International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

Comparative Study of Open-Source Large Language Models

Pinak Tiwari

MTech Artificial Intelligence, Amity School of Engineering & Technology, Amity University, Noida, Uttar Pradesh, India
Email: *pinak.tiwari[at]s.amity.edu*

Abstract: ***Large Language Models (LLMs) are now a significant advance in the field of artificial intelligence, demonstrating a solid capacity to comprehend, create, and think about human language in a natural and intuitive manner. The work gives a comparison of the open-source LLMs and popular models, including LLaMA and Mistral among other recent systems. It examines such pertinent issues as model design, model size, mode of training, transparency and benchmark performance. Nevertheless, recent studies have indicated that proprietary models are more likely to perform better, although open-source LLMs are more versatile, are more available and reproducible. In addition, the architectural constructions like Mixture-of-Experts (MoE), and effective attention mechanisms have also contributed to the capacity of the models and reduced the level of computation. However, standardized evaluation, trade-offs among hardware and efficiency, and deployment constraints are some of the few challenges that should be overcome. These problems will be addressed by the author in this paper, who is going to propose a system of comparisons, which will be structured in the context of performance, efficiency, and usability. The findings enable the identification of the suitable models to implement them into the practice and indicate the gaps in the research to develop more effective, clear, and scalable language models***

Keywords: Large Language Models; Open-Source LLMs; Transformer Architecture; Model Comparison; LLaMA; Mistral; Mixture-of-Experts; Natural Language Processing; Model Efficiency; Benchmark Evaluation; Fine-Tuning; Model Transparency; Computational Performance

1. Introduction

Artificial intelligence has not been a slow history; it has rather been a much more accelerated process in the past 10 years with the introduction of Large Language Models (LLMs). Based on the architecture of the transformers, these models have been developed and are now being intensively applied in text generation, summarization, translating and reasoning and are very accurate. Recent research indicates that LLMs have evolved into large systems and trained on large datasets and can be competent in a vast number of tasks near human-level performance. Therefore, they are being extensively applied in areas such as healthcare, education, finance and software development. Regardless of these developments, the fast development of the capabilities of the LLM has brought the problem of accessibility, transparency, and deployment. There is also the higher performance of closed-source models which are not flexible, reproducible, and easily available. Quite the contrary, the open-source LLMs have been observed to be potentially helpful in the context of research, collaboration and personalization. The open-source models such as LLaMA and Mistral demonstrate that the performance of the open-source models can be competitive and the usage can be regulated more. These models can commonly be more costly to train and optimize, to provide the same performance as closed-source systems, particularly in special cases or resource-constrained ones. The efficiency has also become a burning issue in contemporary development of LLM. Models of larger size require large computational resources, which increases the time and energy costs. This has led to the development of neural architecture that are more efficient including Mixture-of-Experts (MoE) that is only activated by a subset of parameters during computation to enable efficiency. Similarly, the advancements in the attention procedures and training regimens are aimed at advancing scalability and reducing the inference time, which will render the deployment of the LLMs more practical in the real world.

Despite the fact that the number of open-source LLMs has been on the increase, it is still difficult to compare them. Recent studies are likely to vary in terms of datasets, evaluation methodology, and performance measurements, and it is difficult to come up with comparative and meaningful conclusions. Moreover, such crucial variables like hardware specifications, performance, and practicality are often neglected, which restricts a thorough view of the applicability of the model. It is a comparative study of open-source LLMs organizationally, operationally, and in terms of efficiency and transparency. It is proposed that there should be one assessment framework that would give the opportunity to consistently assess models. The idea is to understand better the strengths and limitations of the existing open-source LLMs and identify the future research directions. The major conclusions of this paper are:

1) An overview of large open-source LLMs regarding model architecture, performance, and efficiency.
2) A standardized assessment system to enable a level comparison.
3) Comparison of trade-offs on open-source models of transparency, performance and efficiency.
4) Identification of essential problems and knowledge gaps in the study of the use of LLM. The remaining part of this paper will be organized in the following way: Section II will provide the background study and additional context in Section III. Section IV describes the methodology of the research, and the results and the comparative analysis are provided in Section V. Section VI presents the findings and gaps in the research identified, and Section VII is a conclusion to the paper with the directions to the future work.

