# VALUEDIFF: VALUE-GEOMETRIC KV CACHE EVICTION FOR SINK-SUPPRESSED LLMS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Modern LLMs with QK-normalization, gated attention, learned attention sinks, or logit softcapping exhibit weaker persistent attention sinks, on which existing KV cache eviction methods primarily rely. We observe that across these models, weaker sinks co-occur with greater value-vector dispersion relative to key-vector dispersion. Motivated by this value-side dispersion, we present VALUEDIFF, a value-geometric eviction that ranks tokens by the L2 deviation of their value vectors from the cache mean. The same score arises as the minimal-disturbance eviction under a max-entropy assumption about future attention. We evaluate under fixed cache budgets, with eviction at every block boundary during prefill and at every decoding step during generation. On RULER at a tight 2k token budget, VALUEDIFF retains 88–99% of dense across seven sink-suppressed models (best on 6 out of 7). On LongBench at the 4k budget, VALUEDIFF averages 92% retention across sink-suppressed models versus 83% for the strongest prior baseline. On MATH-500, VALUEDIFF is the strongest non-dense method on every sink-suppressed model tested at the 25% cache budget, outperforming prior methods by up to $\sim$20 points on gated-attention models. Across all three benchmarks, value geometry emerges as the more reliable query-invariant eviction signal for sink-suppressed models.

## 1 INTRODUCTION

Deploying large language models under strict memory constraints (on edge devices with fixed RAM budgets (Gunter et al., 2024; Xu et al., 2024; Chen et al., 2025), or in serving systems where peak GPU memory bounds throughput) requires the KV cache to stay within a bounded footprint throughout inference, not only during generation. *KV cache eviction* addresses this by assigning each cached token a score and discarding the lowest-scoring tokens whenever the cache exceeds a budget $N$. With long prompts of length $T$, eviction must run during prefill as well. The prompt is processed in non-overlapping blocks of size $B \ll T$ and the cache is pruned at each block boundary, a procedure known as *block prompt processing* (Park et al., 2025; Goel et al., 2025). Every eviction decision is *permanent*: a discarded token cannot be recovered from later context and must remain valid against all future queries.

**Attention sinks and architectural trend.** Attention's spikiness is well-studied as the *attention sink* phenomenon (Xiao et al., 2024; Su et al., 2026; Barbero et al., 2025), in which a small set of tokens absorbs disproportionate attention mass. The mechanism is fundamentally key-side: sink tokens receive disproportionate attention because their key vectors are positively aligned with most query vectors, yielding consistently high dot products across the majority of query positions (Park et al., 2025; Datta et al., 2026). Prior eviction methods rely on this key-side structure, whether through attention scores (which measure current key–query alignment) or key-geometry signals (which use key-vector structure as a proxy for token importance) (Park et al., 2025; Oren et al., 2024; Li et al., 2024; Zhang et al., 2023; Devoto et al., 2024; Datta et al., 2026). However, recent LLMs incorporating QK-normalization (Dehghani et al., 2023), gated attention hybrids (Qiu et al., 2025), learned attention sinks (Agarwal et al., 2025), and logit softcapping (Gemma Team, 2024) exhibit lower sink rates: the sink key vectors no longer maintain persistent alignment with query vectors nor geometric distinctiveness from content keys (Table 7). This motivates inspecting eviction signals that do not rely on key-side information.



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.

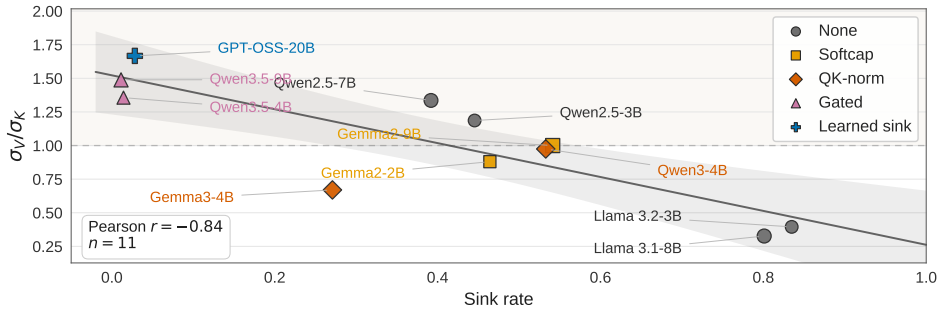



Figure 1: Sink rate versus $\sigma_V/\sigma_K$. Color and marker encode suppression category; shaded band = 95% conditional-mean CI; region above $\sigma_V/\sigma_K = 1$ marks value-dominant geometry

**Increased value-vector diversity.** As the key-side signal weakens, value vectors, which form the "basis" of the attention output, become the natural place to look for eviction signals. Since eviction asks which tokens are redundant and which are distinctive, we characterize both sides by the spread of value and key vectors around their respective means, $\sigma_V$ and $\sigma_K$. Across models spanning four recent architectural families and standard controls, we find a consistent pattern: models where attention sinks are weaker tend toward higher $\sigma_V/\sigma_K$, ranging from key-dominant ($\sigma_V < \sigma_K$) to value-dominant ($\sigma_V > \sigma_K$), as shown in Figure 1.

**Value-diversity-based KV cache eviction.** These observations motivate geometry-dependent eviction: as sinks weaken, value-side signals become more useful. We therefore propose VALUEDIFF, which scores each token by $s_j = \|\mathbf{v}_j - \bar{\mathbf{v}}\|$ where $\bar{\mathbf{v}}$ is the cache mean (Equation (4)). Low-$s_j$ tokens are near-redundant and evicted first, while high-$s_j$ tokens carry distinctive content and are retained. The same score arises from symmetry: when future attention is unknown, VALUEDIFF is the eviction that least disturbs the attention output under a uniform default (Section 3.3).

**Contributions.**

- **Geometry-dependent eviction.** Across four recent model families and standard controls, weaker sinks co-occur with greater relative value dispersion. This motivates VALUEDIFF, $s_j = \|\mathbf{v}_j - \bar{\mathbf{v}}\|$, and a pair of direction/norm ablations, VALUEDIFF-Dir and VALUEDIFF-Norm (Sections 3.1 and 3.2).
- **Max-entropy interpretation.** When future attention is unknown, VALUEDIFF is the eviction that minimally disturbs the attention output under the least-informative assumption, providing a principled derivation of the score from a symmetry argument (Proposition 3.1, Section B).
- **Empirical validation.** VALUEDIFF achieves best or tied-best RULER retention on 6 of 7 sink-suppressed models at a 2k-token budget. On LongBench, it averages 92% retention across five sink-suppressed models at 4k, versus 83% for the strongest prior baseline (Section 4). A two-axis geometric picture combining $\sigma_V/\sigma_K$ and key-space alignment explains these patterns (Section 4.4).

## 2 BACKGROUND

**KV Caching and Eviction.** The attention operator of Vaswani et al. (2017) projects input tokens into queries, keys, and values ($\mathbf{Q} = \mathbf{X}\mathbf{W}_Q$, $\mathbf{K} = \mathbf{X}\mathbf{W}_K$, $\mathbf{V} = \mathbf{X}\mathbf{W}_V$) and computes

$$\mathbf{O} = \mathrm{Softmax}\left(\mathbf{Q}\mathbf{K}^\top/\sqrt{d} + \mathbf{M}\right)\mathbf{V} = \mathbf{A}\mathbf{V}, \tag{1}$$

where $\mathbf{M}$ is the causal mask. During autoregressive generation, previously computed key-value pairs are stored in a *KV cache* $\mathcal{C} = (\mathbf{K}, \mathbf{V})$ and reused. The cache grows linearly with the number of processed tokens and dominates memory for long sequences. Architectural sharing such as grouped-query attention (Ainslie et al., 2023) reduces per-token cache size but leaves the linear-in-length growth unchanged, and IO-efficient attention kernels (Dao et al., 2022) that can accelerate the computation do not reduce the cache footprint.

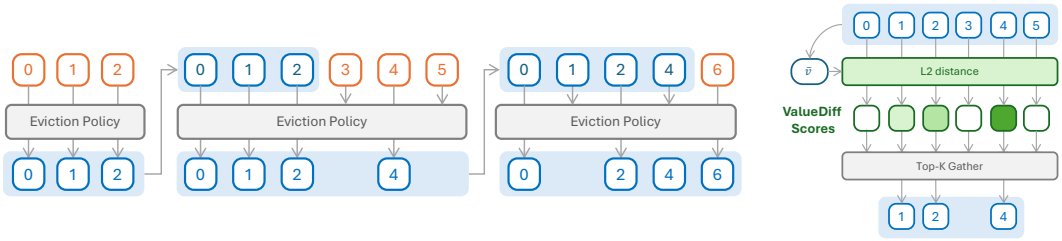



Figure 2: Method overview. **Left:** New tokens (orange) are appended to the retained cache (blue) before each eviction. **Right:** VALUEDIFF scores tokens by distance from the mean value and retains the top scores.

To bound memory usage, an eviction policy $\pi_N(\mathcal{C})$ removes tokens whenever the cache size exceeds a budget $N$. Given a scoring function $s(\cdot)$, eviction retains the $N$ highest-scoring tokens:

$$\mathcal{S} = \text{top-}N(s(\mathcal{C})), \quad \mathbf{K}' = \text{gather}(\mathbf{K}, \mathcal{S}), \quad \mathbf{V}' = \text{gather}(\mathbf{V}, \mathcal{S}). \tag{2}$$

Attention-score methods ($H_2O$ (Zhang et al., 2023), TOVA (Oren et al., 2024), SnapKV (Li et al., 2024), KVZip (Kim et al., 2025)) define $s(\cdot)$ using attention weights, either from observed queries or from reconstruction prompts. Key-geometric methods such as KeyDiff (Park et al., 2025) and ManifoldKV (Datta et al., 2026) use query-invariant properties of the keys.

**Blockwise Prefill and Decode-Time Eviction.** Prior token-eviction work commonly compresses the full-prompt KV cache once before generation, without decoding-time eviction (Li et al., 2024; Kim et al., 2025). Peak cache memory then grows with prompt or output length, contrary to our fixed-memory setting. We instead evict after each $B$-token prompt block (*block prompt processing* (Park et al., 2025; Goel et al., 2025)) and after every decoding step. Let $\Delta_i$ be the new KV pairs: at most $B$ during prefill and one during decoding. Both phases follow the same update, illustrated in Figure 2,

$$\mathcal{C}_i = \pi_N(\mathcal{C}_{i-1} \parallel \Delta_i), \qquad \mathcal{C}_0 = \emptyset, \tag{3}$$

where $\parallel$ denotes concatenation. The cache holds at most $N$ KV pairs after eviction and at most $N + B$ transiently during prefill.

## 3 VALUEDIFF FOR VALUE-GEOMETRIC KV CACHE EVICTION

### 3.1 SINK RATE AND VALUE GEOMETRY

We measure sink rate and the per-token L2 dispersion of key and value vectors across models spanning four recent architectural families and standard controls (Figure 1; full statistics in Table 6). As a general trend, models with lower sink rates tend to exhibit greater value-vector dispersion relative to keys. For head $h$ at layer $\ell$ with $n$ tokens, let $\sigma_K^{(\ell,h)} = \text{std}_i(\|k_i^{(\ell,h)} - \bar{k}^{(\ell,h)}\|_2)$ and $\sigma_V^{(\ell,h)} = \text{std}_i(\|v_i^{(\ell,h)} - \bar{v}^{(\ell,h)}\|_2)$, where $\bar{k}^{(\ell,h)} = \frac{1}{n}\sum_i k_i^{(\ell,h)}$. Each model's position in Figure 1 uses the mean of $\sigma_K^{(\ell,h)}$ and $\sigma_V^{(\ell,h)}$ across all layers and heads. Sink rate—the fraction of heads where the leading token exceeds weight $\tau$=0.3—ranges from 0.80–0.84 on Llama 3 to near zero on Qwen3.5 and GPT-OSS-20B. Models with lower sink rates tend to show higher $\sigma_V/\sigma_K$ (Figure 1). Qwen2.5 is a notable exception with high per-layer geometric heterogeneity; we discuss this separately in Section C.

The sink-rate observation directly challenges sink-preserving eviction. When leading-token sink rates approach zero, fixed-sink policies such as StreamingLLM (Xiao et al., 2024) have less reason to privilege the earliest cache positions. For attention-score methods such as TOVA (Oren et al., 2024) and SnapKV (Li et al., 2024), the implication is subtler. Persistent high-mass sink positions provide a stable shortcut; as those sinks weaken, attention-score methods fall back on query-specific rankings under block prompt processing, where evictions are permanent and future queries are unavailable. Attention-ranking overlap is consistently modest (top-256 Jaccard $\approx 0.4$; Section D), making this fallback an unreliable proxy for future cache utility. The value-geometry observation challenges key-only geometric eviction. Since eviction asks which tokens are redundant and which are distinctive, the rising $\sigma_V/\sigma_K$ points to value geometry as the stronger query-invariant signal for

this discrimination. We revisit this picture after the empirical results through an additional key-space alignment measure that further characterizes the geometric conditions (Section 4.4).

### 3.2 VALUEDIFF SCORE

We define the VALUEDIFF score as

$$s_j \;=\; \|\mathbf{v}_j - \bar{\mathbf{v}}\|, \qquad \bar{\mathbf{v}} \;=\; \tfrac{1}{n}\textstyle\sum_i \mathbf{v}_i, \tag{4}$$

where $n$ is the number of cached tokens. As illustrated on the right of Figure 2, a high $s_j$ marks a token as geometrically distinctive and worth retaining; a low $s_j$ marks it as near-redundant and safely discarded.

