Supplementary Materials

Freshness-Aware LLM Inference Routing across Heterogeneous Cloud-Edge Domains: Empirical Evaluation of SLO-Risk Calibration under Stale Telemetry

Supporting information for the revised manuscript

Scope and evidence policy. Tables S1–S6 document the original model-development, calibration, and 225-run factorial campaign. The separate 225-block controlled-contention experiment is summarized in Section S7 and must not be pooled with the original factorial evaluation. Model settings are taken from the frozen selected-hyperparameters record and final-model manifest, not inferred from unpopulated fields in saved XGBoost JSON files. Numeric outputs are not newly estimated in this document.

Section S1. Frozen Experimental Partitioning and Governance

The split unit was the entire experimental run; random per-row splitting and imputation were prohibited. TRAIN used seeds 20261101 and 20261102; MODEL_VALIDATION used 20261103; CONFORMAL_CALIBRATION used 20261104. Model fitting required valid inference, latency, telemetry treatment, and predictor fields. The prespecified unavailable model-validation execution (order 17) was not imputed or retried.

Partition	Seed(s)	CPU rows	GPU rows	Purpose
TRAIN	20261101, 20261102	1,121	6,322	Model fitting
MODEL_VALIDATION	20261103	430	3,060	Hyperparameter selection
Final refit	TRAIN + MODEL_VALIDATION	1,551	9,382	Frozen B3/B4 fitting
CONFORMAL_CALIBRATION	20261104	565	3,176	P1-only calibration (held out)

Table S1. Development partitions and row counts. Conformal calibration has 3,741 rows (565 CPU and 3,176 GPU); final refit is not an independent data partition.

Section S2. Frozen B3/B4 Predictor Con iguration

The selected configurations were confirmed in modeling-results/MODEL-SELECTION-V1/selected-hyperparameters.json and modeling-results/FINAL-B3-B4-V1/final-model-manifest.json; the corresponding four model JSON files passed SHA-256 verification. All four models use XGBoost 3.4.1, objective reg:squarederror, max_depth=3, learning_rate=0.03, random_state=20261105, subsample=1.0, colsample_bytree=1.0, min_child_weight=1, gamma=0, reg_alpha=0, reg_lambda=1. They were fitted on CPU, including the GPU latency predictors.

Frozen model	Trees	Depth	LR	Features	SHA-256 (full)
b3-cpu-final.json	200	3	0.03	6	989737608ac6ed4bff6641e9b01de869b9a2f82309b3f988ff4cf1d4c5b41fda
b3-gpu-final.json	200	3	0.03	8	e4f075ed60d516324cd0d4b72694d4576b37b92b67dc8ed62bdb76834394ad55
b4-cpu-final.json	200	3	0.03	7	9edede31beb000d2ad5529e0f3c96bac154b3fad12ec8fb6e5bdbfea48b5773a
b4-gpu-final.json	500	3	0.03	9	2c1f42dc782e2f6eaa7f2aa83c39e330c1cce2330678f3cbb6d7d1aed8de96fd

Table S2. Frozen selected B3/B4 configurations and model checksums. LR denotes learning rate; the common random state is 20261105. The checksums are of the original saved model files.

Feature contract. B3-CPU uses input_tokens, active_requests, queue_depth, active_slots, host_cpu_pct, and host_mem_pct. B3-GPU additionally uses gpu_util_pct and gpu_mem_used_frac. Each B4 model adds actual_aoi_ms to the corresponding B3 features. Nominal AoI treatment is not a predictor. The originally selected B4-GPU model has 500 trees versus 200 for B3-GPU, so the original B3–B4 comparison is not capacity-matched; a separate development-validation ablation is discussed in the manuscript.

Section S3. Final-Refit Audit

After model selection, final B3 and B4 models were fitted on TRAIN + MODEL_VALIDATION. CONFORMAL_CALIBRATION was not used for model fitting. The recovered model base scores agree with the combined target means to the precision stored by XGBoost; this is an artifact-level consistency check, not a replacement for the training ledger.

