Journal of Information Technology Review

DLINE JOURNALS

Print ISSN: 0976-3511
Online ISSN: 0976-2922

JITR 2026: 17 (3)
https://doi.org/10.6025/jitr/2026/17/3/105-119

Mapping the Contemporary LLM Landscape: A Descriptive Analysis of Benchmark Performance and Capability Stratification

Pit Pichappan
Digital Information Research Labs
Chennai 600017. Tamil Nadu
India
pichappan@dirf.org

ABSTRACT

The rapid proliferation of Large Language Models (LLMs) has established benchmark evaluations as the primary mechanism for assessing model capability and technological progress. However, growing concerns regarding benchmark validity, data contamination, and the interpretability of aggregate scores highlight a critical gap in understanding how these metrics reflect the broader LLM ecosystem. This study addresses this gap by conducting a comprehensive descriptive analysis of benchmark performance and capability stratification across contemporary LLMs. Utilizing the Comprehensive LLM Benchmark Dataset, comprising 390 model-benchmark observations from 2022 to 2024, we employ descriptive statistics, density estimation, and performance-tier categorization to map the performance landscape. Our findings reveal a negatively skewed distribution with a high median but substantial variability, indicating that while baseline competencies are standardizing, significant capability gaps persist. Furthermore, the analysis identifies distinct capability strata, with Strong and Top Tier models accounting for over 56 per cent of observations, yet a substantial proportion of Weak and Moderate performers remain. These results demonstrate that the contemporary LLM landscape is highly stratified rather than homogeneous. This stratification highlights the need for delicate evaluation. Ultimately, this research underscores that aggregate benchmark scores often obscure underlying heterogeneity in capabilities. We conclude that future evaluation frameworks must evolve toward multidimensional, capability-oriented methodologies to accurately capture model maturity and real-world utility, providing a foundational baseline for subsequent research on scaling laws and architectural effectiveness.

Keywords: Large Language Models, Benchmark Performance, Capability Stratification, Descriptive Analysis, Model Heterogeneity, AI Benchmarking, Evaluation Frameworks, Performance Tiers

Received: 30 December 2026, Revised 1 March 20026, Accepted 20 March 2026

1. Introduction

Since the emergence of early Large Language Models (LLMs) such as GPT-2 and GPT-3 around 2020, rapid advancements in model architecture, scale, and capability have significantly transformed the field of code intelligence. These developments have enabled the automation of a broad range of software engineering tasks, including code generation, program repair, debugging, and software testing. As a result, benchmark performance has become a central mechanism for evaluating and promoting new AI technologies.

2. Early Studies and Background

Historically, claims regarding artificial intelligence achievements have often relied on vague assertions and selective quotations rather than rigorous empirical evidence [1]. More recently, successive generations of AI models have been presented as superior, largely based on their performance on benchmark tasks such as coding proficiency [DeepSeek] and multilingual capabilities [2], and ancient script translation [3] . While these benchmarks provide quantitative measures of progress, growing concerns have emerged regarding their practical relevance and methodological validity.

Several studies have questioned the extent to which benchmark results accurately reflect real-world capabilities. Researchers have highlighted issues related to practical applicability [4] , statistical robustness [5], susceptibility to adversarial manipulation [6], and the tendency of language models to prioritize linguistic fluency over factual correctness [7]. Transparency challenges are further intensified by closed source evaluation procedures [8, 9], while automated scoring systems may penalize useful responses simply because they deviate from predefined reference answers [10, 11].

The rapid evolution of artificial intelligence systems has simultaneously increased the complexity of benchmarking itself. AI models continue to evolve in architecture, scale, deployment settings, and capability, while benchmark datasets and evaluation environments are continuously updated. This dynamic landscape creates a moving target for assessment and necessitates more adaptive benchmarking methodologies. Large language models are particularly susceptible to benchmark memorization, which can produce substantial discrepancies between benchmark performance and real-world effectiveness [12]. To address these challenges, von Laszewski and colleagues developed a benchmark ontology that assists users in identifying appropriate evaluation frameworks for specific contexts [13, 14]. Among the available initiatives, MLCommons [15] (mlcommons.org) has emerged as one of the most comprehensive and standardized AI benchmarking ecosystems, encompassing training, inference, scientific computing, and domain-specific evaluations.

As LLMs become increasingly integrated into coding and software engineering tasks, the need for meticulous and meaningful evaluation becomes even more critical. Although benchmark scores are frequently used as indicators of model quality, such aggregate measures often obscure the diverse range of cognitive and technical skills required to complete benchmark tasks successfully. Consequently, there remains a lack of systematic methodologies for determining whether benchmarks genuinely measure the capabilities they claim to assess [16, 17, 18].

The pursuit of state of the art benchmark performance has exposed a fundamental vulnerability in contemporary evaluation practices [19]. While LLMs continue to achieve unprecedented scores on standardized assessments, concerns persist that such results may reflect dataset exploitation and pattern recognition rather than genuine language understanding [20].

Nevertheless, the introduction of new models is routinely accompanied by performance claims supported by benchmark results, reinforcing a technological landscape in which benchmark scores serve both as indicators of progress and as marketing tools for emerging AI systems [21].