**Two axes of content.** Expanding Equation (4) via the law of cosines isolates two distinct geometric quantities,

$$s_j^2 \;=\; \|\mathbf{v}_j\|^2 - 2\|\mathbf{v}_j\|\|\bar{\mathbf{v}}\|\cos\theta_j + \|\bar{\mathbf{v}}\|^2, \tag{5}$$

where $\theta_j$ is the angle between $\mathbf{v}_j$ and $\bar{\mathbf{v}}$. This L2 score is shaped by **magnitude** ($\|\mathbf{v}_j\|$, tokens with large value norms) and **direction** ($1-\cos\theta_j$, tokens pointing away from the mean regardless of norm). To understand the role of each axis, we further evaluate two VALUEDIFF variants. VALUEDIFF-Dir keeps only the directional score $1-\cos(\mathbf{v}_j, \bar{\mathbf{v}})$, and VALUEDIFF-Norm keeps only the magnitude score $\|\mathbf{v}_j\|$. These variants ask whether value information is carried more by direction or by norm.

### 3.3 A MAX-ENTROPY INTERPRETATION OF VALUEDIFF

Eviction ultimately aims to preserve information that may be useful under unknown future attention distributions. Modeling those distributions is challenging in both block prompt processing and standard full-prefill eviction. We therefore ask which token is safest to remove under the least-informative assumption about future attention.

**Expected output disturbance.** The attention output under a future attention vector $\boldsymbol{\alpha} \in \Delta^{n-1}$ is $\mathbf{o}(\boldsymbol{\alpha}) = \sum_i \alpha_i \mathbf{v}_i$. A natural eviction criterion is the expected disturbance to the attention output when token $j$ is removed and the remaining attention weights are renormalized over the survivors.

$$j^\star \;=\; \arg\min_j \; \mathbb{E}_{\boldsymbol{\alpha}\sim P}\left[\left\|\mathbf{o}(\boldsymbol{\alpha}) - \mathbf{o}_{\setminus j}(\boldsymbol{\alpha})\right\|^2\right], \qquad \mathbf{o}_{\setminus j}(\boldsymbol{\alpha}) \;=\; \sum_{i\neq j} \tfrac{\alpha_i}{1-\alpha_j}\,\mathbf{v}_i, \tag{6}$$

where $P$ is a distribution on $\Delta^{n-1}$ encoding our belief about the future attention. With no information distinguishing tokens at eviction time, the symmetric choice is the uniform attention vector $\boldsymbol{\alpha}^\star = (1/n, \ldots, 1/n)$, the unique point on $\Delta^{n-1}$ invariant under permutations of token indices.

**Reduction at $\boldsymbol{\alpha}^\star$.** Substituting $\boldsymbol{\alpha} = \boldsymbol{\alpha}^\star$, the attention output collapses to the per-head cache mean, and uniform redistribution over the remaining $n-1$ tokens gives the post-eviction output.

$$\mathbf{o}^\star \;=\; \bar{\mathbf{v}}, \qquad \mathbf{o}^\star_{\setminus j} \;=\; (n-1)^{-1}\textstyle\sum_{i\neq j}\mathbf{v}_i \;=\; (n\,\bar{\mathbf{v}} - \mathbf{v}_j)/(n-1). \tag{7}$$

Evaluating the integrand of Equation (6) at $\boldsymbol{\alpha} = \boldsymbol{\alpha}^\star$ reduces the objective to

$$\left\|\mathbf{o}^\star - \mathbf{o}^\star_{\setminus j}\right\|^2 \;=\; \left\|\mathbf{v}_j - \bar{\mathbf{v}}\right\|^2/(n-1)^2, \tag{8}$$

whose minimizer over $j$ is exactly the VALUEDIFF score.

**Proposition 3.1** (VALUEDIFF as the symmetric optimum)**.** *At the symmetric attention vector $\boldsymbol{\alpha} = \boldsymbol{\alpha}^\star = (1/n, \ldots, 1/n)$, the eviction objective in Equation (6) is minimized at*

$$j^\star \;=\; \arg\min_j \left\|\mathbf{v}_j - \bar{\mathbf{v}}\right\|, \qquad \bar{\mathbf{v}} \;=\; \tfrac{1}{n}\textstyle\sum_i \mathbf{v}_i.$$

**Remark.** The derivation above uses the symmetric reference point $\boldsymbol{\alpha}^\star$. The same ranking also arises from the max-entropy version of the reconstruction objective under an unconstrained Dirichlet$(1, \ldots, 1)$ distribution over future attention (Section A). The relevance of this reference point grows as attention distributions flatten; Section B formalizes how the distance from the symmetric limit bounds the approximation error.

Table 1: RULER accuracy retention (% of dense, ↑; ctx=8192). Each model has two budget rows (2048, 4096). **Bold** marks the best non-dense strategy per budget, and underline marks the second best.

| | | *Sink* | *Attn* | | *Key Geometry* | | | *Attn×Val.* | *Value Geometry* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Bud. | Streaming | TOVA | SnapKV | KNorm | KeyDiff | ManifKV | F-CAOTE | V-Norm | V-Dir | VALUEDIFF |
| *Standard* | | | | | | | | | | | |
| Llama 3.1-8B | 2k | 40.1 | 62.0 | 75.5 | **91.8** | 89.6 | 60.9 | 78.8 | 53.1 | 88.7 | 88.1 |
| | 4k | 60.4 | 85.6 | 85.0 | **95.5** | 95.2 | 84.9 | 90.0 | 82.9 | 94.1 | 95.0 |
| Llama 3.2-3B | 2k | 40.2 | 59.7 | 72.2 | 79.6 | **92.9** | 40.5 | 72.3 | 44.5 | 89.3 | 88.8 |
| | 4k | 60.6 | 82.4 | 82.3 | 94.1 | **97.4** | 89.9 | 87.9 | 78.7 | 96.5 | 94.7 |
| Qwen2.5-3B | 2k | 18.0 | 22.1 | 23.7 | 31.4 | 39.8 | 34.4 | 28.8 | 19.6 | 39.7 | **41.1** |
| | 4k | 22.0 | 48.5 | 39.6 | 60.7 | 63.6 | 62.7 | 55.5 | 43.9 | 63.7 | **65.4** |
| Qwen2.5-7B | 2k | 10.9 | 14.0 | 18.7 | 23.1 | 31.9 | 13.1 | 20.3 | 19.3 | **32.4** | 21.8 |
| | 4k | 14.5 | 36.8 | 43.1 | 52.3 | 56.2 | 46.7 | 42.6 | 36.6 | **56.6** | 46.7 |
| *Logit softcapping* | | | | | | | | | | | |
| Gemma2-2B | 2k | 48.7 | 59.2 | 64.5 | 79.8 | **96.4** | 50.6 | 71.1 | 54.9 | 93.7 | 89.6 |
| | 4k | 63.7 | 82.8 | 80.0 | 86.5 | **98.7** | 57.6 | 86.2 | 61.5 | 98.1 | 95.7 |
| Gemma2-9B | 2k | 52.3 | 66.7 | 81.7 | 59.5 | 99.0 | 66.0 | 72.4 | 54.2 | 97.8 | **99.4** |
| | 4k | 66.5 | 69.7 | 85.8 | 67.9 | 99.3 | 80.6 | 77.8 | 72.2 | 99.6 | **100.0** |
| *QK-normalization* | | | | | | | | | | | |
| Qwen3-4B | 2k | 42.4 | 44.5 | 52.8 | 40.5 | 89.6 | 92.7 | 71.7 | 89.7 | 89.8 | **95.2** |
| | 4k | 63.0 | 81.0 | 78.6 | 69.8 | 95.1 | 98.0 | 91.6 | **98.4** | 97.1 | 98.0 |
| Gemma3-4B | 2k | 40.6 | 54.7 | 66.2 | 11.9 | 81.7 | 87.1 | 69.1 | 59.4 | 86.2 | **87.6** |
| | 4k | 63.0 | 76.1 | 80.5 | 19.9 | 91.9 | 94.6 | 91.7 | 94.5 | 95.0 | **96.1** |
| *Gated attention* | | | | | | | | | | | |
| Qwen3.5-4B | 2k | 42.0 | 42.0 | 62.6 | 59.7 | 93.3 | 92.5 | 67.6 | 93.7 | 94.7 | **97.6** |
| | 4k | 61.4 | 66.5 | 82.2 | 96.0 | 96.6 | 97.0 | 84.1 | 98.7 | 99.5 | **99.7** |
| Qwen3.5-9B | 2k | 42.9 | 42.0 | 77.2 | 24.2 | 91.9 | 92.1 | 66.4 | 86.9 | 95.4 | **95.6** |
| | 4k | 61.6 | 61.6 | 84.3 | 70.2 | 96.2 | 96.7 | 81.8 | **98.4** | 98.2 | 98.2 |
| *Learned Sink* | | | | | | | | | | | |
| GPT-OSS-20B | 2k | 39.3 | 54.2 | 80.2 | 64.1 | 88.9 | 68.5 | 75.7 | 50.2 | 90.1 | **91.3** |
| | 4k | 62.8 | 85.0 | 87.1 | 88.0 | 93.1 | 94.9 | 90.1 | 86.6 | 93.2 | **97.8** |

## 4 EXPERIMENTS

**Evaluation protocol.** We evaluate under *block prompt processing* (Equation (3)) with a chunk size of $B = 128$ and eviction after every decoding step, following the protocol of prior work (Park et al., 2025; Goel et al., 2025). This is strictly harder than the full-prefill setting assumed by most token eviction work because decisions are made with only partial context available, and discarded tokens cannot be recovered when later chunks arrive.

**Models.** Models span four architectural mechanisms. **QK-normalization** (per-head RMSNorm on queries and keys; Qwen3-4B, Gemma3-4B (Yang et al., 2025a; Gemma Team, 2025)); **gated attention** (Qwen3.5-4B, 9B (Qwen Team, 2026)); **learned attention sink** (per-head sink logit appended to the attention logits; GPT-OSS-20B); and **logit softcapping** (tanh-bounded attention logits; Gemma2-2B, 9B). Standard architectures (Llama 3.1-8B, 3.2-3B (Grattafiori et al., 2024); Qwen2.5-3B, 7B (Yang et al., 2024)) serve as controls.

**Baselines.** **Attention sink.** StreamingLLM (Xiao et al., 2024) (fixed sink + recent window). **Attention score.** TOVA (Oren et al., 2024) and SnapKV (Li et al., 2024). **Key geometry.** KeyNorm (Devoto et al., 2024) prefers smaller key norms; KEYDIFF (Park et al., 2025) uses cosine distance from the mean key; and ManifoldKV (Datta et al., 2026) uses Euclidean distance from a key centroid. They are key-side counterparts to VALUEDIFF-Norm, VALUEDIFF-Dir, and VALUEDIFF, respectively; KeyNorm reverses the radial preference. **Attention × value.** FastCAOTE (Goel et al., 2025) multiplies VALUEDIFF's value-distance score by $\alpha_j/(1-\alpha_j)$, isolating the attention factor. **Value geometry.** VALUEDIFF-Dir (direction-only), VALUEDIFF (L2), and VALUEDIFF-Norm (magnitude-only). We exclude KVZip (Kim et al., 2025): doubled-length prefill at every eviction is prohibitive under block prompt processing.

**Benchmarks.** We cover both long-context prefill and long decode. **RULER** (Hsieh et al., 2024) evaluates long-context capability in controlled 8k-context tasks spanning retrieval (NIAH (Kamradt, 2023)), variable tracking, aggregation, and multi-hop QA, while **LongBench** (Bai et al., 2024) tests transfer to 21 heterogeneous document tasks with contexts reaching up to 65k tokens. **MATH-500** (Hendrycks et al., 2021; Lightman et al., 2023) stresses decode-time cache growth during chain-of-thought reasoning.

