

What Signal Does Preemption Need? Benchmarking Preemption-Target Selection for SLO-Aware Reasoning LLM Serving

Qi Yang Huang¹ Alay Dilipbhai Shah² Pragaash Ponnusamy² Hoang Anh Duy Le¹
Thomas A. Joseph¹ Adam Seligman¹ Zhaozhuo Xu^{1*}

¹Workato ²Together AI

Abstract

When the key-value (KV) cache fills, a serving engine must choose which request to evict. We present PreemptBench, comparing seven preemption-target policies across seven reasoning-model families and two benchmarks. Our service-level objective (SLO) breach rate counts every accepted request, including late completions, post-acceptance failures, and unfinished work; this survivorship correction prevents deferred backlog from being omitted from the metric. In the low-breach regime, where last-in, first-out (LIFO) breach is below 15%, young-target-biased policies form a lower-breach cluster than policies without this bias, including random and largest-KV, consistent with a stable incumbent cohort; graded eviction depth is secondary. Within a pure remaining-decode rule, replacing the online predictor with a privileged empirical statistical length oracle, constructed from prior temperature-zero runs, yields no significant gain; neither significantly outperforms LIFO. It summarizes a prompt-conditioned length distribution, not the current execution’s realization, and is not SLO-optimal. Finally, the uncensored single-stream completion-time tail defines a conceptual feasibility floor, but our traces provide only a censored low-load proxy. Post-hoc rescoring across 600–1800-second deadlines finds aggregate memory-freeing reversals at tighter SLOs, but none at 1800 seconds. Thus LIFO is a strong conditional baseline, not a universal optimum.

1 Introduction

Reasoning-LLM serving strains finite GPU memory: generations reach 10k–30k tokens, with heavy-tailed lengths unknown at admission. When the key-value (KV) cache fills under load, the system evicts an in-flight request’s cached tensors; the request later resumes by swapping from CPU memory or recomputing its prefix.

Frameworks such as vLLM (Kwon et al., 2023) and Orca (Yu et al., 2022) treat the choice of *which* running request to evict as secondary; vLLM uses LIFO largely for implementation convenience. Sarathi-Serve (Agrawal et al., 2024b) and Splitwise (Patel et al., 2024) study chunked prefill and phase disaggregation, respectively, but do not isolate target selection.

Target selection controls both discarded work and running-cohort evolution. Evicting a request with K computed tokens requires a K -token prefill on resumption, while repeatedly disrupting the same part of the running set can prevent any stable cohort from completing. The former is event-local; the latter is a system-level stability effect. We separate them. A cheap-to-resume victim need not be the best global choice: a fast-changing score can repeatedly select cheap-to-resume victims while the running set makes little progress. Conversely, recency can protect progressed incumbents without predicting their lengths. This distinction motivates our taxonomy and per-request accounting.

Long-chain-of-thought models amplify both costs through large KV footprints, widely varying completion lengths, and frequent preemption (Guo et al., 2025; Yang et al., 2025; Qwen Team, 2025). We also isolate an implementation-sensitive selection livelock in `smallest_kv`.

We therefore ask: *what signal does preemption-target selection need?* **PreemptBench**¹ organizes selectors by signal—from admission recency (LIFO), through prompt features and decoded progress, to learned length estimates and an empirical statistical length oracle. We evaluate seven reasoning-model families across four architecture classes on two benchmarks. Our accepted-request, survivorship-corrected SLO breach rate counts every accepted request still unfinished at the safety stop rather than hiding backlog. We keep

*Corresponding author: zhaozhuo.xu@workato.com

¹<https://github.com/KevinHuangW/preemptbench>

three questions separate. The headline comparison asks which target signal works where LIFO breach is below 15%. The feasibility analysis asks how much breach any selector could plausibly remove. The SLO-sensitivity analysis then asks when tighter deadlines change the objective enough for a memory-freeing policy to win (§4.4). This separation prevents a conditional reversal from being read as a universal ranking, or our censored feasibility proxy as an exact lower bound.

Contributions.

- A signal taxonomy and vLLM v1 benchmark of seven target selectors, using five-replicate paired experiments across seven model families and two datasets.
- **Finding 1:** in the low-breach regime, policies that evict progressed incumbents disrupt the active cohort and increase breach; protecting incumbents matters more than graded eviction depth.
- **Finding 2:** within the pure remaining-decode family, replacing a noisy online predictor with an empirical statistical length oracle does not improve breach in the low-breach regime.
- **Finding 3:** the uncensored single-stream completion-time tail sets a conceptual feasibility floor and selector headroom. We report a censored low-load proxy, a within-model truncation intervention, and a non-distilled Llama boundary control.

2 Background

Serving systems cache per-token key/value tensors (the KV cache) to avoid recomputation; it grows with sequence length and batch size, so GPU memory is the bottleneck. When it fills, the scheduler makes two coupled decisions: an *admission* order (e.g. FCFS) and, when the cache is full, a *preemption-target selection* over the running set, i.e. which in-flight request to evict. An evicted request is either *swapped* to CPU DRAM or *recomputed* on resume; we study recomputation (vLLM’s default), where the resume cost scales with the evicted prefix length. vLLM (Kwon et al., 2023) pairs FCFS with a recency-based preemption-target rule, tuned for throughput rather than per-request deadlines. Admission ordering and the swap-versus-recompute mechanism are well studied, but preemption-target selection is not, even though it sets how much decode work each preemption discards and which requests are repeatedly delayed, the focus of this work.

3 Preemption-Target Selectors

Rather than propose a policy, we benchmark KV-cache *preemption-target selectors* by the signal each consumes, from admission recency to privileged offline statistical length information, and ask which signals reduce breach across seven reasoning-model families.

A running set R occupies GPU memory up to cap C . Requests arrive open-loop at Poisson rate λ ; when memory fills, the scheduler evicts an incumbent and later recomputes its $K_r = P_r + N_r$ -token prefix. We track prompt length P_r , decoded tokens N_r , and unknown true total T_r ; policy π selects $r \in R$ at each event to minimize SLO breach (§4.1).

Table 1 organizes selectors by signal; each extremizes its score. Exact min/max ties follow running-list order, with no request-ID secondary key. **default** (vLLM’s LIFO) evicts the newest request, implicitly biasing toward *young* targets with small recompute cost. **smallest_kv** (artifact key `shortest_first`) uses $\arg \min K_r$ and evicts the smallest computed prefix. **input_length** uses $\arg \max P_r$; **largest_kv** uses $\arg \max K_r$ and typically selects the most progressed request; seeded **random** samples uniformly and has no progress key.

The length-aware family evicts the request with the largest score $\max(0, \hat{T}_r - N_r)$. **entropy** estimates $\hat{T}_r$ online from cumulative output-entropy spikes using fixed pre-sweep, model-specific coefficients and no serving-time labels. **oracle_remaining** (the *empirical statistical length oracle*) substitutes $\hat{T}_r$, a prompt-conditioned reference derived from prior temperature-zero executions. It does not observe the evaluated execution’s realized T_r . It maximizes estimated remaining *decode*, not SLO attainment, and is therefore neither a perfect-information nor an SLO-optimal oracle; our finding does not concern exact realized-length information or length-aware scheduling generally. Predictor, calibration, and score details are in App. B.

The selectors are not an assumed ladder from “less” to “more” intelligent. Comparing entropy with the empirical statistical length oracle holds the remaining-decode rule fixed and replaces an online estimate with a privileged offline prompt-conditioned statistic; comparing either with LIFO also changes the selection rule. This distinction is the scope of Finding 2.

Table 1: Preemption-target selectors by the signal each consumes. The length-aware family (bottom) shares the pure remaining-decode rule (Eq. 3) and differs only in how the total length $\hat{T}_r$ (or empirical $\tilde{T}_r$) is obtained. The entropy predictor needs the maturity gate (Appendix B); the empirical statistical length oracle is a non-deployable, prompt-conditioned reference, not the current execution’s realized length or an oracle for SLO-optimal scheduling.

Selector	Signal (when avail.)	Evicts
default (LIFO)	admission time (<i>runtime</i>)	most recently admitted
smallest_kv	current KV $K_r = P_r + N_r$ (<i>runtime</i>)	smallest KV, $\arg \min K_r$
input_length	prompt P_r (<i>admission</i>)	longest prompt, $\arg \max P_r$
largest_kv	KV $K_r = P_r + N_r$ (<i>runtime</i>)	largest KV, $\arg \max K_r$
random	none (seeded PRNG)	uniform random member
entropy_remaining	entropy-spike rate $s_r / C_r \rightarrow \hat{T}_r$ (<i>runtime</i>)	most remaining, $\arg \max (\hat{T}_r - N_r)$
oracle_remaining	empirical $\tilde{T}_r$ from prior temperature-0 runs (<i>offline</i>)	most remaining, $\arg \max (\tilde{T}_r - N_r)$

4 Experiments

4.1 Settings

We evaluate two reasoning benchmarks with contrasting decode-length profiles: **AIME** (competition mathematics; long heavy-tailed completions, median 6–14k tokens with a tail beyond 30k) and **HLE** (Center for AI Safety et al., 2026) (Ph.D.-level questions across domains, more heterogeneous, which destabilizes the entropy signal on several families). We serve **seven model families across four architecture classes** on a single H200 under our vLLM-v1 fork: dense transformers (Qwen3-32B, QwQ-32B, GLM-Z1-32B), an MoE (gpt-oss-120b), two R1-distilled models (DeepSeek-R1-Distill-Qwen-32B and -Llama-70B), and a Mamba/Transformer hybrid (Nemotron-H-47B); the set spans the axes that govern preemption (parameter count, per-token KV footprint, base architecture). Full model citations and the load protocol are in App. A.

Requests arrive open-loop (Poisson rate λ , swept per-cell by calibration); reported paired estimates use **5 audited Poisson seeds** (42–46; App. G), a 900s tracked window, and a safety cap that grants every request its full 1800s deadline, so a non-completing request is by construction an SLO breach. `max_tokens` is each model’s native context except in the truncation study (§4.3). During collection, a legacy screen for a rare, policy-independent vLLM *engine stall* triggered same-seed reruns. Our primary analysis uses one audited canonical trace per replication: reproduced stalls remain data, while non-reproduced stalls use the healthy execution from the same-seed pair. We apply no automatic outcome-dependent deletion;

Appendix D documents the adjudication and Appendix E reports the legacy filtered result only as a sensitivity.

Metric. We report the **accepted-request, survivorship-corrected SLO-breach rate**: among requests accepted by the server in the tracked window, the fraction that complete late, fail after acceptance, or remain in flight at the safety cap (Eq. 1, App. A). Requests whose prompt plus configured output cap exceeds the context window are rejected before scheduling; we exclude these validation failures and report them separately. An infrastructure-failed replication is invalidated rather than scored. Paired runs replay the same scheduled arrival and task trace. Counting all accepted unfinished work prevents a policy from looking good by deferring it into an untracked backlog. For example, in a QwQ AIME diagnostic, the completer-only rate is 0.0% for both `largest_kv` and LIFO, whereas counting unfinished accepted requests yields 9.9% versus 1.3% breach (App. H). Breach is our primary axis. We report throughput, goodput, JCT, and fairness as orthogonal axes (§4.3, Table 16). Our non-streaming workload evaluates completion-time SLOs rather than TTFT or interactive responsiveness; those objectives may expose different admission and prefill-queueing tradeoffs.

4.2 Results

How to read the results. Table 1 defines what each selector observes; it does not imply a ranking. Table 2 compares policies at one reproducible point: λ^* , the first seed-verified shared-grid load where LIFO reaches 10% breach. Figure 1 instead follows policy-minus-LIFO breach over load. Here *young-target-biased* means that a selector tends to evict

Table 2: Accepted-request SLO-breach percentage (SLO = 1800 s) at λ^* , the first seed-verified shared-grid arrival rate at which the mean audited-canonical LIFO breach reaches 10%. A breach is an accepted request that finishes late, reports an accepted deadline failure, or remains unfinished at the safety cap. Point estimates average every audited-valid canonical trace (Appendix E). Bold marks the best of LIFO, entropy, and oracle. * denotes a policy significantly worse than LIFO and $\dagger$ significantly better, using a paired two-sided 95% confidence interval over common seeds; no $\dagger$ occurs. All displayed estimates use five canonical reps. Full intervals and MDEs are in Appendix C.

Model	DS	λ^*	LIFO	SKV	InLen	LKV	Ent	Orac	Rnd
Qwen3	AIME	0.100	16.9	27.5*	26.1*	32.0*	27.0	24.3	33.2*
	HLE	0.140	12.1	22.2	21.5	25.1*	19.7	15.7	21.8*
QwQ	AIME	0.100	10.1	21.0*	24.4*	22.4*	15.3	24.7*	29.4*
	HLE	0.110	12.0	19.4*	25.3*	22.1*	24.8*	18.3*	30.6*
GLM-Z1	AIME	0.800	21.1	25.9*	20.3	18.8	18.0	19.5	23.3
	HLE	0.900	14.0	18.3	15.8	14.8	18.9	13.5	16.2
R1D-Llama (20k)	AIME	0.150	11.2	22.3*	22.3*	25.9*	17.2*	15.8*	27.8*
	HLE	0.235	13.8	20.9	26.6*	20.6	16.3	13.4	18.4
R1D-Qwen (32k)	AIME	0.075	15.3	20.6	26.8	27.7*	27.1	16.5	26.8*
	HLE	0.075	19.4	26.3	32.1*	28.5	28.5*	29.1*	37.1
Nemotron-H	AIME	0.200	12.9	13.9	14.4	14.5	14.8	15.0*	12.5
	HLE	0.300	10.2	7.4	9.1	11.3	8.9	11.2	7.7

requests with little computed progress, allowing progressed incumbents to remain stable; *not-young* means it lacks that tendency and instead selects random or more-progressed victims. These are empirical tendencies, not age cutoffs. Findings 1–3 interpret the cohort mechanism, prediction, and feasibility floor; Section 4.3 checks special mechanisms and secondary objectives. Finally, Table 3 re-scores the traces at five deadlines using separate regime-assignment and effect-estimation seeds. It asks where the fixed-point ranking generalizes and when the objective changes. Throughout, policy-minus-LIFO deltas are percentage points, not relative percentages; positive means worse than LIFO. We also keep three levels distinct: the signal a selector reads, the victim it chooses, and the outcome the run produces. Saving prefix recomputation at one event need not improve the run, and equal total breach can hide different mixtures of late, failed, and unfinished requests. This vocabulary is used consistently below.

