Int. j. adv. multidisc. res. stud. 2026; 6(6):719-731

Received: 28-09-2026
Accepted: 08-11-2026

International Journal of Advanced Multidisciplinary Research and Studies

ISSN: 2583-049X

Evidence-Constrained Performance Evaluation of GPU-as-a-Service Platforms for AI Workloads: A Cross-Release Analysis of MLPerf Inference Benchmarks

[1] Sahal Ilyas Dadabhoy, [2] Sameer Abbasi
[1, 2] Detecon Al Saudia Co. Ltd. (Detasad), Riyadh, Saudi Arabia

Corresponding Author: **Sahal Ilyas Dadabhoy**

Abstract
The increasing use of GPU-as-a-Service (GPUaaS) platforms for AI inference has intensified the need for reliable performance evaluation. However, identical GPU models and accelerator counts do not necessarily establish whole-system comparability, potentially limiting the interpretation of published benchmark rankings. This study proposes an evidence-constrained framework to assess the technical comparability and ranking robustness of cloud-associated MLPerf Inference submissions. The dataset for the primary analysis was MLPerf v6.1 and the data set for cross-release methodological replications was MLPerf v6.0. Published throughput results have been matched based on workload, inference scenario, performance units, GPU model, and the number of accelerators. Candidate cross-provider comparisons were subsequently evaluated against eight reported hardware and software metadata fields. Progressive ablation, hypothetical ranking perturbations, scenario comparisons, and submitter-selection sensitivity analyses were performed. In v6.1, 77 eligible throughput records yielded 18 comparison groups and 34 candidate provider pairs. Only four pairs (11.8%), all involving Nebius–Oracle 72-GPU GB300 configurations, satisfied the eight-field matching criterion, with reported throughput differences ranging from 2.73% to 8.12%. Under an assumed ±1% adverse score perturbation, 13 of 18 initial group leaders retained their rankings. An external replication in v6.0 found 36 eligible records, and 4 candidate pairs, with no records that matched well enough for the absolute matching requirement. The findings illustrate that there is a significant practical compromise between the level of technical comparability and level of comparison coverage. The proposed framework will allow clear, evidence-based reading of published performance rankings for GPUs.

Keywords: GPU-as-a-Service, MLPerf Inference, GPU Cloud Computing, AI Inference, Performance Benchmarking, Technical Comparability

1. Introduction

Advancements in artificial intelligence (AI) workloads are spurring the need for scalable infrastructure for accelerators, with the growing number of inference requests for generative and reasoning models further intensifying the demand. When evaluating GPU-as-a-Service (GPUaaS) offerings, the nominal capacity of the circle of GPUs isn't enough to make a choice. Commercial platforms differ in accelerator topology, host processors, runtimes, provisioning conditions and service characteristics [1]. Those factors make it complicated to take the results of published work to represent the suitability of cloud services.

Modern inference systems show how much application performance depends on the computing environment. PagedAttention improves key–value-cache management in model serving [2], while ORCA, DistServe and related architectures exploit request scheduling and the different resource demands of prefill and decoding [3, 4]. Consequently, matching two submissions by GPU model and quantity alone does not make their complete systems equivalent. High throughput is partly due to accelerator hardware as well as optimised software, node topology or workload specific execution, all of which increases throughput.

Industry benchmarking provides more of a structured underpinning for assessment. The MLPerf Suite includes representative procedures to measure everything related to machine-learning systems [3] and machine-learning inference [5] scenarios and accuracy requirements to help compare heterogeneous inference implementations. Its Datacenter suite includes Offline and Server scenarios, which exercise different request patterns and quality-of-service constraints [6]. Nevertheless, benchmark rules allow substantial differences among participant systems. Even Closed Division submissions may use distinct processors,

frameworks, operating systems and deployment topologies [7], so the same benchmark score unit does not establish identical technical conditions.

Recent work has compared cloud GPU offerings, analysed MLPerf-based throughput and proxy energy efficiency, and constructed performance–energy–cost optimisation tools [1, 8, 9]. Such studies demonstrate the value of public benchmarks while also highlighting the difficulty of drawing simple purchasing recommendations from heterogeneous infrastructure. A distinct methodological question remains: how many apparent cross-provider throughput comparisons retain sufficiently similar reported technical characteristics once reproducible additional matching rules are applied? Moreover, do leaderboard positions remain stable under alternative assumptions, and can the same evidence-screening procedure be applied to another published benchmark round?

This investigation relies mainly on the results obtained for the MLPerf Inference v6.1 release (September 2026) and the results obtained for the MLPerf Inference v6.0 (April 2026), which only contains a separate release for replication [10-13]. Throughput observations are analysed per release only because their workload semantics can be different from each release and/or they can report differently. The study focuses on published submissions by a prespecified set of cloud-associated organisations; their reported results are not assumed to represent independently provisioned production services.

The objective of the research is to assess the comparability and the task-specific robustness of the submissions of the inference benchmarks related to the field of GPUaaS within an evidence constrained framework. Four research questions are used to analyze: (RQ1) what throughput differences emerge between those workloads, scenarios and GPU configurations that were submitted initially; (RQ2) what changes in eligible comparison coverage does progressive hardware/software matching induce; (RQ3) what stability does the set of published rankings have with respect to scenario and explicitly hypothetical score changes; and (RQ4) how does the framework behave in another release of MLPerf, and under alternative submitter populations? The study uses descriptive performance differences, eight reported system characteristics, progressive ablation, ranking-sensitivity checks and cross-release replication.

The contribution is a transparent method for separating an identifiable published score leader from the evidence supporting a strict comparison. The framework records insufficient evidence rather than issuing an unqualified recommendation when required metadata do not match. It quantifies the coverage–comparability trade-off and makes screening decisions reproducible. This is a methodological contribution to interpreting benchmarks, not an assertion that strict metadata equality ensures controlled infrastructure or that the framework predicts actual GPUaaS performance more accurately than existing techniques. These distinctions are essential when public rankings inform research or cloud procurement.

This distinction matters for both infrastructure research and practical interpretation. A vendor may publish several valid results for one accelerator family, but those results can describe different server footprints, precision modes and inference stacks. Where comparison groups are constructed from different submitters' best available scores, selecting the apparent winner risks conflating hardware capability with optimisation choices. Conversely, demanding identical descriptions on every possible system property may leave too little evidence for comparative claims. The appropriate task is therefore not simply to rank more systems, but to disclose which comparability assumptions are being made and what coverage is sacrificed when those assumptions are tightened [14-16].

A two-round evaluation is valuable for precisely this reason. The secondary MLPerf release is not treated as a hold-out sample for predictive model testing: the proposed framework contains no trained estimator or independently measured performance target. Instead, the release supplies a differently composed set of published submissions through which the reproducibility of evidence decisions can be examined. A strict match in one release and no strict match in another should not be presented as improvement or deterioration in provider technology. It indicates differences in what the two sets of available, submitted records can support. Framing the research questions around that documentary evidence makes the empirical conclusions testable without promising controlled measurements that the study did not conduct.

The resulting evidence hierarchy is deliberately conservative. Initial matching provides a transparent description of where throughput contrasts arise; subsequent field-level screening explains why many of those contrasts do not meet a stricter definition of reported similarity. Separate analyses of sensitivity and external replication then show the dependence of comparison coverage on score assumptions and submission composition. These outputs are interpreted as descriptive comparisons, not independent experimental replications.