Table 2: LongBench accuracy retention (% of dense, ↑) at **4k / 8k** budgets. **Bold** marks the best non-dense strategy by average retention over the two reported budgets for each model.

| | *Sink* | *Attn* | | *Key Geometry* | | | *Attn×Val.* | *Value Geometry* | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | Streaming | TOVA | SnapKV | KeyNorm | KeyDiff | ManifoldKV | F-CAOTE | V-Norm | V-Dir | VALUEDIFF |
| *Standard* | | | | | | | | | | |
| Llama3.1-8B | 55.0 / 74.4 | 69.0 / 88.4 | 70.6 / 91.0 | 73.5 / 90.3 | **82.6 / 93.6** | 71.7 / 91.9 | 71.4 / 90.0 | 65.1 / 85.2 | 79.7 / 91.5 | 73.1 / 88.0 |
| Llama3.2-3B | 70.0 / 79.4 | 74.8 / 85.5 | 75.9 / 85.6 | 76.6 / 86.3 | **85.4 / 89.2** | 76.9 / 86.8 | 76.4 / 85.3 | 69.6 / 82.2 | 79.7 / 87.5 | 75.9 / 85.1 |
| *QK-normalization* | | | | | | | | | | |
| Qwen3-4B | 78.0 / 89.4 | 81.3 / 92.0 | 83.8 / 93.7 | 71.9 / 87.4 | 85.0 / 94.6 | 89.5 / 97.1 | 84.0 / 94.9 | 95.2 / 98.8 | 91.2 / 97.4 | **95.3 / 98.7** |
| Gemma3-4B | 72.1 / 83.7 | 77.0 / 90.5 | 72.7 / 86.6 | 49.5 / 58.2 | 76.7 / 90.9 | 73.0 / 79.8 | 78.2 / 92.5 | 82.6 / 93.0 | **90.5 / 96.5** | 88.1 / 94.5 |
| *Gated attention* | | | | | | | | | | |
| Qwen3.5-4B | 77.7 / 87.9 | 84.0 / 91.2 | 83.1 / 93.0 | 79.4 / 92.0 | 78.6 / 90.7 | 80.5 / 93.1 | 86.2 / 94.1 | **92.8 / 97.6** | 88.4 / 95.8 | 92.2 / 95.5 |
| Qwen3.5-9B | 74.1 / 86.7 | 84.6 / 95.2 | 85.6 / 95.1 | 82.2 / 92.4 | 81.3 / 92.9 | 80.9 / 92.8 | 87.1 / 95.7 | 92.4 / 96.7 | 90.0 / 95.5 | **92.1 / 97.2** |
| *Learned Sink* | | | | | | | | | | |
| GPT-OSS-20B | 60.7 / 79.1 | 75.3 / 91.3 | 74.4 / 90.2 | 25.3 / 28.5 | 83.4 / 94.5 | 31.0 / 33.2 | 77.8 / 93.6 | 81.9 / 96.3 | 84.4 / 95.9 | **91.0 / 96.6** |

### 4.1 RULER

On RULER, we evaluate long-context capability at 8k context under tight cache budgets of 2k and 4k tokens. Table 1 reports retention versus each model's dense score across 13 RULER tasks. The results split across three geometry regimes consistent with Section 3.1. On value-dominant models (Qwen3.5, GPT-OSS-20B; $\sigma_V/\sigma_K > 1.3$), VALUEDIFF is the strongest. On mixed-geometry models ($\sigma_V/\sigma_K \approx 1$; Qwen3, Gemma2, Gemma3), VALUEDIFF generally performs best, with key-geometry methods such as KeyDiff and ManifoldKV remaining competitive. In this regime neither signal has a clear geometric advantage; VALUEDIFF leads on most models here, with Gemma2-2B the exception where KeyDiff leads—consistent with its $\sigma_V/\sigma_K$=0.88 placing it toward the key-dominant side (Section 4.4). On key-dominant non-suppressed controls ($\sigma_V/\sigma_K < 0.5$; Llama), KeyDiff and KeyNorm lead. Qwen2.5 is an outlier: both value- and key-geometry methods degrade substantially, consistent with the high per-layer geometry variance discussed in Section C.

**Sink and attention baselines.** The non-geometric baselines expose two different failure modes. Streaming's fixed-window design fails broadly at tight budgets regardless of sink behavior; on sink-suppressed models it additionally loses the stable BOS anchor that partially compensates on sink-dominated architectures. TOVA and SnapKV face a different problem. The attention scores available at a block boundary must predict token importance for later blocks and decoding, and the low rank overlap in Section D suggests that this proxy is unreliable.

**Key vs. value geometry.** Matched key/value pairs generally favor the value side on sink-suppressed models: VALUEDIFF usually matches or improves on ManifoldKV (centroid), and VALUEDIFF-Dir often improves over KeyDiff (directional). The norm-only comparison is less diagnostic because both KeyNorm and VALUEDIFF-Norm underperform their directional or L2 counterparts, indicating norm alone is not dominant. Overall, value geometry is the more robust query-invariant signal; we examine the underlying geometric conditions in Section 4.4.

**Comparison with FastCAOTE.** FastCAOTE keeps the same centroid-distance signal as VALUEDIFF but weights it by observed attention. The multiplier consistently hurts relative to VALUEDIFF (Table 1), matching the broader failure mode of attention-based eviction: observed attention is query-local, not a stable estimate of future cache utility (Section D).

**Direction and magnitude ablations.** Within the value-side family, VALUEDIFF-Dir is often close to the default score, suggesting that much of the useful signal is directional. However, VALUEDIFF-Norm is much less reliable on its own, and the default L2 score is the most stable variant overall. This suggests that magnitude is not a strong standalone eviction criterion, but it helps calibrate the directional signal when combined through centroid distance.

### 4.2 LONGBENCH

Through RULER, we confirmed that the value-geometry advantage holds in controlled, synthetic long-context conditions. In LongBench, we further evaluate the methods on 21 heterogeneous tasks spanning bilingual QA, multi-document summarization, code completion, and open-ended retrieval. In this more diverse setting, VALUEDIFF retains 88–95% of dense performance on sink-suppressed models at the tighter 4k budget (see Table 2). Note that logit softcapping models (Gemma2-2B

Table 3: MATH-500 reasoning retention (% of dense flex@5 except Qwen3.5; pass@1 greedy; ↑; $n = 100$) at **25% / 50%** cache budgets, set per model as fractions of average sequence length (see Section F). **Bold** marks the best non-dense strategy by average retention over the two budgets. †Sink on Qwen3.5-4B collapses output formatting; see Section H.

| | *Sink* | *Attn* | | *Key Geometry* | | | *Attn×Val.* | *Value Geometry* | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | Streaming | TOVA | SnapKV | KeyNorm | KeyDiff | ManifoldKV | F-CAOTE | V-Norm | V-Dir | VALUEDIFF |
| *Standard* | | | | | | | | | | |
| DS-R1-Qwen-1.5B | 74.2 / 89.7 | 70.3 / 88.8 | 81.1 / 91.1 | 45.7 / 75.4 | 89.2 / 99.0 | **95.9 / 97.1** | 75.1 / 93.1 | 85.4 / 94.5 | 79.2 / 93.1 | 89.0 / 97.6 |
| DS-R1-Llama-8B | 63.1 / 73.9 | 69.8 / 79.1 | 76.0 / 83.7 | 64.3 / 79.1 | **79.4 / 84.9** | 77.0 / 84.9 | 74.1 / 84.9 | 77.7 / 83.2 | 75.8 / 84.7 | 77.9 / 85.1 |
| *QK-normalization* | | | | | | | | | | |
| Gemma3-4B | 81.1 / 94.6 | **94.3 / 104.9** | 88.1 / 98.1 | 75.5 / 90.0 | 87.1 / 97.8 | 87.3 / 100.0 | 95.4 / 103.2 | 92.7 / 101.3 | 91.9 / 100.5 | 96.8 / 102.2 |
| Qwen3-4B | 71.3 / 87.9 | 81.7 / 92.0 | 83.8 / 93.1 | 67.2 / 86.4 | 79.5 / 90.9 | 87.3 / 94.2 | 84.9 / 93.5 | 87.9 / 95.7 | 85.3 / 94.2 | **90.5 / 96.3** |
| *Gated attention (pass@1 greedy)* | | | | | | | | | | |
| Qwen3.5-4B | 1.1† / 31.5 | 39.3 / 80.9 | 24.7 / 84.3 | 40.4 / 75.3 | 65.2 / 86.5 | 62.9 / 89.9 | 34.8 / 83.1 | 76.4 / 94.4 | 79.8 / 96.6 | **84.3 / 100.0** |
| Qwen3.5-9B | 2.2 / 27.0 | 42.7 / 82.0 | 19.1 / 73.0 | 48.3 / 80.9 | 68.5 / 89.9 | 71.9 / 92.1 | 49.4 / 89.9 | 87.6 / 94.4 | 87.6 / 97.8 | **92.1 / 100.0** |
| *Learned Sink* | | | | | | | | | | |
| GPT-OSS-20B | 73.9 / 87.1 | 71.9 / 86.5 | 74.9 / 88.4 | 61.7 / 84.4 | 89.1 / 96.1 | 90.1 / 95.9 | 75.6 / 90.1 | 91.2 / 96.8 | 76.7 / 91.0 | **93.1 / 96.6** |

and Gemma2-9B) are excluded from the evaluation as their trained context length is 8k (Gemma Team, 2024). Qwen2.5 models are also excluded: severe per-layer geometry heterogeneity causes all eviction strategies to degrade substantially regardless of signal type (see Section C), making their inclusion uninformative for the geometric comparison.

Magnitude-based key selection (KeyNorm, ManifoldKV) shows architecture-dependent failures in LongBench. On Gemma3-4B (QK-norm), KeyNorm collapses to 49.5% retention at the 4k budget; on GPT-OSS-20B (learned sink), both KeyNorm (25.3%) and ManifoldKV (31.0%) fail catastrophically. KeyDiff, insensitive to key magnitude, avoids these collapses and remains the strongest key-side method overall, though it is still outperformed by VALUEDIFF on sink-suppressed models.

### 4.3 MATH-500

On MATH-500, we examine a complementary pressure point, decode-time cache growth during chain-of-thought reasoning. We evaluate the models in Table 3 with model-specific generation protocols. Budgets are model-specific, set at approximately 25% and 50% of each model's average sequence length, so compression ratios remain comparable despite different decode lengths; dense decode lengths and truncation rates are reported in Section F. VALUEDIFF remains at or near the top across budgets on sink-suppressed models and is most clearly separated on Qwen3.5 at the tight 25% budget, where the strongest prior baselines trail by roughly 19–20 points. On the 100-problem subset, across sink-suppressed models, VALUEDIFF retains at least 84% of dense performance even at the tight budget, showing that the value-geometry signal transfers to reasoning. On the full 500-problem set at the 25% budget, VALUEDIFF leads the tested baselines on all five models (Table 13).

### 4.4 DISCUSSION

The empirical results above show that VALUEDIFF often matches or outperforms key-geometry baselines across architectures and benchmarks. We now characterize the geometric conditions underlying this pattern through a lens that completes the picture established in Section 3.1.

**BOS cosine similarity.** Sink rate (Section 3.1) is a post-softmax measure. It summarizes the attention output after temperature scaling and normalization, compressing the underlying key geometry into a threshold-crossing fraction. We instead examine a more direct geometric quantity, the cosine similarity between the BOS key vector and the mean of content key vectors,

$$\text{bos-cosim} \;=\; \cos\!\big(k_{\text{BOS}},\; \tfrac{1}{n-1}\textstyle\sum_{i>0} k_i\big).$$

This query-invariant, pre-softmax key-space measure captures whether BOS is geometrically distinct from the content keys: a negative bos-cosim is the structural signature of attention sinks, with BOS anti-aligned with the content mean, while a positive value indicates no such distinction.

**Two-axis picture.** The KeyDiff eviction score implicitly depends on this distinction. KeyDiff's theoretical analysis (Park et al., 2025) favors tokens whose keys diverge from the cache mean $\bar{k}$. When bos-cosim $< 0$, $\bar{k}$ is pulled toward the content direction and BOS is a clear geometric outlier:

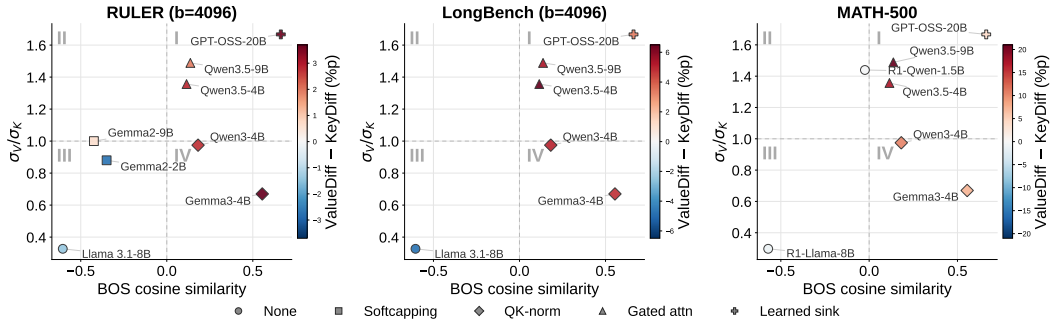



Figure 3: bos-cosim vs. $\sigma_V/\sigma_K$ for all models, colored by VALUEDIFF−KeyDiff gap, across RULER (left), LongBench (center), and MATH-500 (right).

Table 4: RULER 8k accuracy (%), $N$=2048: one-shot / blockwise prefill.

| Method | Llama-3.2-3B | Gemma3-4B | Qwen3-4B |
|---|---|---|---|
| TOVA | 86.5 / 55.0 | 88.2 / 50.3 | 94.4 / 42.6 |
| SnapKV | 89.7 / 66.5 | 91.2 / 60.8 | 94.2 / 50.6 |
| KeyDiff | 90.6 / **85.6** | 88.3 / 75.1 | 89.6 / 85.8 |
| FastCAOTE | 89.3 / 66.6 | 90.4 / 63.5 | 95.4 / 68.7 |
| VALUEDIFF | 80.4 / 81.8 | 85.8 / **80.5** | 93.4 / **91.2** |

Table 5: Qwen3-4B RULER (% dense), $N$=4096; 16k/32k/64k: 25/12.5/6.25%.

| Method | 16k | 32k | 64k |
|---|---|---|---|
| SnapKV | 65.8 | 55.4 | 46.8 |
| KeyDiff | 90.7 | 85.8 | 82.0 |
| ManifoldKV | 92.8 | 89.5 | 86.3 |
| FastCAOTE | 80.6 | 71.7 | 61.7 |
| VALUEDIFF | **94.6** | **90.9** | **88.0** |

the anchor is meaningful. When bos-cosim $> 0$, $\bar{k}$ is no longer a clean reference and the anchor dissolves, degrading KeyDiff's discriminative signal.

