QATTRIB-MOE: RUNTIME-CONTROLLED ATTRIBUTION OF CROSS-PRECISION TRAJECTORY DIVERGENCE IN MIXTURE-OF-EXPERTS INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-precision mixture-of-experts inference is usually compared by endpoint accuracy, yet similar accuracy can conceal substantial changes in the autoregressive trajectory, while runtime variability can confound attribution to numerical precision. We introduce **QAttrib-MoE**, a pre-frozen, runtime-controlled, noise-floor-aware protocol for attributing cross-precision trajectory divergence. Same-path fresh-server variability defines an empirical control floor, cross-path divergence is the observed contrast, and excess divergence measures the contrast remaining above the two path-specific floors. The protocol freezes the evaluation cohort and decision rule before generation, counterbalances execution order, and reconstructs multiple decoding budgets by censoring a single stored trajectory rather than regenerating requests. We instantiate QAttrib-MoE on a shared GLM-5.2 checkpoint executed through FP8 and online-NVFP4 routed-expert paths. At the pre-frozen 256-token prefix, both paths exhibit 0.00%/0.00% within-path mismatch on 64 controls, while cross-path mismatch and excess divergence reach 98.44% (95% CI: 95.31–100.00%); the effect is positive on both GSM8K and MMLU-Pro and grows from 50.00% at 32 tokens to 98.44% at 256 tokens. In contrast, strict 1024-token accuracy is 64.45% versus 66.41% (difference 1.95 pp, 95% CI: -1.95 to 5.86 pp; McNemar $p = 0.442$). A gated teacher-forced diagnostic preserves token-ID alignment yet finds top-1 prediction flips in all 48 anchored inputs, with median first flip position 2. The results show that individually reproducible precision paths can be strongly non-equivalent at the reasoning-trajectory level without a statistically significant endpoint-accuracy shift in the evaluated setting.

1 INTRODUCTION

Quantization is usually validated through memory, throughput, latency, and endpoint accuracy. For reasoning models, however, endpoint agreement does not imply process equivalence: small numerical changes can alter token choices and persistent autoregressive state while leaving aggregate accuracy almost unchanged. Prior work has shown low-bit reasoning degradation, token inflation, precision-sensitive behavioral disagreement, FP4 component sensitivity, kernel-dependent divergence, and numerical-state propagation in sparse mixture-of-experts (MoE) inference (Li et al., 2025; Lian et al., 2026; Wang et al., 2026; Cim et al., 2026; Chen, 2026; Zhu, 2026). At the same time, cache state and replay/runtime effects can create their own trajectory variability (Patodiya, 2026; Gonuguntla, 2026). Therefore, observing different outputs under two precision configurations is not by itself sufficient evidence that the precision path caused the difference.

We ask a narrower question: *when does changing the precision execution path produce a trajectory difference that exceeds the variability of either path by itself?* QAttrib-MoE treats same-path fresh-server variability as an empirical control floor rather than assuming the serving stack is perfectly repeatable. Path B executes the fixed GLM-5.2 FP8 source path; Path C uses the same checkpoint with online NVFP4 quantization for MoE expert weights. Cross-path divergence is therefore not sufficient by itself: the central attribution statistic is the *excess* remaining after subtracting the two path-specific control floors. Cohorts, thresholds, prefix checkpoints, model/image identities, and execution order are frozen before generation so that attribution is not tuned after observing the effect.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Our contributions are deliberately scoped to this cross-precision MoE setting. First, we formulate a pre-frozen attribution contract whose core statistic is excess cross-path divergence, with task-stratified uncertainty and a within-prompt path-label randomization test. This is more than detecting that FP8 and NVFP4 produce different strings: evidence is discounted whenever either path is itself unstable. Second, on 64 held-out controls, both paths show 0.00%/0.00% within-path mismatch while excess B–C divergence reaches 98.44% (95% CI 95.31–100.00%) and replicates across GSM8K and MMLU-Pro. Third, endpoint accuracy does not mirror trajectory divergence: the 1024-token accuracy difference is only 1.95 pp and its confidence interval crosses zero. Finally, a post-confirmatory teacher-forced diagnostic shows precision-path-dependent next-token rankings under identical anchored contexts. We do not claim that matched control floors are new in general or that NVFP4 universally induces divergence; the novelty is the frozen cross-precision MoE attribution contract and its process-versus-endpoint evidence in the evaluated setting.

