RANKING IS NOT SUPPORT SELECTION: EXACT MARGINALS AND CERTIFIED HEADROOM FOR SPARSE ATTENTION IN LLMS

Anonymous author
Paper under double-blind review

ABSTRACT

Long-context large language model (LLM) inference often reduces attention memory or bandwidth by retaining or fetching only a limited subset of cached tokens. A common abstraction is to score candidates independently, keep the Top-K, and renormalize attention over the retained set. We ask a basic question: if those per-token scores were exact, would Top-K recover the best K-token support? In general, it does not. Even when every candidate is scored by its *exact downstream singleton-deletion Kullback–Leibler (KL) divergence* under the same next-token objective used to evaluate the retained set, an explicit Top-2 example incurs $133.2\times$ the KL of the optimal pair. The reason is structural: softmax renormalization makes the value of adding or replacing a token depend on which other tokens are retained. We derive the exact exchange identity and prove support-conditioned preference reversals for every $K \geq 2$. We then test this distinction by exhaustive support enumeration in softmax-attention heads of SmolLM2, Gemma-2-2B, and Qwen3.5-4B. At the primary $K = 4$ budget, exact-singleton Top-K leaves at least 10% relative headroom to the exact optimum within the same $M = 24$ candidate universe in 55/60, 49/60, and 60/60 prompt–head cells, respectively. The effect persists across alternative candidate-universe constructions, while its magnitude varies substantially across models. Direct mechanism analysis observes the predicted support-conditioned reversals in nearly every substantial-headroom case. Finally, SWAPCERT searches the same-budget one-swap neighborhood for a concrete witness of remaining headroom and detects 52/55, 48/49, and 57/60 exact positives at $\varepsilon = 0.10$. These results separate two questions that can be conflated in sparse LLM inference: which tokens are important individually, and which fixed-budget set best preserves the downstream next-token distribution.

1 INTRODUCTION

Autoregressive large language models (LLMs) repeatedly attend to previously processed tokens during generation, while serving systems cache their key and value representations so that these states can be reused across subsequent decoding steps rather than reconstructed from the full prefix (Vaswani et al., 2017; Kwon et al., 2023; Xiao et al., 2024). As context length grows, the key–value (KV) cache grows with the number of stored past positions and can become a major memory and data-movement bottleneck (Zhang et al., 2023; Ribar et al., 2024; Kwon et al., 2023). A broad family of long-context inference methods therefore reduces the active history by retaining, fetching, or processing only selected cached information, for example through heavy-hitter retention, selective history fetching, prompt-derived KV compression, or query-aware cache-page selection (Zhang et al., 2023; Ribar et al., 2024; Li et al., 2024; Tang et al., 2024).

At the level of an attention head, these approaches motivate a basic fixed-budget question: if only a limited part of the past can remain active, which candidates should be kept? A common design pattern is to assign each candidate an importance score and preferentially retain the highest-ranked candidates (Zhang et al., 2023; Li et al., 2024; Tang et al., 2024; Geng et al., 2025). We call the fixed-cardinality form of this abstraction *rank then retain*. If the scores are only approximate heuristics, a suboptimal retained set is unsurprising. The more fundamental question is whether ranking becomes sufficient when the individual scores themselves are exact.

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

There is a structural reason it need not. Softmax attention normalizes over the active keys (Vaswani et al., 2017); after part of the support is removed, the surviving weights are renormalized over the retained subset (Shi and Gu, 2026). Removing or replacing one token therefore changes not only that token's contribution, but also the normalized contribution of every survivor. The usefulness of a candidate can depend on which other candidates are retained. In ordinary terms, knowing exactly how important every item is *by itself* need not identify the best group of items. We isolate this distinction from errors in the importance estimator. Suppose an oracle reveals the exact downstream effect of removing each candidate individually, measured by the same next-token objective later used to evaluate complete retained sets. Would retaining the Top-K candidates under these exact singleton effects recover the best size-K support? We show that the answer is no. An objective-matched counterexample produces more than two orders of magnitude difference in loss between exact-singleton Top-K and the jointly optimal support. More generally, we show that the preference between two candidates can reverse when only the surrounding retained context changes. Thus independent marginal importance and jointly renormalized support quality are different mathematical objects.

This boundary is relevant because recent KV-selection work has increasingly improved the fidelity of individual importance estimates. AnDPro incorporates value-space geometry into token scoring, while DropKV derives residual-output perturbation scores and proves a constant-factor guarantee under an explicit geometric condition (Geng et al., 2025; Zhang et al., 2026). At the same time, subset structure is already recognized explicitly: BumbleBee formulates KV summarization as subset selection, and *Trust the Mass* studies retained-set optimization under forced attention renormalization (Kumari et al., 2024; Shi and Gu, 2026). So we do not say that token importance is useless or that set-aware KV selection is new. The missing boundary is more specific: *without additional structural assumptions, even exact objective-matched singleton importance does not generally reduce jointly renormalized fixed-budget downstream support selection to independent ranking*.

We then ask whether this theoretical distinction is merely pathological or also visible inside real LLMs. Exhaustive support evaluations across SmolLM2, Gemma-2-2B, and Qwen3.5-4B show that exact-singleton ranking can leave substantial room for improvement even though every individual score is already matched to the downstream objective. The effect persists across multiple support budgets and alternative candidate-universe constructions, although its magnitude varies substantially across models. Direct mechanism analysis also observes the support-conditioned preference reversals predicted by the theory.

Finally, we introduce SWAPCERT, a support-level diagnostic for a different practical question: given a support produced by any ranking rule, can we exhibit a concrete same-budget replacement that proves meaningful headroom remains? SWAPCERT searches the one-swap neighborhood and returns such a witness when one exists above a chosen threshold. A positive witness is mathematically sound, while a non-fire is deliberately inconclusive. In the exact audits, this inexpensive local test recovers most supports with substantial remaining candidate-universe headroom.

Contributions. We address five gaps: **(i) Objective-matched impossibility:** we show that even exact downstream singleton-deletion effects do not, in general, reduce jointly renormalized fixed-cardinality support selection to scalar Top-K ranking; **(ii) support-conditioned mechanism:** an exact exchange identity and an every-$K \geq 2$ preference-reversal construction identify retained context as the source of the mismatch, and direct LLM audits observe the predicted reversals; **(iii) exact multi-model evidence:** exhaustive experiments reveal substantial exact-singleton support headroom across three LLM families, multiple support budgets, and alternative candidate-universe constructions, while exposing strong model heterogeneity; **(iv) constructive headroom auditing:** SWAPCERT provides a mathematically sound positive witness of remaining same-budget headroom and high empirical recall against exact candidate-universe labels; and **(v) behavioral boundary:** one-step next-token decision crossings occur and are rescued in the observed cases, but the experiments do not establish population-level top-1 or free-running generation improvement.