Figure 3 plots bos-cosim against $\sigma_V/\sigma_K$ for all evaluated models, colored by the VALUEDIFF−KeyDiff gap, across RULER, LongBench, and MATH-500. Two dashed lines divide the space into four quadrants. VALUEDIFF generally leads in Quadrants I, II, and IV, with R1-Qwen-1.5B on MATH-500 a near-parity exception. KeyDiff's clearest advantage is in Quadrant III (bos-cosim $< 0$ and $\sigma_V/\sigma_K < 1$), where the BOS anchor is intact and key geometry dominates. Gemma2-9B is particularly informative, sitting close to the $\sigma_V/\sigma_K$=1.0 boundary (bos-cosim $\approx -0.4$, $\sigma_V/\sigma_K \approx 1.0$) with near-parity between the two methods (+0.6 pt), consistent with neither signal being dominant.

### 4.5 PREFILL PROTOCOL AND CONTEXT LENGTH

**One-shot versus blockwise prefill.** Blockwise prefill (Equation (3)) makes eviction an online decision: after each $B$-token chunk, a method must commit without seeing the rest of the prompt, whereas one-shot prefill sees the full context before its first eviction. We therefore measure how much accuracy each method loses in this harder setting. At $N$=2048, the attention-based TOVA, SnapKV, and FastCAOTE drop 22.7–51.8 points from one-shot to blockwise prefill, while the query-invariant KeyDiff and VALUEDIFF change by at most 13.2 points (Table 4). This larger degradation is consistent with the modest cross-query overlap of attention rankings (top-256 Jaccard $\approx 0.4$; Section D): under blockwise prefill, attention-derived scores from observed queries must anticipate unseen later queries, whereas one-shot prefill can use the full prompt. Both settings use the same per-step decoding eviction. One-shot also uses 939.5 MB of peak KV-cache memory versus 249.6 MB for blockwise (3.76×).

**Longer contexts and tighter budgets.** With $N$=4096 fixed on Qwen3-4B, increasing RULER context length from 16k to 64k reduces the retained fraction from 25% to 6.25%. VALUEDIFF remains the strongest eviction method at all three lengths (Table 5).

### 4.6 LATENCY AND MEMORY

We focus on memory-constrained long-context inference, measuring peak memory during long-prefill processing and time-to-first-token (TTFT). Decode throughput depends on kernel-level cache management orthogonal to the eviction scoring rule.

**Peak memory.** Eviction caps the cache at $O(\text{budget})$ regardless of input length, while dense scales linearly (Figure 9). On Qwen3.5-4B at context length 128k, peak GPU memory drops from 15.0 GB to 8.6 GB ($-43\%$); on GPT-OSS-20B at context length 65k the reduction is 29%. These savings follow from the budget itself and are shared by token eviction methods.

**Prefill latency.** Under block prompt processing ($B$=128), each chunk after cache fill attends only to the post-eviction budget-sized cache. In our implementation, VALUEDIFF reduces TTFT relative to dense on both long-context settings and stays within 3–4% of the cheapest eviction baseline. These measurements are intended to characterize scoring overhead under a fixed block-processing implementation; optimized serving performance also depends on kernel-level cache management and implementation details. The full TTFT table is reported in Table 11. Among the evaluated eviction rules, VALUEDIFF has the same asymptotic online cost as TOVA, SnapKV, and KeyDiff; fixed-window sink eviction is cheaper but significantly less accurate on sink-suppressed models.

## 5 RELATED WORK

**Attention sinks.** Xiao et al. (2024) identified attention sinks and proposed StreamingLLM, which preserves fixed sink tokens plus a recent window. Follow-up work examines sink emergence and control (Barbero et al., 2025; Sun et al., 2024; Godey et al., 2024; Ruscio et al., 2025). Modern architectures also include QK-normalization (Dehghani et al., 2023), gated attention and hybrid linear/softmax models (Qiu et al., 2025; Yang et al., 2025b), logit softcapping (Gemma Team, 2024), and learned attention sinks (Agarwal et al., 2025), which are associated with weaker attention sinks. Our work connects this architectural trend to KV eviction. As sink behavior weakens, value geometry becomes a more useful query-invariant signal.

**Attention-score eviction.** $H_2O$ (Zhang et al., 2023), TOVA (Oren et al., 2024), and SnapKV (Li et al., 2024) score cached tokens using observed attention mass or local attention windows. KVZip (Kim et al., 2025) is query-agnostic but still attention-scored. It uses a repeat prompt to induce context reconstruction and scores KV pairs by the maximum cross-attention they receive. These methods rely on attention as a useful token-importance signal; our work studies how that signal changes under modern sink-suppression mechanisms.

**Key-geometric eviction.** Key-geometric methods rank tokens without using attention scores. Key-norm scoring ranks keys by $L_2$ norm (Devoto et al., 2024), KeyDiff (Park et al., 2025) scores cosine distance from the mean key, and ManifoldKV (Datta et al., 2026) scores Euclidean distance from a key centroid, the key-side counterpart of our L2 value score. These methods implicitly assume key geometry is the dominant eviction signal; our two-axis analysis (Section 4.4) shows this is architecture-dependent: key geometry leads where sinks are strong, value geometry leads where they weaken. VALUEDIFF uses the value side of the cache, motivated by our observation that value geometry becomes more discriminative than key geometry in sink-suppressed models.

**Attention-weighted value eviction.** FastCAOTE (Goel et al., 2025) is closest to VALUEDIFF in form, combining the same centroid-distance value score with an additional query-specific attention multiplier $\frac{\alpha_j}{1-\alpha_j}$. Under block prompt processing, this multiplier is computed from past queries and may not reflect future cache utility, introducing noise. Section A derives both scores from the same reconstruction objective: VALUEDIFF arises as the consistent max-entropy limit.

## 6 CONCLUSION

We revisited KV cache eviction for sink-suppressed LLMs and showed that weaker sinks co-occur with greater value-vector dispersion relative to key-vector dispersion. This motivates VALUEDIFF, the score $s_j = \|\mathbf{v}_j - \bar{\mathbf{v}}\|$, which also arises as the symmetric max-entropy solution to minimizing attention-output disturbance when future queries are unknown. Across RULER, LongBench, and MATH-500, VALUEDIFF is strongest on sink-suppressed models while remaining competitive on mixed and key-dominant models. Our analysis is observational rather than causal; the training-time origin of this geometric shift and adaptive hybrid scores for mixed-geometry models remain promising directions for future work.

### AI USE STATEMENT

We used generative AI tools to assist with experimental code, manuscript drafting and editing, mathematical derivations and proof checking, and table formatting.

### REPRODUCIBILITY STATEMENT

The VALUEDIFF score and eviction schedule are specified in Sections 2 and 3.2, and Section 4 describes the evaluation protocol, including the chunk size, cache budgets, models, baselines, and benchmarks. Complete derivations and assumptions appear in Sections A and B. The remaining appendices document the geometry and attention-stability measurements, latency and memory setup, dense decode statistics, full MATH-500 results, and failure analysis.

## REFERENCES


Sandhini Agarwal et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Federico Barbero, Álvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do LLMs attend to the first token? *arXiv preprint arXiv:2504.02732*, 2025.

Zhiyang Chen, Daliang Xu, Haiyang Shen, Chiheng Lou, Mengwei Xu, Shangguang Wang, Xin Jin, and Yun Ma. Accelerating mobile language model via speculative decoding and NPU-coordinated execution. *arXiv preprint arXiv:2510.15312*, 2025.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Debajyoti Datta, Trishala Neeraj, Bibek Paudel, Vyom Sharma, and Subhabrata Mukherjee. ManifoldKV: Training-free KV cache compression via euclidean outlier detection. *arXiv preprint arXiv:2602.08343*, 2026.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning (ICML)*, 2023.

Alessio Devoto, Yu Zhao, Simone Scardapane, and Pasquale Minervini. A simple and effective $L_2$ norm-based strategy for KV cache compression. *arXiv preprint arXiv:2406.11430*, 2024.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Nathan Godey, Éric de la Clergerie, and Benoît Sagot. Anisotropy is inherent to self-attention in transformers. *arXiv preprint arXiv:2401.12143*, 2024.

Raghavv Goel, Junyoung Park, Mukul Gagrani, Dalton Jones, Matthew Morse, Harper Langston, Mingu Lee, and Chris Lott. CAOTE: KV cache selection for LLMs via attention output error-based token eviction. *arXiv preprint arXiv:2504.14051*, 2025.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Tom Gunter et al. Apple intelligence foundation language models. *arXiv preprint arXiv:2407.21075*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

Edwin T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957.

Greg Kamradt. Needle in a haystack — pressure testing LLMs. https://github.com/gkamradt/LLMTest_NeedleInAHaystack, 2023.

Jang-Hyun Kim, Jinuk Kim, Sangwoo Kwon, Jae W. Lee, Sangdoo Yun, and Hyun Oh Song. KVzip: Query-agnostic KV cache compression with context reconstruction. *arXiv preprint arXiv:2505.23416*, 2025.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Matanel Oren, Michael Hassid, Nir Yarden, Yossi Adi, and Roy Schwartz. Transformers are multi-state RNNs. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Junyoung Park, Dalton Jones, Matthew J Morse, Raghavv Goel, Mingu Lee, and Chris Lott. KEY-DIFF: Key similarity-based KV cache eviction for long-context LLM inference in resource-constrained environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. *arXiv preprint arXiv:2505.06708*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Valeria Ruscio, Umberto Nanni, and Fabrizio Silvestri. What are you sinking? a geometric approach on attention sink. *arXiv preprint arXiv:2508.02546*, 2025.

Zunhai Su, Hengyuan Zhang, Wei Wu, Yifan Zhang, Yaxiu Liu, He Xiao, Qingyao Yang, Yuxuan Sun, Rui Yang, Chao Zhang, Keyu Fan, Weihao Ye, Jing Xiong, Hui Shen, Chaofan Tao, Taiqiang Wu, Zhongwei Wan, Yulei Qian, Yuchen Xie, and Ngai Wong. Attention sink in transformers: A survey on utilization, interpretation, and mitigation. *arXiv preprint arXiv:2604.10098*, 2026.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *Conference on Language Modeling (COLM)*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024.

Daliang Xu, Hao Zhang, Liming Yang, Ruiqi Liu, Gang Huang, Mengwei Xu, and Xuanzhe Liu. Fast on-device LLM inference with NPUs. *arXiv preprint arXiv:2407.05858*, 2024.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

An Yang et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations (ICLR)*, 2025b.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. $H_2O$: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

# A VALUEDIFF UNDER DIRICHLET MAX-ENTROPY, AND RELATIONSHIP TO FASTCAOTE

Proposition 3.1 in the main body establishes VALUEDIFF as the minimizer of the CAOTE reconstruction objective Equation (6) evaluated at the symmetric attention vector $\boldsymbol{\alpha}^\star = (1/n, \dots, 1/n)$, using only the observation that $\boldsymbol{\alpha}^\star$ is the mean of any exchangeable prior on $\Delta^{n-1}$. This appendix formalizes the prior via Jaynes' max-entropy principle (Jaynes, 1957). On $\Delta^{n-1}$ without further constraints, the maximum-entropy distribution is Dirichlet$(1, \dots, 1)$, a specific exchangeable prior with strictly positive density throughout the simplex. We (i) recover the VALUEDIFF ranking under the full expectation $\mathbb{E}_P[\|\mathbf{o}(\boldsymbol{\alpha}) - \mathbf{o}_{\setminus j}(\boldsymbol{\alpha})\|^2]$ with $P = \text{Dirichlet}(1, \dots, 1)$, and (ii) clarify the precise relationship to FastCAOTE (Goel et al., 2025), whose score corresponds to a partial max-entropy approximation of the same objective.

**Claim.** Let $\boldsymbol{\alpha} \sim \text{Dirichlet}(1, \dots, 1)$ on $\Delta^{n-1}$ and $\bar{\mathbf{v}} = \frac{1}{n}\sum_i \mathbf{v}_i$. Then

$$\mathbb{E}_{\boldsymbol{\alpha}}\left[\left(\frac{\alpha_j}{1-\alpha_j}\right)^2 \|\mathbf{v}_j - \mathbf{o}(\boldsymbol{\alpha})\|^2\right] \;=\; \frac{2}{n(n-1)}\,\|\mathbf{v}_j - \bar{\mathbf{v}}\|^2 \;+\; C(n), \tag{A.1}$$

where $\mathbf{o}(\boldsymbol{\alpha}) = \sum_i \alpha_i \mathbf{v}_i$ and $C(n)$ depends only on $n$ and the cache (not on $j$). The integrand is the per-query squared reconstruction error after evicting $j$ and renormalizing the softmax, with $\|\mathbf{o} - \mathbf{o}'\|^2 = (\alpha_j/(1-\alpha_j))^2\|\mathbf{v}_j - \mathbf{o}\|^2$ and $\mathbf{o}' = (\mathbf{o} - \alpha_j\mathbf{v}_j)/(1-\alpha_j)$. The minimizer is $j^\star = \arg\min_j \|\mathbf{v}_j - \bar{\mathbf{v}}\|$, matching the main-body proposition and confirming that VALUEDIFF is robust to the specific form of max-entropy adopted.

