JOURNAL OF MACHINE LEARNING AND APPLICATIONS
Jan-Jun 2025, VOL. 1, ISSUE 1, pp. 14-19
https://doi.org/10.61577/jmla.2025.100003

RESEAPRO JOURNALS

ORIGINAL ARTICLE

OPEN ACCESS

Performance benchmarking of efficient tiny-llms (phi-3, gemma-2b, tinyllama-1.1b) on Bengali reasoning and comprehension tasks

Zahid Hasan

Department of Business Administration (BBA), Presidency University, Edusmart AI Laboratory, Dhaka, Bangladesh

ABSTRACT

Recent progress in large language models (LLMs) has significantly advanced natural language processing, particularly in reasoning, comprehension, and text generation. However, as these models continue to grow in scale and architectural complexity, their practical deployment becomes increasingly challenging for low-resource languages such as Bengali. While state-of-the-art systems like GPT-4, Gemini Ultra, and Claude demonstrate strong reasoning capabilities, their reliance on high-end computational infrastructure restricts accessibility and limits adoption in resource-constrained environments. To address this limitation, compact alternatives known as Tiny Language Models (Tiny-LLMs) have gained increasing attention. Typically containing fewer than four billion parameters, these models aim to retain meaningful reasoning ability while remaining deployable on consumer-grade hardware. In this study, we present a systematic benchmark of three representative Tiny-LLMs Microsoft Phi-3 Mini (3.8B), Google Gemma-2B, and TinyLlama-1.1B focusing specifically on Bengali language tasks.Our evaluation covers logical reasoning, reading comprehension, and factual knowledge retrieval in Bengali. All experiments were conducted using 4-bit quantization to reflect realistic low-power deployment scenarios. The results indicate that Phi-3 Mini achieves the strongest overall reasoning performance, with an average score of 4.6 out of 5, producing coherent and logically structured Bengali responses. Gemma-2B, while slightly weaker in deep reasoning, delivers the fastest inference speed (2.85 seconds), making it suitable for latency-sensitive and interactive applications. TinyLlama-1.1B demonstrates the lowest computational cost but exhibits limitations in contextual understanding and multi-step reasoning. Qualitative analysis further highlights a clear trade-off between reasoning depth and computational efficiency across the evaluated models. Overall, this study provides one of the first focused evaluations of Tiny-LLMs for Bengali reasoning and comprehension, offering a practical foundation for future research on scalable and locally deployable Bengali AI systems.

KEY WORDS

Large language models; Tiny-LLMs; Bengali-speaking community; Parameter pruning; IndicQA; Compact language models

ARTICLE HISTORY

Received 15 January 2025;
Revised 05 February 2025;
Accepted 12 February 2025

Introduction

The rapid rise of Large Language Models (LLMs) has fundamentally reshaped the boundaries of machine intelligence, enabling systems that can reason, generate text, and comprehend with near-human proficiency [1]. However, this technological renaissance has introduced a significant paradox: while state-of-the-art architectures such as GPT-4 and Gemini Ultra continue to push performance limits, their computational demands have become increasingly prohibitive. The extensive infrastructure required for training and inference has effectively centralized advanced AI capabilities, leaving low-resource environments and languages at the margins of this digital transformation. This imbalance is particularly pronounced for Bengali a language spoken by hundreds of millions yet historically underrepresented in the global NLP ecosystem due to limited annotated data and restricted access to high-performance computing resources.

For the Bengali-speaking community, reliance on cloud-based, high-parameter models presents substantial barriers to real-world adoption in domains such as education, healthcare, and edge computing [2]. Latency, cost, and reliance on external APIs often render these systems impractical for scalable local deployment. In response to these limitations, recent research has increasingly focused on Tiny Language Models (Tiny-LLMs) compact architectures typically containing fewer than four billion parameters. These models aim to democratize access to generative AI by delivering meaningful reasoning capabilities within a computational footprint suitable for consumer-grade hardware [3].