This dynamic has encouraged selective reporting practices among organizations developing frontier models. Companies often emphasize benchmarks that showcase their models' strengths while minimizing attention to weaker performance areas. For instance, Anthropic's Claude 2 highlighted achievements on graduate-level examinations and coding tasks [22], whereas Microsoft and Nvidia's Megatron-Turing NLG 530B emphasized strong zero-shot learning performance [23]. Such practices underscore the need for more comprehensive and transparent evaluation frameworks.

A significant shift is also occurring in the philosophy of LLM evaluation. Traditional assessments focused on isolated tasks, whereas contemporary approaches increasingly emphasize broader capability based evaluations. As language models unify numerous natural language processing tasks within a common generative framework, conventional task boundaries have become less distinct. Consequently, each prompt can now be viewed as a unique task instance, shifting evaluation priorities toward measuring the underlying capabilities required to address practical real-world problems [24].

Within this context, benchmark ecosystems have become the principal measurement instruments of modern machine learning [25]. Open evaluation platforms such as HELM [26] and the Open LLM Leaderboard [27]consolidate model performance across diverse benchmarks, including MMLU [28] and BBH [29] in to a limited set of benchmark scores and composite rankings. Although such aggregation facilitates comparison, it also conceals important assumptions regarding the interpretation and meaning of benchmark performance [30, 31, 32, 33].

Multiple methodologies have been proposed for evaluating LLM performance, including statistical metrics such as BLEU (Bilingual Evaluation Understudy), ROUGE (Recall-Oriented Understudy for Gisting Evaluation), perplexity, Mean Reciprocal Rank, and BERTScore (Bidirectional Encoder Representations from Transformers Score). Additional approaches include human evaluation, model based evaluation through LLM-as-a-judge frameworks, and comprehensive evaluation platforms. Despite this diversity, benchmark-based assessment remains the dominant methodology within the field [34].

Recent research has sought to improve the interpretability of benchmark outcomes. Dongjun Kim introduced a diagnostic framework that decomposes benchmark performance into ten cognitively grounded abilities. The approach combines gradient based importance scoring with targeted parameter ablation to calculate an Ability Impact Score (AIS), which quantifies the contribution of individual abilities to overall benchmark success [Dongjun Kim]. Such methods represent an important step toward understanding not only how well models perform, but also why they achieve particular outcomes.

The implications of benchmark evaluation extend beyond computer science into high-stakes domains such as clinical medicine. As AI systems increasingly support decision-making processes, practitioners must be able to critically assess benchmark results, understand the limitations of evaluation methodologies, and make informed decisions about integrating AI tools into patient care [35, 36].

Research examining benchmark datasets further illustrates the rapid expansion of evaluation resources. In a survey of coding intelligence benchmarks, Mohammad Abdollahi [37] found that Python was the dominant programming language, appearing in 77% of datasets, while GitHub served as the primary data source in 46% of cases. The study also documented a substantial increase in the number of benchmark datasets published over the previous three years. Complementing this work, Danyang Cao [38] proposed a four-stage evaluation framework consisting of generation, execution, evaluation, and compliance, which organizes existing research along capability, scenario, and system dimensions.

Similarly, Weerasinghe conducted a systematic review of 313 studies focused on LLM-based test generation. The review demonstrated that while architectural innovations such as retrieval augmentation and agent-based systems can improve benchmark coverage, they also introduce significant costs associated with reasoning, repair, and semantic alignment. The findings suggest that transitioning from experimental benchmarks to enterprisescale software testing requires deterministic, uncontaminated evaluation pipelines capable of handling complex and stateful software architectures [39].

Despite substantial progress in benchmark development, significant challenges remain. Existing evaluation frameworks continue to suffer from limited task coverage, inconsistent metrics, inadequate reproducibility, and unresolved concerns regarding safety and alignment [40]. These limitations highlight the need for more transparent, capability oriented, and practically relevant benchmarking methodologies that support meaningful comparisons across the rapidly evolving landscape of contemporary large language models.

2.1 Conceptual Perspective on Benchmark Performance

Benchmark scores are commonly interpreted as proxies for model capability; however, benchmark performance reflects the interaction of multiple underlying factors including model architecture, parameter scale, training data quality, instruction tuning, reinforcement learning strategies, and benchmark design characteristics. Consequently, benchmark outcomes should be viewed not merely as performance indicators but as observable manifestations of latent model capabilities. From this perspective, variations in benchmark scores can reveal the extent to which capability concentration, performance stratification, and competitive differentiation exist within the contemporary LLM ecosystem. This conceptual lens provides the foundation for examining benchmark distributions as indicators of broader technological maturity and capability hierarchy.

3. Research Statement

The rapid advancement of Large Language Models (LLMs) has led to an increasing reliance on benchmark evaluations as indicators of model capability, technological progress, and competitive positioning. However, growing concerns about benchmark validity, capability representation, benchmark contamination, and the interpretability of aggregate performance measures have raised important questions about how accurately benchmark scores reflect real-world model capabilities. Despite the widespread use of benchmark leaderboards and composite rankings, relatively little research has systematically examined the overall structure, distribution, and stratification of benchmark performance across contemporary LLMs.