The proof proceeds in three steps. First, a renormalization identity rewrites the integrand in a more tractable form. Second, a Dirichlet independence property factors the expectation. Third, elementary moment calculations finish the reduction.

**Step 1. Renormalization identity.** Define the post-eviction output $\mathbf{o}'_j := \sum_{i\neq j}(\alpha_i/(1-\alpha_j))\mathbf{v}_i$. Direct algebra gives

$$\mathbf{v}_j - \mathbf{o} \;=\; (1-\alpha_j)\mathbf{v}_j - \sum_{i\neq j}\alpha_i\mathbf{v}_i \;=\; (1-\alpha_j)(\mathbf{v}_j - \mathbf{o}'_j),$$

so that

$$\left(\frac{\alpha_j}{1-\alpha_j}\right)^2 \|\mathbf{v}_j - \mathbf{o}\|^2 \;=\; \alpha_j^2\|\mathbf{v}_j - \mathbf{o}'_j\|^2. \tag{A.2}$$

**Step 2. Dirichlet independence.** A standard property of the Dirichlet distribution states that if $\boldsymbol{\alpha} \sim \text{Dirichlet}(a_1, \dots, a_n)$, then the marginal $\alpha_j \sim \text{Beta}(a_j, \sum_{i\neq j} a_i)$ is *independent* of the renormalized other weights $\boldsymbol{\beta} = (\alpha_i/(1-\alpha_j))_{i\neq j} \sim \text{Dirichlet}(a_{-j})$. In the symmetric case $a_i = 1$ this gives

$$\alpha_j \sim \text{Beta}(1, n-1), \qquad \boldsymbol{\beta} \sim \text{Dirichlet}(1, \dots, 1) \text{ on } \Delta^{n-2}, \qquad \alpha_j \perp \boldsymbol{\beta}.$$

Since $\mathbf{o}'_j$ is a function only of $\boldsymbol{\beta}$ (and the fixed cache values), $\mathbf{o}'_j \perp \alpha_j$. Combined with Equation (A.2),

$$\mathbb{E}\left[\left(\frac{\alpha_j}{1-\alpha_j}\right)^2 \|\mathbf{v}_j - \mathbf{o}\|^2\right] \;=\; \mathbb{E}[\alpha_j^2]\cdot\mathbb{E}[\|\mathbf{v}_j - \mathbf{o}'_j\|^2]. \tag{A.3}$$

**Step 3. Moments.** *First factor.* For $\alpha_j \sim \text{Beta}(1, n-1)$, standard Beta moments give $\mathbb{E}[\alpha_j] = 1/n$ and

$$\mathbb{E}[\alpha_j^2] \;=\; \frac{2}{n(n+1)}. \tag{A.4}$$

*Second factor.* Since $\sum_{i\neq j}\beta_i = 1$, we can rewrite

$$\mathbf{v}_j - \mathbf{o}'_j \;=\; \sum_{i\neq j}\beta_i(\mathbf{v}_j - \mathbf{v}_i) \;=\; \sum_{i\neq j}\beta_i\,\mathbf{d}_i, \tag{A.5}$$

where $\mathbf{d}_i := \mathbf{v}_j - \mathbf{v}_i$. Standard Dirichlet(1,...,1) moments on $\Delta^{n-2}$ give $\mathbb{E}[\beta_i^2] = 2/((n-1)n)$ and $\mathbb{E}[\beta_i\beta_k] = 1/((n-1)n)$ for $i \neq k$, which can be combined as $\mathbb{E}[\beta_i\beta_k] = (1+\mathbf{1}[i=k])/((n-1)n)$. Therefore

$$\mathbb{E}\Big[\big\|\textstyle\sum_{i\neq j}\beta_i\mathbf{d}_i\big\|^2\Big] \;=\; \frac{1}{(n-1)n}\Big(\textstyle\sum_{i\neq j}\|\mathbf{d}_i\|^2 \;+\; \big\|\sum_{i\neq j}\mathbf{d}_i\big\|^2\Big). \tag{A.6}$$

Both sums reduce to cache-level quantities. Using $\sum_{i\neq j}\mathbf{v}_i = n\bar{\mathbf{v}} - \mathbf{v}_j$,

$$\Big\|\textstyle\sum_{i\neq j}\mathbf{d}_i\Big\|^2 \;=\; \big\|(n-1)\mathbf{v}_j - (n\bar{\mathbf{v}} - \mathbf{v}_j)\big\|^2 \;=\; n^2\|\mathbf{v}_j - \bar{\mathbf{v}}\|^2. \tag{A.7}$$

Expanding the squared distances with $S_2 := \sum_i\|\mathbf{v}_i\|^2$ and using $\|\mathbf{v}_j\|^2 - 2\mathbf{v}_j^\top\bar{\mathbf{v}} = \|\mathbf{v}_j - \bar{\mathbf{v}}\|^2 - \|\bar{\mathbf{v}}\|^2$,

$$\sum_{i\neq j}\|\mathbf{d}_i\|^2 \;=\; n\|\mathbf{v}_j\|^2 - 2n\mathbf{v}_j^\top\bar{\mathbf{v}} + S_2 \;=\; n\|\mathbf{v}_j - \bar{\mathbf{v}}\|^2 + \big(S_2 - n\|\bar{\mathbf{v}}\|^2\big). \tag{A.8}$$

Substituting Equations (A.7) and (A.8) into Equation (A.6),

$$\mathbb{E}[\|\mathbf{v}_j - \mathbf{o}'_j\|^2] \;=\; \frac{n+1}{n-1}\,\|\mathbf{v}_j - \bar{\mathbf{v}}\|^2 \;+\; \frac{S_2 - n\|\bar{\mathbf{v}}\|^2}{(n-1)n}. \tag{A.9}$$

The second term depends only on $n$ and the cache, not on $j$.

**Conclusion.** Substituting Equations (A.4) and (A.9) into Equation (A.3),

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\alpha_j}{1-\alpha_j}\right)^2\|\mathbf{v}_j - \mathbf{o}\|^2\right] &= \frac{2}{n(n+1)}\cdot\left(\frac{n+1}{n-1}\|\mathbf{v}_j - \bar{\mathbf{v}}\|^2 + \frac{S_2 - n\|\bar{\mathbf{v}}\|^2}{(n-1)n}\right)\\ &= \frac{2}{n(n-1)}\|\mathbf{v}_j - \bar{\mathbf{v}}\|^2 + \underbrace{\frac{2\,(S_2 - n\|\bar{\mathbf{v}}\|^2)}{n^2(n-1)(n+1)}}_{=:\,C(n)},\end{aligned}$$

which is Equation (A.1). Since $C(n)$ does not depend on $j$, the minimizer of the expected reconstruction error is $j^\star = \arg\min_j\|\mathbf{v}_j - \bar{\mathbf{v}}\|$. ■

**Relationship to FastCAOTE.** The renormalization identity in Step 1, taken in unsquared form, reads

$$\big\|\mathbf{o}(\boldsymbol{\alpha}) - \mathbf{o}'_j\big\| \;=\; \frac{\alpha_j}{1-\alpha_j}\,\|\mathbf{v}_j - \mathbf{o}(\boldsymbol{\alpha})\|, \tag{A.10}$$

where $\mathbf{o}'_j = \sum_{i\neq j}(\alpha_i/(1-\alpha_j))\mathbf{v}_i$ is the post-eviction softmax-renormalized output. FastCAOTE (Goel et al., 2025) evaluates Equation (A.10) at an observed attention vector $\tilde{\boldsymbol{\alpha}}$ and applies a single approximation, $\mathbf{o}(\tilde{\boldsymbol{\alpha}}) \to \bar{\mathbf{v}}$, yielding the score

$$s_j^{\text{FastCAOTE}} \;=\; \frac{\tilde{\alpha}_j}{1-\tilde{\alpha}_j}\,\|\mathbf{v}_j - \bar{\mathbf{v}}\| \tag{A.11}$$

(Eq. 23 of FastCAOTE (Goel et al., 2025)). The substitution $\mathbf{o} \to \bar{\mathbf{v}}$ is itself max-entropy in character. Under $\boldsymbol{\alpha} \sim$ Uniform, $\mathbb{E}[\mathbf{o}(\boldsymbol{\alpha})] = \bar{\mathbf{v}}$, and the approximation is exact at $\boldsymbol{\alpha}^\star$. The factor $\tilde{\alpha}_j/(1-\tilde{\alpha}_j)$, in contrast, retains the observed attention weight without a corresponding max-entropy substitution.

Applying max-entropy consistently to *both* quantities, as VALUEDIFF does, additionally substitutes $\alpha_j \to 1/n$, giving $\alpha_j/(1-\alpha_j) \to 1/(n-1)$. This collapses Equation (A.11) to $\|\mathbf{v}_j - \bar{\mathbf{v}}\|/(n-1)$, and the $j$-independent factor $1/(n-1)$ drops out of the argmin, recovering VALUEDIFF. From this angle, VALUEDIFF is the consistent max-entropy limit of the same reconstruction error in which FastCAOTE is rooted, differing only in that the attention weighting is also max-entropy approximated and thus vanishes from the ranking.

## B Distance from the Symmetric Limit

The symmetric reference point in Section 3.3 is useful because it makes the absence of future query information explicit. In this appendix we analyze the eviction error in the more general setting where $\boldsymbol{\alpha}$ may deviate from $\boldsymbol{\alpha}^*$, and quantify how that deviation controls the gap from the exact rule. For any realized attention vector, the attention output decomposes as

$$\mathbf{o}(\boldsymbol{\alpha}) = \sum_i \alpha_i \mathbf{v}_i = \bar{\mathbf{v}} + \sum_i \left(\alpha_i - \tfrac{1}{n}\right) \mathbf{v}_i, \qquad \bar{\mathbf{v}} = \tfrac{1}{n} \sum_i \mathbf{v}_i. \tag{A.12}$$

Substituting Equation (A.12) into Equation (A.10) makes the dependence on the attention weights explicit:

$$e_j(\boldsymbol{\alpha}) = \frac{\alpha_j}{1-\alpha_j} \left\| (\mathbf{v}_j - \bar{\mathbf{v}}) - \textstyle\sum_i \left(\alpha_i - \tfrac{1}{n}\right) \mathbf{v}_i \right\|. \tag{A.13}$$

When $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$, the sum $\sum_i (\alpha_i - 1/n)\mathbf{v}_i$ vanishes and $\alpha_j = 1/n$, so Equation (A.13) reduces to the ValueDiff score $s_j = \|\mathbf{v}_j - \bar{\mathbf{v}}\|$ scaled by $1/(n-1)$, recovering Proposition 3.1 exactly. To bound Equation (A.13) for general $\boldsymbol{\alpha}$, we introduce a scalar measuring how far $\boldsymbol{\alpha}$ departs from $\boldsymbol{\alpha}^*$:

$$\delta(\boldsymbol{\alpha}) = \left\| \boldsymbol{\alpha} - \tfrac{1}{n}\mathbf{1} \right\|_1, \tag{A.14}$$

the total attention mass displaced from uniform — zero when $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$ and growing as attention concentrates. $\delta(\boldsymbol{\alpha})$ captures the attention weight departure, but the resulting displacement of the attention output also depends on how spread out the cached value vectors are. The following bound makes this connection explicit, combining the attention deviation $\delta$ with the value geometry to translate one into an output error.

For the sum in Equation (A.13), let $\varepsilon_i = \alpha_i - 1/n$ and let $\mathcal{V} = \{\mathbf{v}_i\}_{i=1}^n$, $\operatorname{diam}(\mathcal{V}) = \max_{i,k} \|\mathbf{v}_i - \mathbf{v}_k\|$. Since $\sum_i \varepsilon_i = 0$, the total positive mass $D^+ = \sum_{i:\varepsilon_i>0} \varepsilon_i$ equals the total negative mass $\sum_{i:\varepsilon_i<0} |\varepsilon_i|$, and both equal $\delta/2$. For $\delta > 0$ (the $\delta = 0$ case is immediate), define convex combinations $\mathbf{p} = \sum_{i:\varepsilon_i>0} (\varepsilon_i/D^+)\mathbf{v}_i$ and $\mathbf{q} = \sum_{i:\varepsilon_i<0} (|\varepsilon_i|/D^-)\mathbf{v}_i$; then $\sum_i \varepsilon_i \mathbf{v}_i = \frac{\delta}{2}(\mathbf{p} - \mathbf{q})$ and $\|\mathbf{p} - \mathbf{q}\| \le \operatorname{diam}(\mathcal{V})$. Therefore

$$\left\| \sum_i \left(\alpha_i - \tfrac{1}{n}\right) \mathbf{v}_i \right\| \le \frac{\delta(\boldsymbol{\alpha})}{2} \operatorname{diam}(\mathcal{V}). \tag{A.15}$$

For $\delta(\boldsymbol{\alpha}) < 1 - 1/n$, applying the triangle inequality to Equation (A.13) with Equation (A.15) and $\alpha_j \le 1/n + \delta(\boldsymbol{\alpha})$ gives

$$e_j(\boldsymbol{\alpha}) \;\le\; \underbrace{\varphi\big(\delta(\boldsymbol{\alpha})\big)}_{\text{attention factor}} \Big( \underbrace{s_j}_{\text{ValueDiff score}} + \tfrac{\delta(\boldsymbol{\alpha})}{2} \underbrace{\operatorname{diam}(\mathcal{V})}_{\text{value spread}} \Big), \qquad \varphi(\delta) \;=\; \frac{\frac{1}{n} + \delta}{1 - \frac{1}{n} - \delta}. \tag{A.16}$$

The attention factor $\varphi(\delta)$ bounds $\alpha_j/(1-\alpha_j)$ and equals $1/(n-1)$ at uniform attention. The exact eviction cost at the symmetric point is $s_j/(n-1)$.