2. Background and Related Work

2.1 GPU Infrastructure and AI Inference

GPU cloud platforms expose a heterogeneous mix of accelerator generations, host resources and distributed-node architectures. Even nominally similar services can differ in software libraries, kernel implementations, memory management and scheduling [1, 17]. The transformer-serving literature explains the importance of the surrounding system, for example by ORCA introducing iterative level scheduling [18], PagedAttention preventing key value cache fragmentation [2], or DistServe breaking the prefill and decoding phases to meet different serving constraints [4]. Splitwise also deals with phase separation [19] and Sarathi-Serve with the optimization of throughput–latency trade-off via scheduling [20]. These mechanisms can influence measured throughput without changes to nominal accelerator count.

The execution stack offers further sources of variation. FlashAttention and FlashAttention-2 optimise data movement and GPU attention parallelism [21, 22]; SGLang targets structured generation and cache reuse [23]. Punica considers multi-tenant LoRA serving [24], FlexGen examines constrained single-GPU inference [25], and system-level transformer scaling studies describe interactions among parallelism, memory and arithmetic intensity [26]. Quantisation methods such as AWQ alter weight representations and associated computational demands [27]. This literature establishes technical reasons to record runtimes and numerical representations when interpreting

GPU inference results; it does not establish that any particular software stack caused a difference in the present secondary data.

2.2 Benchmark Scenarios, Metrics and MLPerf

Inference workloads come in varying sizes and batches as well as have varying latency requirements. Throughput is about the work done in an aggregate, latency is about the experience to an individual request. These measures are not scalable or combined across different models or across different scores. MLPerf Inference standardises model definitions, quality targets and load-generation rules [3, 5]. Its Offline scenario emphasises bulk processing, whereas Server tests operate under defined request-arrival and latency conditions [6]. Thus a throughput comparison is meaningful only within the relevant workload and scenario, with reported units preserved.

MLPerf Closed and Open Divisions permit different levels of model and implementation flexibility. Closed Division rules improve workload-level comparability, yet submissions need not use identical CPUs, frameworks or multi-node topologies [7]. The April 2026 v6.0 and September 2026 v6.1 benchmark rounds document evolving workloads and participating systems [11, 12]. Their official repositories include summary performance and system descriptions [10, 13]. The current study uses those public records as a basis for documentary analysis rather than rerunning MLPerf or certifying submission compliance.

MLPerf result interpretation also depends on the type of comparison intended. Closed Division constrains model-level compliance, not every aspect of implementation [7]. A ratio of Offline token throughput is not a statement about token latency for an interactive application, and a Server score does not by itself reveal cloud billing or service-level availability. Published inference benchmarks are consequently best understood as system-level measurements under a common set of benchmark requirements. The present analysis deliberately excludes incompatible non-throughput outcomes from higher-is-better rankings and retains benchmark-model identities as part of its matching key. This procedural restriction avoids converting unlike scores into apparently comparable units.

2.3 Prior Research on Cloud and GPU Benchmarking

Ahuja *et al*. [1] compared major cloud GPU offerings across hardware, software services, pricing and practical operational considerations. Their contribution is a broad infrastructure perspective, but it does not directly quantify how technical metadata restrictions alter eligibility for individual published throughput comparisons. Beitsayadeh and Darbyshire [8] analysed 739 MLPerf v5.0 submissions to examine inference throughput and TDP-normalised efficiency across cloud categories. Large-sample analysis shows the utility of public records for systematic assessment but does not provide a direct assessment of the system power because the energy is not measured directly, nor is the specific assessment of matched pairs of submissions in the same cloud-type category.

Castaño *et al*. [9] applied enriched data from MLPerf v4.1/v5.1, predictive modelling, and Pareto optimisation for their study to achieve a balance between performance, estimated energy and indicative cost for different hardware configurations. The present study explores a narrower issue concerned with the documented contexts within which specific score comparisons can be made: their work provides conditions-aware decision support and tackles ranking trade-offs. It does not attempt to outperform their recommender or provide an alternative predictive cost model. Earlier system-level public-cloud evaluations [17] similarly motivate full-environment benchmarking rather than accelerator-only ranking.

Research on benchmarking methodology further cautions against treating unreplicated cloud measurements as interchangeable. Papadopoulos *et al*. [16] formalised methodological principles for reproducible cloud-performance evaluation. Laaber *et al*. [15] examined the difficulty of cloud microbenchmarking, and Gillam *et al*. [14] demonstrated substantial variation even among apparently similar virtual resources. More recent studies document timing-related and cross-cloud performance variability [28, 29]. Their controlled measurements differ from this paper's documentary analysis, but they motivate transparent configuration reporting, careful handling of missing information and restraint in assigning causal interpretations.

Methodologically, the studies differ in the types of evidence they produce. Ahuja *et al*. [1] review what a buyer may consider across offerings; Beitsayadeh and Darbyshire [9] investigate aggregate association in a much larger release; and Castaño *et al*. [9] use predictive models for constrained optimisation, with grouped validation designed to mitigate system-level leakage. This paper instead evaluates a rule-based screening process whose outputs are counts, eligibility labels and reported score contrasts. The methods are complementary rather than directly competing. An evidence-abstention decision can identify an unsupported strict contrast, but cannot automatically replace cost modelling, performance forecasting or independent execution trials. These limits help locate the contribution without implying that a simple eight-field rule solves the broader cloud procurement problem.

An additional distinction concerns the statistical unit. A single submitting organisation may contribute multiple system entries, and a benchmark group may contain several pairs sharing one submission. Such observations are useful for descriptive pairwise screening but are not independent draws from a cloud-provider population. An analysis that uses each pair as a separate independent replicate risks overstating precision. Here, each group is defined before the pairwise decomposition; an organisation's reported scores within a group are summarised by a median where necessary, and provider pairs are used to characterise documentation and score contrasts. Inferential claims require repeated execution or a sampling design that the published summary files do not supply.

Documentation quality should not be equated with system performance. Matching operating systems, processor characteristics, and accelerator topology provides a transparent and reproducible basis for assessing technical comparability; however, the performance implications of individual configuration differences may vary across inference workloads. Furthermore, systems satisfying all eight reported metadata criteria may still differ in unreported characteristics, including firmware versions, driver configurations, interconnect bandwidth, and optimisation settings. Accordingly, the eight-field matching procedure is treated as an operational measure of reported system comparability rather than evidence of complete infrastructure equivalence or causal control. The study

therefore explicitly defines the matching criteria, evaluates their individual effects through progressive ablation, and classifies the surviving comparisons as eight-field metadata-matched comparisons. This terminology ensures that the interpretation of performance differences remains consistent with the technical evidence available in the published benchmark records.

There is also a practical selection issue. Some organisations submit reference or laboratory systems, while commercial GPUaaS offerings are specified and priced through separate product catalogues. A submission attributed to a cloud-associated organisation cannot be interpreted automatically as an orderable instance with identical customer software and deployment guarantees. For this reason, the findings support evaluation of published system evidence; any extension to rental-cost superiority, provisioned-instance availability or customer latency would require direct linkage to a verified service configuration. That distinction protects the scope of the study while retaining the practical relevance of transparent benchmark comparison.

2.4 Comparability, Ranking Robustness and Research Gap

Benchmark rankings are useful summaries but can obscure distinctions between a high score, a comparably measured score and a deployable service recommendation. GPU count and model describe only part of the technical environment. Likewise, an Offline score need not imply identical relative performance under Server request patterns. Green AI highlights the value of resource-aware assessment [30], while hardware-selection research shows the relevance of workload-specific objectives [9]. Neither removes the requirement to establish what the published submission metadata actually support.