2 Related Work

Quantized reasoning and FP4 sensitivity. Low-bit reasoning studies have localized early error cascades and shown that quantization can increase reasoning-token usage even when endpoint accuracy is preserved (Li et al., 2025; Lian et al., 2026). For FP4 specifically, Cim et al. (2026) isolate layer- and block-level sensitivity of NVFP4 and MXFP4, finding strong sensitivity in MLP projections. These studies characterize degradation or component sensitivity. Our question is instead whether two deployed precision execution paths are trajectory-equivalent after accounting for each path's own runtime variability.

Cross-precision disagreement. PrecisionDiff generates inputs that expose behavioral disagreements across numerical precisions (Wang et al., 2026). It demonstrates that precision-induced output differences can be security-relevant and widespread. We complement this differential-testing perspective with a fixed held-out cohort, pre-frozen decision rules, fresh-server same-path controls, and trajectory-depth analysis. The goal is attribution and measurement of process non-equivalence rather than adversarial input generation.

Kernel, MoE, cache, and replay divergence. Cross-kernel INT8 experiments show that two individually reproducible kernel arms can diverge end-to-end, with teacher-forced replay localizing token-level flips (Chen, 2026). Controlled sparse-MoE reduction changes can propagate numerical state into different routes and continuations (Zhu, 2026). Cache state can also create reproducible-but-different serving paths, especially under quantization (Patodiya, 2026). Matched same-model controls have been used to separate intervention effects from replay/runtime variability in agentic model switching (Gonuguntla, 2026). These results motivate our cache-off, fresh-server, counterbalanced design and constrain our novelty claim: our contribution is the pre-frozen cross-precision MoE attribution protocol and the observed separation between trajectory divergence and endpoint accuracy, not a general "noise-floor" principle.

3 Method

3.1 Precision paths and trajectory divergence

For input x, precision path $p \in \{B, C\}$, and fresh-server replica r, let

$$T_{p,r}(x) = (y_1^{(p,r)}, \ldots, y_{L_{p,r}}^{(p,r)}). \tag{1}$$

For prefix length k, define a trajectory mismatch indicator

$$d_k(T, T') = \mathbf{1}[T_{1:k} \neq T'_{1:k}], \tag{2}$$

with truncation at the available sequence length. With R fresh-server replicas per path,

$$W_B(x,k) = \binom{R}{2}^{-1} \sum_{r<s} d_k(T_{B,r}, T_{B,s}), \tag{3}$$

$$W_C(x,k) = \binom{R}{2}^{-1} \sum_{r<s} d_k(T_{C,r}, T_{C,s}), \tag{4}$$

$$X(x,k) = R^{-2} \sum_{r=1}^{R} \sum_{s=1}^{R} d_k(T_{B,r}, T_{C,s}). \tag{5}$$

We define excess cross-path divergence as

$$E(x,k) = X(x,k) - \tfrac{1}{2}\left(W_B(x,k) + W_C(x,k)\right). \tag{6}$$

For a control cohort $\mathcal{C}$ of size N_c, the confirmatory statistic is

$$\bar{E}(k) = \frac{1}{N_c} \sum_{x \in \mathcal{C}} E(x,k). \tag{7}$$

Raw cross-path mismatch X alone cannot distinguish a precision-path effect from a serving stack that changes across fresh starts. W_B and W_C are path-specific empirical control floors; subtracting their mean makes evidence weaker whenever either path is internally unstable. Thus $\bar{E}(k)$ is the quantity on which the attribution decision is based. It is an operational attribution statistic under the matched, frozen protocol, not a claim of causal identification against arbitrary unmeasured confounders. We also report a stable-cross indicator, equal to one when all B replicas agree, all C replicas agree, and the two consensus prefixes differ. This separates "the two paths differ" from "one path is internally unstable."

3.2 Pre-frozen confirmatory protocol

Before the first model generation, the protocol locks the package digest, held-out/control hashes, two counterbalanced shards, dataset manifest, model/image identities, seed, token budgets, prefix checkpoints, decision thresholds, tensor-parallel degree, and maximum generation length. Historical pilot/dev/control prompts are excluded before deterministic sampling, followed by an independent overlap audit. Primary trajectories run shard 0 as B→C and shard 1 as C→B. Controls use four fresh-server replicas per path in BC/CB/BC/CB blocks, reversing request order in even blocks. Radix caching and CUDA graphs are disabled. Each completed request is fsync-persisted; recovery fills only missing request IDs after a server restart.