## C KV Geometry Across Architectures

### Aggregate statistics

Table 6 reports sink rate and key/value L2 diversity for all 11 models in the same order as Table 1. Statistics are computed on 20 calibration prompts spanning diverse content categories (narrative, QA, code, math, retrieval), each processed at context length 2048. $\sigma_K$ and $\sigma_V$ are the standard deviation of per-token L2 distances from the cache mean, averaged over layers, heads, and the 20 prompts.

Table 6: Sink rate and K/V L2 diversity across 11 models. $\sigma_V/\sigma_K > 1$ indicates value-dominant geometry.

| Model | Mechanism | Sink rate | $\sigma_K$ | $\sigma_V$ | $\sigma_V/\sigma_K$ |
|---|---|---|---|---|---|
| Llama 3.1-8B | None | 0.801 | 1.783 | 0.583 | 0.33 |
| Llama 3.2-3B | None | 0.835 | 1.478 | 0.584 | 0.40 |
| Qwen2.5-3B | None | 0.446 | 1.561 | 1.851 | 1.19 |
| Qwen2.5-7B | None | 0.392 | 1.597 | 2.133 | 1.34 |
| Gemma2-2B | Softcapping | 0.464 | 1.807 | 1.590 | 0.88 |
| Gemma2-9B | Softcapping | 0.542 | 1.808 | 1.808 | 1.00 |
| Qwen3-4B | QK-norm | 0.548 | 2.322 | 2.263 | 0.97 |
| Gemma3-4B | QK-norm | 0.271 | 12.506 | 8.377 | 0.67 |
| Qwen3.5-4B | Gated attn | 0.015 | 1.596 | 2.164 | 1.36 |
| Qwen3.5-9B | Gated attn | 0.012 | 2.344 | 3.488 | 1.49 |
| GPT-OSS-20B | Learned sink | 0.029 | 2.191 | 3.652 | 1.67 |

### BOS TOKEN Q-K ALIGNMENT AND KEY-SPACE GEOMETRIC DISTINCTIVENESS

Table 7 reports two complementary signals that characterize the attention sink mechanism across our 11 evaluation models. *BOS Q-K weight* ($\bar{\alpha}_{\text{BOS}}$) is the mean attention mass received by the BOS token from non-BOS query positions ($t \geq 1$),

$$\bar{\alpha}_{\text{BOS}} = \frac{1}{T-1} \sum_{t=1}^{T-1} \alpha_{t,0},$$

where $\alpha_{t,0}$ is the attention weight query position $t$ assigns to position 0 (BOS), and the sum excludes $t{=}0$ since the BOS token can only attend to itself under causal masking. High values indicate the BOS key vector aligns strongly with most query vectors, yielding disproportionate attention regardless of query content. *BOS key cosim* is the cosine similarity between the BOS key vector and the mean of non-BOS key vectors ($i \geq 1$),

$$\text{bos-cossim} = \cos\Big(k_0,\ \tfrac{1}{T-1}\textstyle\sum_{i=1}^{T-1} k_i\Big),$$

where $k_0$ is the BOS key and $k_i$ ($i \geq 1$) are the remaining key vectors. Negative values indicate the BOS key is a geometric outlier in key space (Figure 6). Together, these two signals capture the dual mechanism that prior eviction methods rely on: high Q-K alignment makes BOS a reliable attentional anchor, while geometric distinctiveness from content keys makes BOS a reliable key-geometry outlier. Both signals weaken in architectures where sinks are suppressed.

Table 7: BOS token Q-K alignment and key-space geometric distinctiveness across 11 models. High $\bar{\alpha}_{\text{BOS}}$ indicates strong persistent Q-K alignment (attentional anchor). Negative BOS key cosim indicates the BOS key is anti-aligned with the content key mean (geometric outlier).

| Model | Mechanism | $\bar{\alpha}_{\text{BOS}}$ | BOS key cos. sim. |
|---|---|---|---|
| Llama 3.1-8B | None | 0.591 | −0.604 |
| Llama 3.2-3B | None | 0.612 | −0.586 |
| Qwen2.5-3B | None | 0.317 | 0.091 |
| Qwen2.5-7B | None | 0.279 | 0.337 |
| Gemma2-2B | Softcapping | 0.333 | −0.350 |
| Gemma2-9B | Softcapping | 0.396 | −0.424 |
| Qwen3-4B | QK-norm | 0.408 | 0.182 |
| Gemma3-4B | QK-norm | 0.199 | 0.555 |
| Qwen3.5-4B | Gated attn | 0.034 | 0.115 |
| Qwen3.5-9B | Gated attn | 0.031 | 0.137 |
| GPT-OSS-20B | Learned sink | 0.030 | 0.663 |

To further characterize the geometric regime difference, we compare $\sigma_V/\sigma_K$ against two additional metrics. The *weight-space spectral norm ratio* $\|\mathbf{W}_V\|_2/\|\mathbf{W}_K\|_2$ is entirely data-independent: it

requires no calibration text and captures only how much each projection amplifies a unit-norm input, a property fixed in the model weights. The *activation-space spectral norm ratio* $\sigma_1(V)/\sigma_1(K)$, where $\sigma_1(\cdot)$ denotes the largest singular value of the stacked activation matrix, is data-dependent like $\sigma_V/\sigma_K$ but measures a different geometric property (dominant variance direction rather than radial spread).

Table 8: Three geometric metrics across architectures. The weight ratio is data-independent (projection matrix structure only); the other two require forward passes on calibration text. All three show the same broad regime separation.

| Model | Mechanism | $\|W_V\|_2/\|W_K\|_2$ | $\sigma_1(V)/\sigma_1(K)$ | $\sigma_V/\sigma_K$ |
|---|---|---|---|---|
| Llama 3.1-8B | None | 0.192 | 0.057 | 0.33 |
| Llama 3.2-3B | None | 0.195 | 0.071 | 0.40 |
| Qwen2.5-3B | None | 0.668 | 0.216 | 1.19 |
| Qwen2.5-7B | None | 0.743 | 0.194 | 1.34 |
| Gemma2-2B | Softcapping | 0.550 | 0.257 | 0.88 |
| Gemma2-9B | Softcapping | 0.632 | 0.288 | 1.00 |
| Qwen3-4B | QK-norm | 0.651 | 0.320 | 0.97 |
| Gemma3-4B | QK-norm | 0.497 | 0.397 | 0.67 |
| Qwen3.5-4B | Gated attn | 1.174 | 1.358 | 1.36 |
| Qwen3.5-9B | Gated attn | 1.134 | 1.339 | 1.49 |
| GPT-OSS-20B | Learned sink | 1.120 | 0.815 | 1.67 |

Two observations follow from Table 8. First, all three metrics show the same broad regime separation: Llama at the key-dominant extreme, Qwen3.5 and GPT-OSS at the value-dominant extreme, with intermediate architectures between. Second, and more importantly, the weight ratio and the activation ratio diverge substantially in magnitude. For Llama, $\|\mathbf{W}_V\|_2/\|\mathbf{W}_K\|_2 = 0.19$ but $\sigma_1(V)/\sigma_1(K) = 0.06$—the activation-level key dominance is roughly three times stronger than the weight-level bias alone predicts. For Qwen3.5 the gap runs in the opposite direction (1.17 vs 1.36). This systematic divergence shows that token content contributes an independent geometric signal on top of the projection structure: $\sigma_V/\sigma_K$ measures a genuine property of how tokens distribute in the projected spaces, not an artifact of how the weight matrices are scaled.

### PER-LAYER SINK RATE

Figure 4 shows the per-layer sink rate ($\tau = 0.3$) for all 11 models. Non-suppressed models (Llama) maintain high sink rates consistently across all layers. Sink-suppression models split into two patterns: gated-attention and learned-bias models (Qwen3.5, GPT-OSS) show near-zero sink rates throughout, while QK-norm and softcapping models (Qwen3, Gemma2, Gemma3) show intermediate rates with moderate layer-to-layer variation. Qwen2.5 is the sole exception with sharp, alternating oscillations discussed further in Section C.




Figure 4: Per-layer sink rate ($\tau = 0.3$) for all 11 models. Llama maintains high sink rates throughout; Qwen3.5 and GPT-OSS show near-zero; mixed-regime models show intermediate rates. Qwen2.5-3B uniquely exhibits sharp oscillations attributable to its 2 KV heads.

### PER-LAYER $\sigma_K$ AND $\sigma_V$

Figure 5 shows $\sigma_K^{(\ell)}$ and $\sigma_V^{(\ell)}$ at each attention layer for all 11 models. Error bars reflect per-head standard deviation within each layer, averaged over samples. For Qwen3.5, only the 8 full-attention layers are shown; the x-axis indexes attention layers sequentially.

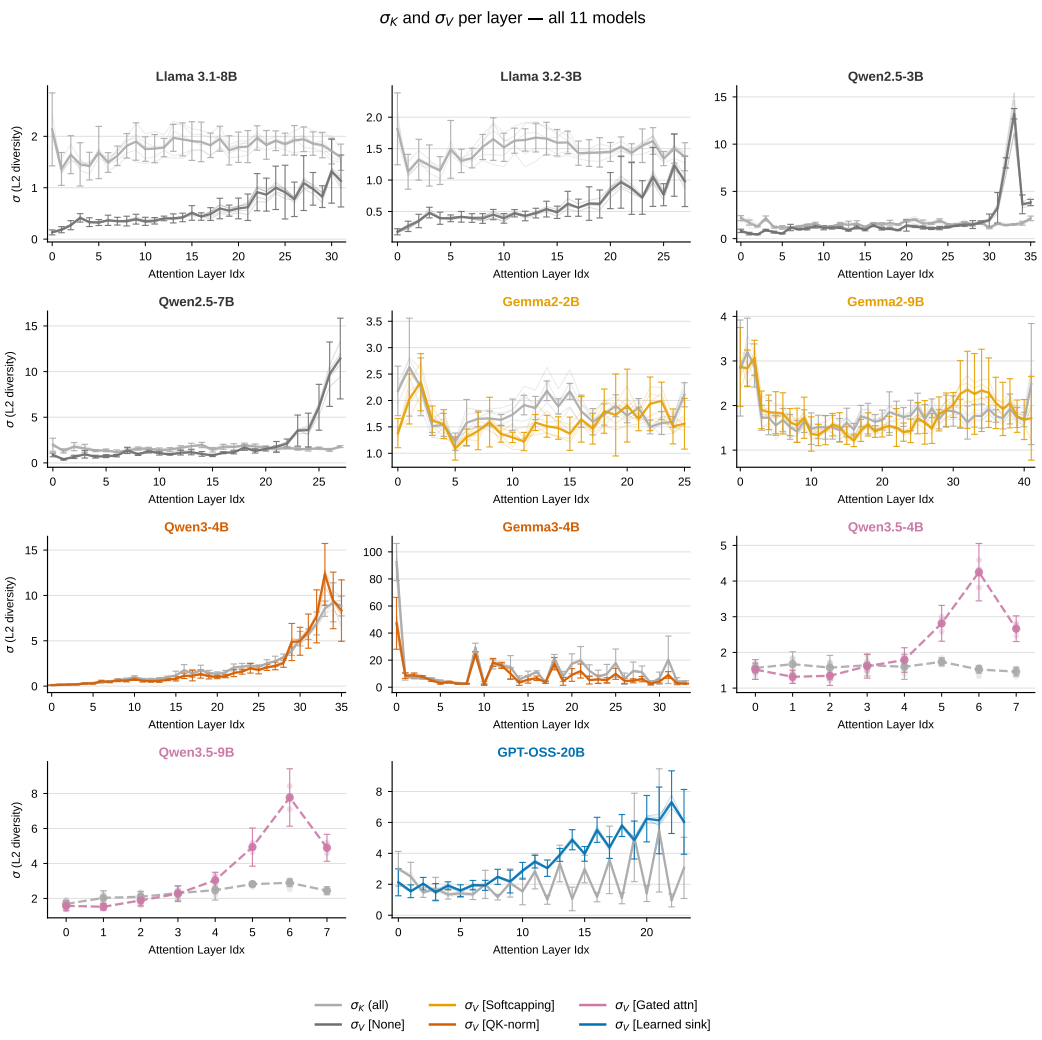



Figure 5: Per-layer key and value L2 diversity ($\sigma_K^{(\ell)}$, grey; $\sigma_V^{(\ell)}$, mechanism color) for all 11 models. Error bars show per-head standard deviation. Qwen3.5 shows only 8 full-attention layers.

## PER-LAYER BOS KEY COSINE SIMILARITY

Figure 6 shows the mean BOS key cosine similarity $\text{bos-cossim}^{(\ell)}$ at each attention layer. Negative values indicate the BOS key vector is a geometric outlier relative to the content key mean—the structural signature of an attention sink. Positive values indicate BOS is not geometrically distinct from content keys. The pattern mirrors the aggregate picture in Figure 1: Llama maintains consistently negative values throughout all layers (strong geometric sink), Qwen3.5 and GPT-OSS are near zero or positive (sink suppressed), and QK-norm and softcapping models (Qwen3, Gemma2, Gemma3) sit in between.