This study seeks to address this gap by conducting a comprehensive descriptive analysis of benchmark performance within the contemporary LLM ecosystem. Using the Comprehensive LLM Benchmark Dataset, the research investigates how benchmark scores are distributed across models, identifies patterns of capability stratification, and examines the extent of performance heterogeneity among current-generation language models. Rather than focusing on a single benchmark or model family, the study adopts an ecosystem level perspective to understand the broader landscape of benchmark performance and capability differentiation.

The findings are intended to provide an empirical foundation for future investigations into scaling laws, architectural effectiveness, benchmark specialization, model efficiency, and capability oriented evaluation frameworks. By characterizing the distributional properties of benchmark outcomes, the study contributes to ongoing discussions regarding the interpretation, reliability, and practical significance of benchmark-based assessments in contemporary artificial intelligence research.

3.1 Research Questions

The study is guided by the following research questions:

RQ1: What are the overall distributional characteristics of benchmark performance among contemporary Large Language Models?

RQ2: To what extent does benchmark performance vary across models, and what level of heterogeneity exists within the current LLM ecosystem?

RQ3: Do benchmark outcomes indicate the presence of distinct capability strata or performance tiers among contemporary language models?

RQ4: How are benchmark scores distributed across low, medium, high, and frontierperformance categories?

RQ5: What insights do descriptive performance distributions provide regarding the maturity, competitiveness, and capability concentration of the contemporary LLM landscape?

3.2 Research Design

This study adopts a quantitative descriptive research design. The analysis utilizes the Comprehensive LLM Benchmark Dataset containing 390 model–benchmark observations representing a diverse set of contemporary Large Language Models developed between 2022 and 2024. Descriptive statistical techniques are employed to examine benchmark outcomes, including summary statistics, frequency distributions, density estimation, violin plot analysis, boxplot analysis, and performancetier categorization.

The analytical framework focuses on characterizing central tendency, variability, distributional shape, performance concentration, and capability stratification. By combining numerical summaries with graphical analyses of distributions, the study provides a comprehensive overview of benchmark performance patterns across the contemporary LLM ecosystem.

4. Dataset and Benchmark Performance

4.1 Dataset Characteristics and Analytical Context

The empirical investigation is based on the Comprehensive LLM Benchmark Dataset (Dhiman, 2024), which provides a structured evaluation framework for contemporary Large Language Models (LLMs). The dataset comprises 390 model benchmark observations spanning fifteen variables that capture model specifications, benchmark characteristics, and performance outcomes. The unit of analysis is the model benchmark dyad, enabling detailed comparisons across model architectures, organizational developers, benchmark domains, and performance levels.

The dataset encompasses models released between 2022 and 2024, a period characterized by rapid advances in foundation model development. The sample includes models developed by major artificial intelligence organizations such as Anthropic, Meta AI, Mistral AI, and Cohere. Furthermore, the dataset represents both open-source and proprietary development paradigms, thereby providing a comprehensive view of the contemporary LLM ecosystem.

4.2 Variable Definitions and Data Preparation

Prior to analysis, benchmark scores were standardized to facilitate comparison across heterogeneous evaluation tasks. The dataset contains variables describing model identity, developer organization, benchmark category, raw benchmark score, and normalized performance metrics. Benchmark scores represent reported performance outcomes collected from publicly available evaluations. Normalized performance percentages were calculated to place benchmark outcomes on a common scale ranging from 0 to 100. Data were screened for missing values, duplicate observations, and extreme anomalies prior to statistical analysis.

4.3 Analytical Procedure

The analytical strategy proceeded in four stages. First, descriptive statistics were computed to summarize central tendency and variability. Second, distributional characteristics were examined using histograms and kernel density estimation. Third, violin plots and boxplots were used to identify concentration patterns, spread, and outliers. Finally, benchmark outcomes were categorized in to performance tiers to evaluate capability stratification within the LLM ecosystem. Together, these complementary approaches provide a

multidimensional characterization of benchmark performance.

5. Analysis

To establish a foundation for subsequent analyses of scaling behavior, architectural effectiveness, and benchmark specific performance differences, the study first examines the overall distribution of benchmark outcomes. Specifically, descriptive statistics, distributional analyses, density estimation, violin plots, boxplots, and performancetier distributions are employed to characterize the underlying structure of model perfo-rmance.

Metric	Mean	Median	Std. Dev	Min	Max
Benchmark Score	64.5882	72.4	23.3901	5.2	98.1
Normalized Performance (%)	67.3643	74.1	20.6733	5.2	98.1

Table 1. Summary Statistics of Performance Measures

The descriptive statistics reveal substantial variability in benchmark performance across the evaluated models. Benchmark scores range from 5.2 to 98.1, while normalized performance values exhibit a similar range. The relatively high standard deviations indicate considerable heterogeneity among models, reflecting differences in architectural design, parameter scale, training methodologies, and benchmark specialization.

