

Research on DeepSeek-Oriented Benchmarking Methods for AI Inference Accelerators

Hongfei Qiu *

China Telecom Corporation Limited Research Institute

Guangzhou, China

qiu hongf@chinatelecom.cn

Abstract—With the rapid shift of large model technology focus from training to inference, evaluating the performance of domestic AI accelerators in inference scenarios has become critical for building intelligent computing infrastructure. To tackle this, this paper proposes a multi-dimensional evaluation methodology for benchmarking AI accelerator inference performance, based on the DeepSeek-R1 series of models. Through quantitative analysis of how hardware parameters (computing power, memory bandwidth, interconnect bandwidth) and quantization schemes impact inference efficiency, this study provides empirical evidence to support the selection of domestic hardware options. Experimental results indicate that memory bandwidth is a primary bottleneck for large model decoding throughput, while INT8 quantization enables approximately 2x throughput gains compared to FP16 precision. This research offers valuable insights for hardware selection and performance optimization in intelligent computing centers, thereby enhancing the competitiveness of domestic AI accelerators in the inference market.

Keywords—LLM; Test; Inference; Accelerators; DeepSeek

I. INTRODUCTION

The explosive growth of generative AI has accelerated the migration of large model technologies from training-intensive to inference-intensive paradigms. Through architectural innovations such as MoE sparsification, MLA attention compression, and GRPO unsupervised optimization, the open-source model DeepSeek significantly reduces inference costs, triggering exponential growth in industrial demand. Projections indicate that the computing power ratio for AI training versus inference will reverse from 8:2 to 2:8, revealing a 10-fold incremental market potential for inference [1].

DeepSeek large models have been deployed at scale in fields such as cloud computing, communications, healthcare, education, and automotive, achieving deep integration with leading enterprises. Among them, the general-purpose model DeepSeek V3 adopts a Mixture-of-Experts (MoE) architecture, demonstrating outstanding performance in long-context comprehension and high-precision output scenarios; the specialized reasoning model DeepSeek R1 is optimized for mathematical deduction, programming, and logical reasoning tasks, supporting a long-chain reasoning mechanism (Chain-of-Thought). Its reasoning performance significantly surpasses that of V3 and international competitors like OpenAI o1 in benchmark tests such as AIME 2024 and MATH-500[2]. The lightweight models DeepSeek-R1-Distill-Llama-70B and DeepSeek-R1-Distill-Qwen-32B, generated through direct distillation fine-tuning techniques (with base architectures

derived from the Llama/Qwen open-source series), achieve performance comparable to OpenAI o1-mini in multiple benchmarks[2]. Driven by the domestic substitution strategy, domestic GPU manufacturers such as Huawei Ascend, Moore Threads, and Tianshu Zhixin have completed hardware adaptation for DeepSeek in inference scenarios and plan to extend this to the full fine-tuning and training pipeline. Currently, systematic evaluation standards are lacking to assess the adaptation effectiveness of DeepSeek models in software-hardware systems, including AI hardware chips, computing devices, and intelligent computing clusters. Therefore, establishing standardized AI accelerator card performance testing methods is of strategic urgency for intelligent computing infrastructure construction.

II. CHALLENGES IN AI ACCELERATOR CAPABILITY ASSESSMENT

AI accelerators are categorized into training chips and inference chips. Training chips focus on constructing and optimizing models through massive data, emphasizing high-precision computing and high performance. Inference chips are used for model prediction in practical applications, pursuing low latency and low power consumption. Regarding inference precision, typical inference tasks generally require only FP16, BF16, FP8, INT8, and INT4 to ensure inference accuracy. Currently, a single inference card can support the operation of small-scale models. Medium-scale models usually require single-node multi-card configurations, while large-scale models demand multi-node multi-card hardware setups; for example, 7B and 14B models adopt single-card inference, 32B and 70B models adopt single-node multi-card configurations, and 405B and 671B models adopt multi-node multi-card configurations. This study investigates the primary AI accelerator products of four manufacturers based on the configuration requirements of DeepSeek large models, evaluating computing power, memory capacity, memory bandwidth, and interconnect bandwidth, as shown in Table 1:

Table 1. Hardware Specifications of AI Accelerators

Manufacturer	FP16 (TFlops)	INT8 (Tops)	Memory bandwidth (GB/s)	Memory Capacity (GB)	Interconnect bandwidth (GB/s)
A	280	560	1.8TB/s	64GB/HBM2e	PCIe5.0/128 GB/s
B	140	280	1.4TB/s	64GB/HBM2	PCIe5.0/128 GB/s
C	129	258	840GB/s	48GB/GDDR6	PCIe5.0/128 GB/s
D	280	560	1.8TB/s	64GB/HBM2e	OAM/896GB/s

TTFT (Time to First Token), TPOT (Time Per Output Token), and decoding throughput constitute the three core metrics for evaluating large model inference performance, respectively indicating the system's initial response latency and continuous generation efficiency. However, current hardware selection and performance assessment face significant challenges: existing metrics struggle to authentically reflect the actual efficacy of inference accelerators, while architectural disparities across hardware platforms and the absence of standardized testing protocols result in ineffective cross-platform comparisons. These issues manifest concretely in the following four categories:

A. Disconnection between metrics and real-world scenarios

inflated utilization rates occur as FP16 computing power promoted by manufacturers is based on idealized benchmarks, whereas operational environments suffer from inefficient resource scheduling, yielding actual utilization rates consistently below 40% [3]. Simultaneously, the memory wall bottleneck intensifies as system constraints shift from a compute-bound to a memory-intensive paradigm; insufficient memory bandwidth causes computational units to idle due to data supply delays, exacerbating inefficiencies during high-concurrency long-sequence generation [4].

B. Blind spots in system-level evaluation

unquantified interconnect bandwidth disparities are evident since traditional inference benchmarks rely on single-node single-card setups, yet modules like NVLink provide 5× higher bandwidth than PCIe 5.0, substantially impacting distributed inference efficiency. Moreover, inadequate modeling of multi-card communication overhead leads to scalability prediction distortions—experimental data reveals only 68% efficiency at 8-card scaling versus theoretical projections [5].

C. Divergent framework adaptations

the vLLM framework leverages PagedAttention for dynamic memory management, achieving triple the throughput of traditional frameworks (e.g., KTransformers) in high-concurrency contexts, whereas KTransformers demonstrates superior speed during the pre-filling stage [6].

D. Fragmented benchmark standards

while MLPerf represents the most authoritative international AI benchmark adopted by numerous enterprises, its inference suite omits large language model long-sequence generation workloads; additionally, limited domestic participation restricts tested models to GPT/LLaMA families, largely excluding domestic models such as DeepSeek [7].

III. TESTING PLAN

A. Hardware environment

Fragmented benchmark standards: while MLPerf represents the most authoritative international AI benchmark adopted by numerous enterprises, its inference suite omits large language model long-sequence generation workloads; additionally, limited domestic participation restricts tested models to GPT/LLaMA families, largely excluding domestic models such as DeepSeek [6].

B. Model and data sets

The purpose of this test is to investigate the inference performance under single-node single-card and multi-node multi-card configurations, with the test models being DeepSeek-R1-671B, DeepSeek-R1-Distill Llama-70B, and DeepSeek-R1-Distill Qwen-32B. The accuracy evaluation tool used is OpenCompass, and the dataset selected is MMLU.

C. Equations Testing Indicators

Research indicates that text generation speed directly impacts user experience. In daily conversation scenarios, a speed of 10 to 20 tokens per second suffices for general users, meaning generating a 100-token response takes only 5 to 10 seconds, fully aligning with human conversational expectations. However, advanced users or enterprise applications requiring rapid responses may demand higher speeds. Key inference test metrics are as follows:

1) Time to First Token (TTFT, unit: ms): a critical metric for large language model inference latency, refers to the time required for the model to generate and return the first output token after receiving a user prompt. TTFT primarily reflects computational efficiency during the Prefill stage—a compute-intensive task centered on self-attention computation for input sequences. As computational complexity and memory usage in Prefill scale proportionally with input token length, longer sequences significantly increase TTFT. To ensure smooth user experience, TTFT is typically recommended to remain below 3 seconds. However, factors like input length, hardware limitations, and system load may extend TTFT to 3–5 seconds or higher[9].