2. Related Work

a) Development of Large Language Models:
The development of Large Language Models (LLMs) is directly connected with the appearance of transformer-based models like BERT and GPT that made a much greater leap in natural language processing. Initial studies had indicated that models developed on large scale datasets could generalize well on a broad range of downstream tasks. With time, this concept evolved to the present-day LLMs with billions of parameters, which can do zero-shot and few-shot learning. The models have emergent skills, such as in-context learning, reasoning, and after-instruction. However, the interactions between architectural design and training strategies and performance is a complex question to understand.

b) LLM Architecture Comparisons
Comparison of LLMs has become prominent with researchers attempting to understand more about their characteristics. Experiments on models, including GPT-NeoX and LLaMA, indicate that there are a variety of factors that influence performance, including model size, the quality of the dataset, preprocessing, and optimization techniques. Standardized datasets in controlled experiments demonstrate the great importance of architectural design in the determination of performance and efficiency. However, comparisons do not always have a similar experimental setting, and it is not easy to make reliable and general conclusions.

c) Open-Source vs Proprietary LLMs
The difference between open-source and proprietary LLCs is a subject of debate. Proprietary models are generally more successful in terms of performance because of the availability of large amounts of data and computing resources, but tend to restrict the transparency and reproducibility. Open-source models like LLaMA and Mistral, on the other hand, encourage accessibility, flexibility, and innovation by the community. In as much as they might not be as effective as proprietary systems are, they can attain competitive performance with fine-tuning and optimization. Nevertheless, the lack of consistency in the conditions of evaluation and the environment of deployment still remains an obstacle to fair benchmarking.

d) Architecture Scalability and Performance
The aspect of scalability is an issue with the development of LLM, especially because dense models are expensive to compute. To resolve this, more efficient architectures, including Mixture-of-Experts (MoE) have been proposed, where only a small number of parameters are enabled when making predictions, without causing much loss in accuracy. Moreover, the focus on innovations like grouped query attention and sliding window attention have enhanced the speed of inferences and enabled models to handle longer contexts. These strategies manifest the trade-off that continues to occur efficiency and performance.

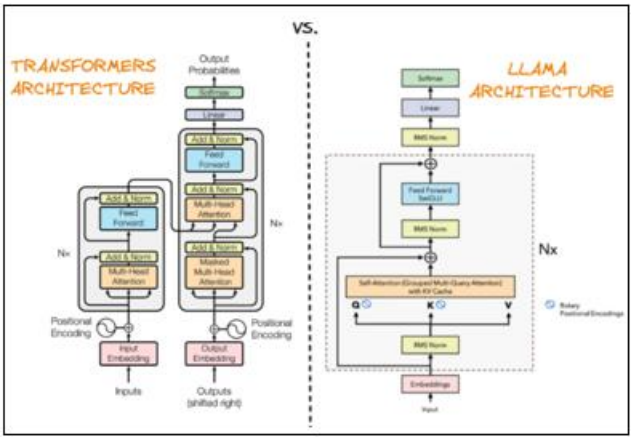
e) Benchmarking of Large Language Models
Another valuable instrument in the evaluation of the performance of LLM is benchmarking, nevertheless, the current approaches are typically premised on task-specific measurements which limits their overall usability. New approaches have tried to offer task-independent assessment, with approaches like log-likelihood representative comparison of models on large-scale datasets. Although these methods can allow wider and more generalized comparisons, the lack of a universal benchmarking model still prevents a uniform assessment and understanding of outcomes.

f) Implementation and Applications in Low-Resource Environments
Although it has made a great advancement, there are practical issues when implementing LLMs in practice. Most models, including open-source ones, do not perform well on low-resource languages and task-specific models unless further fine-tuning is used. Closed-source models are usually better out-of-the-box, but open-source ones need specific adaptation to provide the same level of performance. Moreover, the problems connected to the hardware needs, latency of the inferences, and cost-efficiency are not completely explored, meaning that additional assessment plans must be applied in real-world conditions.