This study therefore operationalises technical comparability through eight host-system, software and topology fields, introduced after workload/GPU matching. The framework quantifies how many candidate pairs survive progressively stricter rules, whether reported leaders are sensitive to hypothetical score changes, and whether strict comparison evidence appears in another release. Its abstention rule prevents an unqualified strict ranking where evidence is insufficient. Unlike accuracy-prediction validation, this evaluation tests transparency, coverage and portability of a screening procedure. It cannot prove that retained comparisons are causally controlled or that rejected pairs have inaccurate results. This focused gap positions the work alongside, rather than above, prior MLPerf-based performance and efficiency studies.

The proposed gap is thus bounded and falsifiable. We test whether formally matching reported technical descriptors materially reduces cross-provider coverage, whether initial score rankings exhibit sensitivity under declared perturbations, and whether equivalent processing steps yield interpretable evidence decisions in another release. We do not claim that filtered pairs are experimentally controlled, that excluded pairs are invalid, or that a provider identified as higher-scoring has greater expected performance when rented commercially. This distinction between a published score and a decision supported by comparable reporting is the central analytical premise.

3. Research Methodology

3.1 Research Design

This study adopted a quantitative, secondary empirical research design to evaluate the technical comparability and ranking robustness of published GPU-as-a-Service (GPUaaS)-related AI inference benchmark submissions. The analysis was based on publicly available results from the MLPerf Inference benchmarking programme, which provides standardised workload definitions, performance measurements, and system-level submission information [5, 6]. Rather than conducting new benchmarking experiments on rented cloud infrastructure, the study independently processed and analysed published performance records using a reproducible Python-based analytical workflow.

The methodological approach consisted of two complementary components. First, MLPerf Inference v6.1 was used to investigate performance differences, technical comparability, and ranking robustness among selected cloud-associated GPU submissions. Second, the same conceptual screening and comparison framework was applied to MLPerf Inference v6.0 to examine its reproducibility and evidence coverage across a separate benchmark release. This cross-release analysis was designed to assess methodological portability rather than to establish predictive accuracy or infer changes in cloud-provider performance over time.

3.2 Data Sources and Acquisition

The primary dataset comprised official MLPerf Inference v6.1 results, released in September 2026 [12, 13]. The external replication dataset consisted of MLPerf Inference v6.0 results, released in April 2026 [10, 11]. Both datasets were obtained from the official MLCommons repositories.

The v6.1 dataset was distributed through a structured summary.csv file, whereas the v6.0 dataset was provided through summary_results.json. These differences required explicit schema harmonisation before the same analytical procedures could be applied across releases. The original source files were retained without modification, while cleaned analytical datasets were generated separately.

The initial source files contained 492 records for v6.1 and 520 records for v6.0. Data acquisition was performed through Python-based retrieval procedures, followed by source-integrity verification using SHA-256 cryptographic hashes. The workflow archived the original source files, retrieval information, source identifiers, and analytical outputs. This procedure enabled the reported results to be traced to the corresponding official data snapshots and distinguished the study's processed records from the original benchmark submissions.

3.3 Study Population, Sampling, and Eligibility Criteria

The broader study population consisted of organisations submitting results to the MLPerf Inference benchmark. The primary analytical sample was purposively restricted to seven cloud-associated submitting organisations: Azure, CoreWeave, Crusoe, Google, Lambda, Nebius, and Oracle. These organisations were selected to examine published performance evidence relevant to commercial GPU cloud infrastructure. Their inclusion did not imply that the selected submissions represented all available GPUaaS services or a

statistically representative sample of cloud platforms.

The analytical screening procedure applied the following eligibility conditions:

- The record belonged to the predefined cloud-associated submitter population.
- The submission was classified as Available, Closed Division, and Datacenter.
- A positive performance result and positive reported accelerator count were present.
- No non-zero error was recorded in the published summary.
- The accelerator was identifiable as a GPU through an explicit model-name classification rule.
- The performance result was directly reported rather than marked as inferred.
- The performance unit represented a throughput rate suitable for higher-is-better comparisons.

Following screening, the v6.1 dataset contained 78 eligible directly reported GPU scores, of which 77 represented throughput measurements. The corresponding v6.0 sample contained 36 eligible cloud-associated throughput records. The different sample sizes reflected release-specific submission availability and application of the predefined eligibility criteria.

3.4 Data Extraction, Harmonisation, and Quality Assessment

The principal extracted variables comprised submitting organisation, benchmark model, MLPerf model identifier, inference scenario, performance result, measurement unit, reported accuracy information, accelerator model, total accelerator count, host-system characteristics, software framework, model weight data types, and system topology. Supplementary variables included submission identifiers, reported compliance information, error indicators, and source-record locations.

For the v6.0 dataset, fields were mapped to the corresponding v6.1 analytical schema. For example, the source variables Submitter, Performance_Result, Total Accelerators, and Software were harmonised to organisation, performance result, total accelerator count, and inference framework, respectively. An explicit organisation-name normalisation was also applied where equivalent submitter labels differed between releases.

Missing values, unavailable metadata, and conflicting accelerator-count information were retained as audit findings rather than silently replaced. Textual metadata were normalised through case conversion and whitespace standardisation for matching purposes. In contrast, missing or placeholder values were classified as unknown and did not qualify as evidence of technical equivalence.

Reported accuracy values were processed separately from benchmark-matching identifiers. Specifically, the metric type was extracted from the published accuracy field, while the achieved numerical accuracy was preserved for documentary auditing. This distinction prevented small differences in achieved accuracy values from incorrectly splitting otherwise eligible benchmark comparison groups.

3.5 Construction of Initial Benchmark Comparison Groups

Initial cross-provider comparisons were constructed by grouping eligible records according to seven matching characteristics:

- Published workload or model identifier.
- MLPerf model identifier.
- Inference scenario.
- Performance measurement unit.
- Reported accuracy metric type.
- Accelerator model.
- Total accelerator count.

Where an organisation contributed more than one eligible score within the same matching group, the median reported score was used as its descriptive performance value. This aggregation reduced dependence on a single selected score within that organisation's reported records but did not convert the observations into independent experimental replications.

For each pair, the relative performance difference was calculated as:

$$\Delta_{AB} = (S_A / S_B - 1) \times 100$$

Where S_A and S_B denote the positive published throughput scores for organisations A and B, respectively.

3.6 Eight-Field Technical Comparability Assessment

To evaluate whether initially matched GPU benchmark submissions also shared additional reported system characteristics, the study implemented an eight-field technical comparability assessment.

Table 1: Technical metadata fields used in strict comparability assessment

Field	Operational relevance
Host processor model	Identifies the reported CPU architecture or processor family
Host processor core count	Captures reported host processing capacity
Host processors per node	Describes the processor arrangement within each node
Operating system	Identifies the reported operating environment
Inference framework	Captures the primary reported inference software framework
Model weight data types	Identifies the reported numerical representation of model weights
Number of nodes	Captures the scale of distributed system deployment
Accelerators per node	Describes GPU allocation across the node topology

For each provider pair, the corresponding fields were compared using normalised reported values. A field was classified as matching only when both records contained non-missing values that were equivalent after normalisation. Where both values were available but different, the field was classified as different. Where either value was missing or uninformative, the field was classified as missing.

For pairs involving multiple records from the same organisation and initial benchmark group, the record whose score was closest to that organisation's median was selected for the source-level metadata comparison, with source-row ordering used to resolve ties. This deterministic rule supported reproducibility, although the selected representative record could not capture all configuration variation associated with multiple submissions.

The strict matching procedure was intended to establish a transparent threshold of reported technical similarity. It was not assumed to guarantee complete infrastructure

equivalence, as other potentially influential characteristics, including system firmware, drivers, interconnect details, and runtime optimisation settings, could remain different or insufficiently documented.

3.7 Progressive Ablation and Comparison-Coverage Analysis

Progressive ablation analysis was used to determine how additional comparability requirements affected the number of eligible provider-pair comparisons.