We use task-stratified bootstrap with 10,000 resamples for 95% intervals. A balanced path-label randomization test repeatedly repartitions the eight control trajectories per prompt into four-vs-four pseudo-paths and compares their excess statistic with the observed value (20,000 draws). Under the null that the path labels are exchangeable within a prompt, this directly asks whether the observed separation is unusual relative to pseudo-path partitions of the same trajectory pool. The frozen GO contract requires aggregate within-path mismatch ≤ 0.10 for both paths, stable-cross rate ≥ 0.25, $\bar{E}(256) \geq 0.20$, a strictly positive lower bootstrap bound, and positive task-level excess on both task families.

3.3 Endpoint and teacher-forced diagnostics

Primary generation stores one trajectory up to 1024 new tokens. Budgets 128/256/512/1024 are reconstructed by censoring the same trajectory, avoiding independent re-sampling. Answers are scored only when the required explicit final-answer format is present. Accuracy differences use a paired task-stratified bootstrap and exact McNemar tests.

After the confirmatory GO decision, we select 24 task-balanced, same-path-stable/cross-path-divergent prompts and construct B-anchor and C-anchor token sequences, yielding 48 teacher-forced inputs. Identical anchors are evaluated under both precision paths. At scored position t,

$$\Delta\ell_t = \left|\log P_B(y_t \mid h_t) - \log P_C(y_t \mid h_t)\right|. \tag{8}$$

We also record whether the next-token top-1 predictions differ. Missing observed-token logprobs are treated as unavailable rather than zero; token-ID alignment is checked independently.

Table 1: Prespecified design choices and the specific validity concern each addresses.

Choice	Rationale
4 replicas per path	Gives six within-path pairs per path and 16 cross-path pairs per control, while enabling balanced 4-vs-4 label randomization. It is intended to detect routine fresh-server instability, not to estimate extreme-tail nondeterminism.
64 controls (32/task)	Provides a task-balanced cohort dedicated to the runtime control floor while bounding the cost of eight fresh-server trajectories per control. Prompt-level task-stratified inference avoids treating pairwise comparisons as independent samples.
256-token primary prefix	Fixed before generation as the single confirmatory checkpoint; 32/64/128 are prespecified sensitivity checkpoints, preventing post-hoc selection of a favorable prefix.
B→C / C→B	Counterbalances precision path with wall-clock/order effects; control blocks additionally alternate BC/CB and reverse request order in even blocks.
20,000 label draws	Provides a nonparametric within-prompt null aligned with the same excess statistic used for the confirmatory decision.

Table 2: Confirmatory attribution at the pre-frozen 256-token prefix.

Task	W_B	W_C	Cross	Excess
GSM8K	0.00%	0.00%	100.00%	100.00%
MMLU-Pro	0.00%	0.00%	96.88%	96.88%
Combined	0.00%	0.00%	98.44%	98.44%

4 Experimental Setup

We run on eight NVIDIA B200 GPUs (183,359 MiB each), driver 595.71.05, tensor parallelism 8, and SGLang (Zheng et al., 2023). The host reports CUDA 13.2; the serving image contains CUDA 13.0.1. The fixed GLM-5.2 FP8 checkpoint is loaded from 141 safetensors shards. Path C applies online NVFP4 quantization to MoE expert weights while Path B uses the source FP8 path. Both use temperature zero, identical prompts, maximum generation length 1024, cache-off serving, and the same eight-GPU placement. The confirmatory cohort contains 256 examples split evenly between GSM8K and MMLU-Pro; the control set contains 64 examples, 32 per task, with 4 fresh-server replicas per path. The exact image/model digests, cohort hashes, execution ledger, and recovery logs are included in the supplementary artifact.