Because the grid is discrete, realized LIFO breach at λ^* can overshoot 10%; Table 2 alone therefore does not define the feasible regime. It covers twelve headline cells across six families; gpt-oss and its power limits are reported separately in Appendix C.

Finding 1: Evicting progressed incumbents is the dominant anti-pattern. The local intuition that evicting a progressed request wastes more re-

computation is not itself surprising. Our contribution is to show its system-level magnitude under a metric that counts the backlog it creates, and to separate per-event depth from cohort dynamics. At λ^* , *largest_kv* and *input_length* are worse than LIFO in 11/12 and 10/12 cells, respectively (mean +7.9 and +8.0 pp; up to +15 pp), and each is significantly worse in six cells (Table 2).

The mechanism is not a smooth penalty in eviction depth. Figure 3 shows young-target-biased policies evicting at ~ 0.3 – 1.3 k computed-prefix tokens, random at ~ 2 – 9 k, and *largest_kv* at ~ 5 – 42 k. Yet random is worst on average at λ^* (+9.7 pp): merely avoiding the single most-progressed target does not recover most of LIFO’s benefit. Instead, the principal split is between young-target-biased selectors, which protect a stable incumbent cohort, and not-young selectors, which keep perturbing that cohort; additional depth within a cluster is secondary. We quantify this independently binned cohort contrast in §4.4.

The one evict-young policy that still fails, *smallest_kv*, does so for a separate reason treated as a secondary result in §4.3.

Finding 2: Privileged length information does not improve pure remaining-decode selection in the low-breach regime. A natural hypothesis is that remaining-length information can identify long or deadline-infeasible jobs whose suspension preserves capacity for savable work. We test the pure rule that evicts the request with the *largest* predicted decode remaining. Across the twelve headline cells, neither the learned entropy predictor nor the empirical statistical length oracle significantly beats LIFO at λ^* (Table 2). The statistical oracle is closest to LIFO (mean +4.0 pp, worse in 9/12 cells, best point estimate -1.6 pp and not significant) but is significantly worse in five cells; entropy averages +5.6 pp and is significantly worse in three. Thus, holding the selection rule fixed, the substitution directly tests whether privileged prompt-conditioned length information improves scheduling

We do not treat a nonsignificant cell as evidence of equivalence. Appendix C reports each paired interval together with its minimum detectable effect and separates the low-headroom gpt-oss and Nemotron-H cells. The independently binned, variant-level entropy–statistical-oracle comparison in §4.4 instead measures the benefit of replacing the noisy predictor with empirical statistical length information. Finding 2 is therefore a power-aware

statement about *pure remaining-decode scoring*, not about length-aware scheduling in general.

The score $\max(0, \hat{T}_r - N_r)$ has an explicit youth bias: conditional on the same predicted total, it decreases as a request progresses. Variation in $\hat{T}_r$ can still reorder requests, so the selector does not mathematically collapse to LIFO. Empirically, however, both length-aware variants occupy the same young-target cluster as LIFO on these workloads (Figure 3); improving prediction within that cluster yields little change in the low-breach regime. Under a tighter deadline, the objective can change because identifying long or already-doomed work becomes valuable (§4.4).

Finding 3: The feasibility floor determines scheduling headroom. A selector cannot avert a breach when single-stream completion time S_i already exceeds deadline d . Our traces do not identify the ideal floor $P(S_i > d)$ exactly: low-load end-to-end times are censored by the safety deadline and decode caps. We therefore report an empirical low-load tail, with sample sizes, preemptions, censoring, and five-deadline sensitivity in Appendix L (Table 21). At 1800 s, Qwen3, QwQ, and GLM-Z1 have no observed low-load misses, Nemotron-H has one AIME miss in a 15-request pilot, and native R1-distilled models retain 21–43% tails. By contrast, non-distilled Llama-3.3-70B has a large context but short outputs, no preemptions in ten single-seed cells, and tails of 0.35% (AIME) and 0.09% (Appendix L.2).

Truncating the R1D-Qwen and R1D-Llama decode budgets removes observed 1200–1800 s misses in the measured minimum-load cells. Because 13–51% of accepted requests then stop at the cap, this changes the deployed workload; it does not recover the uncapped natural tail. The matched Qwen contrast also confounds recipe with context length, so training recipe is a correlate, not an identified cause. The proxy nevertheless exposes how much scheduling headroom remains, subject to its sampling and censoring limits.

4.3 Analysis

smallest_kv is prone to selection livelock. *smallest_kv* evicts young requests yet can diverge sharply from LIFO under contention (GLM-Z1 AIME at λ^* : 25.9% vs. 21.1%). Although $\arg \min K_r$ minimizes immediate recomputation, its fast-changing global key can reselect a churning low-progress set and collapse goodput (a 94% dump rate in one seed). We report this seed-

sensitive *selection livelock*, rather than a universal mean penalty, with per-seed and re-eviction evidence in Appendix K. It evaluates our isolated $\arg \min K_r$ rule, not a production system’s distinct fewest-decoded-tokens rule and backpressure.

Truncation as a designed intervention. Capping generation near $0.7 \times \text{SLO} \times \text{tok/s}$ removes the observed runtime floor but truncates outputs, so it is a mechanism check rather than a quality-free recommendation. For R1D-Llama, a 20k cap removes measured 1200–1800 s low-load misses (Figure 7); LIFO remains best at λ^* , although the oracle wins by 7.9 pp in one substantially higher-load cell. A 32k R1D-Qwen cap produces the same low-load result, with neither predictor beating LIFO at λ^* (Table 2).

Why behavior differs across models. Two orthogonal axes (App. M) explain most of the variation. The natural completion-time tail relative to the deadline sets the feasibility floor; training recipe and context budget are correlated with that tail in our matched contrasts but are not independently identified. Architecture modulates the selection effect: Nemotron-H’s Mamba layers carry no KV, compressing the KV-size spread and reducing the leverage available to a KV-based selector.

Throughput, latency, and fairness. At λ^* , LIFO has the highest mean successful throughput and on-time goodput, with the oracle closest, but higher median successful-completion JCT than memory-freeing policies. LIFO and *smallest_kv* have the largest repeated-victim tails, whereas *prompt/KV* rules have the highest eviction-count Gini. Thus LIFO is best for completion goodput here, not every objective; Table 16 and full per-cell results are in Appendix I.

4.4 SLO Sensitivity and the Conditional Reversal

The 1800 s headline deadline does not make every cell feasible. We therefore rescore the traces at 600–1800 s and index each cell by LIFO’s corrected breach (Table 3; Figure 2 in the appendix). To avoid defining a regime with the same noise used to estimate its policy effect, our primary analysis excludes 53 fixed-schedule exploratory loads, then uses replicate seeds 42–43 to assign model–workload–deadline–load cells to breach strata and the independent seeds 44–46 to estimate paired policy differences. We first average SLO/load cells within each model–dataset variant and compute uncertainty across variants. As a robustness

Accepted-request breach relative to LIFO — AIME (SLO = 1800s)

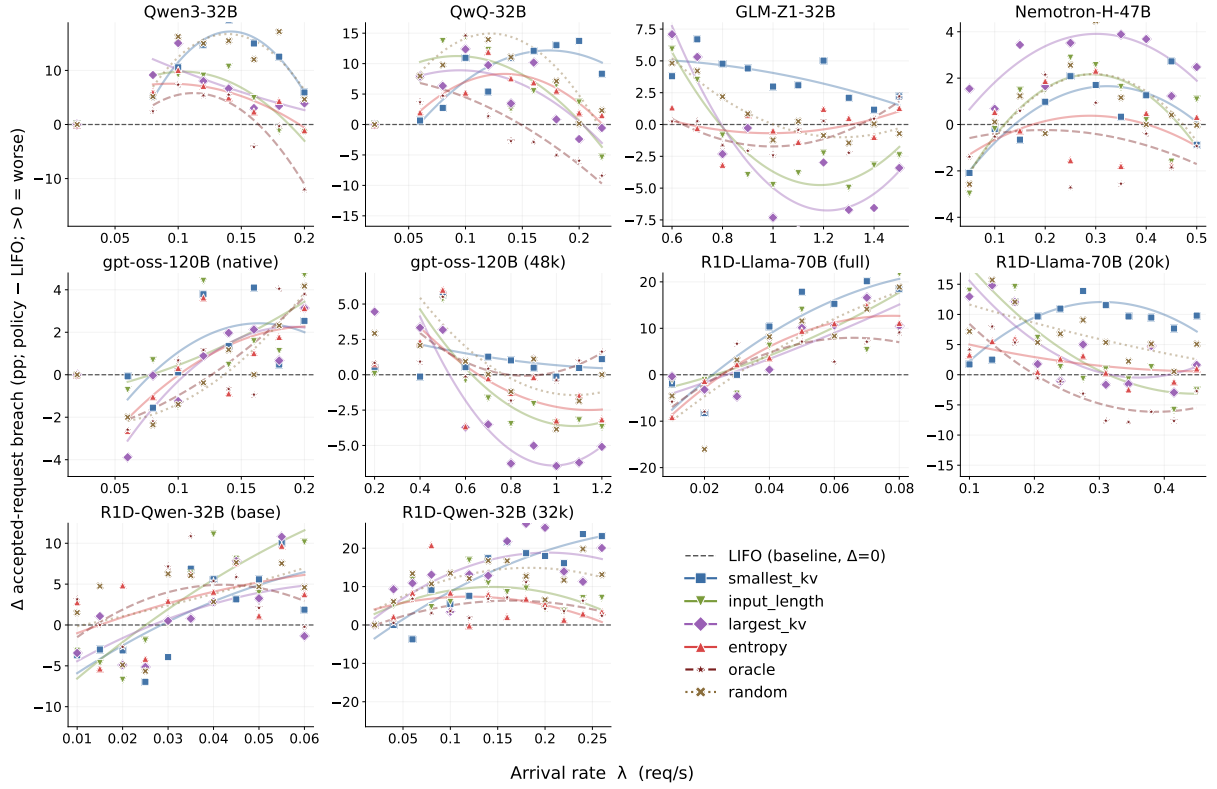


Figure 1: Accepted-request breach relative to LIFO (policy – LIFO; > 0 is worse) over load on AIME. Each panel is one model and decode budget; the dashed $y=0$ line is LIFO. The not-young selectors (input_length, random, and largest_kv) generally fan above the young-target cluster (entropy and the empirical statistical length oracle), while smallest_kv breaks upward on GLM-Z1. No policy is consistently below LIFO in the feasible range. Load windows differ because preemption onset is model-dependent. Only complete load points with five audited arrival schedules are shown.

Table 3: Accepted-request breach difference from LIFO by independently assigned regime (+ is worse). LIFO seeds 42–43 assign regimes; paired seeds 44–46 estimate effects. Cells are averaged within variant, then variants equally. * denotes an exact variant-level sign test with $p < 0.05$. The $> 40\%$ stratum is a subset of $> 30\%$; Table 14 separates the deadline interaction. Cell/variant counts are feasible: 185/18; tie: 176/18; failed: 874/20; deep: 752/19.

Policy – LIFO (pp)	$< 15\%$	15–30%	$> 30\%$	$> 40\%$
smallest_kv	+2.1	+3.4*	+1.8*	+1.6*
input_length	+4.4*	+2.4	–3.3*	–3.7
largest_kv	+4.5*	+2.4	–4.2*	–4.7*
Learned remaining	+2.9*	+2.0	–0.6	–0.9
Decode-length oracle	+2.4*	+0.3	–2.5	–3.0
random	+4.6*	+2.9	–1.8*	–2.3

check, Appendix G repeats the analysis over all ten choices of two binning seeds and three complementary effect seeds.

In cells where LIFO’s breach is below 15%, the tested alternatives remain worse on average:

the cohort-stability result persists, and the empirical statistical length oracle does not rescue pure remaining-work selection. On the fixed split, the not-young cohort is +2.66 pp worse than the young cohort (95% CI [+1.42, +3.90]; positive in 16/18 variants), whereas the more graded largest_kv–random contrast is –0.12 pp ([–1.66, +1.42]). Replacing the learned predictor with the empirical statistical length oracle has a point estimate of –0.49 pp (equivalently, learned minus oracle is +0.49 pp, [–0.96, +1.95]); this interval describes uncertainty, not a declared equivalence margin. Cells above 30% breach do not, however, have a universal ordering. Pooling deadlines can make memory-freeing policies appear to reverse relative to LIFO, but separating the deadline shows that load alone is insufficient. At the 1800 s headline deadline, aggregate largest_kv remains worse than LIFO across the load sweep, although an isolated high-load oracle cell can win. The systematic memory-freeing reversal appears at

tighter deadlines, where late completions form a substantial part of breach. `smallest_kv` remains a poor choice because repeated selection of the same victims creates a high-variance failure mode (§4.3). An accepted request can breach by completing *late*, remaining *unfinished*, or returning a post-acceptance deadline error. `largest_kv` trades fewer late completions for more unfinished work, so it helps only when the former dominates. The late-component difference is zero at 1800s and grows as the deadline tightens. Thus the crossover is deadline- and workload-dependent, as the independent-split, cross-fit, and component estimates in Tables 3 and 14 confirm.

4.5 Practical Guidance for Production Serving

(1) Measure the feasibility floor. Estimate the uncensored low-load completion-time tail with a safety stop beyond the deadline. Near the breach budget, admission control, routing, an acceptable decode cap, or a faster model has more leverage (our censored proxy is Table 21).

(2) Start with LIFO, then diagnose the breach source. No tested predictor reliably beats recency at low breach. Memory-freeing can help when a tight deadline makes late completions dominate; load or a 30% breach rate alone does not predict that crossover (§4.4).

(3) Monitor the induced cohort. Victim progress, repeated selection, completion tails, and late-versus-dump breach expose cohort churn.

5 Related Work

Serving work addresses continuous batching (Orca (Yu et al., 2022)), PagedAttention (vLLM (Kwon et al., 2023)), chunked prefill (Sarathi-Serve (Agrawal et al., 2024b)), phase disaggregation (Splitwise (Patel et al., 2024)), job-level preemption (FastServe (Wu et al., 2026)), and KV migration (Llumnix (Sun et al., 2024)); none isolates which request to evict within an event. Completion-length predictors are typically offline (Zheng et al., 2023; Jin et al., 2023; AlpaServe (Li et al., 2023)); ours uses runtime token statistics with fixed, mostly calibration-informed model coefficients, yet Finding 2 shows that replacing the online estimate with privileged statistical length information does not necessarily improve eviction outcomes. Reasoning workloads (Guo et al., 2025; Yang et al., 2025; Qwen Team, 2025) bring extreme length variance

and routine KV saturation. Vidur (Agrawal et al., 2024a) simulates inference but not target selection; Nemotron-H (Blakeman et al., 2025) cuts KV via Mamba layers (App. M); and token-level KV eviction (Xiao et al., 2024) complements our which-request question. To our knowledge, PreemptBench is the first request-level preemption-target benchmark spanning multiple reasoning-model families.