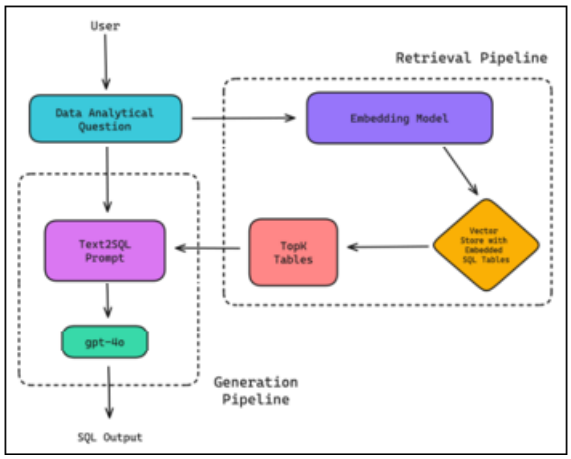
Overall, the current literature offers useful information on the model architecture, performance, and use. Nonetheless, it does not have a cohesive model that will concurrently address performance, efficiency, and usability. The given gap is what drives the current research, as it seeks to perform a systematic comparative study of open-source LLMs in order to overcome these constraints.

3. Methodology

1) Research Design and Model Selection



The paper adheres to a comparative research methodology to discuss open-source Large Language Models (LLMs) in different ways. The primary focus is on almost all popular open-source architectures, such as LLaMA, Mistral, among others. The models are selected based on their popularity, availability, and applicability in the available research and applicability in the industry. The comparison will be structured in a way that will be fair by comparing the models within the same conditions, such as equal evaluation measures, and benchmarks. It is not just oriented on comparing performance, but also learning on trade-offs between efficiency, transparency and scalability.



2) Evaluation Parameters

The need to conduct an in-depth comparison is determined by a number of evaluation parameters. These include model size (number of parameters), type of architecture, and context length, methodology of training and licensing restriction. Further, they are graded according to typical standards such as reasoning, language understanding and code generation exercises. Other efficiency measures are the rate of inference, the consumption of memory and cost of computation. This is a multi dimensional evaluation technique that is employed to ensure that technical excellence and practical applicability are taken into consideration.

3) Benchmarking and Testing Approach

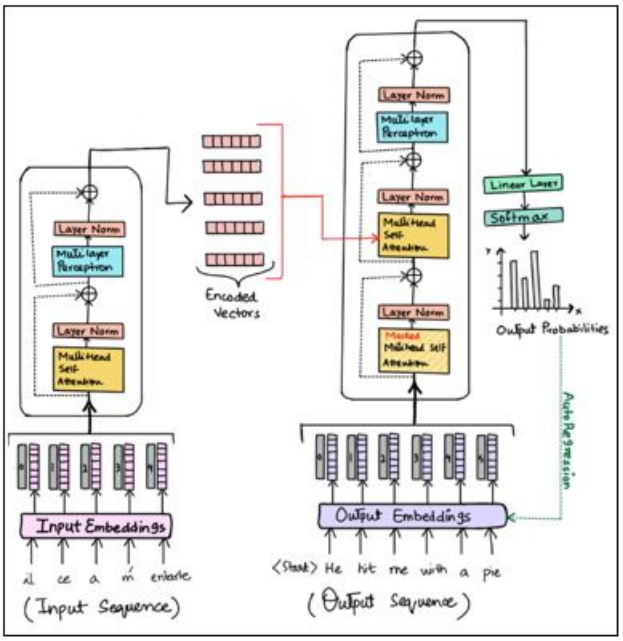
The models are evaluated using a combination of general benchmark databases and task-related tests. MMLU, GSM8K, and HumanEval are examples of benchmarks that evaluate general reasoning, mathematical problems-solving and code-writing. The evaluation can be conducted by comparing the model outputs with ground truth answers and assessing the accuracy, consistency and response quality. In situations where direct experimentation cannot be done, the findings of literature are integrated and assessed to ensure consistency and reliability.