2 RELATED WORK

Sparse attention and KV-cache selection in LLMs. Efficient long-context inference has motivated methods that avoid processing or retaining the complete attention history. H_2O retains heavy-hitter and recent KV entries (Zhang et al., 2023); StreamingLLM preserves attention sinks together with

a recent window (Xiao et al., 2024); SparQ selectively fetches cached history to reduce attention bandwidth (Ribar et al., 2024); SnapKV compresses per-head KV states using prompt attention patterns (Li et al., 2024); and Quest performs query-aware selection of critical KV-cache pages (Tang et al., 2024). These methods differ substantially in scoring, granularity, and systems implementation, but collectively motivate the fixed-budget question studied here: once only a subset can remain active, how should the retained set be chosen?

Beyond attention-weight ranking. Several recent methods improve the estimate of which individual KV entries matter. Output-perturbation analyses show that values and downstream projections can matter in addition to attention weights (Feng et al., 2026); AnDPro incorporates value-space geometry through anchor-direction projections (Geng et al., 2025); and DropKV derives independent residual-output perturbation scores together with a constant-factor approximation guarantee under an explicit cone condition (Zhang et al., 2026). These results motivate stronger marginal scores. Our result addresses a different question: even if the marginal score is replaced by the *exact downstream singleton effect*, what does that imply about the jointly renormalized fixed-budget support?

Set-aware support selection. The importance of set structure itself is established in prior work. BumbleBee formulates KV retention as submodular subset selection with support-dependent gains (Kumari et al., 2024). AnDPro begins from a retained-set/value-geometry optimization before deriving efficient scoring approximations (Geng et al., 2025). Trust the Mass studies forced-weight renormalization and exact restricted subset selection under a head-output reconstruction objective (Shi and Gu, 2026). Ada-KV addresses the distinct problem of allocating retention budgets adaptively across attention heads (Feng et al., 2025). Accordingly, our contribution is not the observation that KV selection can be combinatorial. We establish the complementary negative boundary: *without additional assumptions, exact objective-matched singleton ranking is still insufficient to recover the optimal jointly renormalized support under downstream next-token KL*, and we quantify the resulting headroom by exhaustive LLM experiments.

3 Problem Formulation and Theory

3.1 Fixed-budget support selection in an LLM attention head

We isolate the support-selection primitive in one standard softmax-attention head at one next-token prediction step (Vaswani et al., 2017). Let D denote the set of valid past-token positions available to the dense head. For position $i \in D$, let s_i be its pre-softmax attention logit, v_i its value vector, and

$$a_i = \frac{e^{s_i}}{\sum_{j \in D} e^{s_j}}$$

its dense attention probability. The dense head output is

$$h_D = \sum_{i \in D} a_i v_i. \tag{1}$$

A sparse selector retains a support $S \subseteq D$. Because ordinary softmax attention renormalizes over the surviving positions, the corresponding head output is

$$h_S = \frac{\sum_{i \in S} a_i v_i}{A_S}, \qquad A_S = \sum_{i \in S} a_i. \tag{2}$$

Thus selecting a support changes both which values remain and the normalized weight assigned to every survivor.

Let p_D be the model's dense next-token distribution. For any retained support S, let p_S be the next-token distribution obtained by replacing the target head output by h_S and propagating through the otherwise unchanged downstream network. We measure support distortion by

$$L(S) = D_{\mathrm{KL}}(p_D \| p_S). \tag{3}$$

Let $U \subseteq D$ be a candidate universe containing $M = |U|$ positions, and let $K < M$ be the retention budget. The feasible supports are

$$\mathcal{F}_K(U) = \{S \subseteq U : |S| = K\},$$

and the exact candidate-universe optimum is

$$S_U^\star \in \arg\min_{T \in \mathcal{F}_K(U)} L(T).$$

The common ranking abstraction instead assigns each candidate a scalar score and keeps its Top-K. To isolate the ranking abstraction from score-estimation error, we define each candidate's exact objective-matched singleton effect:

$$d_j = D_{\mathrm{KL}}\big(p_D \| p_{D\setminus\{j\}}\big), \qquad j \in U, \tag{4}$$

where only candidate j is deleted from the otherwise dense head and the remaining positions are renormalized. The central theoretical question is therefore

$$\mathrm{TopK}_K(\{d_j : j \in U\}) \overset{?}{\in} \arg\min_{S \in \mathcal{F}_K(U)} L(S).$$

The answer is no in general, as the following counterexample establishes.

3.2 An objective-matched counterexample

Take $a = (0.2, 0.1, 0.2, 0.5), \quad v = (4, -2, 1, 2.7), \quad h = 2.15$, with downstream logits $z(h) = (1.8h, 0)$. Exact singleton deletion gives

$$d_4 > d_1 > d_2 > d_3,$$

so exact-singleton Top-2 selects $\{1, 4\}$. Exhaustive evaluation of all six size-2 supports instead gives the joint optimum $\{3, 4\}$, with

$$L(\{1,4\}) = 0.01720822267, \qquad L(\{3,4\}) = 0.000129168300. \tag{5}$$

Thus exact-singleton Top-2 incurs $\frac{L(\{1,4\})}{L(\{3,4\})} \approx 133.2$ times the KL of the exact size-2 oracle despite using objective-matched, exact singleton scores.

Proposition 1 (Exact singleton effects are insufficient)**.** *Even an oracle providing every exact downstream singleton-deletion effect d_j does not in general reduce jointly renormalized fixed-cardinality support selection to scalar Top-K ranking.*

Appendix A gives the complete singleton calculations and exhaustive evaluation of all six size-2 supports.

3.3 Exchange value is support-conditioned

For $S' = S - \{i\} + \{j\}$,

$$h_{S'} = \frac{A_S h_S - a_i v_i + a_j v_j}{A_S - a_i + a_j}, \tag{6}$$

$$h_{S'} - h_S = \frac{a_j(v_j - h_S) - a_i(v_i - h_S)}{A_S - a_i + a_j}. \tag{7}$$

Thus the effect of exchanging i for j depends on the removed item i, the retained mass A_S, and the current support representation h_S, rather than on j through a context-independent scalar alone.

Proposition 2 (Support-conditioned preference reversal)**.** *For every $K \geq 2$, the same two candidate positions can reverse downstream-KL preference when only the other retained positions change. Hence a context-independent scalar ordering cannot generally represent all fixed-budget exchange preferences.*

Proof sketch. Use uniform attention on $K+2$ values consisting of $K-2$ zeros and the four special values $-2, +2, +1, -1$. Let C_- contain the shared zeros and -2, and let C_+ contain the same zeros and $+2$. Under a binary downstream map with dense output zero, the KL loss is strictly increasing in the magnitude of the retained-support output. In context C_-, adding $+1$ gives output $-1/K$ whereas adding -1 gives $-3/K$, so $+1$ is preferred. In context C_+, the corresponding outputs are $3/K$ and $1/K$, so the preference reverses and -1 is preferred. Thus the same two candidates reverse order solely because the retained context changes, for every $K \geq 2$. Appendix A gives the full calculation.