The analysis began with the baseline set of workload- and GPU-matched comparisons. Technical metadata conditions were then introduced sequentially, allowing the effect of stricter screening to be quantified.

Table 2: Progressive technical comparability stages

Stage	Additional matching requirements
S0	Initial workload, scenario, performance-unit, accuracy-metric type, GPU model, and GPU-count matching
S1	Host processor model
S2	Inference framework
S3	Model weight data types
S4	Node-count and accelerator-per-node topology
S5	Full eight-field matching, additionally including CPU-core count, host processors per node, and operating system

Comparison coverage was defined as:

$$C_k = (N_k / N_0) \times 100$$

Where N_0 represents the number of baseline provider pairs and N_k represents the number remaining after screening stage k.

The interpretation of the ablation is conditional on field availability. Matching both of two missing values was prohibited, because absence of information is not affirmative evidence of equivalence. Nor were differences automatically treated as evidence that one system was better: they were used to indicate when the researcher's stricter comparison definition was not met. Each stage was applied as an explicit reproducible rule, and the counts remaining after each step were exported so readers can inspect the cumulative effect instead of accepting an unqualified final ranking.

3.8 Performance Ranking and Sensitivity Analysis

Ranking analysis evaluated the relative order of published throughput scores within each eligible benchmark group. Because different benchmark models and scenarios may employ different performance definitions, rankings were calculated within matching groups rather than across unrelated workload categories.

To investigate ranking sensitivity, hypothetical symmetrical score perturbations of ±0.5%, ±1%, ±2%, ±3%, ±5%, and ±10% were evaluated. For each assumed perturbation level, the highest score was decreased while the runner-up score was increased, representing an adverse scenario for the original leader.

The leader was classified as robust under perturbation level δ when:

$$S_{leader}(1 - \delta) > S_{runner\text{-}up}(1 + \delta)$$

These calculations were deterministic stress tests. They did not represent experimentally measured variability, sampling distributions, confidence intervals, or estimated probabilities of ranking reversal.

Scenario-related ranking consistency was evaluated separately by identifying provider pairs with comparable published observations in both Offline and Server scenarios. The relative ordering of the two providers was then examined across the scenarios. A ranking reversal was recorded when the higher-scoring provider differed between them.

The robustness analysis used a conservative two-sided perturbation, decreasing the published leader and increasing the runner-up by the same proportion. These calculations define deterministic sensitivity boundaries; they are neither a fitted probability model nor bootstrap confidence intervals.

3.9 Evidence-Constrained Decision Framework

The evidence-constrained framework integrated published performance comparisons with technical metadata screening to distinguish between numerically identifiable leaders and comparisons supported by the predefined strict evidence criterion.

A score-only baseline identified the organisation with the highest published throughput within each initial benchmark group. The evidence-constrained procedure then examined whether any pairwise comparison within that group satisfied the eight-field matching rule.

Where at least one such pair was available, the analysis reported the higher observed published score within that eligible comparison. Where no pair satisfied the strict condition, the framework recorded an evidence-abstention outcome rather than making an unqualified comparative recommendation.

Exploratory break-even price-ratio calculations were also available in the analytical workflow. However, because verified whole-configuration commercial rental prices were not established, these calculations were treated as hypothetical thresholds rather than observed cost-efficiency results. Actual monetary cost-performance comparisons were excluded from the study's primary empirical conclusions.

3.10 Cross-Release Replication and Provider-Selection Sensitivity

The analytical framework was applied separately to MLPerf Inference v6.0 to examine whether its eligibility rules, metadata assessment, and evidence-abstention procedures could be reproduced outside the primary dataset.

Cross-release analysis focused on methodological outcomes, including eligible record counts, baseline provider-pair coverage, five-field and eight-field matching rates, and the number of groups lacking strict comparable evidence.

Provider-selection sensitivity was examined using two approaches. First, each eligible cloud-associated submitter was removed in turn, and the remaining comparison-group and pair counts were recalculated. Second, the matching procedure was applied to the wider population of eligible published submitters.

External replication assessed within-release eligibility and evidence-based abstention, rather than changes in absolute throughput across benchmark rounds. The wider submitter population was examined only as a selection-sensitivity analysis because an MLPerf submitter does not necessarily offer an equivalent commercial GPUaaS service.

3.11 Reported Accuracy and Compliance Audit

The MLPerf Inference framework incorporates model-quality requirements and submission procedures intended to support valid benchmark comparisons [5, 7]. The present study performed a documentary audit of published accuracy and compliance information associated with eligible records. The analytical workflow extracted source identifiers, reported accuracy measurements, metric names, compliance labels, and error indicators. Quality values were parsed where their source format allowed numerical extraction, while the original published values were retained for traceability.

Reported accuracy information was not treated as evidence of independent validation by the authors of the present study. Reference-model accuracy tests, MLPerf LoadGen compliance procedures, and the underlying benchmark executions were not independently rerun.

3.12 Statistical Analysis, Reproducibility, and Validity Controls

All numerical analyses were implemented using Python, primarily through Pandas and NumPy, with Matplotlib used for graphical presentation. The final consolidated notebook integrated the primary v6.1 analysis and v6.0 external replication into a common reproducibility workflow.

The statistical strategy was primarily descriptive. Relative score differences, median published values, comparison counts, retention percentages, and ranking-consistency indicators were calculated directly from eligible records. Formal hypothesis-testing procedures were not applied because the available observations were published benchmark submissions rather than independent repeated measurements collected under a controlled sampling design. In particular, multiple provider-pair comparisons could share the same underlying hardware systems or benchmark submissions, making conventional assumptions of independent experimental observations inappropriate.

4. Results

This section reports the quantitative findings from the official MLPerf Inference v6.1 (primary) and v6.0 (external-replication) submissions [10, 13]. All reported percentages and comparisons were calculated from archived published-source records using the consolidated Python workflow. The results concern performance reported by submitting organisations; they do not constitute new controlled cloud benchmarks, inferential estimates of population-wide provider performance, or verified rental-cost comparisons.

4.1 Dataset screening and composition

The v6.1 source comprised 492 published summary records containing 33 distinct entries in the source `Organization` field. These field entries include joint or variant submitter labels and are not equivalent to the 30 participating organisations reported by MLCommons [12]. The prespecified cloud-associated sample contained 118 records from seven organisations. Sequential restrictions to Available, Closed Division, Datacenter records; positive and valid hardware/performance fields; zero recorded errors; GPU-labelled accelerators; and directly reported results left 78 eligible GPU scores. One of these used a non-throughput performance measure and was not included in higher-is-better throughput ranking, yielding 77 throughput records. Application of the separately harmonised v6.0 screening procedure yielded 36 eligible cloud-associated throughput records from 520 official source entries.

Table 3: Source screening and analytical sample by release

Indicator	MLPerf v6.1	MLPerf v6.0
Official source records	492	520
Prespecified cloud submitters represented after screening	7	3
Eligible direct GPU results, all metrics	78	—*
Eligible cloud throughput records	77	36
Groups with ≥2 cloud-associated submitters	18	4
Initial cross-provider pairs	34	4

Source: Processed official MLPerf records [13, 14]. *The v6.0 workflow exports throughput eligibility; an equivalent all-metrics direct-GPU count was not used in this cross-release table. Counts describe records, not independent trials.

In the primary release, the 18 candidate comparison groups were evenly divided between Offline and Server scenarios (nine each). Eight groups involved DeepSeek-R1, eight involved GPT-OSS-120B, and two involved Llama 2 70B (99.9 accuracy target). Ten groups contained two organisations and eight contained three, giving 34 distinct cross-provider score contrasts. Four accelerator families appeared in these groups: NVIDIA B200, B300, GB200 and GB300. Comparisons remained within matching model, scenario, metric, accelerator type and GPU-count conditions; scores were not aggregated across different benchmarks.