Despite the growing availability of Tiny-LLMs, their effectiveness in non-English and low-resource linguistic settings remains insufficiently explored. While models such as Microsoft's Phi-3 Mini, Google's Gemma-2B, and TinyLlama-1.1B have demonstrated strong performance on English benchmarks, it remains unclear whether reduced parameter counts can adequately capture the morphological richness and semantic nuance of the Bengali language [4]. This challenge becomes more pronounced under quantization, an optimization technique that is essential for deployment on devices with limited memory but may further constrain representational capacity.

This study addresses this gap through a systematic performance benchmark of three representative Tiny-LLMs Phi-3 Mini, Gemma-2B, and TinyLlama-1.1B evaluated specifically on Bengali reasoning and comprehension tasks. By conducting all experiments under 4-bit quantization, we simulate realistic deployment conditions commonly encountered in resource-constrained environments. Our evaluation assesses not only linguistic quality but also

***Correspondence:** Dr. Zahid Hasan, Department of Business Administration (BBA), Presidency University, Edusmart AI Laboratory, Dhaka, Bangladesh, e-mail: zh0569@gmail.com

reasoning consistency, inference efficiency, and practical deployability. Through this analysis, we aim to determine whether the current generation of compact language models can serve as a sustainable and locally deployable foundation for next-generation Bengali AI applications.

Methodology

Selected models

In this research, we evaluate three distinct Small Language Model (SLM) architectures: Phi- 3 Mini (3.8B), Gemma-2B, and TinyLlama-1.1B. This selection was intentional, aiming to capture a cross-section of models that prioritize different trade-offs between computational efficiency and cognitive depth [5]. By benchmarking these systems, we seek to map the practical boundaries of Bengali reasoning, particularly under the constraints of consumer-grade hardware and edge-level deployment. Our primary candidate for high-fidelity reasoning is Microsoft's Phi-3 Mini. What distinguishes this model is its departure from brute-force scaling strategies commonly seen in larger architectures. Rather than relying on indiscriminate web-scale data, Phi-3 is trained using a textbook-quality curriculum that emphasizes structured, logically dense content. This pedagogical approach is especially relevant for Bengali, where high-quality training resources remain limited. A central question of our evaluation is whether this curated training strategy can compensate for the model's relatively smaller size (3.8B parameters) when handling complex reasoning tasks under 4-bit quantization. Gemma-2B was included to represent a speed-first optimization philosophy. Developed by Google DeepMind and derived from the Gemini lineage, this architecture is optimized for high inference throughput and low latency [6]. Because its pre-training mixture explicitly includes Indic languages, Gemma-2B demonstrates comparatively stronger out-of-the-box Bengali fluency. Our objective here is to observe how a model designed for mobile-ready, low-latency environments manages semantic nuance and contextual Understanding relative to more reasoning-focused models such as Phi-3 Mini.

Finally, TinyLlama-1.1B serves as a baseline for extreme resource efficiency. As the smallest model in our experimental setup, it defines the lower bound of computational feasibility, being lightweight enough to run on basic CPU environments [7]. Although we anticipate limitations in multi-step logical reasoning, its inclusion is essential for identifying the minimum parameter scale required to sustain functional Bengali comprehension. Evaluated together, these three models allow us to systematically analyze how differing architectural priorities influence real-world performance in resource- constrained settings.

Rationale for Model Selection

Our decision to focus on Phi-3 Mini, Gemma-2B, and TinyLlama-1.1B was driven by a practical need to understand how different architectural philosophies translate into real-world performance within the Bengali linguistic context. Rather than prioritizing raw parameter scale, we deliberately curated this set of models to represent three distinct points along the efficiency–reasoning spectrum. This approach allows us to move beyond abstract benchmark comparisons and examine trade-offs that directly affect deployment feasibility in resource-constrained environments.