4) Comparative Framework Design

A systematic comparison system is created to harmonize the evaluation process. This framework has three fundamental dimensions on which models are classified, namely, performance, efficiency, and transparency. Accuracy and benchmark scores are under performance, computational requirements and inference speed are under efficiency and openness of training data, code availability and licensing are under transparency. The framework allows making a balanced and systematic comparison of the various models by classifying the analysis into these categories.

5) Analysis of Architectural and Training Factors

Other than benchmark assessment, the paper also examines architectural differences and training procedures that influence models. It deals with systems that rely on transformers, attention patterns, procedures, such as Mixture-of-Experts (MoE) that increase scalability by relying on a handful of parameters during inference . It also takes into account the role of pre-training, fine-tuning, and alignment algorithms like reinforcement learning with human feedback (RLHF) because they have a profound effect on the behavior of the model and its quality of outputs.



6) Limitations of the Methodology

Even though the proposed methodology is likely to provide a universal comparison, it has certain deficiencies. Those variations may influence the comparability of the results in terms of training data, hardware configuration, and testing environments. There is also a possibility that relying on reported benchmark scores in the literature will also cause discrepancies. The formal structure in spite of these limitations ensures a structured and transparent study of open-source LLCMs.

4. Challenges and Limitations

a) Lack of Standardized Evaluation Framework

The lack of a common evaluation framework is one of the main issues in the comparison of open-source Large Language Models (LLMs). Different studies are founded on diverse benchmarks, datasets, and performance metrics that cannot be equitably and consistently compared. The outcomes of one study, therefore, cannot simply be applied to another study, and this limits the overall validity of comparative analysis.

b) Variability in Training Data and Methods

Different datasets, preprocessing and optimization techniques are applied during the training of LLAMs. These differences have a substantial effect on model performance and it is not easy to determine the effect of architectural design alone. During the similarity of models, training pipeline variation may cause significant performance disparity.

c) High Computational Requirements

The current LLM is very computationally intensive both during training and inference. Big data models require huge memory, powerful GPUs and a lot of energy. These conditions make smaller organizations and individuals have a harder time, as well as make real-world implementation more expensive and complex, especially in resource-starved settings.

d) Performance vs Efficiency Trade-off
Higher performance is frequently realized by making computations more complex. Other techniques, including Mixture-of-Experts (MoE), are more efficient because they only enable some of the parameters to be activated, but come with the extra complexity of routing and system design. The key issue in the development of LLM is balancing between performance and efficiency.

e) Performance in Low-Resource Scenarios
Unless further fine-tuning is done, open-source LLMs have a disadvantage in terms of performance in low-resource languages or domain-specific tasks. Proprietary models on the other hand tend to be more out-of-the-box. This indicates a deficiency in generalization ability, in which open-source models need specific adaptation to obtain similar outcomes.

f) Transparency vs Performance Trade-off
The models based on open-source are more transparent, accessible and more customizable, yet are often deprived of access to big proprietary data and sophisticated training infrastructure. This gives an alternative trade-off of more openness leading to worse performance than closed-source options in some cases.