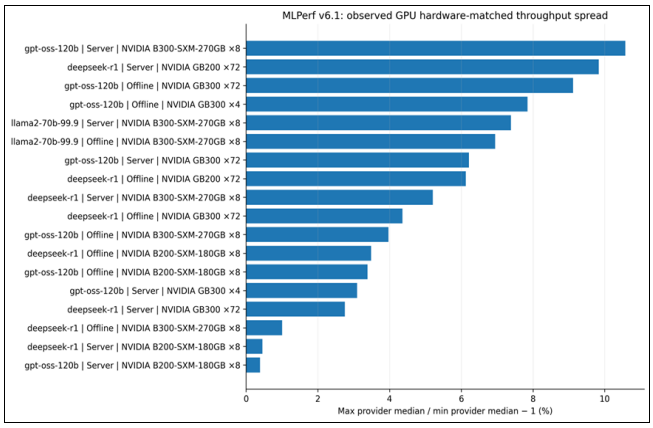


Fig 1: Observed spread in published throughput across the 18 initially GPU- and workload-matched MLPerf v6.1 groups. The spread is descriptive and is not a repeated-run uncertainty estimate

4.2 Baseline throughput contrasts and strict technical comparisons

Initial GPU-and-workload matching identified published throughput differences across cloud-associated organisations. For example, the v6.1 score-only comparison for DeepSeek-R1 on 72 NVIDIA GB200 GPUs under the Offline scenario placed Azure (542,075 tokens/s) above CoreWeave (529,991 tokens/s), a relative difference of 2.28%. This contrast reflects the published complete systems, not a controlled estimate of a provider-specific effect. Across the 18 candidate groups, the reported highest-to-lowest median-score spread varied with workload, scenario and configuration, illustrating why a universal leaderboard would be inappropriate.

After application of all eight system-metadata requirements, only four of the 34 baseline provider pairs remained eligible

(11.76%). These four records comprised the same Nebius–Oracle provider pair on reported 72 × NVIDIA GB300 configurations, evaluated using two AI models and two MLPerf scenarios (Table 5). In each case, Nebius had the higher published throughput, with an advantage ranging from 2.73% to 8.12% relative to Oracle. These are four model/scenario contrasts, not four independently replicated cloud-service experiments.

Table 4: Progressive metadata-screening ablation, MLPerf v6.1

Stage and additional evidence condition	Eligible pairs	Retained
S0 — Initial model/scenario/GPU match	34	100.0%
S1 — Host processor model	14	41.2%
S2 — + Inference framework	8	23.5%
S3 — + Weight data types	8	23.5%
S4 — + Node count and accelerators per node	8	23.5%
S5 — All eight fields (incl. CPU cores, CPUs per node, OS)	4	11.8%

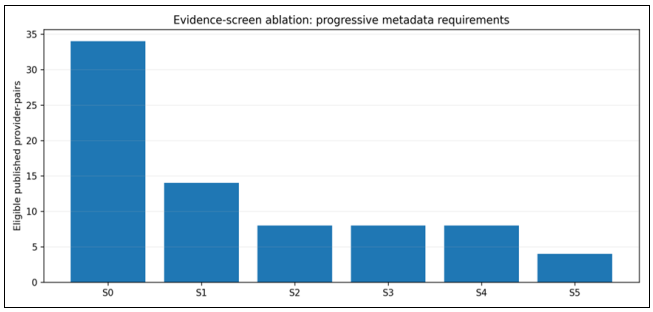


Fig 2: Progressive reduction in eligible provider-pair comparisons as reported system-metadata criteria were added, MLPerf v6.1

Table 5: Four eight-field metadata-matched published throughput comparisons, MLPerf v6.1

Model / scenario	Nebius (tokens/s)	Oracle (tokens/s)	Nebius advantage
DeepSeek-R1 / Offline	689,961	666,613	+3.50%
DeepSeek-R1 / Server	603,023	586,852	+2.76%
GPT-OSS-120B / Offline	1,186,750	1,097,620	+8.12%
GPT-OSS-120B / Server	1,122,490	1,092,610	+2.73%

All four records are 72 × NVIDIA GB300 benchmark submissions and use the same provider pair. Percentage = 100 × (Nebius throughput / Oracle throughput − 1); values rounded to two decimals.

The ablation revealed that software-framework and host-processor characteristics were particularly restrictive within this sample. Omitting the framework criterion from the full eight-field rule increased the eligible-pair count from four to ten; omitting host processor core count increased it to eight. Omitting each of the other six criteria individually did not increase coverage. These differences measure sensitivity to the selected reported-evidence rule, not the causal performance importance of any hardware or software feature.

4.3 Sensitivity of published rankings

The score-only leading submitter was examined against a series of explicitly hypothetical, adverse symmetric perturbations. Fourteen of 18 candidate groups retained their leading submitter when the original highest score was reduced by 0.5% and the runner-up increased by 0.5%; 13 retained the same leader under ±1%. At ±2%, only four remained unchanged, and no group retained its original leader under the ±5% or ±10% assumptions (Table 6). These are deterministic reversal-threshold calculations based on score separation, not confidence intervals or empirical estimates of runtime variation.

Table 6: Rank stability under hypothetical adverse score perturbations, MLPerf v6.1

Assumed change to each competing score	Unchanged leaders / 18	Percentage
±0.5%	14	77.8%
±1%	13	72.2%
±2%	4	22.2%
±3%	3	16.7%
±5%	0	0.0%
±10%	0	0.0%

Hypothetical worst-case perturbation: reported leader × (1 − δ), runner-up × (1 + δ). No sampling-based uncertainty is estimated.

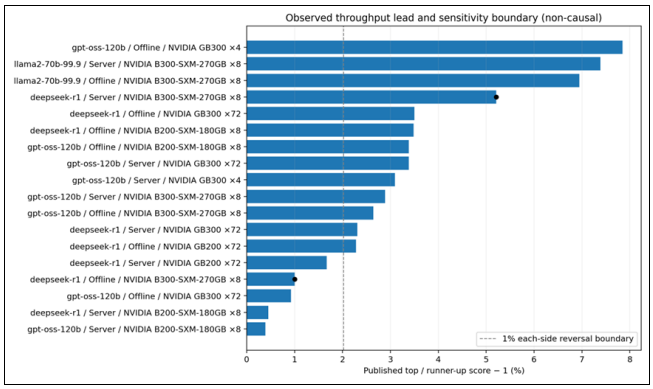


Fig 3: Observed score-leading margins relative to a hypothetical 1% adverse perturbation on each of the two top reported scores

Among 17 provider-pair combinations that were represented in both Offline and Server scenarios, 13 preserved the same directional ranking (76.5%) and four reversed (23.5%). The reversals involved Nebius–Oracle on 8 × B300 for DeepSeek-R1 and GPT-OSS-120B; Azure–Oracle on 72 × GB300 for DeepSeek-R1; and Lambda–Oracle on 4 × GB300 for GPT-OSS-120B. None of these four scenario-reversal pairs satisfied the strict eight-field screen, limiting causal interpretation of the scenario differences.

The sensitivity exercise illustrates why numerical margins and decision evidence should be reported separately. A narrow margin can be fragile to assumed score variation even when benchmark metadata match, while a large margin can arise between differently specified systems. The Offline–Server reversals were observed among comparisons with metadata differences; accordingly, their direction should not be attributed uniquely to the serving scenario. No confidence interval or formal significance test was generated from these published summary contrasts, which are not independent repetitions.