Model	Backend	Train	Validation	Final-refit	Base score / mean (ms)
B3	CPU	1,121	430	1,551	13903.690 / 13903.691
B3	GPU	6,322	3,060	9,382	1344.139 / 1344.139
B4	CPU	1,121	430	1,551	13903.690 / 13903.691
B4	GPU	6,322	3,060	9,382	1344.139 / 1344.139

Table S3. Final-refit sample counts and frozen base-score consistency evidence.

Section S4. One-Sided Conformal Calibration

P1 uses the B4 point predictor with nonconformity score actual_e2e_ms − predicted_b4_e2e_ms. In each prespecified backend-by-request-class stratum, α=0.05, k=ceil((n+1)(1−α)) clamped to n, and q is the kth ascending sorted score. A route is designated safe if B4 prediction + q does not exceed the frozen class-specific SLO. Calibration used 3,741 rows and remained fixed during final evaluation.

Stratum	n	k	q (ms)	Calibration coverage	Minimum n passed
CPU / long	113	109	15704.473	0.9646	Yes
CPU / medium	225	215	13588.644	0.9556	Yes
CPU / short	227	217	13394.989	0.9559	Yes
GPU / long	636	606	826.910	0.9528	Yes
GPU / medium	1,270	1,208	903.533	0.9512	Yes
GPU / short	1,270	1,208	864.468	0.9512	Yes

Table S4. Frozen conformal strata. In-sample calibration coverage is an implementation check, not an independent generalization or post-selection conditional risk guarantee.

Section S5. Original Factorial Evaluation and Statistical Inference

The original campaign included three policies (B3, B4, P1), three offered-load levels, five nominal telemetry-age treatments, and five fixed seeds: 225 canonical runs and 50,565 measured requests. Runs (not requests) were the units of inference. Original run-level OLS models used HC3 robust covariance; Wald-term tests were Holm-adjusted within each outcome model.

Model	Outcome	Term	Wald	Holm p
RUN_OLS_HC3	prediction_mae_ms	load	33.75	3.287e-07
RUN_OLS_HC3	prediction_mae_ms	aoi	5.447	1
RUN_OLS_HC3	prediction_mae_ms	policy	0.02123	1
RUN_OLS_HC3	prediction_rmse_ms	load	29.16	3.263e-06
RUN_OLS_HC3	medium_e2e_p95_ms	load	80.58	2.221e-17
RUN_OLS_HC3	medium_e2e_p95_ms	aoi	0.311	1
RUN_OLS_HC3	medium_e2e_p99_ms	load	112.8	2.236e-24
RUN_OLS_HC3	long_ttft_p99_ms	load	201.2	1.421e-43
P1_RUN_OLS_HC3	safe_fraction	load	58.15	7.066e-13
P1_RUN_OLS_HC3	safe_fraction	aoi	38.46	1.8e-07
P1_RUN_OLS_HC3	safe_fraction	load × aoi	25.94	0.001074
P1_RUN_OLS_HC3	coverage_fraction	load	13.45	0.003596
P1_RUN_OLS_HC3	coverage_fraction	aoi	1.099	1

Table S5. Selected frozen run-level factorial Wald results from the original evaluation. These are not tests of the separate controlled-contention campaign.

The revised manuscript additionally reports adjusted marginal-mean contrasts (Table 6). Those normal-approximation HC3 confidence intervals are exploratory, unadjusted for multiple comparisons, and should not be interpreted as equivalence or causality evidence.

Section S6. Frozen Environment and Artifact Identifiers

The recovered analysis environment was Python 3.12.3, XGBoost 3.4.1, NumPy 2.5.3, pandas 2.3.3, scikit-learn 1.9.0, and SciPy 1.18.1. Inference used llama.cpp 0.7.0-dev (build 11706; commit f74e2fa6a), Qwen2.5-1.5B-Instruct Q4_K_M, 8192-token context, four parallel slots, and continuous batching.