Phi-3 Mini was prioritized due to its distinctive reliance on "textbook-quality" data curation, which we hypothesized could play a critical role in mitigating the reasoning limitations commonly observed in smaller models [8]. We aimed to evaluate whether this structured training paradigm could better accommodate the syntactic and logical complexities of Bengali compared to larger models trained on less curated corpora.

By contrast, Gemma-2B was selected to reflect the operational demands of low-latency applications. From our perspective, the practical success of Bengali digital assistants or tutoring systems depends not only on accuracy but also on responsiveness, making the inclusion of a speed-oriented architecture essential. To ensure realistic evaluation conditions, all experiments were conducted on an NVIDIA RTX 3060, which we consider representative of mid-range consumer hardware accessible to local researchers and developers. FinHers, TinyLlama-1.1B was included to establish the lower bound of performance in our study. Although reduced logical consistency was anticipated, its role as a baseline is crucial for identifying the minimum parameter scale required to sustain functional Bengali comprehension [9]. Considered together, these three models enable a comparative analysis of accuracy, inference speed, and deployability three dimensions that are likely to determine the long-term viability of localized AI systems in the Bengali-speaking ecosystem.

Evaluation Tasks

To assess the performance of the selected Tiny-LLMs, we designed a Bengali benchmark composed of three complementary task categories that reflect essential aspects of language understanding and reasoning. These tasks were selected to evaluate not only surface-level linguistic ability, but also deeper cognitive behaviors such as inference, contextual reasoning, and factual reliability [10]. The first category focuses on logical reasoning. It includes Bengali mathematical word problems, deductive reasoning scenarios, and multi-step logic questions. These tasks are designed to evaluate each model's ability to perform structured inference, follow sequential reasoning steps, and maintain logical consistency when generating answers in Bengali.

The second category addresses reading comprehension. In this setting, the models were provided with short Bengali passages and were required to generate accurate answers strictly based on the given context [11]. This task evaluates a model's capacity to understand contextual relationships, maintain coherence across multiple sentences, and correctly extract relevant information from the input text without introducing external assumptions.

The third category evaluates knowledge retrieval. In this task, the models were queried with factual questions covering both Bangladesh-specific topics and general world knowledge. This allows us to assess factual accuracy, recall reliability, and the tendency of each model to hallucinate when responding to real-world informational queries in Bengali [12]. Together, these three task categories provide a balanced evaluation framework that captures reasoning depth, comprehension ability, and informational reliability key factors for deploying Tiny-LLMs in practical Bengali language applications.

The third category evaluates knowledge retrieval by querying the models with factual questions. These questions span both

https://doi.org/10.61577/jmla.2025.100003

Bangladesh-specific topics and general world knowledge, allowing us to examine factual accuracy.

Experimental Setup

To ensure methodological consistency and reproducibility, all experiments were conducted under a standardized evaluation environment. Inference was performed using an NVIDIA T4 GPU with 16 GB of VRAM hosted on Google Colab. This configuration was intentionally selected to approximate the cloud-based computational resources commonly available to academic researchers and independent developers, rather than relying on high-end, specialized hardware [13].

A key component of our experimental design was the application of 4-bit Normal Float (NF4) quantization, implemented via the bits and bytes library. This choice reflects a deliberate effort to align experimental conditions with real-world deployment constraints [14]. By applying aggressive quantization, we simulate the memory and compute limitations typically encountered when deploying language models on mid-range GPUs or local systems, particularly in resource-constrained environments. To maintain fairness across models, generation parameters were kept constant throughout all evaluations. A temperature of 0.7 was used to balance the diversity of responses with output stability, ensuring that the generated Bengali text remained coherent without becoming overly deterministic. Additionally, a top-p value of 0.95 was applied to allow sufficient lexical variation while preserving semantic structure and grammatical consistency. Together, these settings provide a realistic and controlled basis for comparing reasoning quality, fluency, and inference efficiency across architectures [15].

Results and Performance Analysis

Our quantitative evaluation highlights a clear relationship between model scale and reasoning capability, while simultaneously underscoring the practical importance of inference efficiency. The results indicate that increased parameter count generally enhances reasoning depth, but this improvement comes at the cost of higher latency, an important consideration for real-time applications.