2) Decoding Throughput (unit: tokens/s): measures the total output tokens generated per second by an inference server for all concurrent requests. This metric evaluates parallel processing capability and resource utilization efficiency of large language models, directly reflecting the performance of the Decoding Phase. As a memory-intensive task, decoding performance is constrained by memory bandwidth rather than computational power. Leveraging the Key-Value Cache generated during Prefill, it autoregressively produces output sequences. To maintain interactive fluency, sustained output rates must exceed 10–20 tokens/s, matching average human reading speeds (e.g., visual readers at ~700 words/minute ≈ 18.75 tokens/s).

3) Time Per Output Token (TPOT, unit: ms): denotes the average time for a model to generate a single output token per user request. Its value can be derived from decoding throughput: $\text{Decoding Throughput} = \text{Concurrency} / \text{TPOT}$. TPOT is a core parameter for assessing response fluency, indicating real-time generation efficiency during the Decoding Phase. Together with TTFT, it determines overall responsiveness: TPOT should ideally stay below 100 ms to align with human reading speeds[8].

4) Model Accuracy: OpenCompass is a widely adopted accuracy evaluation tool for domestic LLMs, operating in two modes: PPL (Perplexity) for base models and GEN (Generation)for chat-tuned models. The core dataset MMLU (Massive Multitask Language Understanding) covers 57 disciplines including mathematics, history, and computer

science. Its standardized testing requires models to answer multiple-choice questions under a 5-shot setting, establishing it as a benchmark for evaluating LLM knowledge breadth and reasoning. To balance inference efficiency and accuracy, quantization techniques like W4A16 or W8A16 are permitted. Model accuracy must exceed 95% on MMLU and equivalent benchmarks (using FP32 full-precision models as the baseline).

D. Inference Testing Process

- 1) Deploying the test environment by installing the target AI chip's software stack and drivers on the tested server;
- 2) Downloading the DeepSeek-R1 model and performing model quantization and conversion;
- 3) Utilizing GPTQ/AWQ tools to generate W4A16 and W8A16 quantized versions, meeting the training accelerator's support requirements for INT8/INT4 precision2;
- 4) Testing the model accuracy using the OpenCompass tool with the MMLU_gen dataset under a 5-shot configuration;
- 5) Configuring parameters in the script, including batch size, input sequence length of 3072 tokens, output sequence length of 1024 tokens, and mandating full inference completion for the specified token count even if encountering termination tokens—truncating inputs/outputs exceeding defined lengths;
- 6) Launching the inference script;
- 7) Collecting logs to parse TTFT , TPOT, and decoding throughput, with TPOT required to be under 3 seconds and TTFT under 100 milliseconds.

IV. TEST RESULTS ANALYSIS

Using the aforementioned inference performance testing methodology, inference tests were conducted on AI accelerator cards from four manufacturers using DeepSeek-R1-671B, DeepSeek-R1-Distill-Llama-70B, and DeepSeek-R1-Distill-Qwen-32B, with test results compiled as follows:

A. Correlation Analysis Between Hardware Computing Power and Inference Efficiency (Throughput, TTFT)

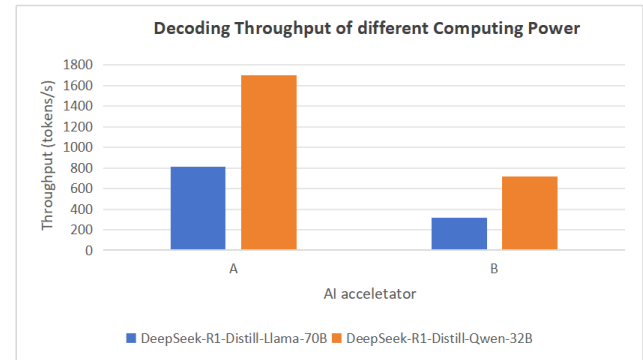


Figure 1: Decoding Throughput of different Computing Power

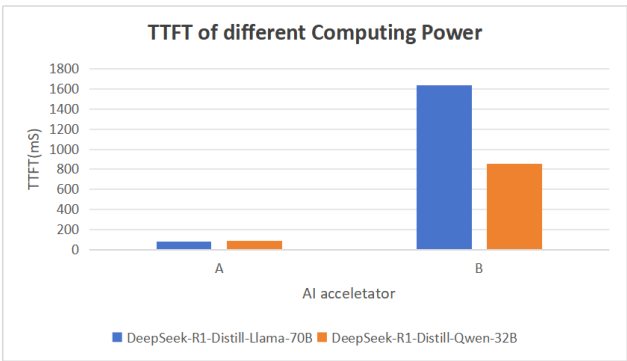


Figure 2: TTFT of different Computing Power

Figure 1 and Figure 2 demonstrate the decoding throughput and TTFT test results of the DeepSeek-R1-Distill-Llama-70B and DeepSeek-R1-Distill-Qwen-32B models, using a single-node 4-card configuration and FP16 inference precision. Experimental results show that Accelerator Card A achieves a 146% higher throughput compared to Accelerator Card B, and the TTFT of Accelerator Card A is significantly lower than that of Accelerator Card B. This indicates a significantly positive correlation between hardware computing power and both decoding throughput and TTFT, and enhancing FP16 computational capability can effectively optimize the inference efficiency of large language models.

B. Correlation Analysis Between Memory Bandwidth and Inference Efficiency

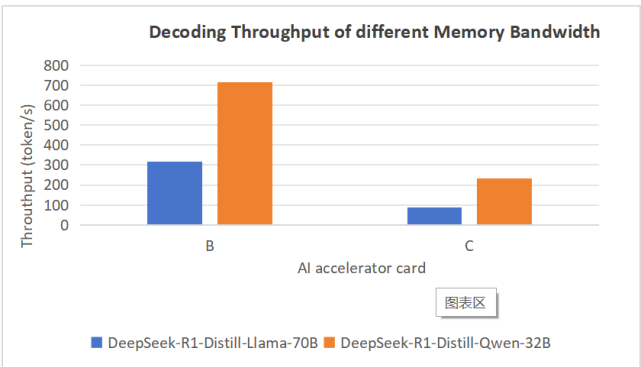


Figure 3: Decoding Throughput of different Memory Bandwidth

Figure 3 demonstrates the decoding throughput test results of the DeepSeek-R1-Distill-Llama-70B and DeepSeek-R1-Distill-Qwen-32B models, using a single-node 4-card configuration and FP16 inference precision. Accelerator Card B achieves a 237% higher throughput compared to Accelerator Card C, validating that memory bandwidth is the core bottleneck constraint during the decoding phase.

C. Correlation Analysis Between Interconnect Bandwidth and Inference Efficiency

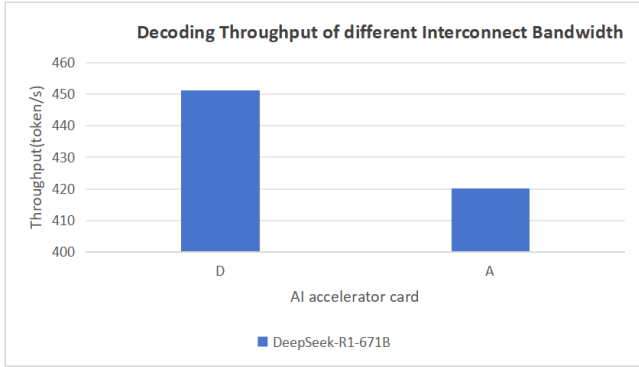


Figure 4: Decoding Throughput of different Interconnect Bandwidth

Figure 4 demonstrates the decoding throughput test results of the DeepSeek-R1-671B model using a dual-node 16-card configuration and INT8 inference precision. Accelerator Card D achieves approximately 10% higher throughput compared to Accelerator Card A. During the inference phase of AI models, the data volume transmitted between GPUs is relatively small. This indicates a positive correlation between interconnection bandwidth and decoding throughput, and enhancing interconnect bandwidth can effectively optimize the multi-card inference efficiency of large language models.