Figure 6: Per-layer BOS key cosine similarity for all 11 models. Values below zero indicate a geometrically distinct BOS key (attention sink); values above zero indicate sink suppression. The dashed line marks bos-cossim $= 0$.

### PER-LAYER $\sigma_V/\sigma_K$

Figure 7 shows the per-layer $\sigma_V/\sigma_K$ ratio. The per-layer profiles reflect the three geometry regimes discussed in Section 4.1. **Value-dominant models** (GPT-OSS-20B, Qwen3.5) tend toward higher $\sigma_V/\sigma_K$ in later attention layers, although some early layers fall below 1. **Mixed-regime models** (Qwen3, Gemma2, Gemma3) vary around the key/value boundary with smaller excursions than Qwen2.5; even here, VALUEDIFF generally comes out ahead as the value signal retains an edge (Section 4.1). **Key-dominant models** (Llama) are consistently below 1 across all layers, where KeyDiff is the strongest method and VALUEDIFF remains competitive. These regime labels summarize aggregate geometry rather than requiring every layer to lie on the same side of 1.

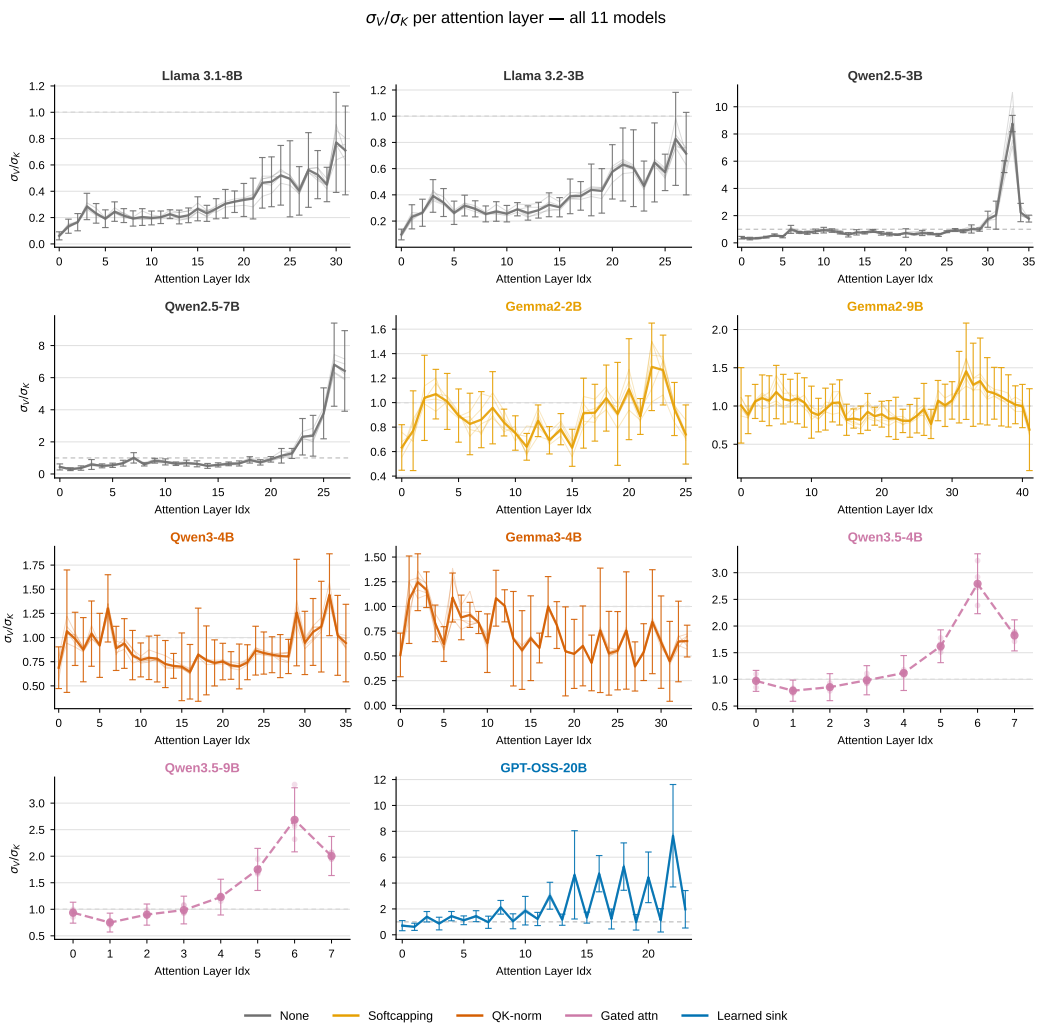



Figure 7: Per-layer $\sigma_V/\sigma_K$ ratio for all 11 models. The dashed horizontal line marks $\sigma_V = \sigma_K$. Aggregate regime labels do not require every layer to remain on the same side of 1.

### THE QWEN2.5 ANOMALY

Qwen2.5 combines substantial layerwise variation in sink rate (Figure 4) with large excursions in $\sigma_V/\sigma_K$. Its ratio crosses the key/value boundary in both directions with extreme amplitude (Figure 7): Qwen2.5-3B ranges from 0.31 to 8.77 (std $= 1.56$) and Qwen2.5-7B from 0.30 to 6.81 (std $= 1.63$). Figure 8 summarizes this joint variability: plotting per-layer sink rate std against $\sigma_V/\sigma_K$ std across all 11 models, Qwen2.5 is the *only* family occupying the high-high quadrant. Qwen3 has comparable sink rate variance (std $= 0.31$) but stable $\sigma_V/\sigma_K$ (std $= 0.18$); GPT-OSS has high $\sigma_V/\sigma_K$ variance (std $= 1.78$) but near-zero sink oscillation (std $= 0.05$). Its ratio is largest in later layers, with some early layers below 1. Thus, large ratio variance alone does not imply the joint instability seen in Qwen2.5, which complicates both attention-based and geometry-based token selection.

We also confirmed that Qwen 2.5 models have extreme grouped-query attention ratio. Table 9 shows the GQA ratios across all 11 models: Qwen2.5-3B and 7B have the highest ratios (8:1 and 7:1) among all-full-attention models. GPT-OSS-20B shares an 8:1 ratio but uses a 128-token sliding window in half its layers, substantially reducing the diversity burden per KV pair. For Qwen2.5, the high GQA compression means each cached KV pair must serve many query heads simultaneously across the full context, leaving the cache dense with non-redundant information and little room for token eviction without significant information loss.

Table 9: GQA compression ratios and effective attention scope across 11 models. Qwen2.5-3B and 7B have the highest GQA ratios among all-full-attention models; GPT-OSS shares an 8:1 ratio but limits the context burden via a 128-token sliding window in half its layers.

| Model | Mechanism | Q heads | KV heads | GQA ratio | Attention scope |
|---|---|---|---|---|---|
| Llama 3.1-8B | None | 32 | 8 | 4:1 | Full |
| Llama 3.2-3B | None | 24 | 8 | 3:1 | Full |
| Qwen2.5-3B | None | 16 | 2 | **8:1** | Full |
| Qwen2.5-7B | None | 28 | 4 | **7:1** | Full |
| Gemma2-2B | Softcapping | 8 | 4 | 2:1 | Local 4096 (alt.) |
| Gemma2-9B | Softcapping | 16 | 8 | 2:1 | Local 4096 (alt.) |
| Qwen3-4B | QK-norm | 32 | 8 | 4:1 | Full |
| Gemma3-4B | QK-norm | 8 | 4 | 2:1 | Local 1024 (alt.) |
| Qwen3.5-4B | Gated attn | 16 | 4 | 4:1 | Linear+full hybrid |
| Qwen3.5-9B | Gated attn | 16 | 4 | 4:1 | Linear+full hybrid |
| GPT-OSS-20B | Learned sink | 64 | 8 | **8:1** | Local 128 (half layers) |

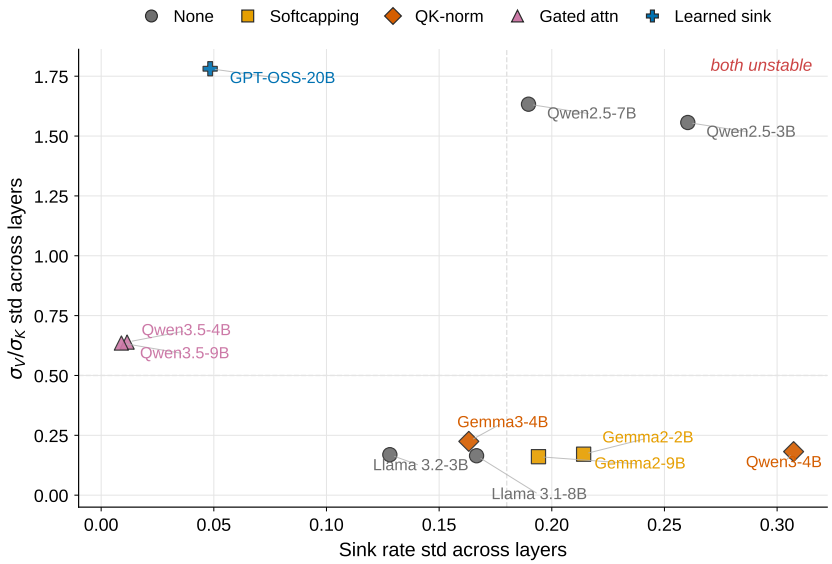



Figure 8: Per-layer instability of sink rate (x-axis) vs. $\sigma_V/\sigma_K$ (y-axis) across the same 11 models as Figure 1; Qwen3-4B denotes Qwen3-4B-Instruct. Qwen2.5 alone occupies the top-right quadrant, where both signals vary strongly across layers.

# D Attention Ranking Stability Across Architectures

## Motivation

Attention-score eviction methods (TOVA (Oren et al., 2024), SnapKV (Li et al., 2024), $H_2O$ (Zhang et al., 2023)) retain tokens with the highest observed attention weight. Under block prompt processing, the eviction decision at block boundary $b$ is made using only the attention computed from queries in blocks $1 \ldots b$; queries in all later blocks and all decode steps are unavailable. For this to work, the top-$k$ set identified by queries in block $b$ must be a good proxy for what future queries will need. We measure how stable the top-$k$ set actually is across different query positions and whether this stability differs between sink-suppressed and non-suppressed models.

### MEASUREMENT PROCEDURE

For each model we run forward passes on 8 calibration sequences at context length 2048, collecting the full attention weight matrix $W^{(\ell)} \in \mathbb{R}^{H \times T \times T}$ at each layer $\ell$. We then sample $n_q = 128$ query positions evenly spaced over the valid range $t \in [k + 32,\ T - 1]$ (requiring a KV prefix of at least $k + 32 = 288$ tokens to form a meaningful top-$k$ set).

For each ordered pair of query positions $(t_i, t_j)$, we compare the top-$k$ attention sets over the *shared* KV prefix of length $\min(t_i, t_j)$.

$$J(t_i, t_j, h, \ell) \;=\; \frac{\left|A^h_{t_i} \cap A^h_{t_j}\right|}{\left|A^h_{t_i} \cup A^h_{t_j}\right|}, \qquad A^h_t = \text{top-}k\big(W^{(\ell)}_{h,\,t,\,:c}\big), \quad c = \min(t_i, t_j),$$

where $k = 256$ (a representative eviction budget). Using the shared prefix ensures both queries rank over the same token set. For each query position $t_i$, we define its *positional stability* as

$$s(t_i, h, \ell) \;=\; \frac{1}{n_q - 1} \sum_{j \neq i} J(t_i, t_j, h, \ell).$$

The model-level aggregate is the mean of $s(t_i, h, \ell)$ over all positions, heads, and layers, averaged over the 8 calibration samples.

### RESULTS

Table 10 reports the mean positional stability and its standard deviation across layers for the 9 models evaluated.

Table 10: Mean top-256 attention Jaccard across query positions (model-level aggregate). Higher = more consistent rankings. Seven of nine models cluster in $[0.33, 0.40]$; Qwen3.5-4B falls slightly below (0.319) and GPT-OSS-20B is the sole high outlier (0.477).

| Model | Mechanism | Mean Jaccard | Std (across layers) |
|---|---|---|---|
| Qwen2.5-7B | None | 0.334 | 0.023 |
| Qwen2.5-3B | None | 0.338 | 0.030 |
| Llama 3.2-3B | None | 0.339 | 0.022 |
| Llama 3.1-8B | None | 0.347 | 0.031 |
| Qwen3-4B | QK-norm | 0.347 | 0.040 |
| Gemma2-9B | Logit-cap | 0.347 | 0.026 |
| Gemma3-4B | QK-norm | 0.398 | 0.036 |
| GPT-OSS-20B | Learned sink | 0.477 | 0.141 |
| Qwen3.5-4B | Gated attn | 0.319 | 0.030 |

### OBSERVED PATTERNS

**Universal instability.** Seven of nine models cluster tightly in $[0.33, 0.40]$ regardless of whether attention is flattened. Qwen3.5-4B falls slightly below this range at 0.319, consistent with near-zero attention sinks providing no stable high-mass anchor. The Jaccard $\approx$0.4 cited in the main body (Section 3.1) is therefore not an artifact of a particular architecture. It reflects a baseline property of attention itself. Attention rankings over the full KV prefix are moderately inconsistent across query positions in *every* architecture tested.