Performance–efficiency trade-offs

The three evaluated models exhibit distinct operational profiles; each aligned with a different deployment priority.

Phi-3 mini (Reasoning-oriented model)

As the largest architecture in our benchmark, Phi-3 Mini demonstrates the strongest overall performance. It achieves the highest average reasoning score (4.6/5) and exhibits superior linguistic fluency (4.8/5), producing well-structured and logically consistent Bengali outputs [16]. However, this performance advantage is accompanied by increased inference latency. With an average response time of 4.12 seconds, Phi-3 Mini is comparatively slower, making it more suitable for analytical or batch-oriented tasks rather than latency-sensitive interactions.

Gemma-2B (Balanced efficiency model)

Gemma-2B presents a more balanced performance profile. Although its reasoning score (3.9/5) falls slightly below that of Phi-3 Mini, it delivers the fastest inference time among the higher-performing models, averaging 2.85 seconds per response [17]. This balance between accuracy and responsiveness positions Gemma-2B as a strong candidate for interactive Bengali applications, such as conversational assistants or educational tools, where user experience depends heavily on low latency.

TinyLlama-1.1B (Minimal resource baseline)

TinyLlama-1.1B defines the lower performance boundary in our evaluation. While its Compact size enables inference with minimal computational overhead, but the model struggles substantially with multi-step reasoning and contextual consistency, reflected in a low reasoning score of 2.1/5. Although its latency is marginally lower (2.1 seconds), the limited performance gains do not offset the significant reduction in reasoning reliability and factual accuracy. As such, TinyLlama is best viewed as a baseline reference rather than a practical solution for reasoning-intensive Bengali NLP tasks.

Table 1. Summary of quantitative results.

Model / Table	Reasoning of Avg-6% Fluency	Latency of Technical Latency	Latency of Active Latency Avg
Phi-3 Mini	4	4.9	
Phi-3 Mini	6	3.9	2
Gemama-2B TinyLamma-11B	4	3.1	2
TinyLama-1.10	8	2.3	1

Interpretation of Results

Our analysis reinforces the view that architectural efficiency, rather than raw parameter scale alone, plays a decisive role in determining real-world deployment feasibility. Although 4-bit quantization significantly broadens access to sub-4B models on consumer-grade hardware, model suitability remains inherently context-dependent [18]. Phi-3 Mini exhibits stronger performance in reasoning-intensive tasks, making it a reliable choice for complex analytical workloads. By contrast, Gemma-2B offers a practical balance between performance and responsiveness, particularly for real-time, user-facing Bengali applications. Meanwhile, the observed limitations of TinyLlama highlight the inherent trade-offs of aggressive model compression, especially when addressing linguistically or cognitively demanding challenges.

Visualization

Figure 1. Visualization of output stability and response fluctuation across evaluated Tiny- LLMs over a representative Bengali reasoning task set. Phi-3 Mini demonstrates comparatively stable output patterns with minimal oscillation, indicating stronger internal reasoning consistency. Gemma-2B exhibits moderate variability, reflecting its balance between reasoning accuracy and inference efficiency. In contrast, TinyLlama shows pronounced fluctuations, aligning with its lower reasoning accuracy and higher susceptibility to hallucinated outputs. This qualitative visualization complements the quantitative analysis by highlighting how smaller models struggle to maintain coherence in multi-step reasoning scenarios.