The proposition does not imply that scalar rankings are always poor; it shows that no context-independent scalar ordering can represent all fixed-budget exchange preferences without additional assumptions.

4 SWAPCERT: A ONE-SIDED HEADROOM DIAGNOSTIC

Assume $1 \le K < M = |U|$ and define the one-swap neighborhood

$$\mathcal{N}_1(S) = \{S - \{i\} + \{j\} : i \in S,\ j \in U \setminus S\}.$$

For $S \in \mathcal{F}_K(U)$ with $L(S) > 0$, define

$$H_U(S) = 1 - \frac{\min_{T \in \mathcal{F}_K(U)} L(T)}{L(S)}, \qquad W_U(S) = 1 - \frac{\min_{T \in \mathcal{N}_1(S)} L(T)}{L(S)}. \tag{8}$$

If $L(S) = 0$, we set $H_U(S) = W_U(S) = 0$. Thus $H_U(S)$ is the relative KL reduction available to the exact candidate-universe oracle, whereas $W_U(S)$ is the relative KL reduction available to the best one-swap neighbor. SWAPCERT $_\varepsilon$ fires when $W_U(S) \ge \varepsilon$ and returns that witness.

Theorem 1 (Positive-certificate soundness)**.** *For any $S \in \mathcal{F}_K(U)$ and $\varepsilon \in (0, 1)$,*

$$W_U(S) \ge \varepsilon \quad \Longrightarrow \quad H_U(S) \ge \varepsilon.$$

Because $\mathcal{N}_1(S) \subseteq \mathcal{F}_K(U)$,

$$\min_{T \in \mathcal{F}_K(U)} L(T) \le \min_{T \in \mathcal{N}_1(S)} L(T),$$

and therefore $H_U(S) \ge W_U(S)$, which proves the implication. At threshold ε, call a cell *headroom-positive* when $H_U(S) \ge \varepsilon$; SWAPCERT recall is the fraction of headroom-positive cells for which $W_U(S) \ge \varepsilon$. A non-fire is inconclusive.

Audit complexity. The complete one-exchange neighborhood contains $K(M-K)$ supports, versus $\binom{M}{K}$ for exhaustive candidate-universe search. At the primary $M = 24, K = 4$ setting, SWAPCERT inspects 80 neighbors while the exact oracle enumerates 10,626 supports—$132.8\times$ fewer support evaluations. This is a diagnostic-complexity comparison, not a serving-latency claim.

5 EXPERIMENTAL DESIGN

Models and intervention. We audit HuggingFaceTB/SmolLM2-1.7B-Instruct, google/gemma-2-2b, and Qwen/Qwen3.5-4B on 30 prespecified LongBench-E examples: 10 each from 2WikiMultihopQA, HotpotQA, and PassageRetrieval-en (Bai et al., 2024). For each model, the target is its deepest layer using ordinary full-softmax attention: zero-indexed layer 23 for SmolLM2, 25 for Gemma-2-2B, and 31 for Qwen3.5-4B. Recurrent or linear-attention blocks in hybrid architectures are excluded because the intervention is defined for softmax attention. The two head indices are prespecified as the first and last attention heads. We capture the dense attention probabilities and values, form the sparse-softmax head from retained original probabilities, map its displacement through the target head's output-projection block, inject that displacement at the attention-module boundary, and propagate through the unchanged remainder of the model. Independent recomputation reproduces the primary $K = 4$ KL values exactly; Appendix H documents a batch-shape numerical correction confined to the later behavioral analysis.

Primary exact $K = 4$ audit. The primary universe is the Top-24 valid (unmasked) token positions by dense attention weight. Each support is formed by the Top-K exact downstream singleton scores d_j from eq. (4). At $K = 4$ there are 60 cells per model, and every cell exhaustively evaluates all $\binom{24}{4} = 10{,}626$ supports: 1,912,680 exact size-4 support evaluations across the 180 primary cells. Thus "global" empirical headroom always means the *global* $M = 24$ *candidate-universe* quantity in eq. (8), never an unrestricted full-context optimum. The substantial-headroom threshold $\varepsilon = 0.10$ was specified before outcome evaluation. Prompt-cluster bootstrap 95% confidence intervals (CIs) resample prompts while keeping the two heads from each prompt together.

Supporting audits. We retain three prespecified extensions. First, candidate-universe sensitivity compares five outcome-independent $M = 24$ constructions, with two seeds for the mixed and random families. Second, exact $K \in \{2, 3\}$ cross-model enumeration provides a lower-budget replication. Third, exact $K = 5$ enumeration uses 12 cells per model and is descriptive only. 2K/4K swap-screen experiments are also reported as supporting evidence in Appendix J; they measure witnessed, not exhaustive, headroom.

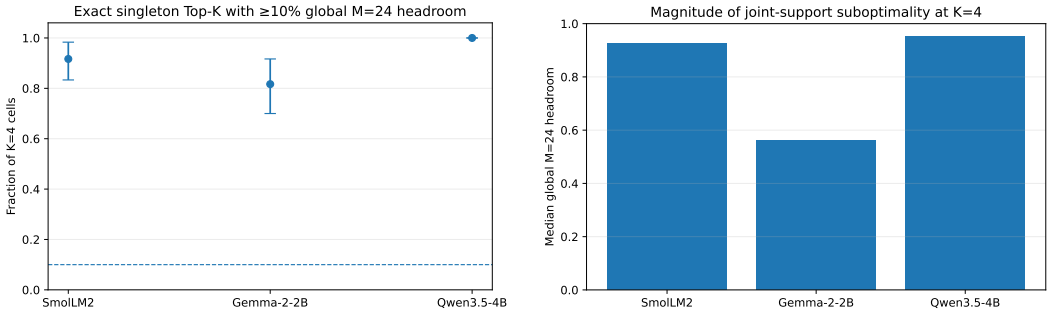


Figure 1: **Primary exact $K = 4$ audit.** Left: fraction of exact-singleton Top-4 supports with at least 10% global $M = 24$ candidate-universe headroom, with prompt-cluster bootstrap 95% CIs. Right: median candidate-universe headroom, retaining model heterogeneity rather than pooling across models.

Table 1: Primary exact $K = 4$ results. CIs are prompt-cluster bootstrap 95% intervals. ΔKL is baseline support KL minus exact candidate-universe oracle KL.