D. Correlation Analysis Between Inference Precision and Inference Efficiency

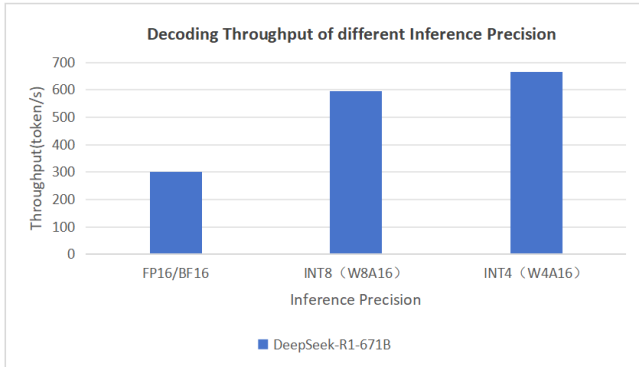


Figure 5: Decoding Throughput of different Inference Precision

Figure 5 demonstrates the decoding throughput test results of Accelerator Card A on the DeepSeek-R1-671B model, covering three experimental configurations: a 4-node 32-card cluster with FP16/BF16 inference precision, a 2-node 16-card cluster with INT8 (W8A16) quantization, and a single-node 8-card cluster with INT4 (W4A16) quantization. INT8 quantization achieves a 100% performance improvement compared to the FP16 baseline, while INT4 quantization shows no significant performance advantage over INT8. This is primarily because INT4 quantization mainly optimizes memory footprint and memory bandwidth, but actual computations are still executed at FP16 precision, and dequantization operations may introduce additional overhead [5]. Based on this analysis, INT8 quantization demonstrates a superior cost-performance

ratio, significantly reducing resource requirements while maintaining high inference throughput.

E. Correlation Analysis Between Reasoning Frameworks and Inference Efficiency

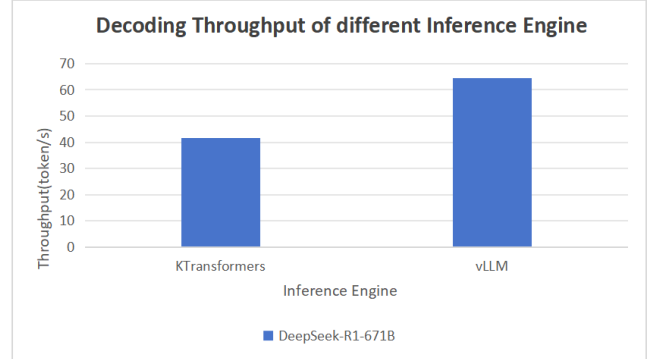


Figure 6: Decoding Throughput of different Inference Engine

Figure 6 demonstrates the decoding throughput test results of Accelerator Card A on the DeepSeek-R1-671B model, employing a single-node 8-card configuration with INT4 (W4A16) quantization inference precision. The test data reveal that the vLLM inference framework achieves a 155% improvement in throughput compared to KTransformers, validating the critical impact of inference framework selection on the efficiency of multi-card inference for large-scale models.