5 Results

5.1 Cross-precision divergence greatly exceeds same-path variability

Across 64 controls, within-path mismatch is zero for both paths, whereas cross-path mismatch and excess divergence are 98.44% (Table 2). The task-stratified 95% bootstrap interval for excess is 95.31–100.00%, stable-cross rate is 98.44%, and the path-label randomization test gives $p \approx 5 \times 10^{-5}$. The effect replicates across tasks: 100.00% excess on GSM8K and 96.88% on MMLU-Pro (95% CI 90.62–100.00%).

Figure 1 shows that excess divergence rises from 50.00% at 32 tokens to 98.44% at 256 tokens. The result is therefore not an artifact of a single hand-picked prefix. The normalized longest-common-prefix fraction simultaneously falls from 0.631 at 32 tokens to 0.193 at 256 tokens, consistent with early differences being amplified by autoregressive continuation. This trend does not, by itself, identify a unique internal causal operator.

5.2 Trajectory divergence does not imply a significant accuracy gap

At 1024 tokens, strict accuracy is 64.45% for B and 66.41% for C, a difference of 1.95 percentage points with 95% CI $[-1.95, 5.86]$ pp and McNemar $p = 0.442$. The same pattern holds at shorter budgets (Figure 2): no evaluated budget supports a statistically significant accuracy advantage for

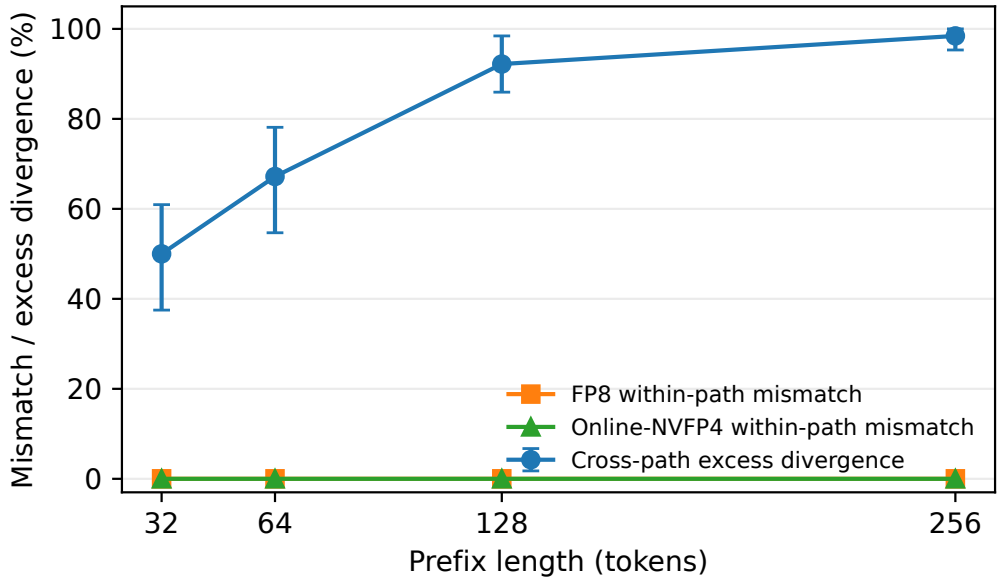

Figure 1: Prefix sensitivity. Same-path mismatch remains zero while cross-path excess divergence rises sharply with trajectory depth. Error bars are task-stratified 95% bootstrap intervals.