6 Conclusion

We benchmarked seven preemption-target selectors across seven reasoning-model families. Below 15% LIFO breach, young-target-biased selectors protect a stable incumbent cohort, while not-young selectors incur more breach; graded per-event recomputation cost is secondary. Within pure remaining-decode scoring, the empirical statistical length oracle yields no significant gain over LIFO or the runtime predictor. This claim is scoped to that scoring family. The uncensored single-stream completion-time tail sets the conceptual feasibility floor, but our traces identify only a censored proxy. Thus practitioners should measure the floor, default to LIFO at low breach, and consider memory-freeing only when tight deadlines make late completions dominate (§4.4). Future work includes a deadline-aware oracle and evaluation of SGLang’s native decode-retraction policy, fewest decoded tokens first, with ties broken toward larger original prefills, with its backpressure mechanism (Zheng et al., 2024). The artifact will provide the harness, pinned scheduler patch, configurations, paired seeds, analysis, and traces.

Limitations

The GPU-hour cost of per-cell calibration followed by seven-policy, five-seed λ sweeps limits our scope to one H200, our vLLM-v1 fork, recompute-only resumption, Poisson arrivals, and two reasoning benchmarks. Multi-SLO results are post-hoc rescoring; isolated controls are identified as such. Our non-streaming workload targets completion-time SLOs rather than TTFT or interactive responsiveness. Broader hardware, backends, swap-based resumption, arrival processes, and the effects of target selection on admission and prefill queueing remain future work.

References

- Amey Agrawal, Nitin Kedia, Jayashree Mohan, Ashish Panwar, Nipun Kwatra, Bhargav S. Gulavani, Ramachandran Ramjee, and Alexey Tumanov. 2024a. **VIDUR: A large-scale simulation framework for LLM inference**. In *Proceedings of Machine Learning and Systems*, volume 6, pages 351–366.
- Amey Agrawal, Nitin Kedia, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav S. Gulavani, Alexey Tumanov, and Ramachandran Ramjee. 2024b. **Taming throughput-latency tradeoff in LLM inference with Sarathi-Serve**. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pages 117–134. USENIX Association.
- Aaron Blakeman, Aarti Basant, Abhinav Khattar, Adithya Renduchintala, Akhiad Bercovich, Aleksander Ficek, Alexis Bjorlin, Ali Taghibakhshi, Amala Sanjay Deshmukh, Ameya Sunil Mahabaleshwar, Andrew Tao, Anna Shors, Ashwath Aithal, Ashwin Poojary, Ayush Dattagupta, Balaram Budharaju, Bobby Chen, Boris Ginsburg, Boxin Wang, and 180 others. 2025. **Nemotron-H: A family of accurate and efficient hybrid Mamba-Transformer models**. *arXiv preprint arXiv:2504.03624*.
- Center for AI Safety, Scale AI, and HLE Contributors Consortium. 2026. **A benchmark of expert-level academic questions to assess AI capabilities**. *Nature*, 649:1139–1146.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. **DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning**. *Nature*, 645(8081):633–638.
- Yunho Jin, Chun-Feng Wu, David Brooks, and Gu-Yeon Wei. 2023. **S³: Increasing GPU utilization during generative inference for higher throughput**. In *Advances in Neural Information Processing Systems*, volume 36, pages 18015–18027. Curran Associates, Inc.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. **Efficient memory management for large language model serving with PagedAttention**. In *Proceedings of the 29th ACM Symposium on Operating Systems Principles (SOSP)*, pages 611–626. ACM.
- Zhuohan Li, Lianmin Zheng, Yinmin Zhong, Vincent Liu, Ying Sheng, Xin Jin, Yanping Huang, Zhifeng Chen, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. **AlpaServe: Statistical multiplexing with model parallelism for deep learning serving**. In *17th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23)*, pages 663–679. Boston, MA. USENIX Association.
- Pratyush Patel, Esha Choukse, Chaojie Zhang, Aashaka Shah, Íñigo Goiri, Saeed Maleki, and Ricardo Bianchini. 2024. **Splitwise: Efficient generative LLM inference using phase splitting**. In *Proceedings of the 51st Annual International Symposium on Computer Architecture (ISCA)*, pages 118–132. IEEE.
- Qwen Team. 2025. QwQ-32B: Embracing the power of reinforcement learning. <https://qwenlm.github.io/blog/qwq-32b/>.
- Biao Sun, Ziming Huang, Hanyu Zhao, Wencong Xiao, Xinyi Zhang, Yong Li, and Wei Lin. 2024. **Llumnix: Dynamic scheduling for large language model serving**. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pages 173–191. Santa Clara, CA. USENIX Association.
- Bingyang Wu, Yinmin Zhong, Zili Zhang, Shengyu Liu, Fangyue Liu, Yuanhang Sun, Gang Huang, Xuanzhe Liu, and Xin Jin. 2026. **FastServe: Iteration-Level preemptive scheduling for large language model inference**. In *23rd USENIX Symposium on Networked Systems Design and Implementation (NSDI 26)*, pages 57–74. Renton, WA. USENIX Association.
- Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. **Efficient streaming language models with attention sinks**. In *The Twelfth International Conference on Learning Representations*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. **Qwen3 technical report**. *arXiv preprint arXiv:2505.09388*.
- Gyeong-In Yu, Joo Seong Jeong, Geon-Woo Kim, Soojeong Kim, and Byung-Gon Chun. 2022. **Orca: A distributed serving system for Transformer-based generative models**. In *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*, pages 521–538. USENIX Association.
- Z.ai Team. 2025. GLM-4-0414 model series. <https://github.com/zai-org/GLM-4>.
- 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. 2024. **SGLang: Efficient execution of structured language model programs**. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Zangwei Zheng, Xiaozhe Ren, Fuzhao Xue, Yang Luo, Xin Jiang, and Yang You. 2023. **Response length perception and sequence scheduling: An LLM-empowered LLM inference pipeline**. In *Advances in Neural Information Processing Systems*, volume 36, pages 65517–65530. Curran Associates, Inc.

A Experimental Setup

Models. Seven families, each on one H200 under our vLLM-v1 fork: dense transformers Qwen3-32B (Yang et al., 2025), QwQ-32B (Qwen Team, 2025) (both 40K context) and GLM-Z1-32B (Z.ai Team, 2025) (32K); the MoE gpt-oss-120b (reported on its 48k decode-budget form, since its ~ 845 k-token KV pool leaves the native model nearly preemption-free, so truncation places it in the steady-preemption regime the others occupy); R1-distilled DeepSeek-R1-Distill-Qwen-32B and -Llama-70B (Guo et al., 2025); and the Mamba/Transformer hybrid Nemotron-H-47B-Reasoning (Blakeman et al., 2025). **Protocol.** Within each replicate, every policy replays the same task order and arrival sequence, so policy comparisons remain paired. The task builder uses `task_seed=42`, but an earlier shuffle is also keyed by the replicate’s Poisson seed; consequently task order, as well as arrival timing, changes across seeds 42–46. We preserve that coupling in the analysis and do not interpret the replicates as varying arrival timing alone. λ is swept per-cell by calibration (windows differ because preemption onset is architectural). Audited paired cells use five Poisson seeds (42–46); early fixed-schedule exploratory points are excluded from five-seed inference (Appendix G). Each run: 60 s warm-up, 900 s tracked window, then untracked pressure until all tracked requests finish, safety cap at window-end + 1800 s; greedy decoding (temperature 0). **Metric.** For the accepted scheduling population, the survivorship-corrected breach rate is

$$\text{breach\%} = 100 \cdot \frac{n_{\text{late}} + n_{\text{fail}} + n_{\text{unfinished}}}{n_{\text{successful}} + n_{\text{fail}} + n_{\text{unfinished}}}, \quad (1)$$

where n_{late} counts successful completions beyond the evaluated SLO, n_{fail} counts deadline/timeout failures after server acceptance, and $n_{\text{unfinished}}$ counts tracked requests still in flight at the safety cap. Immediate HTTP-400 context-window validation rejects are pre-service attempts and are reported but excluded; replication-level infrastructure failures are invalidated rather than scored. Paired runs replay the same scheduled arrival and task trace. In the 1800-s headline traces, no successful completion falls beyond the equal-length safety cap, so breach is one minus the normalized on-time completion fraction (not one minus goodput measured in requests/s).

B Length-Aware Predictor

The entropy predictor estimates remaining length online from the model’s per-token output entropy $H_t = -\sum_v p_{v,t} \log p_{v,t}$. The evaluated `entropy_remaining` implementation is the one-feature version: during a request’s current residency episode it maintains the spike count $s_r = \#\{t : H_t > \theta_m\}$. Let $C_r = \text{num_computed_tokens}$ be the currently materialized prompt-plus-decode prefix and let $N_r = \text{num_output_tokens}$ be the persistent decoded count. Both s_r and C_r reset when discard preemption forces recomputation; N_r does not. Once C_r reaches the model-specific checkpoint g_m , the runner computes

$$\hat{T}_r = \max\left\{\exp\left(a_m + b_m \frac{s_r}{C_r}\right), C_r + 1\right\}; \quad (2)$$

below g_m it returns a model-specific constant F_m . Thus the implemented feature is normalized by `num\_computed\_tokens`, which includes prompt and decoded tokens, rather than by decoded length alone. At serving time it consumes no true-length label. The deployed score then uses the persistent decoded count N_r :

$$\text{score}_r = \max(0, \hat{T}_r - N_r), \quad (3)$$

evicting $\arg \max_r \text{score}_r$; we strip KV-footprint and sunk-cost terms to isolate selection quality, and do not pursue the co-design variant that reintroduces them, since it conflates signal with the `largest_kv` anti-pattern:

$$\text{score}_r = K_r (\hat{T}_r - (1 + \beta) N_r), \quad \beta = 1. \quad (4)$$

`oracle_remaining`, which we call the *empirical statistical length oracle*, replaces $\hat{T}_r$ with a prompt-keyed reference $\tilde{T}_r$ derived from prior temperature-zero default-policy executions. For a key with only uncensored references, the lookup stores their median completion length; if any retained reference lies within 100 tokens of its model’s decode cap, it conservatively stores that cap, and a missing key receives the configured 12,760-token fallback. Because numerically induced greedy decode divergence makes completion length distributional across executions, $\tilde{T}_r$ is a privileged empirical statistic, not the evaluated execution’s realized T_r . The policy is the non-deployable statistical-reference version of this $\arg \max(\hat{T} - N)$ rule, not a perfect-information or SLO-minimizing oracle.

Both remaining-decode selectors exclude requests whose persistent preemption count is already at least four; if this empties the candidate pool, they fall back to all running requests. The other five selectors do not use this eligibility guard. **Calibration scope.** Table 4 records the parameters selected by the harness for the reported model tags; the trace metadata did not snapshot these environment variables. Most are exploratory calibration settings rather than a held-out evaluation, two inherited R1 settings lack a retained fit artifact, HLE reuses the configured one-feature model parameters, and gpt-oss uses prespecified transfer values that were not refit on that model. We explored a separate two-feature entropy_spike_zero policy after observing that spike-only correlation can be weak or negative on HLE, but it is not the entropy arm in the headline tables or figures. We therefore do not report its post-hoc correlations as if they measured the evaluated policy. Finding 2 rests instead on the direct paired substitution of the privileged prompt-conditioned statistic $\hat{T}_r$ for the online estimate $\hat{T}_r$ under the same score. **System design.** We patch the vLLM v1 scheduler to route each preemption through a configurable score; the scheduler updates its entropy counters incrementally from runner outputs. Preemption resumes by recomputation (no CPU offload); each server is tensor-parallel TP=1.

C Paired Confidence Intervals and Detectable Effects

For each policy and cell we pair the accepted-breach difference against LIFO by arrival seed. With sample standard deviation s_d over n paired differences, the two-sided 95% interval is $\bar{d} \pm t_{.975, n-1} s_d / \sqrt{n}$. We report the approximate 80%-power minimum detectable effect $(t_{.975, n-1} + t_{.80, n-1}) s_d / \sqrt{n}$ at $\alpha = .05$. A confidence interval containing zero is not evidence of equivalence when paired variance and small n make the MDE wide. Table 5 gives all six policy comparisons for the dense-model cells; Table 6 keeps Nemotron-H and both gpt-oss treatments separate.

D Decode Nondeterminism and Engine-Stall Robustness

Greedy decoding is not bit-reproducible across runs: vLLM batches all in-flight requests each step, and cuBLAS summation order depends on batch composition, so when the top-2 logits are within $\sim 10^{-5}$ the argmax flips and continuations diverge

(cross-seed median completion-length difference $\approx 1.5k$ tokens). Independently, a server occasionally enters a preempt/restart *engine stall* (a vLLM-internal failure, not the preemption-target policy): it serves far fewer requests than its seed siblings, leaves a large unfinished backlog, and jumps from cohort-typical breach to 50–85%.

During collection, a legacy dump-breach screen flagged outcomes and moved the first execution to quarantine before launching a nominally same-seed rerun. The screen did not require the simultaneous served-count collapse described in the rebuttal and its deletion helper could therefore remove valid, reproducible outcomes. We do not use that automatic subset as the primary analysis.

For the camera-ready audit, an execution is an engine-stall candidate only when (i) its served count is below half the median of the other four replications and (ii) its accepted-request breach differs from the sibling median by more than $\max(15 \text{ pp}, 3 \text{ IQR})$. After integrating the five adjudication reruns, 10,534 rep-cells are present in the full collected trace pool before seed-based exclusions; 10,530 have four available sibling executions and can be screened exactly. Seven meet both criteria, and all seven have a valid same-seed adjudication. Five reproduce the stall and therefore remain data; two do not reproduce the joint collapse, so the healthy execution in each pair is canonical. The audit and integration manifests record every decision, and both executions are preserved. This broader full-pool audit supersedes preliminary counts from the narrower legacy screen.

This rule identifies the Qwen3–HLE LIFO anomaly: the later execution records 23 successful completions and 90 unfinished accepted requests (79.65% breach), whereas its same-seed original records 85 and 28, respectively (24.78% breach). Because the policy is LIFO and the same-seed original does not reproduce the joint collapse, we treat the later trace as an engine stall rather than evidence for the smallest_kv-specific mechanism. The canonical original changes the Qwen3–HLE operating point to $\lambda^* = 0.140$ but does not change any qualitative result. Table 10 gives the complete automatic-subset sensitivity; Table 5 bases inference on paired audited-valid canonical seeds. This separates decode nondeterminism, rare policy-independent engine stall, and the policy-specific selection livelock in Appendix K.