F. Scalability Analysis of Multi-Accelerator Card Systems

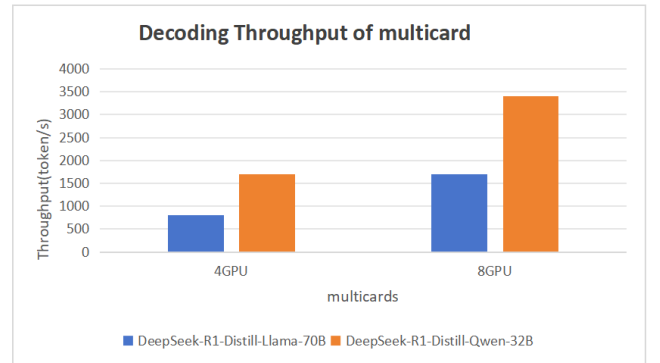


Figure 7: Decoding Throughput of multicard

Figure 7 demonstrates the scalability test results of Accelerator Card A on the DeepSeek-R1-Llama-70B and DeepSeek-R1-Qwen-32B models, employing a single-node 8-card configuration with FP16 inference precision. The test results indicate that when the number of GPUs scales from 4 to 8 cards, model throughput increases by 80%, demonstrating that the expansion of computational resources significantly enhances parallel efficiency. However, the speedup ratio exhibits sublinear growth, suggesting the presence of communication bottlenecks or load imbalance issues in the system. These results provide a configuration basis for distributed inference deployment, where an 8-card setup can be prioritized in throughput-first scenarios to maximize resource utilization.

G. Comparative Analysis of Model Accuracy Rates

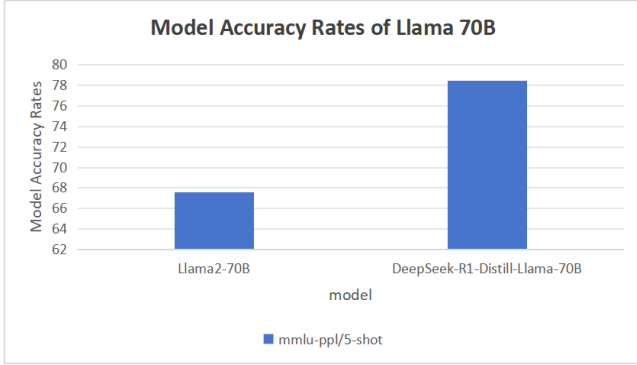


Figure 8: Model Accuracy Test Results

Figure 8 demonstrates the test results of model accuracy for DeepSeek-R1-Llama-70B and Llama2-70B. Testing model accuracy using the OpenCompass tool with the MMLU_gen dataset under a 5-shot configuration. The test data indicate that the model accuracy of DeepSeek-R1-Llama-70B is approximately 15% higher than that of Llama2-70B. This demonstrates that the DeepSeek-R1 distilled model achieves significantly improved accuracy compared to the original model.

V. CONCLUSIONS

The hardware specifications of AI accelerator cards, including computing power, memory bandwidth, and interconnect bandwidth, exert a decisive influence on the inference performance of large-scale models. Based on practical experience, this paper proposes the following systematic selection recommendations and optimization strategies.

A. Computing Power Allocation Strategy

The Prefill stage, serving as the critical initial phase for text generation, constitutes a computationally intensive task. During this stage, user-input prompts are tokenized into token sequences via attention mechanisms to generate KV caches, with its performance primarily determined by the accelerator card's computing power. Theoretical analysis and experiments demonstrate a negative correlation between computing power and TTFT, where higher computing power correlates with shorter TTFT. Therefore, in scenarios requiring optimized Prefill-stage response speed, priority should be given to high-computing-power accelerator cards to ensure TTFT remains within user-experience thresholds.

B. Memory Bandwidth Optimization

The Decoding stage incrementally generates target sequences via autoregression, with its performance constrained by memory bandwidth. Insufficient memory bandwidth causes computational units to idle due to data supply delays, resulting in reduced throughput. Experimental data reveal a significant negative correlation between memory bandwidth and TPOT [10]. Compared to traditional GDDR memory, High Bandwidth Memory (HBM) exhibits substantial advantages in AI computing tasks; its higher bandwidth effectively alleviates the memory wall issue, enhancing decoding efficiency.

C. Interconnect Bandwidth Bottlenecks

During inference, inter-GPU communication primarily encompasses KV caches, intermediate activations, and model parameters, with KV caches dominating communication volume (exceeding 70%) [5]. The bandwidth requirements for large-model inference depend mainly on model size, input/output data volume, and concurrent request quantities. When selecting OAM modules for inference, comprehensive consideration of hardware configuration, model size, and data transfer volume is essential. Sufficient bandwidth ensures smooth operation of model inference services and low-latency responses.