Model	$H_U \geq .10$	Median H_U [95% CI]	Median ΔKL [95% CI]
SmolLM2	55/60 (91.7%)	.926 [.807,.984]	4.018×10^{-4} [2.491×10^{-4},1.126×10^{-3}]
Gemma-2-2B	49/60 (81.7%)	.561 [.346,.658]	1.927×10^{-4} [1.170×10^{-4},3.216×10^{-4}]
Qwen3.5-4B	60/60 (100%)	.954 [.942,.964]	2.473×10^{-3} [1.544×10^{-3},4.360×10^{-3}]

Candidate-universe sensitivity. To test whether the result depends on the Top-24 attention universe, we evaluate five outcome-independent $M = 24$ constructions. U1 is a hybrid universe containing the 16 highest-attention positions plus the 8 highest remaining positions by output-geometric leverage

$$\ell_i = a_i \|W_o(v_i - h)\|_2,$$

where W_o is the target head's output-projection block. U2 contains the 24 highest-attention positions. U3 combines the 12 highest-attention positions with 12 uniformly sampled remaining valid positions, using two prespecified seeds. U4 uniformly samples 24 valid positions, also with two seeds. U5 samples six positions from each quartile of the attention-rank ordering. Full aggregation corresponding cell counts and confidence intervals are reported in Table 2 of Appendix C.

6 Results

6.1 Exact singleton Top-K retains large $K = 4$ headroom

Figure 1 and Table 1 give the primary result. Substantial global $M = 24$ candidate-universe headroom is frequent in all three audited model families: 55/60 SmolLM2 cells, 49/60 Gemma cells, and 60/60 Qwen cells. The effect size is heterogeneous. Median headroom is 0.926 for SmolLM2, 0.561 for Gemma, and 0.954 for Qwen; median absolute KL improvement ranges from 1.93×10^{-4} in Gemma to 2.47×10^{-3} in Qwen. The result therefore supports replication across these three model families, not architecture invariance.

The result is robust to candidate-universe construction. Across the five outcome-independent $M = 24$ constructions U1–U5, the fraction of exact-singleton supports with at least 10% candidate-universe headroom is 88.1%, 87.2%, 86.0%, 96.4%, and 99.2%, respectively. Median headroom is 0.833, 0.828, 0.827, 0.885, and 0.895. U3 and U4 each aggregate two prespecified random seeds; the corresponding cell counts and confidence intervals are reported in Appendix C. The separation therefore persists under substantially different ways of constructing the candidate set.

Cross-model lower-K replication and $K = 5$ stress test. At exact $K \in \{2, 3\}$, the $\geq 10\%$ headroom rates are 107/120 (89.2%) for SmolLM2, 49/120 (40.8%) for Gemma, and 118/120 (98.3%) for Qwen, showing substantial model–budget heterogeneity. In the smaller descriptive

$K = 5$ stress test, all 12/12 audited cells per model exceed 10% candidate-universe headroom. The budget dependence is especially pronounced for Gemma: median candidate-universe headroom is only 0.0041 in the combined $K \in \{2, 3\}$ audit, compared with 0.561 at $K = 4$, reinforcing that the magnitude of the phenomenon is model- and budget-dependent rather than architecture-invariant. Full lower-budget and descriptive $K = 5$ results are reported in Appendices D and I.

6.2 Improving the marginal ranker need not improve the retained support

A separate paired SmolLM2 experiment tests a weaker alternative explanation: perhaps exact downstream singleton scores simply provide a better version of attention ranking and should therefore improve the selected set monotonically. Across 120 $K \in \{2, 3\}$ prompt–head–budget cells evaluated under the same candidate universes, replacing attention Top-K with exact-singleton Top-K produces *lower* joint-support KL in 38 cells, an exact tie in 34, and *higher* joint-support KL in 48. Exact-singleton Top-K still leaves at least 10% candidate-universe headroom in 108/120 cells, compared with 104/120 for attention Top-K. Thus making the scalar score exactly match the singleton downstream intervention does not monotonically improve the jointly renormalized support.

6.3 The observed mechanism is support-conditioned ranking reversal

The theory predicts that a candidate preferred by its independent singleton score can become worse than another candidate once both are evaluated inside the same retained context. We test this prediction directly.

For a ranked support S and retained item $i \in S$, define the one-hole retained context $C_i = S \setminus \{i\}$. For an unselected candidate $j \notin S$, we say that a *strict support-conditioned reversal* occurs when

$$d_i > d_j \qquad \text{but} \qquad L(C_i \cup \{j\}) < L(C_i \cup \{i\}). \tag{9}$$

The first inequality says that independent singleton ranking prefers i to j; the second says that, after holding the rest of the retained support fixed, the downstream objective prefers j to i. This is the empirical counterpart of the context dependence exposed by eq. (7) and proposition 2.

Within each one-hole context, we also compare the singleton-score ordering with the support-conditioned ordering induced by $-L(C_i \cup \{x\})$. We summarize their agreement using Kendall's tie-adjusted rank correlation τ_b and the *inversion rate*, the fraction of candidate pairs whose relative ordering disagrees.

Figure 2 visualizes the relation between support-conditioned ranking instability and candidate-universe headroom. The mechanism analysis contains 216 cells, 900 one-hole retained contexts, and 6,512 improving one-swap witnesses. Among primary $K = 4$ cells with $H_U(S) \geq 0.10$, at least one strict support-conditioned reversal occurs in 100% of SmolLM2 cells, 100% of Gemma cells, and 98.3% of Qwen cells. Median conditioned Kendall τ_b is only .025, .062, and .116, while median inversion rates are 48.8%, 46.9%, and 44.2%, respectively. Thus, in the audited substantial-headroom cases, the independent singleton ordering is almost maximally unstable once candidates are evaluated inside the retained support. For more details, see Appendix F

To quantify how much of the globally available candidate-universe improvement is visible locally, for cells with $H_U(S) > 0$ define the *best-swap capture*

$$C_{\text{swap}}(S) = \frac{\max\{0, W_U(S)\}}{H_U(S)}. \tag{10}$$

Median best-swap capture is 97.2% for SmolLM2, 96.4% for Gemma, and 78.1% for Qwen. A single exchange therefore often exposes most of the available relative improvement.

High local capture does not make one-swap search a complete optimizer. Starting from the ranked support, greedy one-swap descent repeatedly replaces the current support by its lowest-loss strictly improving one-swap neighbor and stops when no such neighbor remains. Among cells with $H_U(S) \geq 0.10$, this procedure reaches the exact candidate-universe optimum in only 41.8% of SmolLM2, 69.4% of Gemma, and 25.0% of Qwen cases. Strong local witnesses are therefore common even when the global support landscape remains nontrivial.