Table 4: One-feature entropy-predictor configuration selected by the harness for the reported entropy_remaining model tags. The trace metadata does not independently snapshot these environment variables. Threshold θ_m is in nats; g_m and F_m are tokens. The same parameters are used for the native and 48k gpt-oss treatments. These are calibration/training parameters rather than a held-out accuracy result.

Model configuration	θ_m	g_m	a_m	b_m	F_m	Parameter source
Qwen3-32B	1.00	256	8.140	10.600	14000	AIME calibration fit
QwQ-32B (40k)	1.00	256	8.550	6.080	14000	prior calibration fit
GLM-Z1-32B	0.05	1024	7.350	3.720	9500	AIME calibration fit
R1D-Llama-70B (20k)	1.00	1536	7.627	6.551	6500	inherited setting; fit artifact unavailable
R1D-Qwen-32B (32k)	2.00	1536	6.816	7.818	6000	inherited setting; fit provenance unavailable
Nemotron-H-47B	1.00	1024	7.670	12.370	12760	AIME calibration fit
gpt-oss-120B	1.00	256	8.000	6.000	15000	prespecified transfer values

Table 5: Per-cell paired policy-minus-LIFO accepted-breach effects for the ten dense-model headline cells. Each entry is mean difference in percentage points [two-sided 95% CI] / MDE. Positive is worse than LIFO. * and † mark intervals wholly above and below zero, respectively. All estimates use five paired audited-valid canonical seeds.

Model	DS	λ^*	SKV	InLen	LKV	Ent	Orac	Rnd
Qwen3	AIME	0.100	+10.6 [+2.5,+18.6]/10.8*	+9.2 [+2.3,+16.0]/9.2*	+15.1 [+3.6,+26.5]/15.3*	+10.1 [-1.0,+21.2]/14.9	+7.4 [-0.9,+15.7]/11.1	+16.2 [+11.2,+21.3]/6.8*
QwQ	AIME	0.100	+11.0 [+3.0,+18.9]/10.7*	+14.4 [+8.9,+19.8]/7.3*	+12.4 [+3.6,+21.1]/11.7*	+5.2 [-4.0,+14.4]/12.3	+14.6 [+7.6,+21.7]/9.5*	+19.3 [+9.1,+29.6]/13.8*
GLM-Z1	AIME	0.800	+4.8 [+0.0,+9.5]/6.4*	-0.9 [-5.3,+3.6]/6.0	-2.3 [-6.3,+1.6]/5.3	-3.2 [-7.5,+1.2]/5.8	-1.6 [-5.3,+2.0]/4.9	+2.2 [-4.0,+8.4]/8.3
R1D-Llama (20k)	AIME	0.150	+11.1 [+3.7,+18.5]/9.9*	+11.1 [+3.3,+18.9]/10.5*	+14.7 [+6.6,+22.9]/10.9*	+6.0 [+3.2,+8.8]/3.8*	+4.6 [+0.7,+8.5]/5.2*	+16.6 [+8.4,+24.8]/10.9*
Qwen3	HLE	0.140	+10.0 [-4.5,+24.6]/19.5	+9.3 [-0.5,+19.2]/13.2	+12.9 [+6.9,+19.0]/8.1*	+7.5 [-1.2,+16.2]/11.7	+3.6 [-2.8,+10.0]/8.6	+9.6 [+2.8,+16.5]/9.2*
QwQ	HLE	0.110	+7.4 [+2.6,+12.2]/6.4*	+13.3 [+7.5,+19.1]/7.8*	+10.1 [+6.5,+13.7]/4.8*	+12.8 [+4.0,+21.6]/11.8*	+6.3 [+1.0,+11.6]/7.1*	+18.6 [+12.6,+24.5]/8.0*
GLM-Z1	HLE	0.900	+4.2 [-1.6,+10.1]/7.8	+1.8 [-3.9,+7.5]/7.7	+0.8 [-7.6,+9.3]/11.3	+4.8 [-2.5,+12.1]/9.8	-0.5 [-5.5,+4.5]/6.6	+2.2 [-5.0,+9.4]/9.6
R1D-Llama (20k)	HLE	0.235	+7.1 [-0.1,+14.3]/9.6	+12.8 [+9.6,+16.0]/4.3*	+6.8 [-2.6,+16.1]/12.5	+2.5 [-10.5,+15.5]/17.5	-0.4 [-10.1,+9.2]/12.9	+4.7 [-2.7,+12.0]/9.9
R1D-Qwen (32k)	AIME	0.075	+5.3 [-4.7,+15.3]/13.4	+11.6 [-1.2,+24.3]/17.1	+12.4 [+3.7,+21.1]/11.6*	+11.9 [-1.9,+25.6]/18.4	+1.3 [-5.9,+8.5]/9.6	+11.5 [+2.6,+20.4]/11.9*
R1D-Qwen (32k)	HLE	0.075	+6.9 [-0.3,+14.1]/9.6	+12.7 [+1.8,+23.6]/14.6*	+9.1 [-2.2,+20.4]/15.1	+9.1 [+0.1,+18.1]/12.1*	+9.7 [+5.1,+14.3]/6.2*	+17.7 [-0.9,+36.3]/24.9

E Trace-Outcome and Rep-Selection Audit

A successful completion is a task row with numeric elapsed time and no error. Maximum-context HTTP 400 responses are pre-admission rejections; accepted deadline or timeout responses remain breaches. Replications with infrastructure failures are invalidated as a whole. Tables 7–9 expose counts, per-policy deltas, rankings, and task-ID-set invariance under the primary and two sensitivity treatments. The automatic dump-breach subset rule is not the documented joint n -collapse plus breach-outlier rerun protocol: at the primary operating points it removes audited-valid outcomes, including reproduced behavior. We therefore use all audited-valid canonical traces for point estimates and inference, and retain the subset only as the complete sensitivity in Table 10.

F Provenance of Reported Numbers

Table 11 distinguishes new executions from re-analysis. The post-submission grid extension adds load points for model configurations already in the study; rescore those traces at additional deadlines does not add GPU executions. Same-seed engine reruns are integrity checks, not extra independent replicates, and only the canonical trace selected by Appendix D enters an estimate. Early

fixed-schedule exploratory executions remain in the archive but are excluded from estimates requiring five distinct seeds (Appendix G).

G Independent Regime Assignment and Cross-Fit Robustness

Poisson-seed audit. Before split-sample inference, we audited each trace’s recorded Poisson seed. At 53 of 301 dataset–model–load points, 756 policy–replicate traces from early exploratory sweeps reused the seed-42 arrival/task schedule in slots later mapped to seeds 43–46. We conservatively exclude each entire load point (1,855 planned policy–replicate cells) rather than treating repeated executions as distinct seeds. The seed-audited grid retains 248 points, and every analyzed cell additionally satisfies the per-split completeness requirement. All 420 traces at the 12 headline operating points pass the same audit.

The matched regime analysis contains 20 model–dataset variants: ten model configurations on each of AIME and HLE. Every included load has traces for all seven selectors, and every trace is rescored at five SLOs (600, 900, 1200, 1500, and 1800 s). The two untruncated R1D-Qwen variants have no feasible cell because their low-load breach already exceeds 15%; hence the fixed-split feasible stratum covers 18 of 20 variants. We require complete

Table 6: The same paired effect, confidence-interval, and MDE audit for Nemotron-H and for the 48k and native-context gpt-oss treatments. These architectural/low-headroom cases are kept separate from the dense-model summary. Each entry is mean difference [95% CI] / MDE in percentage points; positive is worse, and * marks an interval wholly above zero. All entries use five paired seeds.

Model	DS	λ^*	SKV	InLen	LKV	Ent	Orac	Rnd
Nemotron-H	AIME	0.200	+1.0 [-1.5,+3.4]/3.3	+1.5 [-0.3,+3.2]/2.4	+1.7 [-2.0,+5.3]/4.9	+1.9 [-5.1,+8.8]/9.3	+2.2 [+0.9,+3.4]/1.7*	-0.4 [-3.4,+2.6]/4.0
Nemotron-H	HLE	0.300	-2.8 [-11.5,+5.9]/11.6	-1.1 [-9.3,+7.2]/11.0	+1.1 [-5.8,+7.9]/9.2	-1.3 [-7.5,+4.9]/8.3	+1.0 [-1.3,+3.3]/3.1	-2.4 [-9.6,+4.7]/9.6
gpt-oss (48k)	AIME	0.350	-1.1 [-5.8,+3.5]/6.2	+2.7 [-1.9,+7.2]/6.1	+3.4 [-0.0,+6.8]/4.6	+0.4 [-6.1,+6.9]/8.7	-1.1 [-8.6,+6.5]/10.1	+3.4 [-3.2,+10.1]/8.9
gpt-oss (48k)	HLE	0.175	-1.4 [-3.6,+0.9]/3.0	+1.6 [+0.4,+2.8]/1.6*	+8.5 [+2.2,+14.7]/8.4*	+3.1 [+0.9,+5.3]/3.0*	+1.8 [-2.0,+5.6]/5.1	+3.3 [+0.9,+5.6]/3.2*
gpt-oss (native)	AIME	0.140	+1.3 [-2.0,+4.7]/4.4	+0.5 [-0.6,+1.6]/1.5	+2.0 [-3.9,+7.9]/7.9	-0.9 [-3.9,+2.2]/4.1	-0.7 [-6.2,+4.8]/7.3	+1.2 [-4.6,+7.0]/7.7
gpt-oss (native)	HLE	0.080	-0.2 [-5.3,+4.9]/6.8	+4.3 [-5.5,+14.0]/13.1	+0.5 [-5.4,+6.4]/7.9	+1.3 [-5.4,+8.0]/9.0	-0.2 [-5.1,+4.7]/6.6	+0.0 [-6.9,+7.0]/9.3

Table 7: Outcome counts at each primary operating point, aggregated over the five canonical reps. Each entry is S/R/D/I/A: successful completions, pre-admission context rejections, accepted deadline/timeout failures, infrastructure errors, and cap-aborted accepted requests. A replication with an infrastructure error is invalidated as a whole rather than repaired by deleting individual rows.

Model	DS	λ^*	LIFO	SKV	InLen	LKV	Ent	Orac	Rnd
Qwen3	AIME	0.100	383/0/0/0/78	333/0/0/0/128	341/0/0/0/119	311/0/0/0/150	335/0/0/0/126	347/0/0/0/114	307/0/0/0/154
QwQ	AIME	0.100	414/0/0/0/47	362/0/0/0/99	348/0/0/0/113	356/0/0/0/105	389/0/0/0/72	348/0/0/0/113	324/0/0/0/137
GLM-Z1	AIME	0.800	2802/10/0/0/756	2633/10/0/0/925	2834/10/0/0/724	2886/10/0/0/672	2916/10/0/0/642	2860/10/0/0/698	2726/10/0/0/832
R1D-Llama (20k)	AIME	0.150	591/14/0/0/75	518/14/0/0/148	518/14/0/0/148	491/14/2/0/173	551/14/0/0/115	560/14/0/0/106	482/14/0/0/184
Qwen3	HLE	0.140	523/29/0/0/72	460/29/0/0/135	467/29/0/0/128	444/29/0/0/151	475/29/0/0/120	500/29/0/0/95	464/29/0/0/131
QwQ	HLE	0.110	419/24/0/0/59	384/24/0/0/94	355/24/0/0/123	370/24/2/0/106	357/24/0/0/121	389/24/0/0/89	330/24/0/0/148
GLM-Z1	HLE	0.900	3223/247/0/0/533	3062/247/0/0/694	3160/248/0/0/596	3199/247/0/0/557	3039/247/0/0/717	3245/248/0/0/511	3143/247/0/0/613
R1D-Llama (20k)	HLE	0.235	783/143/0/0/124	716/143/0/0/191	667/143/0/0/240	720/143/0/0/187	755/143/0/0/152	783/143/0/0/124	739/143/1/0/167
R1D-Qwen (32k)	AIME	0.075	312/1/0/0/58	293/1/0/0/77	271/1/0/0/99	267/1/0/0/103	270/1/0/0/100	308/1/2/0/60	269/1/0/0/101
R1D-Qwen (32k)	HLE	0.075	281/23/0/0/67	256/23/3/0/89	235/23/0/0/113	247/23/1/0/100	249/23/0/0/99	247/23/5/0/96	216/23/7/0/125
Nemotron-H	AIME	0.200	800/0/0/0/119	791/0/0/0/128	787/0/0/0/132	784/0/0/0/135	783/0/0/0/136	780/0/0/0/139	804/0/0/0/115
Nemotron-H	HLE	0.300	1169/35/0/0/131	1204/35/0/0/96	1180/35/0/0/120	1153/35/0/0/147	1185/35/0/0/115	1156/35/0/0/144	1198/35/0/0/102

binning reps and complete paired effect reps for every policy.

The primary metric is accepted-request breach. It excludes pre-service HTTP-400 maximum-context validation rejects from the eligible offered set, retains scheduler deadline expirations as failed accepted requests, and invalidates replications containing infrastructure errors.

To prevent regression to the mean from manufacturing an ordering, the primary analysis uses only LIFO seeds 42–43 to assign each cell to the feasible ($< 15\%$), transition (15–30%), or failed ($> 30\%$) stratum. It then uses only paired seeds 44–46 to estimate each policy-minus-LIFO difference. The deep-overload ($> 40\%$) stratum is a subset of the failed stratum, not an additional independent group. Within each stratum, we first average all deadline-load cells belonging to a model–dataset variant and then weight the variants equally. Intervals are 95% Student- t intervals over these variant means; sign tests use the same variant as their unit.

As a split-sensitivity check, we repeat the analysis for all $\binom{5}{2} = 10$ choices of two binning seeds and the complementary three effect seeds. Each fold preserves independence between assignment and measurement. We average folds within variant before inference; these folds reuse the five traces and therefore are a robustness analysis, not ten new

independent experiments. Table 12 shows that the main directions and magnitudes are stable.

Immediate request errors. The legacy trace parser classified every row with an elapsed_s value as a completion. Our audit found 80369 HTTP-400 context-limit validation rows, 14722 deadline-timeout rows, and 0 infrastructure-error rows. Deadline timeouts are accepted requests and remain failures under every corrected definition. Among the 633 model–dataset–load–seed groups containing a validation reject, exact rejected task-ID sets are identical across policies in 630 and differ in 3; we therefore do not assume invariance silently. Replications containing infrastructure failures are invalidated rather than scored as completions.