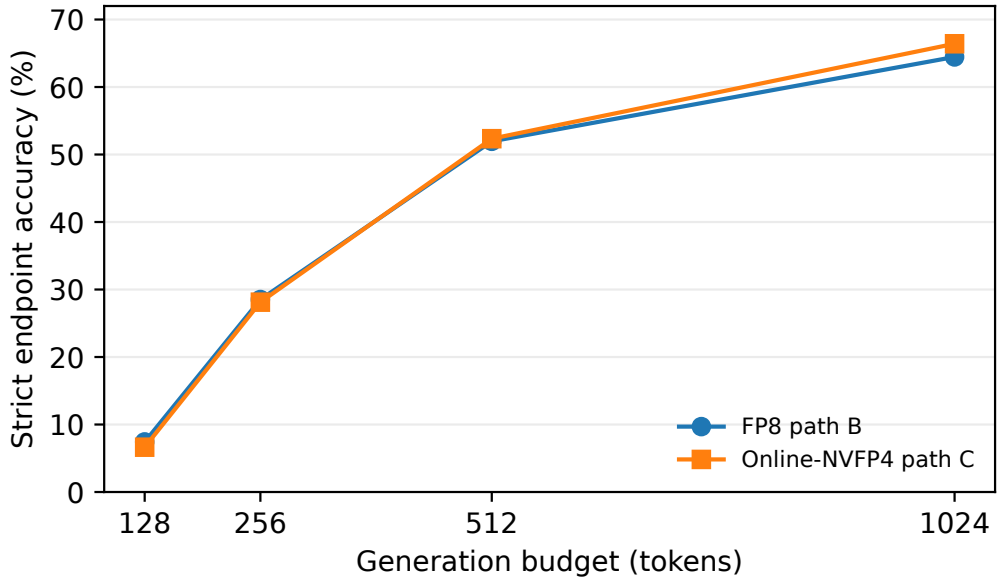

Figure 2: Strict endpoint accuracy reconstructed by censoring a single stored trajectory per example. The two precision paths track closely across budgets despite strong trajectory divergence.

either path. Thus, the data support a process-level non-equivalence claim, not a claim that online NVFP4 improves or harms endpoint accuracy.

5.3 Teacher-forced evidence under identical contexts

The post-confirmatory diagnostic contains 48 paired anchored inputs and 47474 token-ID-comparable positions with zero alignment errors. One observed-token logprob per input is unavailable, leaving 47426 scored positions. Every input exhibits at least one B/C top-1 prediction flip; the median first

flip position is 2, and 1965 of 47426 comparable positions (4.14%) are top-1 flips. Because the token contexts are held identical, this diagnostic shows that the two precision paths can induce different next-token rankings before autoregressive histories themselves diverge. We treat it as explanatory evidence rather than an unbiased population estimate or standalone novelty claim.

6 Discussion and Limitations

The main finding is a separation between trajectory equivalence and endpoint metrics. In our controlled setting, each precision path is individually reproducible under fresh-server repeats, yet the paths are rarely trajectory-equivalent. Aggregate accuracy remains statistically indistinguishable. This matters for evaluations that implicitly treat equal endpoint scores as evidence of interchangeable inference semantics, and for systems requiring trajectory-level reproducibility, auditing, or stateful continuation.

The protocol definition is model-agnostic in the limited sense that it only requires path-labeled trajectories, fresh-server replication, and a frozen comparison contract; the empirical result is not model-agnostic. We evaluate one GLM-5.2 checkpoint, one SGLang/B200 stack, one FP8-source to online-NVFP4 path pair, and two reasoning task families. The task-level replication in Table 2 therefore provides cross-task robustness, not cross-model, cross-engine, or cross-hardware generalization. We make that boundary explicit rather than treating one deployment stack as representative of all low-precision MoE inference.

Stage-2 is selected from already attributable cases and is therefore diagnostic, not an unbiased estimate of flip incidence over the full benchmark. Request-level timing was collected serially at concurrency one and is not a production TTFT/TPOT/throughput benchmark. The strongest next generalization test is a preregistered replication of the same contract on an independent MoE family, quantization pair, or serving engine; the strongest mechanism test is localization to specific operators or persistent-state boundaries. Neither is claimed by the present evidence.

7 Conclusion

A precision path can be deterministic in isolation yet behaviorally non-equivalent to another precision path. Under our pre-frozen experiment, FP8 and online-NVFP4 exhibit zero measured same-path trajectory mismatch on controls but 98.44% excess cross-path divergence at 256 tokens, while the 1024-token endpoint accuracy difference is not statistically significant. These observations argue for reporting trajectory-level compatibility alongside endpoint accuracy when validating low-precision reasoning inference.

Reproducibility Statement

The supplementary materials include the frozen configuration, held-out/control construction, overlap audit, protocol lock, exact trajectory-capture and restart/resume logic, strict answer scoring, analysis scripts, Stage-2 teacher-forced diagnostic, canonical paper-number generator, and SHA256 manifests. Appendix A documents the execution order and statistical estimators, while the anonymous supplementary code reproduces the analysis from the stored evidence and supports a full rerun given the documented model and container environment.