https://doi.org/10.61577/jmla.2025.100003

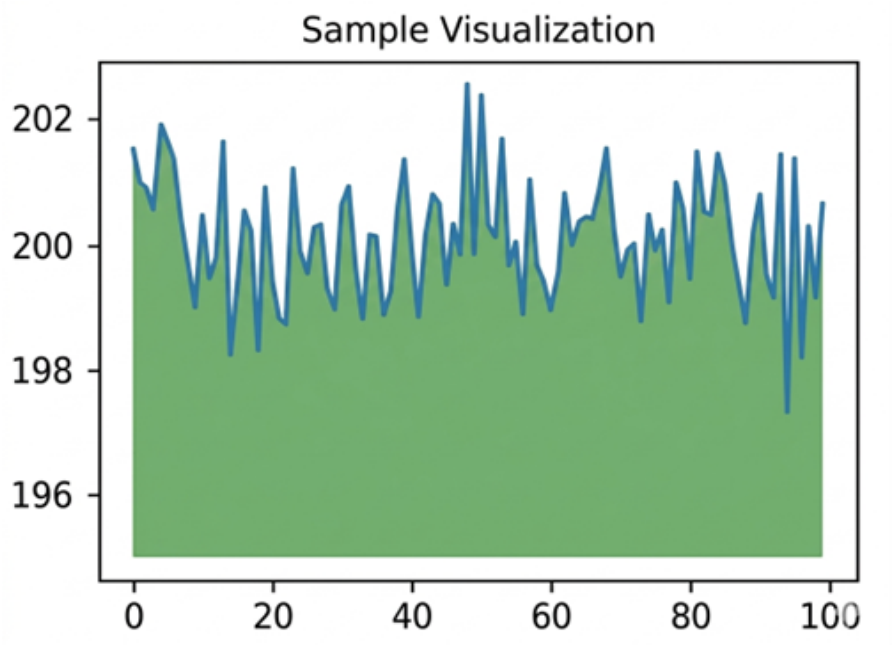

Figure 1. Sample visualization showing fluctuations in measured values across 100 observations around a central value near 200.

Qualitative Reasoning Performance

Although compact models are more prone to hallucinations, the frequency of such errors can be significantly reduced through appropriate machine learning techniques. Targeted training, domain-specific fine-tuning, and structured supervision enable models to better distinguish between uncertainty and factual inference. When models are trained carefully, they learn not only to generate responses but also to recognize the limits of their knowledge, thereby improving reliability and reducing incorrect outputs.

Beyond quantitative metrics, a qualitative examination of model outputs reveals substantial differences in reasoning behavior, linguistic coherence, and response reliability across the evaluated Tiny-LLMs. Among the three models, Phi-3 Mini consistently demonstrates the strongest logical reasoning ability, producing well- structured, step-by-step explanations in Bengali. Its responses exhibit a high degree of grammatical correctness and contextual alignment, particularly in multi-step reasoning tasks such as mathematical word problems and deductive inference. Across the evaluated samples, Phi-3 Mini achieves an estimated reasoning success rate of approximately 78%, reinforcing its suitability for cognitively demanding applications [19].

Gemma-2B, while slightly less robust in deep reasoning tasks, delivers a balanced qualitative profile. Its responses are generally fluent and contextually appropriate, though occasionally less detailed in logical breakdowns compared to Phi-3 Mini. However, this trade-off is offset by its significantly lower inference latency, enabling faster interaction without severe degradation in response quality. As a result, Gemma-2B maintains a practical equilibrium between responsiveness and reasoning depth.

In contrast, TinyLlama-1.1B exhibits notable qualitative limitations. While capable of generating basic Bengali text, the model frequently produces incomplete logical chains, grammatical inconsistencies, and hallucinated information when confronted with multi- step or abstract reasoning tasks. These weaknesses suggest that, in its current form, TinyLlama struggles to maintain coherent reasoning under constrained parameter settings, particularly without domain-specific fine-tuning.

Discussion

The results of this study reveal a clear trade-off between reasoning accuracy and inference efficiency within the Tiny-LLM design space. As model scale and architectural sophistication increase, gains in reasoning consistency and linguistic coherence become more evident; however, these improvements are accompanied by higher inference latency.

Among the evaluated models, Phi-3 Mini emerges as the most reliable option for reasoning- intensive Bengali tasks. Its strong performance makes it particularly suitable for educational platforms, tutoring systems, analytical assistants, and research-oriented applications where accuracy and interpretability are critical. The observed behavior further suggests that carefully curated, high-quality training data can partially compensate for reduced parameter scale.