4.4 Cross-release external replication

The evidence-screening procedure was applied to the separately processed MLPerf v6.0 dataset. The 36 eligible cloud-associated throughput records represented three submitters and produced four benchmark groups containing two organisations, yielding four initially comparable provider pairs. Two pairs shared a reported host processor model, but none satisfied the extended software-framework and five-field criteria. Consequently, no v6.0 cloud-associated provider pair passed the eight-field threshold. In

contrast, four of 34 v6.1 pairs passed this requirement. All four v6.0 candidate groups, and 14 of the 18 v6.1 groups, therefore required evidence-based abstention under the strict criterion.

Table 7: Within-release evidence coverage and external replication outcomes

Measure	v6.1	v6.0
Eligible cloud throughput records	77	36
Submitting organisations in eligible records	7	3
Initial benchmark groups	18	4
Initial cross-provider pairs	34	4
Five-field matched pairs	8	0
Eight-field matched pairs	4	0
Groups with ≥1 strict pair	4	0
Groups without strict pair / abstaining	14	4

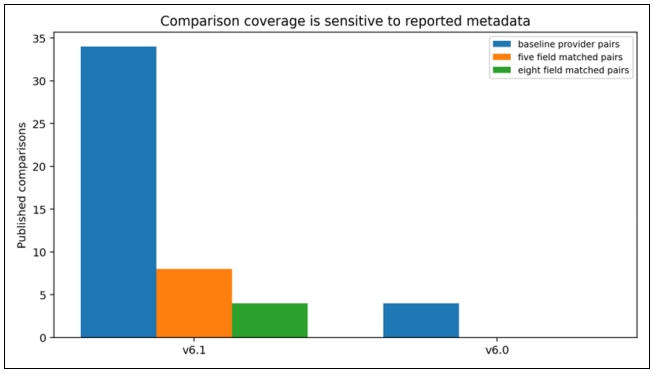


Fig 4: Baseline, five-field and eight-field matched-pair coverage for the v6.1 and v6.0 cloud-associated submitter samples

The contrast between releases shows that the same full strict-evidence rule can be applied to both archived sources while resulting in substantially different comparison availability. A zero count indicates that the available published entries do not satisfy the chosen metadata-equality rule; it does not imply equivalence, inferior performance or benchmark invalidity. The external exercise therefore supports procedural portability and reveals data-coverage limitations, but it is not an independent prediction-accuracy test.

4.5 Submitter selection and evidence coverage

Sensitivity to submitter selection was examined by omitting one included organisation at a time. In v6.1, removing CoreWeave, Nebius or Oracle reduced the 18 initial comparison groups to 12, while removing Lambda reduced the count to 16. Omitting Azure, Google or Crusoe left 18 candidate groups. The presence of all four strict pairs depended on both Nebius and Oracle; removing either eliminated the strict comparison set. In v6.0, omitting Nebius removed all four initial cloud-provider pairs, confirming that the small replication sample was strongly dependent on the participating organisations.

Table 8: Leave-one-submitter-out sensitivity of the published cloud-associated comparison sample

Release / excluded submitter	Pairs remaining	Strict pairs remaining
v6.1 — Azure	26	4
v6.1 — CoreWeave	16	4
v6.1 — Crusoe	34	4
v6.1 — Google	30	4
v6.1 — Lambda	32	4
v6.1 — Nebius	16	0
v6.1 — Oracle	16	0
v6.0 — CoreWeave	2	0
v6.0 — Google	2	0
v6.0 — Nebius	0	0

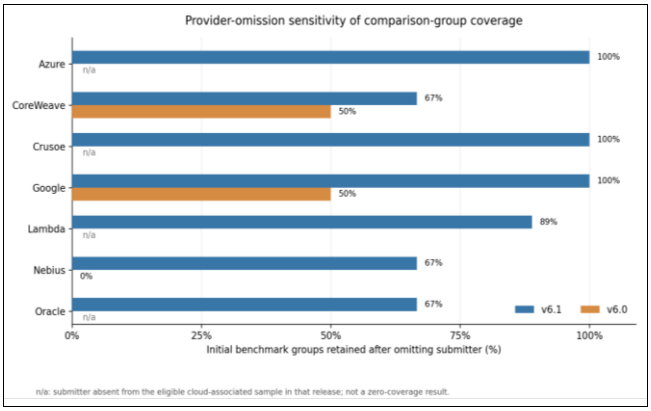


Fig 5: Percentage of initial benchmark groups retained after the omission of each eligible cloud-associated submitter in MLPerf v6.1 and v6.0. Unavailable categories are labelled n/a rather than shown as zero

When eligibility screening was expanded to all published submitters, irrespective of commercial GPUaaS status, the v6.1 records generated 433 initially matched pairs and four eight-field matches; v6.0 generated 308 pairs and two eight-field matches. These exploratory observations show that the sampling frame affects technical comparison coverage. They were not interpreted as an alternative ranking of commercial cloud services, because the wider set contains hardware vendors and other organisations outside the prespecified GPUaaS-related group.

4.6 Reported quality information and source integrity

The primary v6.1 quality audit successfully parsed published numeric quality information for all 78 eligible directly reported GPU records. Within the cross-release throughput samples, quality strings were present for all 77 v6.1 records and all 36 v6.0 records; all retained records also carried a published zero-error indicator. Source-system identifiers and reported quality values were preserved in the export, and both input files were associated with recorded SHA-256 hashes. Additional source-level checks flagged label/count inconsistencies in the broader v6.0 submissions,

which were retained as separate audit entries rather than silently imputed.

These checks establish documentary traceability and reproducibility of the secondary analysis, not independent confirmation that the MLPerf reference-accuracy tests or compliance workloads passed upon rerunning. No new GPUaaS execution measurements, independently observed run-to-run variance, whole-system rental prices or power measurements were collected. Accordingly, the findings support conditional descriptive comparisons and evidence-availability conclusions, rather than statistical significance, cost superiority or universal platform-selection claims.

4.7 Summary of principal findings

The principal empirical result is the marked difference between nominally GPU-matched benchmarking evidence and stricter reported system-metadata comparability. The v6.1 dataset yielded 34 candidate provider pairs, of which four (11.8%) satisfied all eight reported-field checks. The four surviving score contrasts were concentrated in the same Nebius–Oracle GB300 system family. Rank-perturbation and Offline/Server checks further showed that a single published score order may be sensitive to analytical assumptions and scenario selection; these findings do not constitute measured uncertainty or controlled comparisons. External replication on v6.0 yielded four candidate pairs but no eight-field matches, illustrating the operation of evidence-based abstention under more limited submission coverage. These findings support explicit evidence screening but provide limited grounds for generalising provider superiority.

The results also impose a constraint on interpreting evidence-qualified selections. The four strict-pair observations arise from a common provider pair and hardware family across two workloads and scenarios, rather than four independent cloud trials. Similarly, the all-submitter sensitivity results increase the universe of technical submissions, but those organisations do not automatically constitute comparable rented services. These dependencies narrow the inference that can reasonably be drawn from the apparent superiority of any one published score.

5. Discussion

5.1 Interpretation of published performance differences

The analysis demonstrates that a published inference-throughput contrast is not equivalent to a controlled comparison of commercial GPU services. In MLPerf v6.1, 34 cross-provider pairs satisfied the initial matching conditions for workload, scenario, accelerator model and count, yet only four also matched the eight selected technical metadata fields. The four stricter observations concerned Nebius and Oracle systems described as using 72 NVIDIA GB300 accelerators, across two model–scenario combinations. Nebius reported throughput advantages of 2.73–8.12% in those comparisons. These values are descriptive differences between submissions, not four independent replications or evidence that the same advantage would appear in customer-provisioned instances. The importance of whole-system context is consistent with the serving-systems literature: scheduling, attention kernels, phase allocation and inference runtimes can affect achieved performance even where accelerator families are shared [2, 4, 18-23].