D. Quantization Technique Selection

Quantization technique choices necessitate balancing application scenarios, precision requirements, and hardware compatibility. INT8 quantization reduces memory consumption by 50% and decreases communication overhead, significantly improving throughput. While INT4 (W4A16) suits memory-constrained scenarios, its precision loss may degrade model performance by 15%-20% [6]. Note that some hardware architectures lack native INT4 support, necessitating thorough compatibility validation before deployment.

E. Inference Frameworks

Performance variance among inference frameworks can exceed 3× under identical hardware conditions. Taking vLLM as an example, its innovative PagedAttention technology achieves 300% decoding throughput improvement via dynamic memory management, proving particularly suitable for high-concurrency online services. Frameworks supporting optimization techniques (e.g., FlashAttention, Continuous Batching) should be prioritized based on business requirements [8].

F. Distributed Deployment

Experiments indicate that multi-card scaling exhibits sublinear speedup characteristics. In 70B/32B model testing, multi-card deployment achieves 100% inference performance improvement. Intelligent computing centers should adopt cost-performance-balanced configurations, optimizing resource utilization through Tensor Parallelism and Pipeline Parallelism strategies.

G. Evaluation Standard Establishment

An end-to-end evaluation system spanning inference stages is recommended: The Prefill stage should monitor computing power utilization and TTFT; the Decoding stage should track memory bandwidth occupancy and TPOT; multi-machine/multi-card scenarios require additional interconnect bandwidth efficiency metrics. For cases where hardware performance underperforms expectations, the DeepSeek open-source benchmarking tool is recommended to conduct standardized testing, effectively eliminating performance deviations caused by inadequate hardware-software co-optimization and enabling precise performance assessment.

VI. CONCLUDING REMARKS

To address the current lack of unified performance evaluation standards for AI accelerator cards, this paper

proposes a multi-dimensional evaluation system integrating TTFT, TPOT, and decoding throughput. Based on the DeepSeek-R1 series models, comprehensive inference tests are conducted on mainstream domestic AI accelerator cards, quantitatively validating the impact of hardware specifications (memory bandwidth/computing power/interconnect bandwidth), quantization techniques, and framework selection on inference efficiency, and revealing the scaling patterns of multi-card acceleration ratios. Future work will extend to the evaluation of multimodal large models and explore communication-computation co-optimization mechanisms to further enhance system efficiency.

REFERENCES

- [1] Sun B L. Structural Shift in Computing Demand for Large Models[J]. Chinese Journal of Computers, 2025, 48(3): 45-52.
- [2] DeepSeek-V3 Technical Report Team. Multitask Performance and Benchmark Analysis of DeepSeek-V3[J]. Journal of Artificial Intelligence, 2025, 38(2): 45-58.
- [3] Li M Z, et al. Precision-Efficiency Tradeoff in INT8 Quantization for Large Models[J]. Acta Automatica Sinica, 2025, 51(2): 301-310.
- [4] Zhao Y, et al. Impact of Memory Bandwidth on Inference Throughput in AI Chips[J]. Journal of Electronics & Information Technology, 2024, 46(5): 112-120.
- [5] Li X Y. Impact of Multi-Card Interconnect Bandwidth on Distributed Inference[J]. Journal on Communications, 2025, 46(2): 1-10.
- [6] Chen Q H. Computational Overhead of Low-Bit Quantization in LLM Inference[J]. Artificial Intelligence, 2024, 7(4): 88-95.
- [7] MLPerf Consortium. MLPerf Inference Benchmark v3.0 Technical Report[R]. 2024.
- [8] Wang S Y, et al. Performance Comparison of vLLM and KTransformers in Inference Frameworks[J]. Journal of Software, 2025, 36(1): 205-217.
- [9] Zhang W. Memory Bandwidth Optimization Model Based on TPOT[J]. Journal of Computer Research and Development, 2024, 61(9): 1897-1906.
- [10] Liu Z G, et al. Analysis and Optimization of Memory Bandwidth Bottlenecks in Large Language Model Inference[J]. Journal of Computer Research and Development, 2024, 61(6): 130-142.