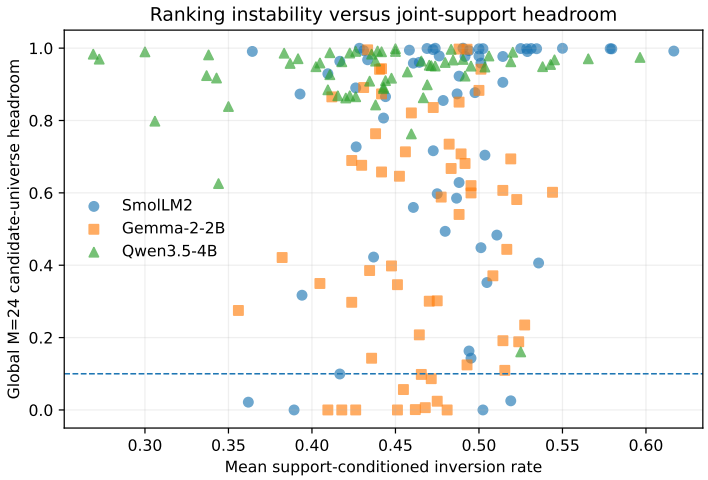


Figure 2: **Support-conditioned instability versus joint-support headroom.** Each point is an audited $K = 4$ cell; the horizontal axis summarizes context-conditioned inversion and the vertical axis is global $M = 24$ candidate-universe headroom. The plot is descriptive: it visualizes the mechanism predicted by eq. (7) and proposition 2 without assigning a causal or confirmatory interpretation to the association.

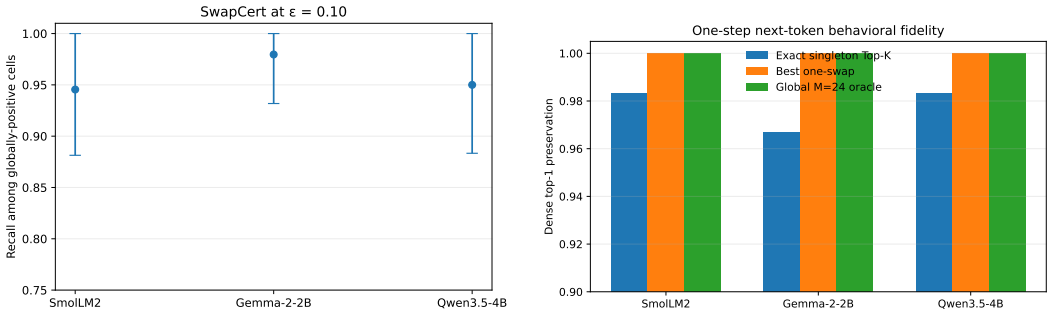


Figure 3: **Diagnostic recall and behavioral boundary.** Left: SWAPCERT recall at $\varepsilon = .10$ among exact $K = 4$ candidate-universe positives. Right: one-step top-1 agreement; ranked-support flips are rare, and all four observed flips are rescued by both the best one-swap support and the exact candidate-universe oracle.

6.4 SWAPCERT IS A HIGH-RECALL POSITIVE DIAGNOSTIC, NOT AN OPTIMALITY CERTIFICATE

At $\varepsilon = .10$, SWAPCERT detects 52/55 SmolLM2 positives (94.5%, CI 88.1–100%), 48/49 Gemma positives (98.0%, CI 93.2–100%), and 57/60 Qwen positives (95.0%, CI 88.3–100%). No soundness violation occurs. The pooled 157/164=95.7% recall is descriptive only. Of the seven misses, six have a positive one-swap improvement below the 10% firing threshold and one is genuinely one-swap-local. These misses are retained; SWAPCERT non-fire is inconclusive. The left panel of Figure 3 summarizes these diagnostic recall rates. More details are in Appendix G

6.5 DISTRIBUTIONAL IMPROVEMENTS DO NOT IMPLY ESTABLISHED BEHAVIORAL GAINS

The right panel of Figure 3 summarizes the corresponding one-step behavioral boundary.Across 180 one-step cells, the exact-singleton ranked support changes the dense next-token top-1 in 1/60 SmolLM2, 2/60 Gemma, and 1/60 Qwen cells. In all four observed crossings, both the best one-swap support and the exact candidate-universe oracle restore the dense top-1. However, prompt-cluster bootstrap 95% intervals for the agreement gain all include zero: [0,5%], [0,8.33%], and [0,5%]. The

behavioral result is therefore an existence/boundary finding. We make no claim of population-level top-1 improvement or free-running generation improvement.

7 Implications for Sparse-Attention Evaluation

The results suggest that sparse selectors should be evaluated along two distinct axes. *Marginal quality* asks whether a scalar score identifies candidates that are individually consequential under a specified intervention. *Support quality* asks whether the final budget-constrained, renormalized set is jointly good under the target downstream objective. The objective-matched counterexample and paired marginal-vs-support comparison show why the first does not imply the second in general: an exact marginal ordering can coexist with a substantially suboptimal support, and replacing a weaker marginal ranker by an exact one need not improve the induced set. This perspective complements conditional guarantees for singleton-based selection. When geometric or structural assumptions couple individual perturbations to set quality, scalar ranking can be justified (Zhang et al., 2026); outside such regimes, residual support headroom should be assessed directly. A natural decomposition is therefore *rank first, audit the support second*: a low-cost scalar selector provides an incumbent support, after which a support-level diagnostic such as SwapCert can test for a concrete same-budget improvement. Our experiments characterize this distinction and SwapCert's behavior relative to the exact candidate-universe oracle; they do not establish SwapCert or exhaustive downstream scoring as production serving algorithms.

8 Discussion and Limitations

The formal result establishes a specific non-equivalence: objective-matched singleton marginals and jointly renormalized fixed-cardinality support quality are different optimization problems. The experiments show that the separation also occurs frequently inside the prespecified candidate universes studied here. Candidate-universe sensitivity, three-model exact replication, exact $K = 4$ enumeration, and direct preference-reversal measurements all support the same explanation based on retained-context dependence. Separate non-exhaustive swap screens provide complementary longer-context evidence: an approximately 2k-token panel spans 1,512 heads and 48,384 prespecified same-cardinality exchanges, and an approximately 4k-token replication spans 504 heads and 16,128 exchanges; these experiments detect support-level downstream-KL-improving exchanges but measure witnessed rather than exhaustive headroom (Appendix J). Several boundaries are equally important. First, every empirical exhaustive statement is relative to $M = 24$; the unrestricted full-context support problem is not solved. The exact $K = 4$ and descriptive $K = 5$ budgets are small enough for enumeration, so prevalence at production cache budgets remains an external-validity question. Second, three model families establish replication rather than architectural invariance, and the exact panel spans 30 prompts rather than a deployment population. Third, exact enumeration and downstream singleton scoring are diagnostic instruments rather than deployment-cost measurements. Fourth, SwapCert certifies only positive witnesses: a non-fire says neither that the support is good nor that larger-radius exchanges are unhelpful. Finally, KL improvements measure distributional fidelity under the specified intervention. The one-step audit shows that large KL headroom can coexist with unchanged argmax decisions, and free-running generation benefit is not established.