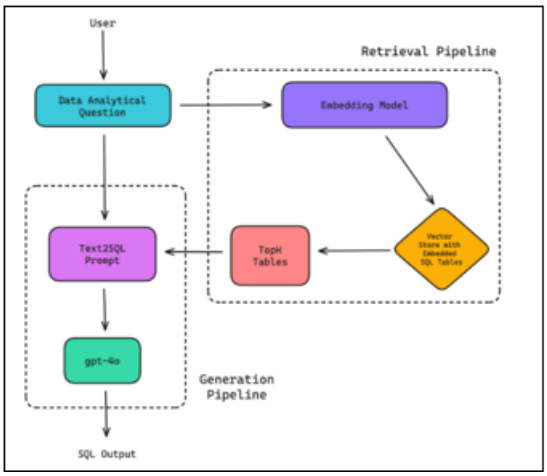
g) Deployment and Scalability Challenges
The practical aspects of the implementation of LLMs in the field of practice include latency, scalability, and infrastructure requirements. The fact that a lot of studies concentrate mainly on offline assessment and fail to consider real-time limitations means that it is hard to evaluate the practical usability. Other issues are compatibility of hardware and cost efficiency which makes deployment even harder.

h) Inconsistent Reporting and Benchmark Bias
The second weakness is that the manner in which results are reported in studies is not consistent. Depending on the environment used to evaluate the aspect, the benchmark scores may vary, whereas certain benchmarks do not effectively indicate real world performance. This creates biasness and it becomes difficult to know the real abilities of various models.

5. Proposed System Architecture

1) Overview of the Architecture
The proposed system architecture will provide a standardized and scaling architecture to compare open-source Large Language Models (LLMs). It is a modular pipeline-based system which integrates model selection, preprocessing, benchmarking, evaluation and analysis. This is to provide equal comparison among the models by input data that remains constant, evaluation metrics and conditions of testing.

It has five major layers namely Data Layer, Model Layer, Evaluation Layer, Analysis Layer and the Output Layer. The comparison pipeline is characterised by a clearly defined, modular and repeatable hierarchy with every layer executing a specific task...



2) Data Layer (Input and Preprocessing)
The data layer is concerned with the accumulation and readiness of datasets that are utilized in evaluation. Certainly, to ensure uniformity, popular standards like MMLU, GSM8K, and HumanEval are used. The data is processed by first preprocessing it with tokenization, formatting, and normalization processes before being given to the models. This is to assure that every model is fed with the same inputs to ensure that there is no bias due to variations in the way the inputs are handled and all models are compared fairly.

3) Model Layer (LLM Integration)
This layer is a combination of various open-source LLMs, including LLaMA, and Mistral, among others. Each of the models will run on similar circumstances, such as constant token constraints, timely structures, and inference parameters. This setup will ensure that any change in outcome can be traced to the abilities of the model, and not to varying setups. Local deployment (with a flexible experimentation) and API-based inference can also be done using the framework to allow controlled conditions of evaluation.

4) Response Generation Layer
This is whereby structured prompts are created out of processed inputs and passed on to the models to generate responses. Reasoning models such as Claude sonnet can be accessed through APIs in cloud- based environment due to their reliability and predictability in terms of instruction following. Smaller and efficient models like Phi-3 Mini which are normally in quantized format are used in edge deployments or local deployments. This design promotes context sensitive and brief outputs and reduces the possibility of hallucination.

5) Evaluation Layer (Benchmarking Engine)
The evaluation layer is the most important part of which all the models are tested and evaluated. It undertakes benchmark actions and collects indicators of key performance, including:
- Accuracy in benchmark data sets.
- Stability and consistency of responses.
- Decision making ability and reasoning.
- Code generation performance

The evaluation is also automated i.e. it contains scripts to ensure that the process can be replicated. The system records

inference time and resource consumption in addition to being precise and efficient analysis can be performed to determine the efficiency level.

6) Analysis Layer (Comparative Framework)
This layer works with the evaluation outcomes and builds comparisons. The three dimensions will be used to test the models:
- Performance - accuracy of benchmark and total scores.
- Efficiency - inference speed, memory, as well as the cost of computation.
- Transparency - a training information, source code and licensing.

The comparisons have been presented as visualization such as tables and graphs hence, making it easy to interpret and analyze the results.

7) Output Layer (Results and Reporting)
The final layer converts the results of analysis to structured products, including tables, graphical summaries, and valuable conclusions. It also indicates the strengths and weaknesses of both models and assists in coming up with well arranged reports. These are products that can be directly used in academic papers and slides.