Table 13 compares the corrected primary with two transparent alternatives. “End-to-end” counts every observed error as a failed offered request. The uncorrected legacy parser treated error rows with elapsed times as completions. Relative to that parser, accepted-request scoring changes the stratum of 17 of 1230 common cells (“1.4%”). We use the corrected accepted-request definition as primary and report legacy values only as an audit sensitivity.

Table 8: Policy-minus-LIFO breach deltas under three explicit error treatments. Each entry is A/C/E in percentage points: accepted-request breach over valid reps (primary), compatibility parsing that treats elapsed error rows as on-time, and end-to-end failure that counts every error as a failure.

Model	DS	λ^*	LIFO	SKV	InLen	LKV	Ent	Orac	Rnd
Qwen3	AIME	0.100	+0.0/+0.0/+0.0	+10.6/+10.6/+10.6	+9.2/+9.2/+9.2	+15.1/+15.1/+15.1	+10.1/+10.1/+10.1	+7.4/+7.4/+7.4	+16.2/+16.2/+16.2
QwQ	AIME	0.100	+0.0/+0.0/+0.0	+11.0/+11.0/+11.0	+14.4/+14.4/+14.4	+12.4/+12.4/+12.4	+5.2/+5.2/+5.2	+14.6/+14.6/+14.6	+19.3/+19.3/+19.3
GLM-Z1	AIME	0.800	+0.0/+0.0/+0.0	+4.8/+4.8/+4.8	-0.9/-0.9/-0.9	-2.3/-2.3/-2.3	-3.2/-3.2/-3.2	-1.6/-1.6/-1.6	+2.2/+2.2/+2.2
R1D-Llama (20k)	AIME	0.150	+0.0/+0.0/+0.0	+11.1/+10.9/+10.9	+11.1/+10.9/+10.9	+14.7/+14.4/+14.4	+6.0/+5.9/+5.9	+4.6/+4.5/+4.5	+16.6/+16.3/+16.3
Qwen3	HLE	0.140	+0.0/+0.0/+0.0	+10.0/+9.6/+9.6	+9.3/+8.9/+8.9	+12.9/+12.4/+12.4	+7.5/+7.2/+7.2	+3.6/+3.4/+3.4	+9.6/+9.2/+9.2
QwQ	HLE	0.110	+0.0/+0.0/+0.0	+7.4/+7.0/+7.0	+13.3/+12.7/+12.7	+10.1/+9.6/+9.6	+12.8/+12.2/+12.2	+6.3/+6.0/+6.0	+18.6/+17.7/+17.7
GLM-Z1	HLE	0.900	+0.0/+0.0/+0.0	+4.2/+4.0/+4.0	+1.8/+1.7/+1.7	+0.8/+0.8/+0.8	+4.8/+4.5/+4.5	-0.5/-0.5/-0.5	+2.2/+2.1/+2.1
R1D-Llama (20k)	HLE	0.235	+0.0/+0.0/+0.0	+7.1/+6.2/+6.2	+12.8/+11.0/+11.0	+6.8/+6.0/+6.0	+2.5/+2.4/+2.4	-0.4/-0.2/-0.2	+4.7/+4.1/+4.1
R1D-Qwen (32k)	AIME	0.075	+0.0/+0.0/+0.0	+5.3/+5.3/+5.3	+11.6/+11.6/+11.6	+12.4/+12.4/+12.4	+11.9/+11.9/+11.9	+1.3/+1.3/+1.3	+11.5/+11.5/+11.5
R1D-Qwen (32k)	HLE	0.075	+0.0/+0.0/+0.0	+6.9/+6.5/+6.5	+12.7/+12.0/+12.0	+9.1/+8.7/+8.7	+9.1/+8.5/+8.5	+9.7/+9.1/+9.1	+17.7/+16.8/+16.8
Nemotron-H	AIME	0.200	+0.0/+0.0/+0.0	+1.0/+1.0/+1.0	+1.5/+1.5/+1.5	+1.7/+1.7/+1.7	+1.9/+1.9/+1.9	+2.2/+2.2/+2.2	-0.4/-0.4/-0.4
Nemotron-H	HLE	0.300	+0.0/+0.0/+0.0	-2.8/-2.7/-2.7	-1.1/-1.0/-1.0	+1.1/+1.1/+1.1	-1.3/-1.3/-1.3	+1.0/+1.0/+1.0	-2.4/-2.4/-2.4

Table 9: Best-to-worst point-estimate rankings under the same three error treatments. ID check reports the number of seeds with exactly identical failing task-ID sets across all seven policies, followed by the minimum all-policy intersection-over-union. Context-rejection sets are therefore not assumed policy-invariant.

Model	DS	Accepted	Compatibility	All errors fail	ID check
Qwen3	AIME	LIFO<Orac<InLen<Ent<SKV<LKV<Rnd	LIFO<Orac<InLen<Ent<SKV<LKV<Rnd	LIFO<Orac<InLen<Ent<SKV<LKV<Rnd	4/5; 0.00
QwQ	AIME	LIFO<Ent<SKV<LKV<InLen<Orac<Rnd	LIFO<Ent<SKV<LKV<InLen<Orac<Rnd	LIFO<Ent<SKV<LKV<InLen<Orac<Rnd	5/5; 1.00
GLM-Z1	AIME	Ent<LKV<Orac<InLen<LIFO<Rnd<SKV	Ent<LKV<Orac<InLen<LIFO<Rnd<SKV	Ent<LKV<Orac<InLen<LIFO<Rnd<SKV	5/5; 1.00
R1D-Llama (20k)	AIME	LIFO<Orac<Ent<SKV<InLen<LKV<Rnd	LIFO<Orac<Ent<InLen<SKV<LKV<Rnd	LIFO<Orac<Ent<InLen<SKV<LKV<Rnd	3/5; 0.50
Qwen3	HLE	LIFO<Orac<Ent<InLen<Rnd<SKV<LKV	LIFO<Orac<Ent<InLen<Rnd<SKV<LKV	LIFO<Orac<Ent<InLen<Rnd<SKV<LKV	5/5; 1.00
QwQ	HLE	LIFO<Orac<SKV<LKV<Ent<InLen<Rnd	LIFO<Orac<SKV<LKV<Ent<InLen<Rnd	LIFO<Orac<SKV<LKV<Ent<InLen<Rnd	3/5; 0.83
GLM-Z1	HLE	Orac<LIFO<LKV<InLen<Rnd<SKV<Ent	Orac<LIFO<LKV<InLen<Rnd<SKV<Ent	Orac<LIFO<LKV<InLen<Rnd<SKV<Ent	4/5; 0.98
R1D-Llama (20k)	HLE	Orac<LIFO<Ent<Rnd<LKV<SKV<InLen	Orac<LIFO<Ent<Rnd<LKV<SKV<InLen	Orac<LIFO<Ent<Rnd<LKV<SKV<InLen	4/5; 0.96
R1D-Qwen (32k)	AIME	LIFO<Orac<SKV<Rnd<InLen<Ent<LKV	LIFO<Orac<SKV<Rnd<InLen<Ent<LKV	LIFO<Orac<SKV<Rnd<InLen<Ent<LKV	4/5; 0.00
R1D-Qwen (32k)	HLE	LIFO<SKV<LKV<Ent<Orac<InLen<Rnd	LIFO<SKV<Ent<LKV<Orac<InLen<Rnd	LIFO<SKV<Ent<LKV<Orac<InLen<Rnd	1/5; 0.31
Nemotron-H	AIME	Rnd<LIFO<SKV<InLen<LKV<Ent<Orac	Rnd<LIFO<SKV<InLen<LKV<Ent<Orac	Rnd<LIFO<SKV<InLen<LKV<Ent<Orac	5/5; 1.00
Nemotron-H	HLE	SKV<Rnd<Ent<InLen<LIFO<Orac<LKV	SKV<Rnd<Ent<InLen<LIFO<Orac<LKV	SKV<Rnd<Ent<InLen<LIFO<Orac<LKV	5/5; 1.00

Mechanism decomposition. Table 14 separates the largest_kv-minus-LIFO effect into successful requests completing after the SLO, tracked unfinished requests, and post-acceptance errors. In deep overload at the tighter 600–1500 s SLOs, fewer late completions outweigh more unfinished requests and post-acceptance errors. At 1800 s the late-completion difference is zero and neither the failed nor deep-overload total differs detectably from zero. Thus the pooled reversal is a tight-deadline mechanism, not a universal consequence of crossing a breach threshold.

H Survivorship Correction

Table 15 contrasts the naive “breaches among completers” rate with the survivorship-corrected rate of Eq. (1). The naive metric hides the dumped backlog created by the deferring anti-patterns (input_length/largest_kv) and is what makes the correction essential to the policy ranking.

I Secondary Performance and Fairness Metrics

At the same accepted-request/audited-canonical λ^* used by the headline, within each rep let $A = N_{\text{accepted}}/900$ be accepted arrival rate, $T = N_{\text{successful}}/900$ successful-completion throughput,

and $G = N_{\text{successful, on-time}}/900$ on-time goodput, all in requests/s. The dimensionless accepted-request breach fraction in that rep is exactly $1 - G/A$; it is not $1 - G$. Reported breach percentages average these per-rep fractions, while reported rates average the per-rep rates. Table 17 reports all three rates for every reported model–dataset–policy cell: the 12 headline cells and the separately analyzed gpt-oss (48k) pair. Table 18 reports p50 and p99 end-to-end JCT over successful completions. TTFT was not logged; target selection can affect later admission and queueing, so we limit the latency claim to measured JCT. For fairness, E_i is the cumulative number of evictions of each accepted tracked request, including $E_i = 0$. Table 19 reports Gini(E) and the repeated-targeting tails $\Pr(E \geq 2)$ and $\Pr(E \geq 5)$.

J Repeated Preemption

Table 20 reports the per-request re-eviction distribution that underlies the smallest_kv selection-livelock analysis (§4.3): the extreme $\geq 10\times$ tail is a useful symptom of repeated targeting. It is not by itself a proof of the mechanism, so we interpret it together with the per-seed completion and dump trajectories.

Table 10: Complete accepted-breach reconciliation. Each entry is the primary mean over all audited-valid canonical traces / the automatic dump-breach-subset sensitivity. The subset sensitivity is not used for point estimates or inference because its exclusions do not implement the documented joint screen. A * or ‡ attached to either number is the paired marker recomputed for that treatment (worse or better than LIFO, respectively); only the first-number marker is used in the headline.

Model	DS	λ^*	LIFO	SKV	InLen	LKV	Ent	Orac	Rnd
Qwen3	AIME	0.100	16.9/16.9	27.5*/27.5*	26.1*/26.1*	32.0*/28.3	27.0/27.0	24.3/24.3	33.2*/33.2*
QwQ	AIME	0.100	10.1/10.1	21.0*/21.0*	24.4*/24.4*	22.4*/22.4*	15.3/15.3	24.7*/24.7*	29.4*/26.4*
GLM-Z1	AIME	0.800	21.1/21.1	25.9*/25.9*	20.3/20.3	18.8/18.8	18.0/18.0	19.5/19.5	23.3/23.3
R1D-Llama (20k)	AIME	0.150	11.2/11.2	22.3*/22.3*	22.3*/22.3*	25.9*/25.9*	17.2*/17.2*	15.8*/15.8*	27.8*/27.8*
Qwen3	HLE	0.140	12.1/9.0	22.2/22.2	21.5/21.5*	25.1*/25.1*	19.7/19.7	15.7/15.7	21.8*/21.8*
QwQ	HLE	0.110	12.0/12.0	19.4*/19.4*	25.3*/25.3*	22.1*/22.1*	24.8*/24.8*	18.3*/18.3*	30.6*/30.6*
GLM-Z1	HLE	0.900	14.0/14.0	18.3/18.3	15.8/15.8	14.8/14.8	18.9/18.9	13.5/13.5	16.2/16.2
R1D-Llama (20k)	HLE	0.235	13.8/13.8	20.9/20.9	26.6*/26.6*	20.6/20.6	16.3/16.3	13.4/13.4	18.4/18.4
R1D-Qwen (32k)	AIME	0.075	15.3/15.3	20.6/20.6	26.8/26.8	27.7*/27.7*	27.1/27.1	16.5/16.5	26.8*/26.8*
R1D-Qwen (32k)	HLE	0.075	19.4/19.4	26.3/26.3	32.1*/32.1*	28.5/28.5	28.5*/28.5*	29.1*/29.1*	37.1/32.1
Nemotron-H	AIME	0.200	12.9/12.9	13.9/13.9	14.4/14.4	14.5/14.5	14.8/14.8	15.0*/15.0*	12.5/12.5
Nemotron-H	HLE	0.300	10.2/10.2	7.4/7.4	9.1/9.1	11.3/11.3	8.9/8.9	11.2/11.2	7.7/7.7

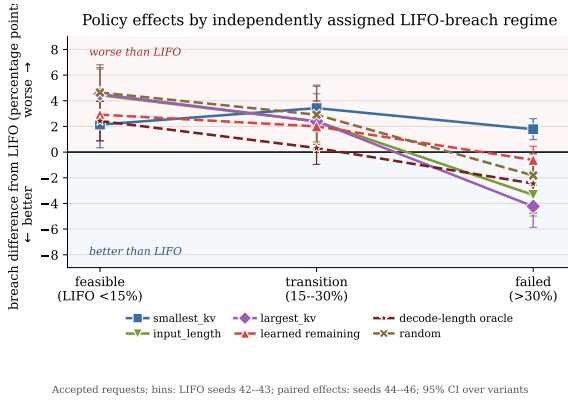


Figure 2: Policy-minus-LIFO accepted-breach effects by independently assigned regime (mean and variant-level 95% CI; positive is worse). LIFO seeds 42–43 assign regimes and paired seeds 44–46 estimate effects. The deep-overload stratum is a subset of the failed stratum. Memory-freeing policies are worse in the feasible regime but can improve under overload; smallest_kv remains worse on average. Fixed-schedule exploratory points are excluded.

K The smallest_kv Selection Livelock

In the diagnostic cells, smallest_kv ($\arg \min K_r$) is close to LIFO on aggregate recompute volume and typical target progress yet completes less work. Its distinctive failure is seed-sensitive: on R1D-Llama-70B-20k one seed serves only ~ 10 requests at a 94% dump rate while a healthy seed and LIFO serve ~ 125 – 135 (Figure 6). The event trace shows the score repeatedly returning to a churning low-progress victim set, which we call a scheduler-level *selection livelock*. The extreme $\geq 10\times$ re-eviction tail (Table 20) is a per-request symptom. Other policies do not show a comparable dump tail in

these cells, but we do not claim they are structurally immune, nor that every smallest_kv seed fails. This policy-specific behavior is retained as data and is separate from the policy-independent engine-install rerun protocol in Appendix D.