9 Conclusion

Sparse attention ranking and support selection answer different questions. Exact singleton effects can rank every candidate correctly for its individual deletion and still select a poor jointly renormalized support because exchange value depends on retained context. This mismatch appears frequently under exact $M = 24$ language-model audits, persists at $K = 4$, and is directly accompanied by support-conditioned preference reversals. SwapCert complements rankings by asking whether a concrete same-budget witness exposes meaningful remaining headroom; its positive implication is exact, while its empirical recall is high and its misses are explicit. The result suggests a simple evaluation principle for sparse selection systems: measure marginal ranking quality and joint support quality separately. The next question could be whether one can characterize the *necessary and sufficient structural conditions* under which exact downstream singleton ranking becomes globally support-optimal despite softmax renormalization.

AI USE STATEMENT

AI tools were used to assist with literature review and writing refinement. All ideas, claims, code, experimental design, and results analysis were done by the author, who takes full responsibility for the scientific content and final submission.

REFERENCES

Bai, Y., Lv, X., Zhang, J., Lyu, H., Tang, J., Huang, Z., Du, Z., Liu, X., Zeng, A., Hou, L., Dong, Y., Tang, J., and Li, J. LongBench: A bilingual, multitask benchmark for long context understanding. *ACL*, pp. 3119–3137, 2024.

Deng, D. and Lindauer, M. Neural Attention Search. *NeurIPS*, 38, 2025.

Feng, Y., Lv, J., Cao, Y., Xie, X., and Zhou, S. K. Ada-KV: Optimizing KV Cache Eviction by Adaptive Budget Allocation for Efficient LLM Inference. *NeurIPS*, 38, 2025.

Feng, Y., Lv, J., Guo, H., Cao, Y., Zhou, S. K., and Xie, X. CriticalKV: Optimizing KV Cache Eviction from an Output Perturbation Perspective. arXiv:2502.03805, 2026.

Gao, Y., Zeng, Z., Du, D., Cao, S., Zhou, P., Qi, J., Lai, J., So, H., Cao, T., Yang, F., and Yang, M. SeerAttention: Self-distilled attention gating for efficient long-context prefilling. *NeurIPS*, 38, 2025.

Geng, Z., Wang, J., Liu, Z., Ju, F., Li, Y., Li, X., Yuan, M., Hao, J., Lian, D., Chen, E., Zhang, Y., and Wu, F. Accurate KV Cache Eviction via Anchor Direction Projection for Efficient LLM Inference. *NeurIPS*, 38, 2025.

Gupta, A. and Berant, J. Value-aware approximate attention. *EMNLP*, pp. 9567–9574, 2021.

Jiang, H., Li, Y., Zhang, C., Wu, Q., Luo, X., Ahn, S., Han, Z., Abdi, A. H., Li, D., Lin, C.-Y., Yang, Y., and Qiu, L. MInference 1.0: Accelerating pre-filling for long-context LLMs via dynamic sparse attention. *NeurIPS*, 37, 2024.

Kumari, L., Wang, S., Zhou, T., Sarda, N., Rowe, A., and Bilmes, J. BumbleBee: Dynamic KV-Cache Streaming Submodular Summarization for Infinite-Context Transformers. *COLM*, 2024.

Li, Y., Huang, Y., Yang, B., Venkitesh, B., Locatelli, A., Ye, H., Cai, T., Lewis, P., and Chen, D. SnapKV: LLM Knows What You are Looking for Before Generation. *NeurIPS*, 37, 2024.

Li, Z., Zhu, L., Gu, H., Hu, X., Wang, Y., Mi, H., Han, S., Liang, L., and Guo, Z. SAS: Simple Attention Sparsification via End-to-End Optimization of Context Ranking. arXiv:2609.13141, 2026.

Ribar, L., Chelombiev, I., Hudlass-Galley, L., Blake, C., Luschi, C., and Orr, D. SparQ Attention: Bandwidth-Efficient LLM Inference. *ICML*, pp. 42558–42583, 2024.

Shi, J. and Gu, J. Trust the Mass: Forced Weights in KV-Cache Eviction. arXiv:2608.25230, 2026.

Tang, J., Zhao, Y., Zhu, K., Xiao, G., Kasikci, B., and Han, S. QUEST: Query-Aware Sparsity for Efficient Long-Context LLM Inference. *ICML*, pp. 47901–47911, 2024.

Xiao, G., Tian, Y., Chen, B., Han, S., and Lewis, M. Efficient Streaming Language Models with Attention Sinks. *ICLR*, 2024.

Zhang, Z., Sheng, Y., Zhou, T., Chen, T., Zheng, L., Cai, R., Song, Z., Tian, Y., Ré, C., Barrett, C., Wang, Z., and Chen, B. H2O: Heavy-Hitter Oracle for Efficient Generative Inference of Large Language Models. *NeurIPS*, 36, 2023.

Zhang, A., Gurses, S., Deng, Y., Wang, N., Liu, C.-C., Wertheimer, D., Liu, D., Li, X., Yang, Z., Ye, F. X.-F., and Yin, P. DropKV: Decoupling Residual-Output Perturbation for Near-Optimal KV-Cache Eviction. *ICML*, PMLR 306, 2026.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008, 2017.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi:10.1145/3600006.3613165.

A Proofs and Exact Counterexample Details

A.1 Exact downstream singleton counterexample

For

$$a = (0.2, 0.1, 0.2, 0.5), \qquad v = (4, -2, 1, 2.7),$$

the dense head output is

$$h = \sum_{j=1}^{4} a_j v_j = 2.15.$$

Deleting candidate j and renormalizing the remaining attention weights gives

$$h_{-j} = \frac{h - a_j v_j}{1 - a_j}.$$

With two-class logits $z(x) = (1.8x, 0)$, let

$$\sigma(t) = \frac{1}{1 + e^{-t}}, \qquad q(x) = \sigma(1.8x),$$

and define the downstream KL relative to the dense prediction by

$$D(x) = q(h) \log \frac{q(h)}{q(x)} + [1 - q(h)] \log \frac{1 - q(h)}{1 - q(x)}.$$

The exact singleton-deletion score is

$$d_j = D(h_{-j}).$$

Direct evaluation gives

j	h_{-j}	$d_j = D(h_{-j})$
1	1.687500	0.00918668148
2	2.611111	0.00536906759
3	2.437500	0.00228497479
4	1.600000	0.01374411597

Hence

$$d_4 > d_1 > d_2 > d_3,$$

so exact-singleton Top-2 selects

$$S_{\text{rank}} = \{1, 4\}.$$

We now evaluate the actual fixed-cardinality objective. For every size-2 support S,

$$h_S = \frac{\sum_{i \in S} a_i v_i}{\sum_{i \in S} a_i}, \qquad L(S) = D(h_S).$$

Exhaustive evaluation of all six size-2 supports gives

Support S	h_S	$L(S)$
$\{1,2\}$	2.000000	0.000796590380
$\{1,3\}$	2.500000	0.003276210803
$\{1,4\}$	3.071429	0.017208222670
$\{2,3\}$	0.000000	0.593430803601
$\{2,4\}$	1.916667	0.002026829576
$\{3,4\}$	2.214286	0.000129168300

Therefore

$$S^\star = \{3,4\}$$

is the unique minimum-loss size-2 support, whereas exact-singleton Top-2 selects $\{1,4\}$. Their loss ratio is

$$\frac{L(\{1,4\})}{L(\{3,4\})} = \frac{0.017208222670}{0.000129168300} \approx 133.2.$$

Thus exact objective-matched singleton ranking can select a jointly renormalized support whose downstream KL is more than two orders of magnitude larger than the exact size-2 optimum. This proves Proposition 1.