5.2 Comparability requirements and the coverage trade-off

The most consequential result is the change in comparison coverage under progressively stricter evidence requirements. Retained pairs fell from 34 at baseline to 14 after the host-processor criterion, eight after the subsequent software-related screening, and four under all eight metadata conditions. Consequently, 88.2% of the initially eligible pairs did not meet the study's strict reporting rule. This reduction should not be described as an improvement in measurement accuracy: it quantifies what can be compared under an explicitly selected set of observable conditions. Studies of cloud benchmarking have repeatedly shown the importance of system heterogeneity, repeated measurement and documentation for interpreting performance differences [14-16, 28, 29]. Requiring more metadata trades breadth for reported technical similarity, but the eight fields do not cover every relevant driver or establish that the surviving systems are functionally identical. Missing fields must therefore trigger an evidence limitation, not an assumption of equality.

5.3 Scenario dependence and ranking robustness

The ranking results also caution against assigning a single provider-wide performance rank. Across 17 eligible Offline–Server provider-pair contrasts, four changed relative ordering; these reversals arose in pairs that were not fully matched on the selected metadata, so scenario effects cannot be isolated from configuration differences. Under an explicitly hypothetical ±1% adverse score perturbation, 13 of 18 initial comparison groups retained their leading submission; at wider perturbations, this retention declined. Such stress tests identify sensitivity to assumed score changes, not the probability of a reversal or a confidence interval. Research on real serving systems illustrates why workload scheduling and latency constraints can create different throughput trade-offs [4, 19, 20]. Interpretation should consequently remain scenario-specific, especially when service procurement decisions also depend on latency, reliability and operating cost that were not measured here.

5.4 Cross-release replication and evidence-based abstention

Application of the same conceptual framework to MLPerf v6.0 yielded 36 eligible cloud-associated throughput records, four initial provider pairs, and no five-field or eight-field matches. The external round thus reproduced the screening process but provided insufficient strict cloud-provider contrasts under its available submissions. An exploratory all-submitter sensitivity analysis recovered two strict pairs in v6.0, although these must not be reclassified as commercial GPUaaS comparisons merely to increase sample size. Together, these outcomes demonstrate portability of the evidence-screening and abstention procedures, not validation of a more accurate performance-prediction algorithm. Release-specific benchmark composition and participating organisations also prohibit interpreting differences in coverage as a chronological performance trend [4, 10-13, 19, 20].

5.5 Relationship to previous research and methodological contribution

Previous work has addressed cloud GPU offerings and practical procurement considerations [1], inference efficiency

based on public MLPerf records [8], and multi-objective hardware selection using performance, energy and cost descriptors [9]. This study asks a narrower question: which of the comparisons underlying a ranking possess sufficient reported technical similarity under transparent, reproducible rules? Its distinctive empirical outputs are the ablation profile, matched-pair coverage, hypothetical ranking-stability diagnostics and cross-release abstention patterns. These complement rather than supersede optimisation studies. In contrast to approaches that estimate efficiency or recommend configurations, the present framework has no verified rental-price evidence, externally measured prediction target or controlled head-to-head platform trials. Its contribution is therefore auditable evidence qualification, not demonstrated causal provider superiority or superior predictive accuracy.

5.6 Implications for researchers and infrastructure decision-makers

For researchers, the results support reporting the full benchmark identity, scoring scenario, accelerator topology and relevant host/software descriptors alongside published throughput; methodological decisions that exclude records should also be reproducible. For infrastructure teams, a high published score is best treated as a screening signal that requires confirmation under the organisation's actual model, concurrency, latency and budget constraints. Equal reported metadata can strengthen comparability but cannot substitute for independently repeated deployment-level tests. Future evaluation should combine controlled within-provider repetitions, uncertainty estimates, matched-instance pricing, regional availability and workload-specific latency measurements. This would test whether evidence-constrained screening improves decisions, a question that cannot be answered by the current secondary dataset alone [15, 16, 29].

More broadly, the framework is best seen as a quality-control layer for published comparisons. Its strongest feature is that it makes the basis for a claim inspectable, including how many candidate contrasts disappear when documented system similarity is required. Its weakest feature is the deliberate absence of an external test showing that these classifications improve actual purchasing decisions. Controlled replication using matched, provisioned GPU instances, measured run-to-run variability and verified whole-instance costs would be needed to test whether additional metadata matching changes predictive utility [15, 16, 28]. Until then, evidence abstention is an interpretive safeguard, not a validated optimisation outcome.

6. Limitations and Threats to Validity

6.1 Internal validity and causal interpretation

The results are based on published MLPerf submissions rather than independently controlled trials of commercial GPUaaS instances. Equal GPU model and accelerator count do not control for the host, software and topology differences evident in many candidate comparisons. Moreover, the four eight-field-matched v6.1 observations involve the same Nebius–Oracle 72-GPU GB300 configuration family across two workloads and scenarios, not four independent system replications. The observed score differences therefore cannot be attributed causally to provider identity.

6.2 Construct validity and comparability criteria

The eight-field rule operationalises similarity in selected reported metadata; it is not a test of complete infrastructure equivalence. Driver versions, firmware, interconnect settings, power limits and runtime optimisations may remain unmeasured. Missing descriptions were classified as insufficient evidence rather than agreement. Conversely, a recorded difference does not establish that the field caused the observed throughput contrast. Aggregating multiple eligible submissions by their median also compresses within-group variation. Throughput alone does not capture latency distributions, provisioning, availability or other service-level outcomes.

6.3 External validity and dataset selection

The purposive sample of seven cloud-associated submitters in v6.1 and three in v6.0 is not representative of all GPUaaS providers or customer-accessible configurations. Release-specific benchmark content and submitter participation further restrict comparisons over time. Replication on v6.0 demonstrated that the screening procedure could be applied to another round, but yielded no strictly matched cloud-provider pair. Thus, cross-release portability of the audit should not be confused with external validation of provider rankings or superior predictive accuracy. The broader all-submitter analysis was considered a selection-sensitivity check, not an expanded GPUaaS sample.

6.4 Measurement reliability and unobserved outcomes

The source-provenance checks confirmed the identity and consistency of archived input files, but benchmark accuracy and compliance procedures were not independently rerun. Published summary scores also did not provide a set of independently repeated, controlled measurements sufficient for sampling-based confidence intervals or statistical significance testing. The ±0.5%–10% ranking perturbations were explicitly hypothetical stress tests, not estimates of observed measurement uncertainty. Finally, independently verified whole-instance rental prices, billed operating costs and measured energy consumption were unavailable, preventing defensible cost- or energy-efficiency conclusions.

6.5 Implications for future validation

Future work should repeat identical workloads on independently provisioned, documented cloud instances across regions and time periods, measure throughput together with latency and variability, and incorporate verified billing and power data. Such trials could test whether evidence-constrained screening improves the reliability of infrastructure decisions compared with score-only selection.

7. Conclusion

This study developed a reproducible, evidence-constrained approach to interpreting published GPUaaS-related inference benchmarks using MLPerf v6.1 and an independent methodological replication on v6.0. The primary analysis identified 34 candidate cross-provider pairs in 18 benchmark groups, yet only four pairs satisfied equality across all eight selected reported system fields. These four observations were confined to one Nebius–Oracle GPU configuration family; they do not establish

general cloud-platform superiority. The external release contained four candidate pairs but no strict cloud-provider matches, demonstrating both the portability of the screening procedure and the limits of available comparable evidence.