5.1 Distribution of Benchmark Scores

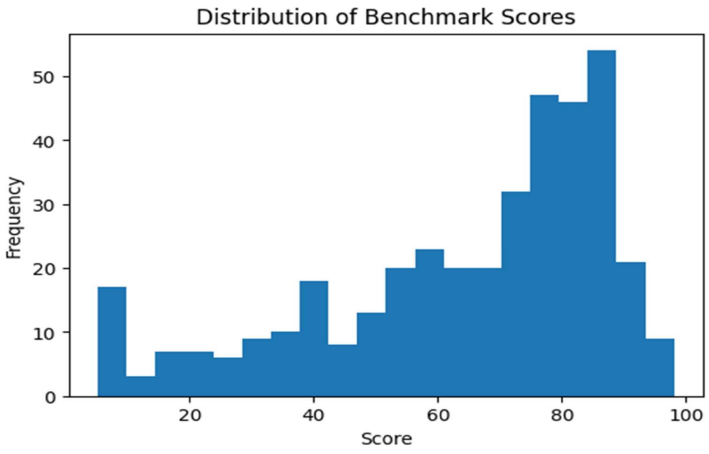

Figure 1. Distribution of Benchmark Scores

The histogram of benchmark scores provides an overview of the performance landscape across the evaluated models. The mean benchmark score is 64.59, while the median score is notably higher at 72.40. This difference suggests a negatively skewed distribution in which most models achieve moderate to high benchmark performance, while a smaller subset of lower performing models pulls the average down.

The substantial range of observed values, spanning 5.2 to 98.1, highlights the diversity of capabilities in the dataset. Such variation reflects the coexistence of frontier scale systems, highly optimized benchmark performers, emerging open source alternatives, and smaller specialized models. The broad dispersion of scores therefore illustrates the fragmented nature of the current LLM landscape, where performance leadership remains concentrated among a limited number of highly capable systems while a larger group of models exhibits varying degrees of task specific proficiency.

The large standard deviation (23.39) further confirms that benchmark outcomes are far from homogeneous. Consequently, benchmark performance cannot be adequately captured by a single measure of central tendency, underscoring the importance of examining the full distribution of outcomes.

5.2 Distribution of Normalized Performance

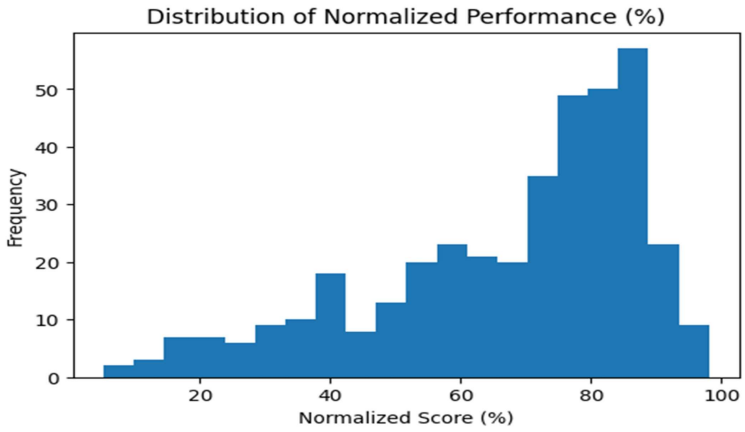

Figure 2. Distribution of Normalized Performance

To facilitate cross benchmark comparability, raw benchmark scores are transformed into normalized performance percentages. The resulting distribution exhibits characteristics similar to those observed for the raw benchmark scores.

The average normalized performance reaches 67.36%, with a median value of 74.10%. The higher median relative to the mean again suggests that lower performing observations exert downward pressure on the overall average. Most evaluated models achieve moderate to high normalized performance, indicating that contemporary LLMs generally demonstrate strong competency across standardized evaluation tasks.

The concentration of observations toward higher performance values reflects the rapid pace of recent advances in language model development. Improvements in parameter scaling, instruction tuning, reinforcement learning techniques, retrieval augmentation, and architectural optimization have collectively contributed to elevated benchmark performance across the industry. Nevertheless, the persistence of lower scoring observations suggests that significant capability gaps remain among different model families.

The observed variability highlights the continuing importance of comparative benchmarking as a mechanism for distinguishing incremental improvements from substantial capability gains.

5.3 Density-Based Examination of Performance Distributions

While histograms provide a discrete representation of performance frequencies, density estimation offers a smoother visualization of the underlying probability distributions. The density curves reveal a pronounced concentration of observations within the medium to high performance region.

The density distributions indicate that relatively few models occupy the lower performance range. Instead, most observations cluster around competitive benchmark scores, suggesting that the industry has reached a stage where baseline language capabilities are becoming increasingly standardized. This finding aligns with recent developments in the LLM ecosystem, where many organizations have adopted similar training strategies, architectural principles, and evaluation practices.

The density curves therefore provide evidence of an increasingly mature benchmark environment in which competitive performance is becoming more common, even though substantial differences remain among leading frontier models.