A.2 Exchange identity and preference reversal

For $S' = S - \{i\} + \{j\}$,

$$h_{S'} = \frac{A_S h_S - a_i v_i + a_j v_j}{A_S - a_i + a_j},$$

and subtracting h_S yields Equation (7). The dependence on h_S, A_S, and the leaving item i is the support-conditioning term used in the mechanism audit.

For Proposition 2, use uniform attention on $K+2$ values containing $K-2$ zeros and four special values $-2, +2, +1, -1$. The dense output is zero. Let C_- be the $K-2$ shared zeros together with -2, and C_+ the same zeros together with $+2$. The two candidates are $+1$ and -1. Each completed support $C_\pm \cup \{+1\}$ or $C_\pm \cup \{-1\}$ has cardinality K. Under logits $(z, 0)$ against dense logits $(0, 0)$,

$$L(z) = D_{\mathrm{KL}}((1/2, 1/2) \| (\sigma(z), 1-\sigma(z))) = \log\cosh(z/2),$$

which is strictly increasing in $|z|$. Renormalized uniform attention gives support outputs $-1/K$ and $-3/K$ after C_-, so $+1$ is preferred to -1; after C_+ the outputs are $3/K$ and $1/K$, so -1 is preferred to $+1$. Thus the same two candidates reverse preference solely with the retained context, for every $K \geq 2$.

A.3 Positive-certificate soundness

For any $S \in \mathcal{F}_K(U)$,

$$\mathcal{N}_1(S) \subseteq \mathcal{F}_K(U).$$

Therefore

$$\min_{T \in \mathcal{F}_K(U)} L(T) \leq \min_{T \in \mathcal{N}_1(S)} L(T).$$

For $L(S) > 0$, dividing by $L(S)$ and subtracting from one gives

$$H_U(S) \geq W_U(S).$$

Hence, for every $\varepsilon \in (0, 1)$,

$$W_U(S) \geq \varepsilon \quad \Longrightarrow \quad H_U(S) \geq \varepsilon.$$

When $L(S) = 0$, both quantities are defined to be zero, so the antecedent cannot hold for $\varepsilon > 0$. This proves Theorem 1.

B Experimental Definitions

Candidate-universe headroom. All empirical exhaustive claims use

$$H_U(S) = 1 - \frac{\min_{T \in \mathcal{F}_K(U)} L(T)}{L(S)}$$

with $|U| = 24$. We use "global $M = 24$ candidate-universe" precisely to distinguish this exhaustive restricted-universe quantity from unrestricted full-context optimization.

Primary panel. The primary $K = 4$ panel contains 30 prespecified LongBench-E prompts (10 each from 2WikiMultihopQA, HotpotQA, and PassageRetrieval-en), two head indices per prompt (the first and last attention heads), and 60 prompt–head cells per model. The target layers are zero-indexed SmolLM2 layer 23, Gemma-2-2B layer 25, and Qwen3.5-4B layer 31, each the deepest layer using ordinary full-softmax attention; recurrent or linear-attention blocks are excluded. The candidate universe is the Top-24 valid token positions by attention weight, and the baseline support is Top-4 under exact downstream singleton-deletion KL. Each cell exhaustively evaluates $\binom{24}{4} = 10{,}626$ supports.

Statistical unit. Prompt-cluster bootstrap intervals resample prompts rather than treating the two heads from one prompt as independent. No null-hypothesis p-values are used. Threshold, task, $K = 5$, and correlation analyses are reported as sensitivity or descriptive analyses according to their prespecified roles.

C Candidate-Universe Sensitivity

Candidate-universe sensitivity compares five outcome-independent $M = 24$ constructions. U1 contains the 16 highest-attention positions plus the 8 highest remaining positions by output-geometric leverage

$$\ell_i = a_i \|W_o(v_i - h)\|_2,$$

where W_o is the target head's block of the attention output projection. U2 contains the 24 highest-attention positions. U3 contains the 12 highest-attention positions plus 12 uniformly sampled remaining valid positions, with two prespecified seeds. U4 uniformly samples 24 valid positions, also with two seeds. U5 sorts valid positions by attention rank, divides them into four contiguous quartiles, and samples six positions from each quartile. The aggregation counts are 360 cells for U1/U2/U5 and 720 cells for U3/U4 because U3 and U4 aggregate both seeds.

Table 2: Candidate-universe sensitivity for exact-singleton supports. Rates are the fraction with at least 10% global $M = 24$ candidate-universe headroom.

Universe	Cells	Rate	Cluster 95% CI	Median H_U
U1 hybrid	360	.8806	[.8250,.9306]	.8329
U2 Top-M attention	360	.8722	[.8167,.9222]	.8275
U3 mixed, 2 seeds	720	.8597	[.8028,.9139]	.8272
U4 random, 2 seeds	720	.9639	[.9389,.9847]	.8851
U5 attention-stratified	360	.9917	[.9750,1.0000]	.8948

All five constructions retain high rates of substantial candidate-universe headroom, showing that the separation is not specific to a single universe-construction rule.

D Cross-Model Exact $K \in \{2, 3\}$ Audit

The final-layer exact $M = 24$, $K \in \{2, 3\}$ audit contains 120 cells per model. Table 3 records the exact-singleton support results. The large Gemma difference between the lower-K and $K = 4$ audits

Table 3: Supporting exact $K \in \{2, 3\}$ replication.

Model	$H_U \geq .10$	Cluster 95% CI	Median H_U	Median ΔKL
SmolLM2	107/120 (89.17%)	[82.5%,95.0%]	.83058	3.7795×10^{-4}
Gemma-2-2B	49/120 (40.83%)	[32.5%,49.17%]	.004078	1.6688×10^{-6}
Qwen3.5-4B	118/120 (98.33%)	[95.0%,100%]	.92994	3.0970×10^{-3}

is a reason to report model and budget heterogeneity explicitly rather than to assert architecture-invariant prevalence.