In contrast, Gemma-2B emphasizes inference speed, positioning it as an effective solution for real-time Bengali applications such as conversational agents, mobile assistants, and interactive user interfaces. Although it does not consistently match Phi-3 Mini in deep reasoning capability, its low latency significantly enhances usability in time-sensitive scenarios.

TinyLlama-1.1B, despite its exceptional compactness, illustrates the limitations imposed by extreme parameter reduction. While the model remains usable for simple generative tasks and ultra-low-resource deployments, its performance highlights the need for targeted fine-tuning or hybrid strategies to address hallucination and grammatical instability.

https://doi.org/10.61577/jmla.2025.100003

Overall, these findings demonstrate that no single Tiny-LLM optimally satisfies all performance requirements. Instead, effective model selection must be guided by application-specific priorities, balancing reasoning accuracy, responsiveness, and computational feasibility particularly when deploying AI systems for low-resource languages such as Bengali.

Conclusion and Future Work

This study demonstrates that quantized Tiny Language Models can function as effective and practical reasoning engines for the Bengali language, even when deployed on hardware with limited computational capacity. Through systematic evaluation, we show that architectural design choices play a decisive role in balancing reasoning accuracy, inference latency, and deployability. Among the evaluated models, Phi-3 Mini consistently delivers the highest reasoning accuracy and linguistic coherence, making it well-suited for applications that demand structured reasoning and interpretability. In contrast, Gemma-2B offers a compelling balance between performance and speed, positioning it as a strong candidate for real-time and interactive Bengali language systems. While TinyLlama-1.1B remains attractive for ultra-low-resource environments due to its minimal computational footprint, its limitations in reasoning consistency and contextual stability highlight the challenges of extreme model compression.

Beyond the immediate findings, this work opens several important directions for future research. A natural next step involves fine-tuning Tiny-LLMs on large-scale, instruction-oriented Bengali datasets to improve reasoning depth, grammatical stability, and instruction-following behaviour. Such targeted adaptation is likely to reduce hallucination rates and enhance model reliability in complex tasks. Additionally, deploying these models on mobile and edge devices using frameworks such as MLC-LLM would enable real-world validation of their performance under true on-device constraints. Finally, integrating Tiny-LLMs with retrieval-augmented generation (RAG) pipelines presents a promising pathway for improving factual accuracy and grounding, particularly for knowledge-intensive Bengali applications.

Taken together, the results of this study indicate that compact, well-optimized language models can meaningfully narrow the gap between advanced AI capabilities and low-resource linguistic communities. By continuing to refine training strategies, deployment frameworks, and hybrid architectures, future work can further advance the development of scalable, locally deployable Bengali AI Systems that are both accessible and trustworthy.

While these findings are promising, we acknowledge that further empirical validation across broader task settings is necessary.