L Feasibility-Floor Sensitivity and Controls

L.1 Low-load floor sensitivity across deadlines

We ask whether the feasibility floor follows $P(T_{\text{natural}} > sr)$ over $s \in \{600, 900, 1200, 1500, 1800\}$, where T_{natural} is uncapped output length and r is single-stream decode throughput. The traces identify a closely related but different quantity: the configured workload’s low-load elapsed-time tail. Table 22 reports that quantity at the minimum recorded arrival rate for every family and both datasets. The SLO is only a post-hoc threshold on elapsed time, so no new run is needed for this re-scoring.

The distinction matters. The task traces contain end-to-end elapsed time, which includes pre-fill, continuous batching, queueing, and recomputation; they do not contain a separate single-stream decode-time measurement. Moreover, a request that hits max_tokens reveals only that its natural stop length is at least the configured cap, and a request killed at 1800 seconds records no completion length. Thus the existing traces do *not* identify the literal uncapped $P(T_{\text{natural}} > sr)$. Treating completion_tokens/elapsed_s as r and treating killed requests as observed lengths would make the proposed agreement circular. We instead report the identifiable low-load runtime tail and expose all

Table 11: Provenance for every numerical display in the camera-ready manuscript. “Mixed” means the displayed grid combines the submission’s operating band with post-submission load extensions; it does not mean that reruns are counted as additional replicates.

Paper material	Trace source	New GPU execution?	Camera-ready operation
Headline breach table and AIME/HLE load figures	420 seed-audited headline traces; seed-verified portions of the submitted operating band and post-submission extension for load figures	Mixed	Recompute the 10%-crossing over seed-verified loads and accepted-request means; use the canonical stall manifest; no outcome-based entropy-arm selection
Five-SLO regime table/figure, cross-fit robustness, and reversal decomposition	The seed-audited 248-point subset of the 20-variant load grid	No additional runs for the SLO axis	Exclude 53 fixed-schedule exploratory points; re-evaluate at 600–1800 s and split regime assignment from effect estimation
Confidence intervals, MDEs, raw/subset audit, and error-accounting tables	The same canonical headline traces	No	Paired re-aggregation and parser/adjudication sensitivity
Throughput, goodput, JCT, and fairness tables	Existing task, prompt, and preemption logs at the headline points	No	Read additional metrics from stored logs; no serving rerun
Naive-versus-corrected, eviction-depth, repeated-victim, and selection-livelock diagnostics	Existing task and preemption-event logs at the named diagnostic loads	No	Re-aggregate stored outcomes or event distributions
Entropy-predictor parameter table	Harness configuration, scheduler code, and calibration notes	No	Configuration audit; parameters were not independently snapshotted in trace metadata
Feasibility-floor sensitivity and capped-model intervention	Existing low-load native and capped-model traces	No for the camera-ready analysis	Re-score low-load elapsed outcomes at five deadlines and audit cap censoring
Llama-3.3-70B boundary control	Ten new default-policy cells (five loads on each dataset), one seed per cell	Yes (post-review)	Report as a negative/boundary control, not as a seven-policy comparison
Engine-stall verification	Same-seed reruns plus preserved original executions	Yes, as integrity reruns	Do not increase n ; retain reproduced stalls as data and use the healthy execution when the same-seed pair does not reproduce the stall

censoring in Table 23.

The multi-SLO pattern remains operationally useful. At native context, R1D-Qwen’s observed tail is 37.5% (AIME) and 43.2% (HLE) even at 1800 s; R1D-Llama’s is 31.1% and 21.2%. Under the 32k/20k serving caps, all four model–dataset cells have zero observed misses at 1200–1800 s in their minimum-load runs, although they can miss tighter 600–900 s deadlines. This supports decode-budget control as an intervention for the *configured* 1800-s workload. It does not establish that truncation removes the model’s natural long-output tail.

The minimum-load evidence is intentionally not over-aggregated. The “runs/seeds” column reports both execution repetitions and distinct Poisson seeds: several HLE pilot cells repeat seed 42, and the Nemotron-H and capped R1D-Llama minima

have only one run. Some R1D minima also contain preemptions (Table 23), so their low-load miss rate is an operational floor estimate, not a contention-free single-stream measurement.

L.2 Decode-cap intervention and the Llama control

A decode cap and a natural length distribution play different roles. A cap above an SLO token budget merely *permits* an over-budget generation; a nonzero floor additionally requires probability mass in the natural runtime tail. Conversely, a cap below a nominal token budget truncates generation but does not prove that the uncapped tail is empty, nor does it guarantee a deadline under queueing or reduced batch throughput. Table 23 makes this explicit. The native R1D cells are deadline-censored

Table 12: Independent-split and all-splits estimates (mean and variant-level 95% CI; positive is worse for the row’s first policy/cohort). Young is {LIFO, smallest_kv, learned remaining, decode-length oracle}; not-young is {input_length, random, largest_kv}. Cohort contrasts are formed within cell before averaging cells within variant.

Stratum	Comparison (pp)	Fixed 42–43 / 44–46	All 10 splits
Feasible	largest_kv–LIFO	+4.52 [+2.46, +6.58]	+3.96 [+2.01, +5.90]
Feasible	input_length–LIFO	+4.43 [+2.40, +6.47]	+3.82 [+1.97, +5.67]
Feasible	learned remaining–LIFO	+2.92 [+0.87, +4.96]	+2.14 [+0.73, +3.54]
Feasible	decode-length oracle–LIFO	+2.43 [+0.88, +3.97]	+1.98 [+0.83, +3.12]
Failed	largest_kv–LIFO	−4.23 [−5.86, −2.59]	−3.99 [−5.69, −2.28]
Failed	smallest_kv–LIFO	+1.78 [+0.96, +2.60]	+2.21 [+1.53, +2.88]
Feasible	Not-young – young	+2.66 [+1.42, +3.90]	+2.42 [+0.93, +3.91]
Feasible	largest_kv – random	−0.12 [−1.66, +1.42]	−0.17 [−1.75, +1.42]
Feasible	learned – decode-length	+0.49 [−0.96, +1.95]	+0.16 [−0.84, +1.16]

Table 13: Request-error sensitivity (fixed independent split; mean and variant-level 95% CI). Accepted-request scoring is primary: it removes only pre-service context-validation rejects, retains deadline timeouts and other accepted failures in both the eligible set and failure numerator, and invalidates replications containing infrastructure failures. The legacy column is included only to quantify the old parser’s error treatment.

Stratum	Comparison (pp)	Accepted (primary)	End-to-end	Legacy parser
Feasible	largest_kv–LIFO	+4.52 [+2.46, +6.58]	+4.19 [+2.26, +6.12]	+4.38 [+2.38, +6.38]
Feasible	decode-length oracle–LIFO	+2.43 [+0.88, +3.97]	+2.37 [+0.83, +3.91]	+2.38 [+0.96, +3.80]
Failed	largest_kv–LIFO	−4.23 [−5.86, −2.59]	−4.06 [−5.69, −2.43]	−4.23 [−5.84, −2.61]
Feasible	Not-young – young	+2.66 [+1.42, +3.90]	+2.53 [+1.31, +3.76]	+2.56 [+1.34, +3.77]

on 21–42% of accepted requests, while the capped variants terminate 13–51% of accepted requests at the length cap. Both mechanisms hide the uncapped natural tail, but only the cap changes the deployed workload and removes its observed 1800-s misses here. All percentages exclude immediate HTTP 400 validation failures. In the HLE minimum-load cells, the excluded counts are one each for native R1D-Qwen and R1D-Llama, three for capped R1D-Qwen, and eleven for capped R1D-Llama; Nemotron-H has none. Table 22 reports the corresponding rejection count for every row.

We also swept Llama-3.3-70B-Instruct, the non-distilled base of R1D-Llama-70B, with default/LIFO on both AIME and HLE at $\lambda \in \{0.1, 0.2, 0.4, 0.8, 1.2\}$. Each of the ten cells is one run with Poisson seed 42 and task seed 42. Median stop lengths range from 811–861 tokens on AIME and 669–770 on HLE; no successful output hits the 130k decode cap, and the server records zero preemptions in all ten cells. Of 2317 offered requests per dataset, 2309 AIME and 2220 HLE requests return an observed stop. Eight AIME and two HLE requests are killed at 1800 s. Excluding immediate admission rejections, the observed tails are therefore $8/2317 = 0.35\%$ and $2/2222 = 0.09\%$: approximately zero, not exactly zero. A further 95 HLE requests are rejected before generation

because max_tokens plus prompt length exceeds the 131,072 context limit; they are not completion outcomes. Accordingly, every *returned generation* stops naturally, but it would be incorrect to say that every offered request does.

This sweep is a boundary-condition check for the length-tail account, not a second causal matched pair for training recipe: the outputs are roughly an order of magnitude shorter, the cache never reaches the preemption path, and the victim selector is never exercised. Its large 130k cap together with its tiny observed deadline tail also directly refutes the stronger claim that “cap above budget” is sufficient to predict a floor.

Provenance. Table 22 is an unfiltered post-hoc re-scoring of the existing default-policy task traces; it adds no GPU runs. The Llama-3.3 sweep is a separate ten-cell control measurement. Exact paths, per-run values, outcomes, and censoring counts are generated by analyses/floor_audit.py and recorded in analyses/floor_audit.json.

M Cross-Model Drivers

The feasibility floor is a deadline-relative completion-time property. A request contributes to the natural floor when its single-stream completion time exceeds the evaluated SLO. Token length divided by effective decode rate is a useful approx-

Table 14: Mechanism decomposition for largest_kv minus LIFO under accepted-request scoring (percentage points; mean [95% CI]). Regimes use LIFO seeds 42–43; components use paired seeds 44–46. Cells are averaged within model–dataset variant and intervals are over equal-weight variant means. “Unfinished” denotes tracked requests aborted at the run cap; post-accept errors include scheduler deadline expirations. Pre-service validation rejects are excluded and replications containing infrastructure errors are invalidated. Positive values are worse for largest_kv.

SLO (s)	LIFO regime	Cells/variants	Δ late	Δ unfinished	Δ post-accept error	Δ total
600	deep >50%	179/18	−9.54 [−12.69, −6.38]	+3.08 [+0.57, +5.59]	+0.67 [−0.01, +1.34]	−5.79 [−7.49, −4.09]
900	deep >50%	158/18	−9.57 [−12.88, −6.25]	+2.87 [+0.05, +5.68]	+0.65 [−0.03, +1.33]	−6.05 [−8.17, −3.93]
1200	deep >50%	136/18	−9.27 [−13.37, −5.17]	+2.36 [−0.63, +5.36]	+0.75 [−0.04, +1.53]	−6.16 [−9.13, −3.18]
1500	deep >50%	102/17	−7.80 [−11.91, −3.70]	+1.21 [−2.31, +4.73]	+0.48 [−0.33, +1.29]	−6.12 [−8.42, −3.81]
1800	failed 30-50%	74/18	+0.00 [+0.00, +0.00]	+1.43 [−2.22, +5.08]	+0.58 [−0.07, +1.23]	+2.01 [−1.71, +5.73]
1800	deep >50%	44/15	+0.00 [+0.00, +0.00]	−0.62 [−5.31, +4.08]	+0.46 [−0.66, +1.57]	−0.16 [−4.74, +4.42]

Table 15: Completer-only late fraction versus accepted-request breach for a separate diagnostic at $\lambda \in \{0.055, 0.060\}$ requests/s, the upper end of this diagnostic window; these are not the headline λ^* cells. Values are unweighted means over five distinct seeds and the two arrival rates. The completer-only metric excludes cap-aborted accepted requests and therefore makes backlog-deferring policies look harmless; accepted breach includes that unfinished backlog. Pre-admission validation rejections are excluded from both metrics.

Cell	Policy	Completer-only%	Accepted breach%
QwQ AIME	LIFO	0.0	1.3
	smallest_kv	0.0	1.6
	input_length	0.0	8.7
	largest_kv	0.0	9.9

imation, but neither a large context window nor a configured decode cap determines the realized tail. Qwen3-32B and R1D-Qwen-32B provide an associated matched-backbone contrast, but recipe and context budget remain confounded; we therefore do not infer that distillation alone causes the floor. The stronger intervention is within model: reducing the R1D decode budget removes configured 1800 s misses in the measured low-load cells. Conversely, the non-distilled Llama-3.3-70B control has a large context but short natural outputs, an approximately zero observed accepted-request 1800-s tail, and no preemption in its measured sweep. Appendix L.1 reports the multi-deadline comparison and Appendix L.2 documents this control. *Architecture modulates the anti-pattern penalty.* The largest_kv–LIFO gap is larger for several dense transformers than for Nemotron-H, whose Mamba layers carry no KV cache. This is consistent with a smaller KV-size lever, but the architecture comparison is observational rather than an isolated causal test.

Table 16: Secondary metrics at the headline operating points, unweighted over 12 model–dataset cells. ΔT , ΔG , and JCT deltas are relative to LIFO; JCT is over successful completions. Fairness includes all accepted requests, including zero evictions. Gini concentration and the repeated-targeting tails answer different questions: LIFO and smallest_kv have the largest $E \geq 5$ tails, while input-length and largest-KV have the largest Gini coefficients. Accepted breach equals one minus normalized goodput G/A , not one minus goodput in requests/s. Full per-cell values are in Appendix I.

Policy	ΔT (req/s)	ΔG	$\Delta p50$ (s)	$\Delta p99$	Gini	$E \geq 2$	$E \geq 5$
LIFO	+0.000	+0.000	+0	+0	0.56	54	35
smallest_kv	−0.013	−0.013	+8	+10	0.56	55	35
input_length	−0.010	−0.010	−135	−73	0.83	22	10
largest_kv	−0.009	−0.009	−136	−71	0.78	12	2
entropy	−0.007	−0.007	−63	+1	0.47	56	13
oracle	−0.003	−0.003	−110	+5	0.54	52	11
random	−0.013	−0.013	−109	−16	0.60	27	1

N Additional Figures

Figure 3 shows the preemption-target progress mechanism behind Finding 1 on AIME. Figures 4 and 5 replicate the AIME breach-vs-load and preemption-target progress panels on HLE; they show the same qualitative cohort pattern on this more heterogeneous workload. Figure 6 details the smallest_kv selection livelock and Figure 7 the R1D-Llama feasibility floor at native context.