Additional analyses showed that reported performance rankings depend on the conditions used to interpret them. Four of 17 available Offline–Server pair comparisons changed ranking direction, although differences in underlying configurations preclude causal attribution. Under a hypothetical ±1% adverse score perturbation, 13 of 18 primary benchmark groups retained their published leader. These diagnostics reveal the sensitivity of score-based comparisons without supplying experimental confidence intervals or predictive validation.

The central contribution is therefore not another universal GPUaaS leaderboard, but a transparent method for documenting comparison eligibility, quantifying the coverage lost through stricter metadata screening and abstaining when sufficient reported equivalence is unavailable. Public MLPerf submissions remain valuable for infrastructure assessment, provided that their scope and technical limitations are made explicit. Controlled cloud deployments and independently measured service-level outcomes are necessary before stronger procurement or provider-performance claims can be supported.

8. Acknowledgements
None.

9. Conflict of Interest
The authors declare there are no competing or financial interests.

10. Funding
None.

11. Data Availability
All of the data is available in the manuscript.

12. References

1. Ahuja SP, Golanakonda M, Reddivari S. A Comparative Study of Cloud GPU Offerings for AI/ML Engineers. International Journal of Cloud Applications and Computing (IJCAC). 2026; 16(1):1-15.
2. Kwon W, Li Z, Zhuang S, Sheng Y, Zheng L, Yu CH, *et al*. Efficient Memory Management for Large Language Model Serving with PagedAttention. Proceedings of the 29th Symposium on Operating Systems Principles; Koblenz, Germany: Association for Computing Machinery, 2023, 611-626.
3. Mattson P, Reddi VJ, Cheng C, Coleman C, Diamos G, Kanter D, *et al*. MLPerf: An Industry Standard Benchmark Suite for Machine Learning Performance. IEEE Micro. 2020; 40(2):8-16.
4. Zhong Y, Liu S-Y, Chen J, Hu J, Zhu Y-B, Liu X-Z, *et al*. DistServe: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving. USENIX Symposium on Operating Systems Design and Implementation, 2024.
5. Reddi VJ, Cheng C, Kanter D, Mattson P, Schmuelling G, Wu C-J, *et al*. MLPerf inference benchmark. Proceedings of the ACM/IEEE 47th Annual International Symposium on Computer Architecture. Virtual Event: IEEE Press, 2020, 446-459.
6. MLCommons. MLPerf Inference: Datacenter [official benchmark documentation]. Available from: https://mlcommons.org/benchmarks/inference-datacenter/
7. MLCommons. MLPerf Inference benchmark rules [official benchmark policy documentation]. Available from: https://github.com/mlcommons/inference_policies/blob/master/inference_rules.adoc
8. Beitsayadeh C, Darbyshire PE. Benchmarking AI Inference Efficiency in Public and Private Clouds: An MLPerf-Based Comparative Study. IEEE Transactions on Cloud Computing. 2026; 14(2):862-876.
9. Castaño J, Bustillo J, Franch X, Martínez-Fernández S. Data-driven multi-objective optimization of ML inference hardware configurations for energy, performance and cost. Sustainable Computing: Informatics and Systems. 2026; 50:101347.
10. MLCommons. MLPerf Inference v6.0: Official benchmark results repository [Available from: https://github.com/mlcommons/inference_results_v6.0
11. MLCommons. MLCommons releases new MLPerf Inference v6.0 benchmark results 2026 [Updated April 1, 2026. Available from: https://mlcommons.org/2026/04/mlperf-inference-v6-0-results/
12. MLCommons. MLCommons sets participation record with new MLPerf Inference v6.1 benchmark results 2026 [Updated September 16, 2026. Available from: https://mlcommons.org/2026/09/mlperf-inference-v6-1-results/
13. MLCommons. MLPerf Inference v6.1: Official benchmark results repository: GitHub, 2026 [Available from: https://github.com/mlcommons/inference_results_v6.1
14. Gillam L, Li B, O'Loughlin J, Tomar APS. Fair Benchmarking for Cloud Computing systems. Journal of Cloud Computing: Advances, Systems and Applications. 2013; 2(1):6.
15. Laaber C, Scheuner J, Leitner P. Software microbenchmarking in the cloud. How bad is it really? Empirical Softw Engg. 2019; 24(4):2469-2508.
16. Papadopoulos AV, Versluis L, Bauer A, Herbst N, Kistowski JV, Ali-Eldin A, *et al*. Methodological Principles for Reproducible Performance Evaluation in Cloud Computing. IEEE Transactions on Software Engineering. 2021; 47(8):1528-1543.
17. Ahuja S. System Level Benchmarking of Public Clouds. International Journal of Cloud Applications and Computing. 2022; 12:1-17.
18. Yu G-I, Jeong JS, Kim GW, Kim S, Chun B-G, editors. Orca: A Distributed Serving System for Transformer-Based Generative Models. USENIX Symposium on Operating Systems Design and Implementation, 2022.
19. Patel P, Choukse E, Zhang C, Shah A, Goiri Í, Maleki S, *et al*. Splitwise: Efficient Generative LLM Inference Using Phase Splitting. Proceedings of the 51st Annual International Symposium on Computer Architecture. Buenos Aires, Argentina: IEEE Press, 2025, 118-132.
20. Agrawal A, Kedia N, Panwar A, Mohan J, Kwatra N, Gulavani BS, *et al*. Taming throughput-latency tradeoff in LLM inference with sarathi-serve. Proceedings of the 18th USENIX Conference on Operating Systems Design and Implementation; Santa Clara, CA, USA: USENIX

Association, 2024, Article 7.
21. Dao T. Flashattention-2: Faster attention with better parallelism and work partitioning. International Conference on Learning Representations, 2024.
22. Dao T, Fu DY, Ermon S, Rudra A, Ré C. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. ArXiv, 2022. abs/2205.14135.
23. Zheng L, Yin L, Xie Z, Sun C, Huang J, Yu CH, *et al*. Sglang: Efficient execution of structured language model programs. Advances in neural information processing systems. 2024; 37:62557-62583.
24. Chen L, Ye Z, Wu Y, Zhuo D, Ceze L, Krishnamurthy A. Punica: Multi-tenant lora serving. Proceedings of Machine Learning and Systems. 2024; 6:1-13.
25. Sheng Y, Zheng L, Yuan B, Li Z, Ryabinin M, Chen B, *et al*. FlexGen: High-Throughput Generative Inference of Large Language Models with a Single GPU. In: Andreas K, Emma B, Kyunghyun C, Barbara E, Sivan S, Jonathan S, editors. Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research: PMLR, 2023, 31094-31116.
26. Pope R, Douglas S, Chowdhery A, Devlin J, Bradbury J, Heek J, *et al*. Efficiently scaling transformer inference. Proceedings of machine learning and systems. 2023; 5:606-624.
27. Lin J, Tang J, Tang H, Yang S, Chen W-M, Wang W-C, *et al*. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. Proceedings of machine learning and systems. 2024; 6:87-100.
28. Ahuja S, Lopez Chalacan VH, Resendez H. Performance Variability in Public Clouds: An Empirical Assessment. Information. 2025; 16(5):402.
29. Henning S, Vogel A, Pérez-Wohlfeil E, Ertl O, Rabiser R. When Should I Run My Application Benchmark? Studying Cloud Performance Variability for the Case of Stream Processing Applications, 2025, 400-410.
30. Schwartz R, Dodge J, Smith NA, Etzioni O. Green AI. Commun ACM. 2020; 63(12):54-63.