E PRIMARY $K = 4$ STATISTICAL AUDIT

Prompt-cluster bootstrap gives the following median headroom intervals: SmolLM2 .92591 [.80678,.98388], Gemma .56094 [.34624,.65773], and Qwen .95385 [.94152,.96414]. Median absolute KL-improvement intervals are SmolLM2 4.0183×10^{-4} [2.4910×10^{-4},1.12597×10^{-3}], Gemma 1.9266×10^{-4} [1.1702×10^{-4},3.2155×10^{-4}], and Qwen 2.4732×10^{-3} [1.5441×10^{-3},4.3597×10^{-3}]. Independent recomputation of headroom in the mechanism, SWAPCERT, and behavioral analyses matches the primary $K = 4$ values exactly (maximum absolute discrepancy 0).

F MECHANISM AUDIT DETAILS

Across the mechanism analysis there are 216 cells, 900 conditioned contexts, and 6,512 improving one-swap witnesses. Table 4 reports the primary $K = 4$ summaries. The contrast between high one-swap capture and incomplete greedy convergence is important: one-swap neighborhoods are strong diagnostics of remaining headroom but are not a general replacement for exact support optimization.

Table 4: Support-conditioning diagnostics at $K = 4$. Reversal rates are among cells with $H_U \geq .10$. Greedy-oracle rates are among those positive cells.

Model	Reversal	Med. τ_b	Med. inversion	Best-swap capture	Greedy reaches oracle
SmolLM2	100%	.0250	48.75%	97.16%	41.82%
Gemma-2-2B	100%	.0619	46.90%	96.40%	69.39%
Qwen3.5-4B	98.33%	.1155	44.23%	78.10%	25.00%

G SWAPCERT AUDIT AND EVERY PRIMARY MISS

At $K = 4$ and $\varepsilon = .10$, exact-oracle positives, hits, and misses are:

Model	Positives	Hits	Misses	Recall [95% CI]
SmolLM2	55	52	3	94.55% [88.14%,100%]
Gemma-2-2B	49	48	1	97.96% [93.18%,100%]
Qwen3.5-4B	60	57	3	95.00% [88.33%,100%]

There are zero soundness violations. Across the seven misses, six supports have an improving one-swap neighbor whose relative gain is positive but below the 10% firing threshold; one Qwen case has no improving one-swap support despite at least 10% global $M = 24$ candidate-universe headroom. This is a genuine local miss. The pooled 157/164=95.7% recall is reported only descriptively.

H BEHAVIORAL BOUNDARY AND BATCH-SHAPE NUMERICAL CORRECTION

The one-step audit contains 180 cells. Exact-singleton ranked supports change the dense next-token top-1 in 1/60 SmolLM2, 2/60 Gemma, and 1/60 Qwen cells. All four observed flips are rescued by both the best one-swap support and the exact candidate-universe oracle. The prompt-cluster bootstrap intervals for agreement gain are [0%,5%], [0%,8.33%], and [0%,5%], respectively, so population-level improvement is not established.

Batch-shape numerical correction. An initial behavioral run used a different batch shape from the primary $K = 4$ enumeration. Under bfloat16 (BF16) matrix-multiplication execution, four Gemma cells differed from the primary KL values beyond the unchanged 2×10^{-7} numerical tolerance, with maximum discrepancy 9.1527×10^{-5}. Behavioral quantities were therefore recomputed with batch size one, while the primary $K = 4$ KL values were independently reproduced using the original exhaustive batching scheme. The latter reproduction has maximum absolute error exactly zero. The numerical amendment and superseded run are included in the supplement for provenance.

I Descriptive $K = 5$ Stress Test

The prespecified $K = 5$ extension contains 12 cells per model, selected independently of outcomes. All 12/12 cells in each model have at least 10% global $M = 24$ candidate-universe headroom. Median headroom is .9751 for SmolLM2, .6179 for Gemma, and .9364 for Qwen. Because this slice was designated descriptive, it is not promoted to a confirmatory result.

J Additional Supporting Experiments

Two additional experiment families provide lower-budget, larger-context, and paired-ranking evidence beyond the primary exact $K = 4$ analysis.

SmolLM2 $K \in \{2, 3\}$ panel. A separate exact panel uses an outcome-independent hybrid $M = 24$ universe (16 highest-attention positions plus eight output-geometric positions), SmolLM2, 30 prompts, two heads, and $K \in \{2, 3\}$. It compares attention ranking, exact head-space singleton-deletion magnitude

$$u_j = \|h_{-j} - h\|_2 = \frac{a_j}{1 - a_j}\|v_j - h\|_2,$$

and exact downstream singleton-deletion KL. All three rankings can leave substantial candidate-universe headroom, providing lower-budget evidence complementary to the primary multi-model $K = 4$ analysis.

Larger-context swap screens. A separate approximately 2k-token context panel spans all three models, three layer depths, two sparsity levels, 1,512 heads, and 48,384 same-cardinality exchanges; an approximately 4k-token last-layer replication contains 504 heads and 16,128 exchanges. Each head is tested with 32 prespecified same-cardinality one-swap candidates, so these experiments estimate *witnessed* headroom rather than exhaustive candidate-universe headroom. They show that near-saturated head-space marginal ranking can coexist with support-level downstream-KL improvements at substantially longer context lengths.

Paired marginal-vs-support comparison. A SmolLM2 panel uses 30 prespecified LongBench-E prompts, two head indices per prompt, $M = 24$, and exact $K \in \{2, 3\}$ enumeration, giving 120 prompt–head–budget cells. Its candidate universe combines 16 highest-attention positions with eight outcome-independent geometric-leverage positions. Comparing attention Top-K with exact downstream-singleton Top-K on the same cells gives 38 lower-KL, 34 equal-KL, and 48 higher-KL supports for the exact-singleton ranker. Candidate-universe headroom of at least 10% occurs in 108/120 exact-singleton supports and 104/120 attention supports. The paired comparison directly illustrates that improved singleton ranking need not monotonically improve joint-support quality.

K Reproducibility and Prespecification

The anonymous supplement contains machine-readable result summaries, experimental protocols, analysis and verification scripts, final figure assets, numerical amendment records. For the primary $K = 4$ experiment, the models, layers, prompts, heads, universe construction, singleton score, loss, threshold, and prompt-cluster bootstrap procedure were specified before outcome evaluation. The candidate-universe sensitivity variants and the descriptive role of the $K = 5$ stress test were also specified in advance. A numerical consistency checker validates manuscript values against the machine-readable result artifacts.

Implementation corrections and provenance. The supplement records outcome-independent candidate-universe wrapper fixes, a canonical support-order correction with an exact SmolLM2 regression check, a logging-only preflight change, and the BF16 batch-shape correction described above. Superseded outputs are retained for provenance but excluded from the reported evidence.