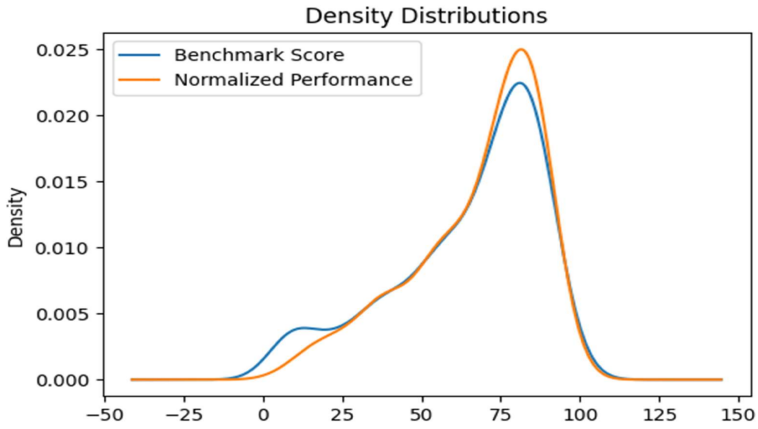

Figure 3. Density Plot of Benchmark and Normalized

5.4 Distributional Characteristics through Violin Plot Analysis

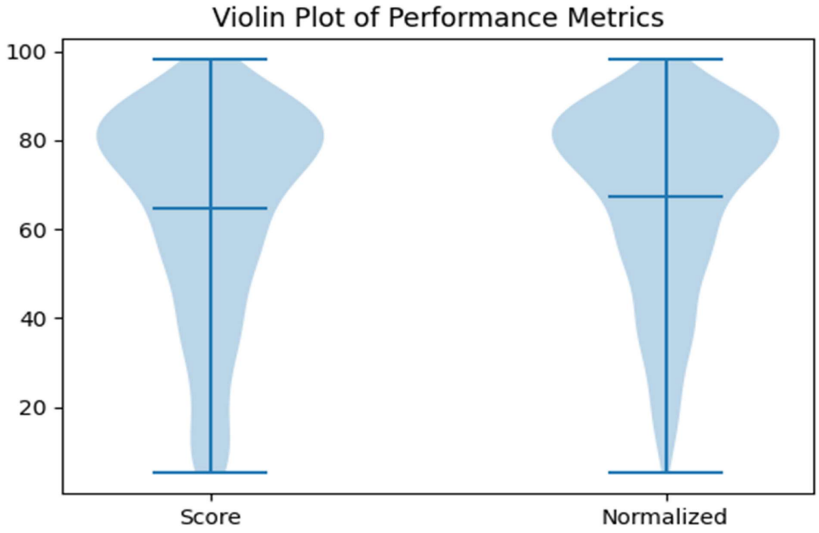

Figure 4. Violin Plot of Performance Metrics

The violin plot provides a more detailed view of the distributional characteristics of benchmark performance by simultaneously displaying density, spread, and central tendency. Compared with traditional summary statistics, violin plots offer enhanced visibility into the concentration of observations across different performance ranges.

The wider sections of the violin plot correspond to regions where many observations are concentrated, while narrower regions indicate relatively sparse performance levels. The plot reveals that both benchmark scores and normalized performance values exhibit substantial clustering within the upper middle performance range. This observation supports the findings derived from the histogram and density analyses.

Furthermore, the shape of the distribution suggests that performance improvements among modern LLMs are not uniformly distributed. Instead, performance appears concentrated around several dominant capability levels, possibly reflecting common architectural configurations and training methodologies adopted across the industry.

The violin plot thus highlights performance stratification within the broader population of language models.

5.5 Boxplot Analysis of Performance Variability

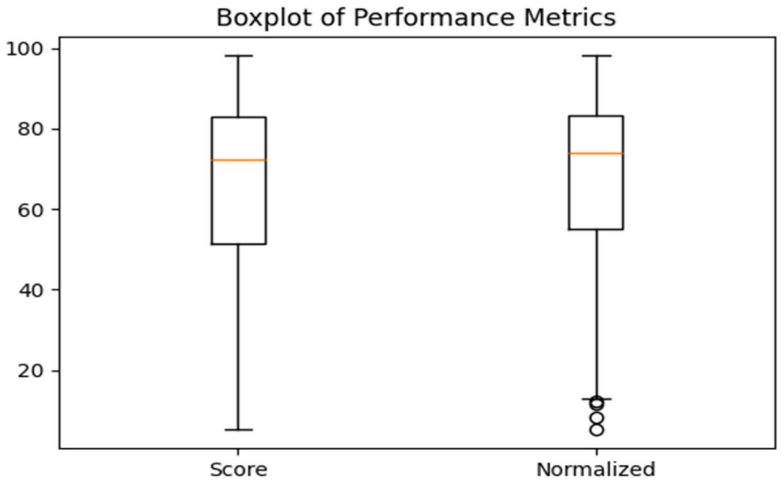

Figure 5. Boxplot of Performance Metrics

Boxplots provide a concise summary of distributional characteristics through the visualization of medians, interquartile ranges, and potential outliers. The boxplots reveal relatively high median values for both benchmark scores and normalized performance, confirming the generally strong performance of contemporary language models.

The interquartile ranges indicate that substantial variation exists even among the middle 50% of observations. This variability suggests that benchmark success remains strongly influenced by differences in model architecture, parameter count, training corpus quality, and optimization strategies.

Additionally, the extended whiskers observed in the boxplots indicate the presence of extreme performance values. These observations likely correspond to highly specialized models positioned at opposite ends of the capability spectrum, including both frontier systems and comparatively limited baseline models.

Consequently, the boxplot analysis reinforces the conclusion that benchmark performance remains highly heterogeneous despite overall improvements in language model capabilities.