AI Use Statement

Generative AI tools were used to polish author-written text, including improving readability and performing language editing, and for literature retrieval, identification, and summarization. Generative AI tools were not used to generate synthetic datasets; develop theoretical models or conceptual frameworks; formulate mathematical claims; provide ingredients for, or assist with, mathematical proofs; propose or refine hypotheses; design or provide feedback on research methodology or experiments; implement research methods; translate research content; clean or reformat datasets; perform qualitative or thematic data analysis; or interpret experimental results. They were also not used to draft substantive sections of the paper. All reported experimental outputs were produced by

the documented executable pipeline, and numerical claims were verified against stored raw artifacts and manifests. The authors reviewed all AI-assisted material and take responsibility for the final content of the submission.

REFERENCES

Zhen Li, Yupeng Su, Songmiao Wang, Runming Yang, Congkai Xie, Aofan Liu, Ming Li, Jiannong Cao, Yuan Xie, Ngai Wong, and Hongxia Yang. Quantization Meets Reasoning: Exploring and Mitigating Degradation of Low-Bit LLMs in Mathematical Reasoning. arXiv:2505.11574, 2025.

Xinyu Lian, Walid Krichene, Beichen Huang, Masahiro Tanaka, Olatunji Ruwase, Li Zhang, and Minjia Zhang. Quantization Inflates Reasoning: Token Inflation as a Hidden Cost of Low-Bit Reasoning Models. arXiv:2606.25519, 2026.

Yifei Wang, Tianlin Li, Xiaohan Zhang, Xiaoyu Zhang, Wei Ma, Mingfei Cheng, and Li Pan. Hidden Reliability Risks in Large Language Models: Systematic Identification of Precision-Induced Output Disagreements. arXiv:2604.19790, 2026.

Musa Cim, Burak Topcu, and Mahmut Taylan Kandemir. Diagnosing FP4 inference: a layer-wise and block-wise sensitivity analysis of NVFP4 and MXFP4. arXiv:2603.08747, 2026.

Teng-Ruei Chen. The Integer Alibi: Localizing Cross-Kernel Divergence in INT8-Quantized LLM Inference. arXiv:2608.13756, 2026.

Tianyang Zhu. From Expert Reduction to Behavioral Divergence: Tracing Numerical State through Sparse MoE Inference. arXiv:2607.28097, 2026.

Aditi Patodiya. Same Request, Different Answer: Quantization Amplifies Cache-Induced Divergence in LLM Serving. arXiv:2609.04748, 2026.

Ashritha Gonuguntla. The Replay Gap: Static Evaluation of Model Switching in LLM Agents Scores the Wrong World. arXiv:2608.08239, 2026.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. SGLang: Efficient Execution of Structured Language Model Programs. arXiv:2312.07104, 2023.

A CONFIRMATORY PROTOCOL AND DECISION CONTRACT

The confirmatory cohort contains 256 examples; controls contain 64 examples and four fresh-server replicas per path. Historical pilot/dev/control prompts are excluded before deterministic held-out sampling, followed by an independent overlap audit. Primary execution is counterbalanced B–C and C–B across two task-balanced shards. Control replicas alternate BC/CB/BC/CB and reverse request order in even blocks. The frozen thresholds are: maximum within-prefix mismatch 0.10, minimum stable-cross rate 0.25, minimum excess divergence 0.20, positive lower bootstrap bound, and positive task-level excess for both task families. These thresholds were locked before generation.

B BUDGET AND SYSTEMS DIAGNOSTICS

The strict accuracy values at budgets 128/256/512/1024 are B: 7.42/28.52/51.95/64.45% and C: 6.64/28.12/52.34/66.41%. At the 1024-token cap, the fraction ending by length limit is 51.17% for B and 59.77% for C. For 256 serial primary requests per path, request-latency P50 is 268.18 s for B and 280.29 s for C; P95 is 372.55 vs. 357.26 s; P99 is 470.43 vs. 432.94 s. These timings are request-level, concurrency-one diagnostics and are not interpreted as production serving throughput or TTFT/TPOT.

C STAGE-2 MISSING-LOGPROB HANDLING

The teacher-forced API returns exactly one unavailable observed-token logprob per input. The analyzer records these values as missing rather than zero, checks token-ID alignment independently, and computes logprob/top-1 statistics only where both paths provide valid values. The raw B/C capture hashes, analyzer hash, diagnostic CSV hash, and summary hash are stored in the analysis-fix manifest.