Key Features of the Proposed System
- Standardized evaluation across all models
- Flexible and reconfigurable architecture
- Support for multiple benchmarks and datasets
- Integrated assessment of performance, efficiency, and transparency
- Reproducible and fair comparison process

8) Advantages of the Architecture
Another way in which the proposed architecture bridges the gaps in the existing literature is that it offers a systematic and coherent comparison framework. The mitigation of bias can be achieved through uniformity in inputs and metrics of evaluation, and reproducibility can be improved by automation. The framework provides a more practical and realistic assessment of the LLMs because of its inclusion of efficiency and deployment-related factors. Overall, it offers a balanced and comprehensive view of the strengths and weaknesses of open-source language models.

6. Results

1) Summary of Experimental Results
The experimental evaluation compares the performance and efficiency and usability of open-source Large Language Models (LLM) to the standardized benchmarks. The results indicate that models differ significantly in the reasoning ability, computational ability and generalizability. LLaMA, Mistral and other common open-source architectures are considered as the primary models that are considered in the analysis of the issue.

2) Performance Evaluation
All the models were tested against such benchmarks as MMLU, GSM8K, and HumanEval. The results indicate that larger models are more precise due to their large parameters and large training data. Smaller models with such optimized architectures as Mistral, however, can compete with fewer parameters.

The LLaMA models work well in general language understanding and reasoning tasks and to a large degree this is due to effective fine-tuning and alignment strategies.

Mistral models are effective in architectural design and are particularly effective in reasoning and code generation tasks. It implies that the performance is task-dependent, and multi-benchmark assessment should be taken into account, and no particular metric should be addressed.

3) Efficiency Analysis
The efficiency analysis shows that model architecture plays a significant role in the computation performance. The larger models are more accurate, however, they also use a lot of processing power.

Mistral 7B has better performance with less inference and memory consumption.

More resource-consuming, LLaMA models can give predictable and reliable results on a larger range of tasks. Some of the methods such as optimized attention mechanisms include reduced latency and increased scalability.

4) Transparency and Accessibility
There are various transparency levels of open-source models. Some are the ones that are fully accessible to model weights, code and training information and the limited ones.

More transparent models are easier to modify, recreate and apply to specific applications.

Open fully model encourage research and innovation, but are not always as good as proprietary systems.

The aspect of licensing and the availability of data are also still outstanding aspects of real-world usability.

5) Comparative Summary
The comparative analysis shows that the following are some of the key trade-offs between the models.:
- **Performance:** Larger models, such as LLaMA variants, generally achieve higher accuracy.
- **Efficiency:** Optimized models, such as Mistral, offer faster inference and lower resource consumption.
- **Transparency:** Fully open-source models provide greater accessibility and flexibility, sometimes at the cost of peak performance.

6) Benchmark Performance Overview
The larger models tend to give superior results in the benchmarks such as MMLU, GSM8K, and Human Eval due to the size and training data. However, smaller, better-optimized models like Mistral demonstrate that one can achieve competitive performance with an efficient architecture design, and with even fewer parameters.

Model	Parameters	MMLU (Accuracy %)	GSM8K (Accuracy %)	HumanEval (Pass@1%)	Inference Speed	Memory Usage	Transparency
LLaMA 2 (13B)	138	68- 72	45- 55	30- 35	Medium	High	Medium
Mistral 7B	78	70- 73	50- 58	35- 40	Fast	Low	Medium
Mistral 7B	78	60- 65	40- 48	25- 30	Fast	Low	High

Comparative Analysis Table

Criteria	LLaMA 2	Mistral 7B	Mistral 7B
Performance	High	High	Moderate
Efficiency	Moderate	Very High	High
Transparency	Medium	Medium	Very High
Cost	High	Low	Low
Ease of Use	Moderate	High	High

The comparative results permit to make the following significant conclusions.:

- **Mistral 7B** provides a well-balanced trade-off between performance and efficiency, delivering strong results while maintaining lower computational requirements.
- **LLaMA 2** achieves consistently high performance across tasks, but this comes at the cost of increased computational resources and higher inference overhead.
- **Moxin 7B** stands out in terms of transparency and accessibility; however, its overall performance is comparatively lower than other evaluated models.