The concentration of observations within the upper-middle performance range suggests that benchmark performance may be approaching a saturation phase for many standardized tasks. As benchmark scores converge, marginal improvements become increasingly difficult to achieve and may provide diminishing insight into meaningful capability differences. This pattern reinforces recent concerns that future evaluations should emphasize robustness, reasoning, adaptability, and real-world task execution rather than incremental gains on established benchmarks.

5.6 Performance-Tier Distribution

To facilitate interpretability, benchmark outcomes were further categorized into performance tiers representing distinct capability levels. The performance tier distribution provides insights into the relative prevalence of Weak, Moderate, Strong, and Top Tier models within the dataset.

The results indicate that the largest proportion of observations falls within the Strong tier (36.92%), followed by the Weak tier (25.13%), Top Tier (20.00%), and Moderate tier (17.95%). Collectively, Strong and Top-Tier models account for 56.92% of all observations, suggesting that advanced benchmark performance has become increasingly prevalent among contemporary LLMs. Nevertheless, the presence of a substantial proportion of Weak and Moderate performers demonstrates that significant capability differences continue to exist across the ecosystem.

Performance Tier	Frequency (n)	Percentage (%)
Weak	98	25.13
Moderate	70	17.95
Strong	144	36.92
Top Tier	78	20.00
Total	**390**	**100.00**

Table 2. Performance Tier Distribution

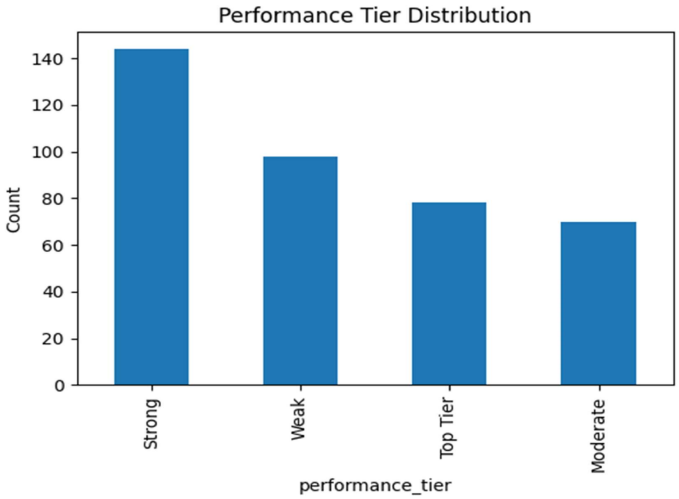

Figure 6. Distribution of Performance Tiers

based representation complements the continuous performance metrics by providing a categorical view of capability stratification within the LLM landscape. The findings provide empirical evidence that contemporary language models are distributed across distinct performance classes rather than forming a homogeneous population.

6. Discussion

Research Question Synthesis

RQ1

Benchmark performance exhibits a negatively skewed distribution characterized by relatively high median values and substantial variability.

RQ2

Considerable heterogeneity exists among models, as evidenced by wide score ranges and large standard deviations.

RQ3

Distributional analyses reveal distinct capability strata, indicating that LLM performance is clustered rather than continuously distributed.

RQ4
Performance tiers demonstrate the concentration of models within advanced and frontier capability categories.

RQ5
The observed distributions suggest increasing ecosystem maturity while simultaneously revealing persistent concentration of leadership among a limited number of frontier systems.

Collectively, the descriptive analyses reveal a benchmark ecosystem characterized by both strong overall performance and substantial heterogeneity. Benchmark scores and normalized performance values indicate that most contemporary LLMs achieve relatively high levels of competency across standardized evaluation tasks. However, the wide performance ranges and large standard deviations demonstrate that significant capability differences persist among models.

The histogram, density, violin, and boxplot analyses consistently indicate a concentration of observations within the medium to high performance range, reflecting the maturation of the LLM landscape and the widespread adoption of effective architectural and training strategies. At the same time, the existence of lower performing models highlights continuing disparities in model development approaches and resource availability.

The performance tier distribution further illustrates the stratification of model capabilities, revealing that the LLM ecosystem consists of distinct performance classes rather than a single homogeneous population. These findings establish an empirical foundation for subsequent analyses investigating scaling laws, organizational performance differences, architectural effects, benchmark specialization, efficiency trade offs, and the broader evolution of large language model capabilities.

7. Conclusion

This study presented a descriptive examination of benchmark performance across contemporary Large Language Models using a comprehensive dataset of model benchmark observations. The analysis revealed that benchmark performance is characterized by both strong overall capability and substantial variability across models. While most models achieve moderate to high benchmark scores, considerable performance differences remain, indicating that the LLM ecosystem is far from homogeneous.

The descriptive statistics demonstrated broad score distributions and significant standard deviations, reflecting differences in model architectures, training methodologies, parameter scales, and organizational development strategies. Histogram, density, violin plot, and box plot analyses consistently revealed a concentration of observations in the medium to high performance range, suggesting that competitive benchmark performance has become increasingly common among modern language models. At the same time, the persistence of lower-performing observations highlights continuing disparities in model capabilities and resource availability.

The performance tier analysis further demonstrated that contemporary LLMs can be grouped into distinct capability strata, ranging from basic and intermediate systems to advanced and frontier models. This stratification suggests that benchmark leadership remains concentrated among a relatively small number of highly capable systems, while the broader ecosystem exhibits varying levels of specialization and performance maturity.