Frozen analysis artifact	SHA-256 fingerprint
FINAL-EVALUATION-V2-STAGE5C-B3-VS-B4	3ecac36e818bfddc5ff7b8113116c2be9a94e333b594d2263a0662fd99e4b064
FINAL-EVALUATION-V2-STAGE5D-P1	0edc05a9df2ae8c955c076dc8fb3537857446e89a384044bd9f9afc254c0b934
FINAL-EVALUATION-V2-STAGE5E-V2-INFERENCE	51a1d8936f69417caa39cfc592bd680856c1567df53b69c4613151729531f148
FINAL-EVALUATION-V2-STAGE5F-PUBLICATION	1bbabf1e69922dea29d1e8e1d8ec4d17490c727a7240b79a61ac4fe53793b8ea

Table S6. Previously recorded frozen analysis fingerprints for traceability. These identify the original archive artifacts and are not checksums of this newly formatted document.

Section S7. Separate Controlled-Contention Comparison (Revision Evidence)

This exploratory live study was separate from the 225-run original factorial experiment. It comprised 225 sequential execution blocks (five policies × three contention conditions × three request classes × five replicates), organized into 45 matched five-policy groups. Each block routed one request; the main metric was backend-dispatch-to-completion inference latency, not the original end-to-end measurement. Conditions C0, C4, and C6 imposed zero, six, and twelve background GPU requests, respectively. All routed requests used a 128-token output budget. The nominal telemetry-age treatment was 0 ms; actual age was logged.

The revised manuscript provides the authoritative 15-cell latency, SLO-miss, and routing summary in Table 7 and the eight paired bootstrap contrasts in Table 8. The corresponding recomputed data are provided as supplementary CSV files, Table S7 and Table S8, respectively. Their headline results are: under C0 all 75 requests met SLO; under C4 and C6 all 150 routed requests missed SLO. Under C4, P1 averaged 4232.62 ms versus Shortest-Queue 7747.99 ms (P1-minus-Shortest-Queue −3515.37 ms; 95% stratified paired-group percentile interval [−3584.35, −3453.65]). Under C6, P1 averaged 11045.86 ms versus Shortest-Queue 7778.76 ms (+3267.10 ms; interval [3192.28, 3340.00]). The pooled difference obscures this reversal.

The controlled-contention package is NETWORK_4685389_CONTROLLED_CONTENTION_225_BLOCKS_AUDIT495A.zip (SHA-256: 9b25d7aa11af2a17d5bb8df86f29cad0e08a1acbefc96019552d0a68de7d3e1d). It includes 225 block-level JSON evidence records, an execution ledger, protocol documents, audit code, and recomputed Table S7/Table S8 CSVs. The ZIP was verified against its sidecar checksum and passed ZIP integrity testing. The separately hashed original execution-plan file is not recovered for independent comparison. The 20,000-replicate paired-group percentile bootstrap (seed 20261117) remains descriptive; condition-specific intervals have no multiplicity correction. Since P1 remained on GPU throughout, the comparison does not establish adaptive backend-choice superiority.

Section S8. Reproducibility and Availability Boundaries

The checked original factorial analysis archive includes run-summary tables, analysis scripts, statistical tables, figures, and SHA-256 manifests. That checked archive does not include the complete original request-level dataset or the four frozen XGBoost JSON files. The four model JSON files were verified locally during preparation against their SHA-256 manifests, but that local availability must not be confused with inclusion in the deposited archive. The separate controlled-contention evidence ZIP includes the 225 validated block records and execution ledger; the separately hashed execution-plan file is not included. The availability and limitations of materials in the journal submission must be assessed according to the files actually deposited. Additional retained research records may be requested from the corresponding author, subject to applicable restrictions.

Editorial note. Tables S1–S6 are explicitly numbered and supplied in this document. All Table S2 hyperparameters have been reconciled with the frozen selected-hyperparameters.json and final-model-manifest.json as reported by the author’s read-only audit. This formatting correction does not retrain models or alter experiment outcomes.