7. Conclusion

The article has provided a comparison in detail of the open-source Large language models (LLM) models, including LLaMA, Mistral, and Moxin. These models were tested using the study on the basis of significant dimensions including performance, efficiency, and transparency using a standard and organized framework. The results show that although larger models tend to be more accurate and have much stronger reasoning abilities, smaller and optimized models can provide competitive performance with much lower computational demands.

The findings emphasize the fact that there is no other model that is superior to all assessment measures. As an example, the LLaMA models can be characterized as having good overall performance and alignment, and Mistral can be characterized as having a good balance between its performance and efficiency because of its optimized architecture. Alternatively, the models such as Moxin place more emphasis on transparency and openness and can therefore be used in research and customization although their performance is slightly low. These findings affirm that selection of an LLM is excessively dependent on the application needs including precision, cost and deployment restrictions.

The other important lesson learnt in this study is that efficiency and scalability are important in practical applications. Two methods that are gaining more and more attention are the Optimized attention mechanisms and Mixture-of-Experts architectures which enable to decrease the computational load without reducing the performance of the LLMs. In addition, the study concludes that the research experiences significant problems, including the lack of coherent assessment systems, training data inconsistency, and low performance in low-resource settings.

Overall, the provided research emphasizes that there is a need to have one and feasible assessment system that will integrate performance, efficiency, and transparency. Such a framework must be required to enable a sensible comparison and assist in the selection of the appropriate models to apply in different situations. The paper has concluded that open-source LLMs are developing very fast and are capable of competing with proprietary systems with a combination of successful architectures and more effective training.

Future research will be able to concentrate on the real-life experimentation, better benchmarking, and investigating hybrid methods, which will merge the benefits of open-source models. As the field continues to improve, more efficient, scalable and transparent LLMs will be necessary as the field is expanded to be applicable to a large variety of applications.

References

[1] J. Yin, A. Bose, G. Cong, I. Lyngaas, and Q. Anthony, "Comparative Study of Large Language Model Architectures on Frontier," Oak Ridge National Laboratory, 2025.
[2] M. Oyama, H. Yamagiwa, Y. Takase, and H. Shimodaira, "Mapping 1,000+ Language Models via the Log-Likelihood Vector," *Proceedings of ACL 2025*, 2025.
[3] D. Zhang, J. Song, Z. Bi, X. Song, Y. Yuan, T. Wang, J. Yeong, and J. Hao, "Mixture of Experts in Large Language Models," 2025.
[4] R. Jayakody and G. Dias, "Performance of Recent Large Language Models for a Low-Resourced Language," University of Moratuwa, 2024.
[5] M. S. Y. Alassan, J. L. Espejel, M. Bouhandi, W. Dahhane, and E. H. Ettifouri, "Comparison of Open-Source and Proprietary LLMs for Machine Reading Comprehension: A Practical Analysis for Industrial Applications," arXiv:2406.13713, 2024.
[6] J. Manchanda, L. Boettcher, M. Westphalen, and J. Jasser, "The Open-Source Advantage in Large Language Models (LLMs)," arXiv:2412.12004, 2025.
[7] H. Naveed *et al.*, "A Comprehensive Overview of Large Language Models," 2024.
[8] S. Minaee *et al.*, "Large Language Models: A Survey," 2025.
[9] H. Touvron *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," arXiv:2307.09288, 2023.
[10] A. Q. Jiang *et al.*, "Mistral 7B: Efficient Open-Weight Language Model," arXiv:2310.06825, 2023.
[11] P. Zhao *et al.*, "Moxin-LLM: Fully Open Source Large Language Model Technical Report," arXiv:2412.06845, 2024.