Overall, the findings indicate that benchmark ecosystems have evolved into important instruments for evaluating and comparing language models, yet benchmark outcomes should be interpreted with caution. Aggregate scores provide useful indicators of comparative performance but may conceal important differences in underlying capabilities and task-specific competencies. Consequently, future evaluation frameworks should move beyond single-score assessments toward more transparent, capability oriented, and multidimensional approaches that better capture the complexity of contemporary LLM behavior.

The present study establishes a descriptive baseline for future research examining scaling relationships,

architectural effects, benchmark specialization, efficiency performance trade offs, organizational competitiveness, and the evolution of capability distributions within the rapidly developing landscape of large language models.

Future research should extend beyond descriptive distributions toward multivariate investigations of capability formation. Potential directions include scaling law analysis, clustering of benchmark specialization patterns, latent capability modeling, organizational benchmarking comparisons, efficiency performance trade off analysis, and longitudinal studies examining the evolution of benchmark ecosystems over time.

7.1 Limitations

Several limitations should be acknowledged. First, the analysis relies on benchmark scores reported within a single dataset and therefore reflects the quality and completeness of those benchmark records. Second, benchmark outcomes may not fully capture real world performance because of benchmark contamination, dataset memorization, and differences in evaluation protocols. Third, the descriptive nature of the study does not permit causal inference regarding the factors driving performance differences. Future studies should incorporate model architecture, parameter size, training strategies, and benchmark domains to investigate the determinants of benchmark success.

References

[1] Miller, J. K., Tang, W. (2025). Evaluating LLM metrics through real-world capabilities. *arXiv preprint arXiv:2505.08253*.

[2] Smith, J., Doe, J. (2024). A comprehensive framework for evaluating multilingual tokenizer quality. *arXiv preprint arXiv:2410.12989*.

[3] Center for AI Safety AI, S. (2025). Humanity's last exam. Retrieved March 4, 2025, from https://lastexam.ai

[4] Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., Lee, J. R., Lee, Y. T., Li, Y., Liu, W., Mendes, C. C. T., Nguyen, A., Price, E., de Rosa, G., Saarikivi, O., Zhang, Y. (2024). Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

[5] Miller, E. (2024). Adding error bars to evals: A statistical approach to LLM evaluations. *arXiv preprint arXiv:2402.03091*.

[6] He, J., Du, J., Neubig, G., Tan, Z., Duh, K. (2023). On the blind spots of model based evaluation metrics for text generation. *In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. https://aclanthology.org/2023.acl-long.871.

[7] Wang, Y., Chen, W., Chen, S., Xu, C., Wang, Y., Liu, Z., Wang, L., Huang, M. (2024). Large language models are not fair evaluators. *In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. https://aclanthology.org/2024.acl-long.152.

[8] Balloccu, S., Schmidtova, P., Lango, M., Dusek, O. (2024). Leak, cheat, repeat: Data contamination and evaluation malpractices in closed source LLMs. In Y. Graham M. Purver (Eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (p. 67–93). *Association for Computational Linguistics*. https://aclanthology.org/2024.eacl-long.5/

[9] Gallifant, J., Cowen, T., Brockman, G., et al. (2024). Peer review of the GPT-4 technical report. *PLOS Digital Health*. https://doi.org/10.1371/journal.pdig.0000291.

[10] Sottana, L., Ribeiro, L. F. R., Gurevych, I. (2023). Evaluation metrics in the era of GPT-4: Can we trust

reference-based scores *In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://aclanthology.org/2023.emnlp-main.99.

[11] Calais, P., Lopes, V., Freire, A., Jariwala, S., Kiyuna, D., Ribeiro, L., et al. (2024). Beyond accuracy: Performance of LLMs on human exams. *arXiv preprint arXiv:2403.05004.*

[12] von Laszewski, G., Brewer, W., Thiyagalingam, J., Papay, J., Foundjem, A., Luszczek, P., Fox, G. C. (2025). AI benchmark democratization and carpentry. *arXiv preprint arXiv:2512.11588.*

[13] Von Laszewski, G., Hawks, B., Colombo, M., Shiraishi, R., Krishnan, A., Tran, N., Fox, G. C. (2025). MLCommons science working group AI benchmarks collection. GitHub. https://mlcommons-science.github.io/benchmark/benchmarks.pdf

[14] Hawks, B., von Laszewski, G., Sinclair, M. D., Colombo, M., Venkataraman, S., Jain, R., Jiang, Y., Tran, N., Fox, G. (2025). An MLCommons scientific benchmarks ontology. *arXiv preprint arXiv:2511.05614.*

[15] MLCommons. (2023). Machine learning innovation to benefit everyone. Retrieved April 13, 2023, from https://mlcommons.org/.

[16] Kim, D., Shim, G., Chun, Y., Kim, M., Park, C., Lim, H. (2025). Benchmark profiling: Mechanistic diagnosis of LLM benchmarks. *In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (p. 15635–15650). *Association for Computational Linguistics*.

[17] Shool, S., Adimi, S., Saboori Amleshi, R., Bitaraf, E., Golpira, R., Tara, M. (2025). A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making, 25*(1), 117.