**Position-dependent decay.** The aggregate Jaccard mixes positions with very different pool sizes. At early query positions ($t \approx 288$–$714$), the shared KV prefix is short and selecting 256 tokens out of $\sim$300 is nearly forced, inflating Jaccard mechanically. At later positions ($t \approx 1600$–$2034$), where the KV pool is large and eviction is most consequential, the per-layer Jaccard falls to 0.25–0.27 on both non-suppressed (Llama 3.1-8B) and sink-suppressed (Qwen3-4B) models. This decay is consistent across architectures. The instability of attention rankings grows with KV pool size, and is equally pronounced in the sink-suppressed models.

Table 11: Long-context inference efficiency on H100 80 GB (budget = 2048). All eviction methods keep peak memory at $O(\text{budget})$ regardless of context length; VALUEDIFF matches the cheapest eviction baseline (Sink) within 3–4% on TTFT.

| | Dense | Sink | TOVA | SnapKV | KeyDiff | VALUEDIFF | V-Dir |
|---|---|---|---|---|---|---|---|
| *Qwen3.5-4B (gated attention), ctx = 128k* | | | | | | | |
| Peak memory (GB) | 15.0 | 8.6 | 8.6 | 8.6 | 8.6 | 8.6 | 8.6 |
| TTFT (s) | 70.6 | **59.7** | 61.9 | 61.9 | 62.6 | 61.5 | 62.2 |
| *GPT-OSS-20B (learned sink), ctx = 65k* | | | | | | | |
| Peak memory (GB) | 19.8 | 14.0 | 14.0 | 14.0 | 14.0 | 14.0 | 14.0 |
| TTFT (s) | 50.2 | **32.7** | 33.3 | 34.4 | 34.8 | 34.0 | 34.6 |

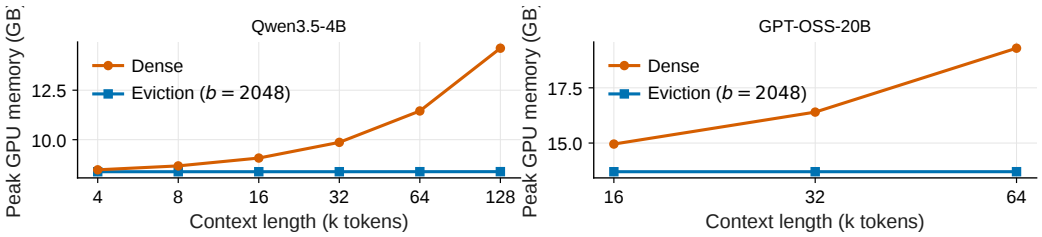



Figure 9: Peak GPU memory vs. context length on Qwen3.5-4B and GPT-OSS-20B (H100 80 GB, budget = 2048). Dense scales linearly; eviction caps at $O(\text{budget})$. All eviction methods overlap within 0.1%; VALUEDIFF is the representative eviction curve.

**GPT-OSS-20B outlier.** GPT-OSS-20B registers a mean Jaccard of 0.477, substantially above the other models. Its learned per-head sink logit absorbs the attention probability that would otherwise concentrate on content-token sink positions, redistributing the remaining probability across content tokens. This redistribution makes the attention distribution closer to uniform over content tokens, which has a dual effect. It removes any stable high-contrast anchor (consistent with near-zero sink rate; Table 6), while also making the top-$k$ set less sensitive to which query is used to evaluate it (many tokens receive similar moderate weight, so the top-$k$ boundary is relatively stable). The high Jaccard therefore does not reflect *reliable discriminative signal* for eviction; it reflects that attention cannot distinguish which tokens matter, a different form of failure.

**No architecture-level separation.** QK-norm, logit-capping, and no-suppression models all fall in the same Jaccard band. This rules out the hypothesis that sink-suppressed models have *worse* ranking stability than sink-dominated ones. Instead, the data support a different account. Attention rankings are universally unreliable at moderate to large KV pool sizes. In sink-dominated models, TOVA is partially compensated by the sink. The leading token(s) are in the top-$k$ of almost every query regardless of content, providing a stable anchor. When sinks are suppressed, this anchor disappears and attention-based eviction is left with only the unstable content-token component, with no compensating mechanism. This is consistent with the large performance gap on sink-suppressed architectures (Section 4.1). The gap reflects the *loss of compensation*, not an increase in ranking instability.

## E LATENCY AND MEMORY SANITY CHECK

The TTFT measurements characterize scoring overhead under the same block prompt processing setup and cache budget across eviction methods. Under this setup, eviction reduces peak memory relative to dense inference, and VALUEDIFF stays within a few percent of the cheapest eviction baseline while substantially improving retention on sink-suppressed models. Further serving optimizations, including decode throughput and kernel-level cache management, are complementary to the eviction scoring rule studied in this paper.

Table 12: MATH-500 dense baselines ($n=100$, first 100 problems). For all models except Qwen3.5, $k=5$ generations per problem, flex@$k$ = mean `grade_answer` accuracy, and score = pass@$k$ ($\geq 1$ correct). For Qwen3.5$^\star$, greedy $k=1$ with thinking disabled; flex@$k$ and score both equal pass@1. Tokens and empty-response rate are per-generation. *Budget* is `max_new_tokens`; Trunc% is the fraction of generations that hit it. $^\ddagger$deepseek-r1-distill-llama-8b aggregated from `dense_decoded/` after a BPE-marker post-processing fix. $^\star$Qwen3.5 uses `enable_thinking=False`, greedy (`do_sample=False`). Average token counts for Qwen3.5 set the reference lengths for eviction budgets $b=256$ and $b=512$ in Table 3.

| Category | Model | Budget | flex@$k$ | exact@$k$ | score | Avg tok | Empty% | Trunc% |
|---|---|---|---|---|---|---|---|---|
| Logit softcapping | Gemma2-2B | 16k | 0.216 | 0.176 | 0.430 | 572 | 45.2 | 0.0 |
| | Gemma2-9B | 16k | 0.526 | 0.446 | 0.680 | 430 | 16.6 | 0.0 |
| QK-normalization | Gemma3-4B | 16k | 0.742 | 0.682 | 0.820 | 932 | 11.2 | 0.0 |
| | Qwen3-4B | 16k | 0.928 | 0.792 | 0.960 | 1 947 | 1.4 | 0.0 |
| | Qwen3-4B-Thinking-2507 | 32k | 0.888 | 0.704 | 0.940 | 6 974 | 6.6 | 0.0 |
| Gated hybrid | Qwen3.5-4B$^\star$ | 32k | 0.890 | 0.780 | 0.890 | 1 124 | 0.0 | 0.0 |
| | Qwen3.5-9B$^\star$ | 32k | 0.890 | 0.790 | 0.890 | 1 228 | 0.0 | 0.0 |
| Learned attention sink | GPT-OSS-20B (reasoning=high) | 16k | 0.934 | 0.772 | 0.960 | 2 432 | 1.4 | 5.0 |
| Reasoning distill | DS-R1-Distill-Qwen-1.5B | 32k | 0.836 | 0.636 | 0.930 | 3 708 | 8.0 | 0.0 |
| | DS-R1-Distill-Llama-8B$^\ddagger$ | 32k | 0.834 | 0.624 | 0.920 | 4 147 | 4.2 | 0.0 |

# F MATH-500 Dense Decode Lengths

**Standard sampling protocol.** For all models except Qwen3.5, we sample $k$=5 generations per problem using the temperature and top-$p$ from each model's official card, with a maximum of 16k or 32k new tokens as shown in Table 12. Cache eviction budgets are set to approximately 25% and 50% of the average dense generation length, rounded to clean multiples of 256/512 tokens for Gemma3-4B and Qwen3.5, 512/1024 tokens for Qwen3-4B and GPT-OSS-20B, and 1024/2048 tokens for the R1-Distill models.

**Qwen3.5 no-think greedy protocol.** With `enable_thinking=True` ($k$=5, recommended sampling parameters), the dense pilot on 100 problems yielded average generation lengths of 21,620 tokens (4B) and 18,611 tokens (9B) with a 32k `max_new_tokens` cap, and 39% and 35% of generations were truncated at that limit. The longest other model in Table 3, R1-Distill-Llama-8B (Guo et al., 2025), averages 4,147 tokens per generation (Table 12); Qwen3.5 think-on traces are 4.5–5.2$\times$ as long. Even at a 6-hour per-sample generation timeout, many traces did not complete.

Beyond the engineering obstacle, the truncation severely degrades accuracy: a generation cut off mid-reasoning cannot produce a correctly-formatted `\boxed{}` answer, inflating the empty-response rate to 19% (4B) and 15.4% (9B). The resulting think-on dense baselines of flex@5 = 0.790 (4B) and 0.828 (9B) are suppressed by this artifact and understate the model's actual capability. We confirmed that `enable_thinking=False` outperforms `enable_thinking=True` in this pilot for both model sizes.

We therefore evaluated Qwen3.5 with thinking disabled (`enable_thinking=False`), greedy decoding (`do_sample=False`, temperature 0), and $k$=1. Despite the setting name, *no-think mode still generates step-by-step mathematical reasoning*: the model works through problems with structured intermediate steps in its response. The resulting responses are complete solutions averaging 1,124 tokens (4B) and 1,228 tokens (9B) with zero truncation. The no-think dense baseline of flex@1 = 0.890 for both models is 10.0 pp (4B) and 6.2 pp (9B) higher than the truncation-affected think-on baseline, making it the more reliable reference for measuring eviction retention.

The eviction budgets $b$=256 and $b$=512 correspond to approximately 23% and 46% of the no-think dense length for the 4B model, and 21% and 42% for the 9B model. Retention in Table 3 is computed relative to the no-think greedy dense baseline for Qwen3.5, and relative to the $k$=5 flex@5 baseline for all other models.

## G FULL MATH-500 RESULTS

Table 13 extends the 25% budget comparison in Table 3 from 100 to all 500 problems. The sampling and model-specific budget protocols are unchanged (Section F). VALUEDIFF ranks first among the tested eviction methods on all five models; its margin over the runner-up ranges from 2.8 to 21.6 points.

Table 13: Full MATH-500 accuracy retention (% of dense; $n$=500) at approximately 25% cache budget. Qwen3.5 uses no-think greedy pass@1; the other models use flex@5 (Section F). Bold marks the best eviction method per model.

| Model | TOVA | SnapKV | ManifoldKV | F-CAOTE | VALUEDIFF |
|---|---|---|---|---|---|
| Gemma3-4B | 91.1 | 86.4 | 84.1 | 90.8 | **93.9** |
| Qwen3-4B | 74.5 | 75.9 | 81.5 | 78.8 | **86.4** |
| GPT-OSS-20B | 67.0 | 68.7 | 86.5 | 71.2 | **90.9** |
| Qwen3.5-9B | 34.5 | 16.7 | 59.2 | 39.3 | **80.8** |
| Qwen3.5-4B | 24.8 | 17.6 | 56.5 | 22.5 | **77.7** |

## H SINK FORMATTING COLLAPSE ON QWEN3.5-4B

Sink eviction on Qwen3.5-4B retains only 1.1% of dense accuracy at $b$=256 and 31.5% at $b$=512, far below every other method. We analyze all 100 MATH-500 generations at each budget to identify the failure mechanism.

**Math is largely correct; output format is not.** Of the 99 failures at $b$=256, 67% contain the ground-truth answer string somewhere in the response (stripping whitespace and LaTeX delimiters); at $b$=512 this rises to 74%. The model solves the problem but fails to wrap the final answer in `\boxed{}`, which is required for flex_match scoring. Table 14 breaks down the failure categories.

Table 14: Failure categories for Sink eviction on Qwen3.5-4B, MATH-500 ($n$=100). "Format (no box)" = correct reasoning, answer present but not in `\boxed{}`. "Truncated" = generation ends mid-reasoning without conclusion. "Context lost" = model states the problem was not provided (input fidelity failure). "Wrong answer" = `\boxed{}` present but mathematically incorrect.

| Category | b256 | % | b512 | % |
|---|---|---|---|---|
| Format (no `\boxed{}`) | 59 | 60% | 40 | 56% |
| Truncated mid-reasoning | 37 | 37% | – | – |
| Context lost (input fidelity) | 25 | 25% | <5 | – |
| Wrong answer (`\boxed{}` present) | 2 | 2% | 1 | 1% |

**Three failure categories (not mutually exclusive).** **(1) Format regression** (60% of b256 failures). The model produces a fully correct solution ending with `**Final Answer:** $...$` or `Answer: $...$` but omits `\boxed{}`, which is required for flex_match scoring.

**(2) Truncated generation** (37% of b256 failures). The response ends abruptly mid-reasoning, without a concluding answer line. The model's state for tracking generation progress is degraded by aggressive context compression, producing premature termination.

**(3) Context collapse** (25% of b256 failures). The model outputs "*It appears that the math problem you intended to ask about was not included in your message. Please provide the problem statement.*" The KV cache at $b$=256 (25% retention) is too compressed to reconstruct the question, which is a pure input-fidelity failure.

**Contrast with other b256 methods.** FastCAOTE, TOVA, and SnapKV also show low accuracy retention at b256 (34.8%, 39.3%, 24.7% of dense, respectively), and their failures also exhibit format regression and truncation at similar rates (55–68% of failures contain the correct answer). However,

all three *recover strongly at b512* (83.1%, 80.9%, 84.3%), demonstrating that their b256 failures are budget-pressure effects. A 25% retention budget is simply too little capacity for reliable formatting and extended reasoning chains. Sink's failure is qualitatively different. It persists at b512 (31.5%), indicating that the problem is not resolved by increasing budget alone.