Table 17: Per-cell rates at the accepted-request/audited-canonical operating point. Each entry is accepted arrival rate / successful-completion throughput / on-time goodput, all in requests/s and averaged over reps. Throughput and goodput coincide here because no successful completion in these selected cells exceeded 1800 s; accepted deadline failures and cap-aborted requests still lower normalized goodput. All displayed entries use five canonical reps.

Model	DS	λ^*	LIFO	SKV	InLen	LKV	Ent	Orac	Rnd
Qwen3	AIME	0.100	0.102/0.085/0.085	0.102/0.074/0.074	0.102/0.076/0.076	0.102/0.069/0.069	0.102/0.074/0.074	0.102/0.077/0.077	0.102/0.068/0.068
QwQ	AIME	0.100	0.102/0.092/0.092	0.102/0.080/0.080	0.102/0.077/0.077	0.102/0.079/0.079	0.102/0.086/0.086	0.102/0.077/0.077	0.102/0.072/0.072
GLM-Z1	AIME	0.800	0.791/0.623/0.623	0.791/0.585/0.585	0.791/0.630/0.630	0.791/0.641/0.641	0.791/0.648/0.648	0.791/0.636/0.636	0.791/0.606/0.606
R1D-Llama (20k)	AIME	0.150	0.148/0.131/0.131	0.148/0.115/0.115	0.148/0.115/0.115	0.148/0.109/0.109	0.148/0.122/0.122	0.148/0.124/0.124	0.148/0.107/0.107
Qwen3	HLE	0.140	0.132/0.116/0.116	0.132/0.102/0.102	0.132/0.104/0.104	0.132/0.099/0.099	0.132/0.106/0.106	0.132/0.111/0.111	0.132/0.103/0.103
QwQ	HLE	0.110	0.106/0.093/0.093	0.106/0.085/0.085	0.106/0.079/0.079	0.106/0.082/0.082	0.106/0.079/0.079	0.106/0.086/0.086	0.106/0.073/0.073
GLM-Z1	HLE	0.900	0.835/0.716/0.716	0.835/0.680/0.680	0.835/0.702/0.702	0.835/0.711/0.711	0.835/0.675/0.675	0.835/0.721/0.721	0.835/0.698/0.698
R1D-Llama (20k)	HLE	0.235	0.202/0.174/0.174	0.202/0.159/0.159	0.202/0.148/0.148	0.202/0.160/0.160	0.202/0.168/0.168	0.202/0.174/0.174	0.202/0.164/0.164
R1D-Qwen (32k)	AIME	0.075	0.082/0.069/0.069	0.082/0.065/0.065	0.082/0.060/0.060	0.082/0.059/0.059	0.082/0.060/0.060	0.082/0.068/0.068	0.082/0.060/0.060
R1D-Qwen (32k)	HLE	0.075	0.077/0.062/0.062	0.077/0.057/0.057	0.077/0.052/0.052	0.077/0.055/0.055	0.077/0.055/0.055	0.077/0.055/0.055	0.077/0.048/0.048
Nemotron-H	AIME	0.200	0.204/0.178/0.178	0.204/0.176/0.176	0.204/0.175/0.175	0.204/0.174/0.174	0.204/0.174/0.174	0.204/0.173/0.173	0.204/0.179/0.179
Nemotron-H	HLE	0.300	0.289/0.260/0.260	0.289/0.268/0.268	0.289/0.262/0.262	0.289/0.256/0.256	0.289/0.263/0.263	0.289/0.257/0.257	0.289/0.266/0.266
gpt-oss (48k)	AIME	0.350	0.351/0.304/0.304	0.351/0.309/0.309	0.351/0.295/0.295	0.351/0.292/0.292	0.351/0.303/0.303	0.351/0.308/0.308	0.351/0.292/0.292
gpt-oss (48k)	HLE	0.175	0.177/0.155/0.155	0.177/0.158/0.158	0.177/0.152/0.152	0.177/0.140/0.140	0.177/0.150/0.150	0.177/0.152/0.152	0.177/0.150/0.150

Table 18: Per-cell successful-completion JCT. Each entry is p50/p99 in seconds, computed within each rep over successful task rows and then averaged over reps. Validation and infrastructure errors are not completions. TTFT was not logged, so these results characterize end-to-end JCT only. All displayed entries use five canonical reps.

Model	DS	λ^*	LIFO	SKV	InLen	LKV	Ent	Orac	Rnd
Qwen3	AIME	0.100	1079/1748	1017/1771	664/1661	577/1593	930/1766	610/1751	713/1741
QwQ	AIME	0.100	1010/1732	974/1757	718/1679	655/1603	771/1722	747/1730	766/1715
GLM-Z1	AIME	0.800	935/1769	902/1776	795/1724	775/1756	815/1777	880/1776	794/1764
R1D-Llama (20k)	AIME	0.150	949/1758	973/1768	620/1675	699/1638	792/1745	675/1707	784/1727
Qwen3	HLE	0.140	925/1714	910/1712	666/1682	753/1660	839/1778	844/1729	838/1729
QwQ	HLE	0.110	916/1714	865/1711	856/1688	694/1761	698/1724	779/1773	776/1708
GLM-Z1	HLE	0.900	812/1765	803/1764	730/1737	681/1615	824/1769	754/1751	759/1771
R1D-Llama (20k)	HLE	0.235	835/1763	766/1753	540/1608	569/1698	733/1747	825/1743	698/1749
R1D-Qwen (32k)	AIME	0.075	824/1750	995/1769	843/1579	785/1544	856/1729	561/1740	846/1684
R1D-Qwen (32k)	HLE	0.075	686/1746	814/1761	789/1480	1075/1664	916/1675	940/1766	639/1638
Nemotron-H	AIME	0.200	738/1699	765/1750	819/1759	729/1743	715/1736	747/1745	746/1729
Nemotron-H	HLE	0.300	735/1714	758/1696	781/1724	817/1750	797/1713	753/1722	776/1726
gpt-oss (48k)	AIME	0.350	520/1750	549/1725	360/1743	374/1617	373/1745	480/1736	362/1733
gpt-oss (48k)	HLE	0.175	468/1755	425/1749	420/1768	376/1650	487/1750	460/1764	436/1761

Table 19: Per-cell eviction fairness over all accepted tracked requests, including requests with zero evictions. Each entry is $\text{Gini}(E) / 100 \Pr(E \geq 2) / 100 \Pr(E \geq 5)$, where E is cumulative eviction count; percentages and Gini coefficients are computed per rep and then averaged. A high Gini means eviction burden is concentrated, while the two tail fractions distinguish repeated targeting. All displayed entries use five canonical reps.

Model	DS	λ^*	LIFO	SKV	InLen	LKV	Ent	Orac	Rnd
Qwen3	AIME	0.100	0.47/68/46	0.48/68/49	0.75/32/16	0.69/13/2	0.37/68/12	0.56/49/5	0.54/31/1
QwQ	AIME	0.100	0.46/66/53	0.48/65/44	0.78/29/14	0.73/13/2	0.39/68/12	0.51/53/6	0.53/33/1
GLM-Z1	AIME	0.800	0.49/68/59	0.48/69/63	0.86/20/14	0.88/8/2	0.41/70/27	0.46/65/16	0.57/32/2
R1D-Llama (20k)	AIME	0.150	0.46/74/46	0.46/67/37	0.79/29/10	0.66/20/2	0.38/71/18	0.51/60/25	0.53/37/2
Qwen3	HLE	0.140	0.45/70/50	0.44/73/53	0.76/31/17	0.76/14/2	0.35/73/31	0.37/74/21	0.54/35/1
QwQ	HLE	0.110	0.58/56/33	0.53/62/38	0.76/30/15	0.74/17/3	0.42/61/11	0.42/64/10	0.54/32/1
GLM-Z1	HLE	0.900	0.53/65/55	0.54/65/56	0.90/15/10	0.90/9/3	0.37/71/20	0.42/66/6	0.60/30/2
R1D-Llama (20k)	HLE	0.235	0.49/73/33	0.49/69/43	0.76/28/10	0.76/16/2	0.39/68/12	0.41/73/27	0.58/34/3
R1D-Qwen (32k)	AIME	0.075	0.57/53/25	0.55/53/23	0.81/23/10	0.72/16/1	0.49/58/5	0.57/47/6	0.62/25/1
R1D-Qwen (32k)	HLE	0.075	0.60/45/17	0.56/53/18	0.79/20/4	0.71/12/1	0.52/50/3	0.46/61/12	0.63/26/0
Nemotron-H	AIME	0.200	0.79/6/0	0.84/7/0	0.99/2/2	0.92/6/0	0.78/6/0	0.90/10/0	0.74/6/0
Nemotron-H	HLE	0.300	0.81/4/0	0.85/6/0	0.99/2/1	0.95/4/1	0.80/5/0	0.87/7/0	0.81/5/0
gpt-oss (48k)	AIME	0.350	0.80/25/11	0.82/23/14	0.93/9/5	0.86/2/0	0.79/16/0	0.72/29/0	0.83/6/0
gpt-oss (48k)	HLE	0.175	0.37/3/0	0.00/0/0	0.95/6/2	0.80/2/0	0.87/6/0	0.73/7/0	0.87/3/0

Table 20: Repeated preemption (GLM-Z1 AIME, $\lambda = 0.700$): fraction of requests evicted at least k times. random spreads evictions (almost none re-evicted); largest_kv concentrates on the largest computed prefixes (high Gini). LIFO and smallest_kv match at $\geq 5\times$, but the extreme tail diverges ($\geq 10\times$: 45% vs. 35%; the gap widens with k), the signature of the $\arg \min K_r$ starvation cycle that LIFO’s admission-timestamp criterion avoids.

Policy	$\% \geq 5\times$	$\% \geq 10\times$	$\% \geq 15\times$	$\% \geq 20\times$	max	Gini
default (LIFO)	79	35	12	2	29	0.31
smallest_kv	80	45	17	3	31	0.31
random	2	0	0	0	7	0.29
largest_kv	9	2	1	0	26	0.42
entropy	47	9	1	0	43	0.28
oracle	38	11	2	0	30	0.26

Table 21: Empirical 1800-s low-load tail (default/LIFO at the minimum recorded λ). The tail is an unfiltered mean over available executions; N is the number of accepted generations after excluding immediate HTTP 400 context-limit rejections. A † marks one pilot run; ‡ marks five executions of the same Poisson seed (42), not five arrival seeds. Rejection counts, full provenance, and five-SLO sensitivity appear in Appendix L.1. It estimates the configured workload’s low-load miss rate, not the uncapped natural-length tail: native R1D requests are deadline-censored, while the 32k/20k variants can end at their decode cap (Appendix L.2). Thus truncation removes the observed 1800-s misses in these cells, but does not show that the model’s natural long-output tail vanished.

Model	DS	Low-load tail%	N
Qwen3	AIME	0.0	84
QwQ	AIME	0.0	103
GLM-Z1	AIME	0.0	460
R1D-Llama (20k)	AIME	0.0 †	57
R1D-Qwen (32k)	AIME	0.0	102
Nemotron-H	AIME	6.7 †	15
Qwen3	HLE	0.0 ‡	75
QwQ	HLE	0.0 ‡	110
GLM-Z1	HLE	0.0 ‡	240
R1D-Llama (20k)	HLE	0.0 †	48
R1D-Qwen (32k)	HLE	0.0	100
Nemotron-H	HLE	0.0 †	15
<i>Untruncated bases (not in the main policy table):</i>			
R1D-Qwen (untrunc.)	AIME	37.5	58
R1D-Qwen (untrunc.)	HLE	43.2	57
R1D-Llama (untrunc.)	AIME	31.1	58
R1D-Llama (untrunc.)	HLE	21.2	57

Table 22: Observed low-load deadline-miss rate (% , default/LIFO), post-hoc re-scored at five SLOs. For each model–dataset we use the minimum recorded arrival rate and report execution replicas/distinct Poisson seeds. N is the accepted-generation denominator; “Reject” counts immediate HTTP 400 context-limit validation failures, which are excluded rather than treated as zero-latency completions. Entries are unfiltered means over replicas: we do not apply the engine-stall filter because a persistent low-load tail is the quantity being measured. These are tails of *configured low-load elapsed time*, not estimates of the uncapped natural-length tail; cap and deadline censoring are audited in Table 23.

Model	Data	$\lambda_{\min}$	Runs/seeds	N	Reject	600 s	900 s	1200 s	1500 s	1800 s
Qwen3-32B	AIME	.015	5/5	84	0	21.8	9.7	0.0	0.0	0.0
QwQ-32B	AIME	.020	5/5	103	0	15.9	6.4	0.0	0.0	0.0
GLM-Z1-32B	AIME	.100	5/5	460	1	4.5	0.0	0.0	0.0	0.0
gpt-oss-120B	AIME	.020	5/5	103	0	10.5	5.5	5.5	2.8	0.0
Nemotron-H-47B	AIME	.010	1/1	15	0	13.3	6.7	6.7	6.7	6.7
R1D-Qwen-32B (native)	AIME	.010	5/5	58	0	37.5	37.5	37.5	37.5	37.5
R1D-Llama-70B (native)	AIME	.010	5/5	58	0	37.7	35.1	32.4	31.1	31.1
R1D-Qwen-32B (32k cap)	AIME	.020	5/5	102	1	38.9	13.1	0.0	0.0	0.0
R1D-Llama-70B (20k cap)	AIME	.060	1/1	57	2	47.4	15.8	0.0	0.0	0.0
Qwen3-32B	HLE	.010	5/1	75	0	5.3	2.7	0.0	0.0	0.0
QwQ-32B	HLE	.020	5/1	110	0	15.5	3.6	0.0	0.0	0.0
GLM-Z1-32B	HLE	.050	5/1	240	20	6.7	0.0	0.0	0.0	0.0
gpt-oss-120B	HLE	.010	5/1	75	0	4.0	0.0	0.0	0.0	0.0
Nemotron-H-47B	HLE	.010	1/1	15	0	0.0	0.0	0.0	0.0	0.0
R1D-Qwen-32B (native)	HLE	.010	5/5	57	1	43.2	43.2	43.2	43.2	43.2
R1D-Llama-70B (native)	HLE	.010	5/5	57	1	21.2	21.2	21.2	21.2	21.2
R1D-Qwen-32B (32k cap)	HLE	.020	5/5	100	3	50.9	24.7	0.0	0.0	0.0
R1D-Llama-70B (20k cap)	HLE	.060	1/1	48	11	0.0	0.0	0.0	0.0	0.0

Table 23: Outcome and censoring audit. “Stop” is an observed natural stop; its p50 is in output tokens. “Cap-cens.” is finish_reason=length, so the natural stop length is unknown. “Deadline-cens.” is a request killed at the 1800-s harness deadline, so its natural completion time and length are unknown. N_{accepted} is Stop + cap-censored + deadline-censored; “Reject” reports immediate HTTP 400 context-length validation failures, which are outside that denominator and do not observe a generation. The R1D rows use the minimum-load cell from Table 22; the control rows aggregate its five-load, single-seed sweep.

