they match up well against industry-leading closed-source models like ChatGPT and PaLM.

Currently, Llama-2 models represent the gold standard for commercially viable open-source LLMs. Yet, it's crucial to underscore Meta's licensing stipulation which mandates or-

[5]https://lmsys.org/blog/2023-03-30-vicuna/

ganizations with a monthly active user base exceeding 700 million to seek a dedicated license. Hence, it's imperative for large-scale enterprises to closely scrutinize these licensing stipulations before adopting Llama-2 models for business use.

VI. Conclusions

This paper provides a comprehensive assessment of Llama 2's potential in the realm of in-context learning. Our findings highlight the comparable performance of Llama-2 models to leading OpenAI models, particularly in terms of answer quality. Moreover, we underscore Llama 2's potential to meet the requirements of commercial applications while emphasizing the importance of understanding and adhering to licensing terms, especially for enterprises with substantial user bases. Llama 2's emergence as a formidable open-source LLM signifies a promising future for NLP and AI-driven applications.

References

[1] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[2] C. Wang, S. X. Liu, and A. H. Awadallah, "Cost-effective hyperparameter optimization for large language model generation inference," *arXiv preprint arXiv:2303.04673*, 2023.

[3] V. Adlakha, P. BehnamGhader, X. H. Lu, N. Meade, and S. Reddy, "Evaluating correctness and faithfulness of instruction-following models for question answering," *arXiv preprint arXiv:2307.16877*, 2023.

[4] Y. Wang, Z. Yu, Z. Zeng, L. Yang, C. Wang, H. Chen, C. Jiang, R. Xie, J. Wang, X. Xie *et al.*, "Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization," *arXiv preprint arXiv:2306.05087*, 2023.

[5] E. Kamalloo, N. Dziri, C. L. Clarke, and D. Rafiei, "Evaluating open-domain question answering in the era of large language models," *arXiv preprint arXiv:2305.06984*, 2023.

[6] O. Sharir, B. Peleg, and Y. Shoham, "The cost of training nlp models: A concise overview," *arXiv preprint arXiv:2004.08900*, 2020.

[7] Y. Xu, H. Cao, W. Du, and W. Wang, "A survey of cross-lingual sentiment analysis: Methodologies, models and evaluations," *Data Science and Engineering*, vol. 7, no. 3, pp. 279–299, 2022.

[8] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer, "Rethinking the role of demonstrations: What makes in-context learning work?" *arXiv preprint arXiv:2202.12837*, 2022.

[9] S. Cai, S. Venugopalan, K. Tomanek, A. Narayanan, M. R. Morris, and M. P. Brenner, "Context-aware abbreviation expansion using large language models," *arXiv preprint arXiv:2205.03767*, 2022.

[10] Y. Chen, R. Wang, H. Jiang, S. Shi, and R. Xu, "Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study," *arXiv preprint arXiv:2304.00723*, 2023.

[11] D. Huang and Z. Wang, "Evaluation of orca 2 against other llms for retrieval augmented generation," in *The 28th Pacific-Asia Conference on Knowledge Discovery and Data Mining Workshops (PAKDDW)*. Springer, 2024.

[12] Z. Zheng, X. Ren, F. Xue, Y. Luo, X. Jiang, and Y. You, "Response length perception and sequence scheduling: An llm-empowered llm inference pipeline," *arXiv preprint arXiv:2305.13144*, 2023.

[13] Z. Zhang, L. Lei, L. Wu, R. Sun, Y. Huang, C. Long, X. Liu, X. Lei, J. Tang, and M. Huang, "Safetybench: Evaluating the safety of large language models with multiple choice questions," *arXiv preprint arXiv:2309.07045*, 2023.

[14] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[15] A. Joshi, E. Fidalgo, E. Alegre, and L. Fernández-Robles, "Deepsumm: Exploiting topic models and sequence to sequence networks for extractive text summarization," *Expert Systems with Applications*, vol. 211, p. 118442, 2023.

[16] S. Dey, V. Vinayakarao, M. Gupta, and S. Dechu, "Evaluating commit message generation: to bleu or not to bleu?" in *Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: New Ideas and Emerging Results*, 2022, pp. 31–35.