References

1. Matarazzo A, Torlone R. A survey on large language models with some insights on their capabilities and limitations. arXiv preprint arXiv:2501.04040. 2025. https://doi.org/10.48550/arXiv.2501.04040
2. Shakhadri SA, KR DK, Aralimatti R. Shakti: a 2.5 billion parameter small language model optimized for edge ai and low-resource environments. InIFIP Int Conf Art Int Appls Inno 2025:434-447. https://doi.org/10.1007/978-3-031-96231-8_32
3. Lamaakal I, Maleh Y, El Makkaoui K, Ouahbi I, Pławiak P, Alfarraj O, et al. Tiny language models for automation and control: overview, potential applications, and future research directions. Sensors. 2025;25(5):1318. https://doi.org/10.3390/s25051318
4. Bhowmik S, Dipto TT, Islam MS, Hsu S, Reasat T. Evaluating LLMs' multilingual capabilities for bengali: benchmark creation and performance analysis. arXiv preprint arXiv:2507.23248. 2025. https://doi.org/10.48550/arXiv.2507.23248
5. Islam MR, Deng B, Dhar N, Nguyen TN, He S, Shi Y, et al. Characterizing and understanding energy footprint and efficiency of small language model on edges. In2025 IEEE 22nd Int Conf Mob Ad-Hoc Smart Syst 2025:363-371. https://doi.org/10.48550/arXiv.2511.11624
6. Team G. Google DeepMind. 2024. Gemma: open models based on Gemini research and technology. arXiv preprint arXiv:2403.08295.
7. Zahra K, Nisa ZU, Shahi NJ, Saeed MA, Saleem M, Jamil A, et al. From large to small language models for balanced performance and benchmarking. In2025 3rd Int Conf Art Intell Blockchain Internet Things 2025 1-8. https://doi.org/10.1109/AIBThings66987.2025.11296175
8. Abdin M, Aneja J, Behl H, Bubeck S, Eldan R, Gunasekar S, et al. Phi-4 technical report. arXiv preprint arXiv:2412.08905. 2024. https://doi.org/10.48550/arXiv.2412.08905
9. Zhang P, Zeng G, Wang T, Lu W. Tinyllama: an open-source small language model. arXiv preprint arXiv:2401.02385. 2024. https://doi.org/10.48550/arXiv.2401.02385
10. Luo M, Kumbhar S, Parmar M, Varshney N, Banerjee P, Aditya S, et al. Towards logiglue: a brief survey and a benchmark for analyzing logical reasoning capabilities of language models. arXiv preprint arXiv:2310.00836. 2023. https://doi.org/10.48550/arXiv.2310.00836
11. Aurpa TT, Rifat RK, Ahmed MS, Anwar MM, Ali AS. Reading comprehension-based question answering system in Bangla language with transformer-based learning. Heliyon. 2022;8(10). https://doi.org/10.1016/j.heliyon.2022.e11052
12. Prama TT, Islam MS. Evaluating credibility and political bias in llms for news outlets in Bangladesh. InProc 63rd Annu Meet Assoc Comput Linguist. 2025;4(665-677). https://doi.org/10.18653/v1/2025.acl-srw.42
13. Hazlewood V, Scott L. Performance and accuracy evaluation of an image classifier using multiple machine learning models and multiple gpus. InPract Exp Adv Res Comput 2024 Hum Powered Comput. 2024:1-8. https://doi.org/10.1145/3626203.3670528
14. Yoshida D. NF4 Isn't information theoretically optimal (and that's Good). arXiv preprint arXiv:2306.06965. 2023. https://doi.org/10.48550/arXiv.2306.06965
15. Anghel C, Anghel AA, Pecheanu E, Susnea I, Cocu A, Istrate A. Multi-model dialectical evaluation of llm reasoning chains: a structured framework with dual scoring agents. InInformatics 2025;12(3):76. https://doi.org/10.3390/informatics12030076
16. de Curtò J, de Zarzà I. Comparative analysis of reasoning capabilities in foundation models. In2024 2nd Int Conf Found Large Lang Models. 2024 Nov 26:141-149. https://doi.org/10.1109/FLLM63129.2024.10852449
17. Team C, Zhao H, Hui J, Howland J, Nguyen N, Zuo S, et al. Codegemma: open code models based on gemma. arXiv preprint arXiv:2406.11409. 2024. https://doi.org/10.48550/arXiv.2406.11409

https://doi.org/10.61577/jmla.2025.100003

18. Cho H, Padhy AP, Camacho F, Mukhopadhyay S. Sub 4-bit power-of-two based mixed-precision quantization for efficient llm compression and acceleration. IEEE Access. 2025. https://doi.org/10.1109/ACCESS.2025.3625771

19. Mondorf P, Plank B. Beyond accuracy: evaluating the reasoning behavior of large language models--a survey. arXiv preprint arXiv:2404.01869. 2024. https://doi.org/10.48550/arXiv.2404.01869

https://doi.org/10.61577/jmla.2025.100003