Configuration	Data	λ	N_{accepted}	Stop	Stop p50	Cap-cens.	Deadline-cens.	Reject	Evictions
R1D-Qwen-32B (native)	AIME	.010	58	36	5,118	0	22	0	51
R1D-Qwen-32B (native)	HLE	.010	57	33	2,690	0	24	1	130
R1D-Llama-70B (native)	AIME	.010	58	41	6,381	0	17	0	16
R1D-Llama-70B (native)	HLE	.010	57	45	4,406	0	12	1	38
R1D-Qwen-32B (32k cap)	AIME	.020	102	62	4,695	40	0	1	0
R1D-Qwen-32B (32k cap)	HLE	.020	100	49	2,968	51	0	3	10
R1D-Llama-70B (20k cap)	AIME	.060	57	40	6,448	17	0	2	278
R1D-Llama-70B (20k cap)	HLE	.060	48	42	6,053	6	0	11	0
Llama-3.3-70B control	AIME	.1–1.2 (5)	2317	2309	842	0	8	0	0
Llama-3.3-70B control	HLE	.1–1.2 (5)	2222	2220	688	0	2	95	0

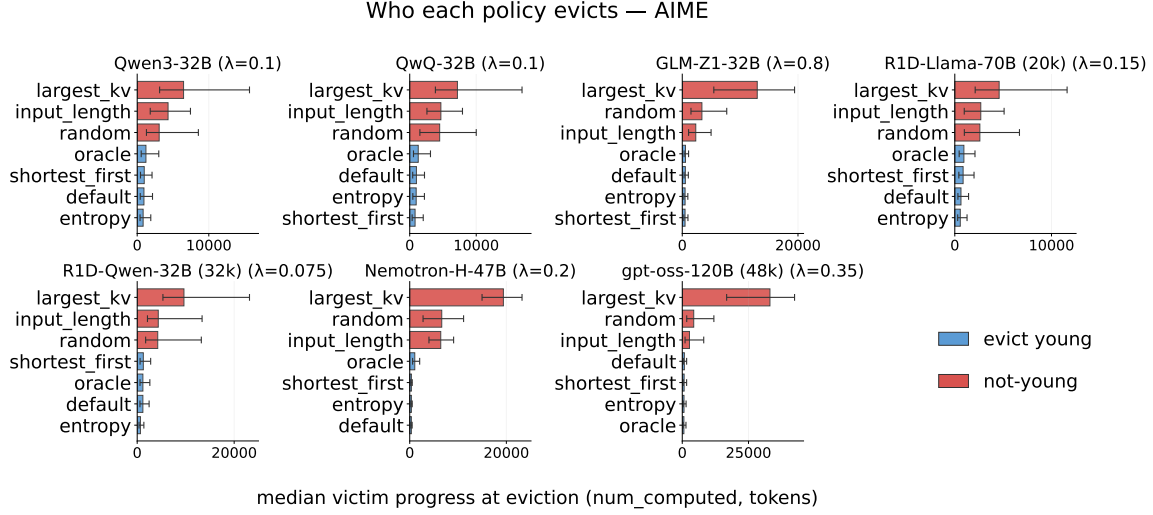


Figure 3: Median preemption-target progress at preemption per policy, all seven families, AIME (reported λ^* for each model); bars are the median preemption target’s computed-prefix length (logged num_computed) and whiskers span the 25th–75th percentiles. Bars are colored by *cohort*: blue = evict-young (LIFO/entropy/oracle/smallest_kv), red = not-young (input_length/random/largest_kv). Bar *height* is eviction depth: within the not-young cohort random evicts at intermediate depth and largest_kv/input_length deepest, yet all breach worse than LIFO (Table 2)—it is the young-vs-not-young split, not graded eviction depth, that tracks breach. The anti-pattern is muted on Nemotron-H, where most layers carry no KV.

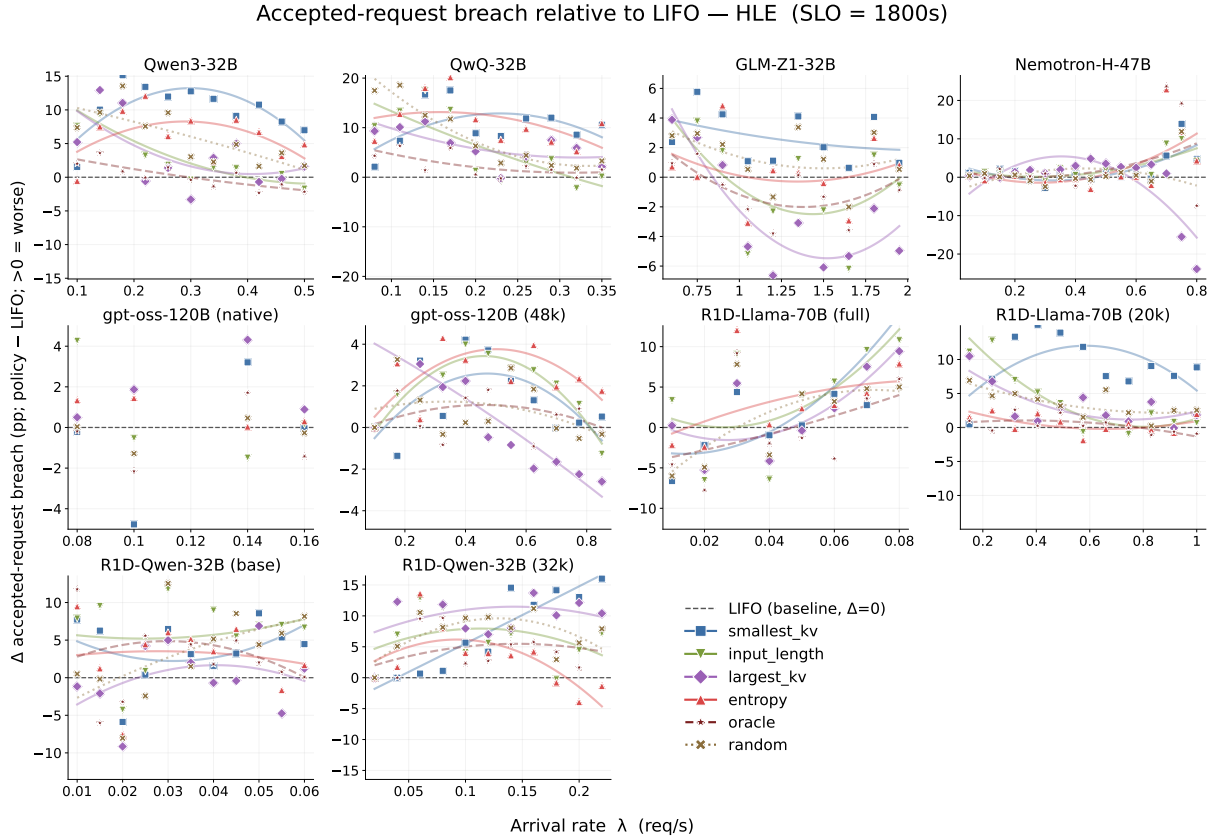


Figure 4: Accepted-request breach *relative to LIFO* (policy – LIFO; > 0 worse) vs. load, HLE, all model families and their truncated variants; LIFO is the dashed $y=0$ baseline. Companion to Figure 1; the same qualitative cohort split is visible on this more heterogeneous workload. Only complete load points with five audited arrival schedules are shown.

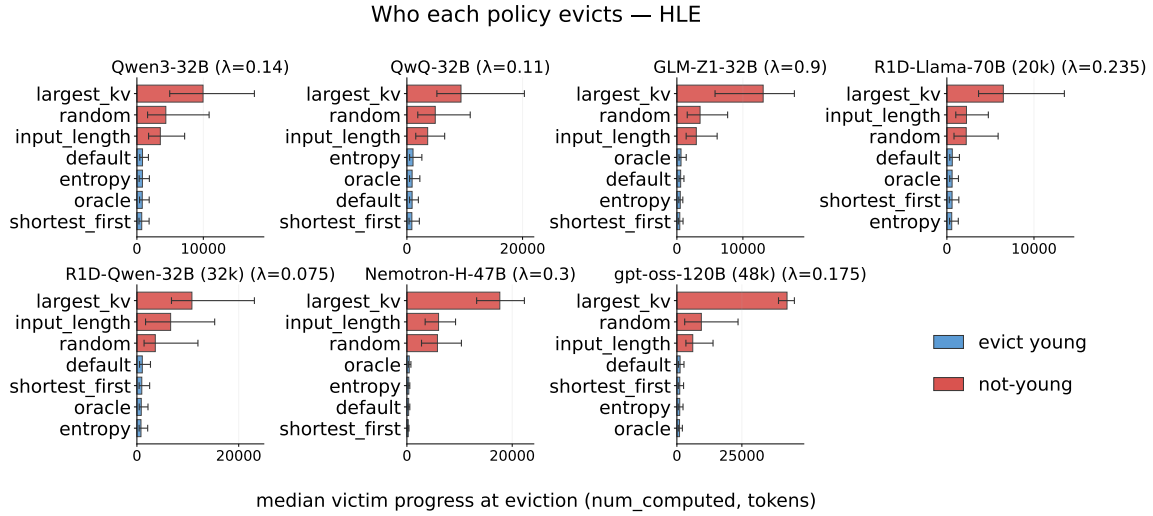


Figure 5: Median preemption-target computed-prefix length per policy, HLE (reported λ^* for each model; bar = median, whiskers = 25th–75th percentiles, color = eviction position). Companion to Figure 3.

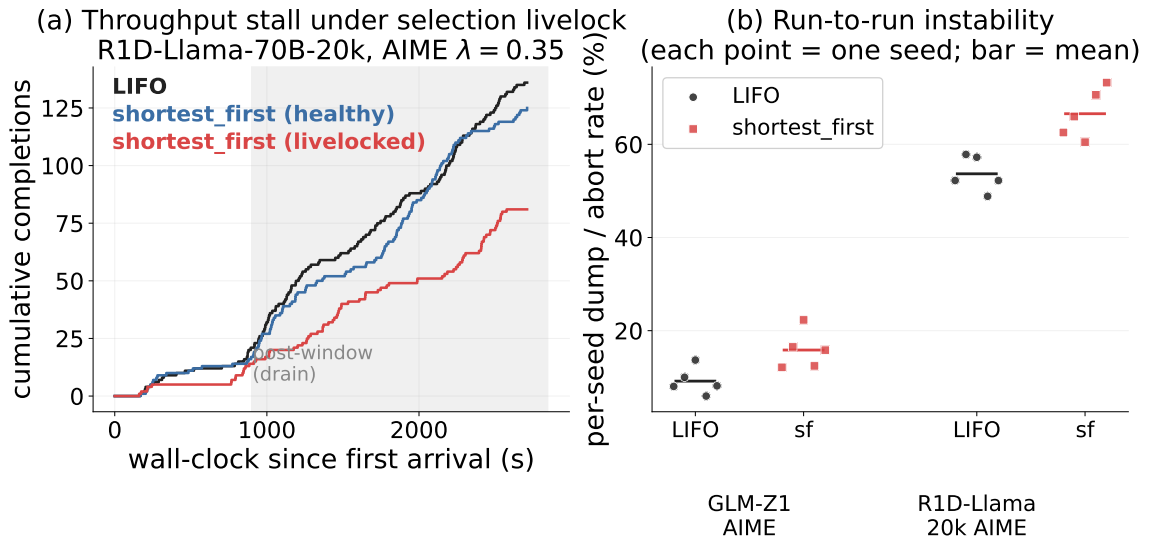


Figure 6: The smallest_kv selection livelock (R1D-Llama-70B, 20k budget, AIME). **(a)** Cumulative tracked completions vs. wall-clock: one smallest_kv seed enters selection livelock early (completion progress collapses to near zero) and finishes only ~ 10 requests, while a healthy smallest_kv seed and LIFO progress steadily to ~ 125 – 135 . **(b)** Per-seed abort (dump) rate: LIFO is tight across seeds whereas smallest_kv has a heavy upper tail (a seed at 94%), in both a dense (GLM-Z1) and a memory-tight (R1D-Llama-20k) cell. The failure is run-to-run instability, not a per-event difference.

R1D-Llama-70B, AIME: the feasibility floor and its removal by truncation

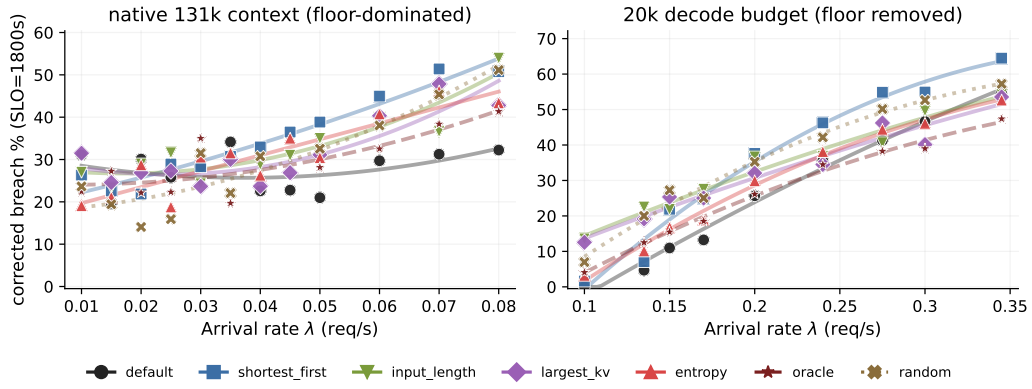


Figure 7: R1D-Llama-70B (AIME): the feasibility floor and its removal by truncation, in *absolute* corrected breach. Values are raw; the engine-stall protocol is not used to erase reproduced low-load breach. **Left, native 131k context:** breach is substantial already at near-idle load and roughly flat across the sweep—no policy escapes it, so the floor swamps any scheduling signal. **Right, 20k decode budget:** configured 1800 s misses disappear at low load and a scheduling regime emerges. LIFO is best at the 10%-crossing point; an isolated high-load oracle win and the broader deadline sensitivity are separated in §4.4.