[18] Gong, E. J., Bang, C. S., Lee, J. J., Baik, G. H. (2025). Knowledge practice performance gap in clinical large language models: Systematic review of 39 benchmarks. *Journal of Medical Internet Research, 27*, e84-120. https://doi.org/10.2196/84120.

[19] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, £., Polosukhin, I. (2017). Attention is all you need. *In Advances in Neural Information Processing Systems (NeurIPS 2017)*. Curran Associates.

[20] Banerjee, S., Agarwal, A., Singh, E. (2024). The vulnerability of language model benchmarks: Do they accurately reflect true LLM performance *arXiv preprint arXiv:2412.03597.*

[21] Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Zaremba, W. (2021). Evaluating large language models trained on code.

[22] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. (2018). Improving language understanding by generative pretraining. *OpenAI*, 1–12.

[23] Raji, I., Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. *In* Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19) (p. 429–435). ACM. https://doi.org/10.1145/3306618.3314244.

[24] Cao, Y., Hong, S., Li, X., Ying, J., Ma, Y., Liang, H., Jiang, Y. G. (2025). Toward generalizable evaluation in the LLM era: A survey beyond benchmarks. *arXiv preprint arXiv:2504.18838.*

[25] Hardy, M., Reuel, A., Zhang, L., Casabianca, J. M., Truong, S., Dave, Y., Koyejo, S. (2026). AI cartography: Mapping the latent landscape of AI benchmark ecosystems. *arXiv preprint arXiv:2605.25272.*

[26] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta Navas, D., Hudson, D. A., Koreeda, Y. (2023). Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110.*

[27] Fourrier, C., Habib, N., Lozovskaya, A., Szafer, K., Wolf, T. (2024). Open LLM Leaderboard v2. Hugging Face. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

[28] Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., Steinhardt, J. (2021). Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874.*

[29] Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., Wei, J. (2022). Challenging BIG-Bench tasks and whether chain ofthought can solve them. *arXiv preprint arXiv:2210.09261.*

[30] Salaudeen, O., Reuel, A., Ahmed, A., Bedi, S., Robertson, Z., Sundar, S., Domingue, B., Wang, A., Koyejo, S. (2025). Measurement to meaning: A validity centered framework for AI evaluation. *arXiv preprint arXiv:2505.10573.*

[31] Reuel, A., Hardy, A., Smith, C., Lamparth, M., Hardy, M., Kochenderfer, M. J. (2024). BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. *arXiv preprint arXiv:2411.12990.*

[32] Truong, S., Tu, Y., Hardy, M., Reuel, A., Tang, Z., Burapacheep, J., Perera, J., Uwakwe, C., Domingue, B., Haber, N., Koyejo, S. (2025). Fantastic bugs and where to find them in AI benchmarks. *arXiv preprint arXiv :2511.16842.*

[33] Casabianca, J. M. (2025). Psychometrics is all you need. https://osf.io/preprints/edarxiv/7w6pz_v1/.

[34] Hu, T., Zhou, X. H. (2024). Unveiling LLM evaluation focused on metrics: Challenges and solutions. *arXiv preprint*. (Preprint posted April 14, 2024).

[35] Shah, N. H., Entwistle, D., Pfeffer, M. A. (2023). Creation and adoption of large language models in medicine. *JAMA, 330*(9), 866–869.

[36] Gong, E. J., Bang, C. S. (2024). Evaluating the role of large language models in inflammatory bowel disease patient information. *World Journal of Gastroenterology, 30*(29), 3538–3540.

[37] Abdollahi, M., Zhang, R., Harzevili, N. S., Shin, J., Wang, S., Hemmati, H. (2026). Surveying the benchmarking landscape of large language models in code intelligence. *ACM Transactions on Software Engineering and Methodology*. Advance online publication. https://doi.org/10.1145/38009.

[38] Cao, D., Yu, B. (2025). Survey of emerging trends in LLM agent benchmarking. In *BDNNDL '25: Proceedings of the 2025 2nd Symposium on Big Data, Neural Networks, and Deep Learning* (p. 31–35). https://doi.org/10.1145/3784013.3784018.

[39] Weerasinghe, S., Li, X., Uddin, M. A., Silva-Junior, D., Neto, M. V. dos S., Ribeiro, M. E. S., Mazhar, H. B., Akbarsharifi, M., Akbarsharifi, R., Mani, N., Graciano Neto, V. V., Mäntylä, M., Galvão, A. (2026). Beyond the prompt: An analysis of the current state of automated test generation with LLMs (SSRN Scholarly Paper No. 6843081). http://dx.doi.org/10.2139/ssrn.6843081.

[40] Mohammadi, M., Li, Y., Lo, J., Yip, W. (2025). Evaluation and benchmarking of LLM agents: A survey. *In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2* (p. 6129–6139).

[41] Dhiman, A. (n.d.). Comprehensive LLM benchmark dataset. Kaggle. Retrieved from https://www.kaggle-.com/datasets/aishricadhiman/comprehensive-llm-benchmark-dataset.

[42] DeepSeek. (2024). DeepSeek-Coder-V2: Advancements in AI coding capabilities. Retrieved March 4, 2025,from https://www.deepseek.com/deepseek-coder-v2.