OpenAttention: One Percent Is Enough with Tail Proxies

Paper discussion **arXiv:2610.30951v1**

Hyunwoo Yu, Jongbeom Lee, Hyunbin Kim, Sungjun Park, Jaemin Choi, Jincheol Yang, Seunghun Moon, Bongjoon Hyun, Kyomin Sohn, Kyeongbo Kong, Suk-Ju Kang

22 October 2026 arxiv.org/abs/2610.30951v1

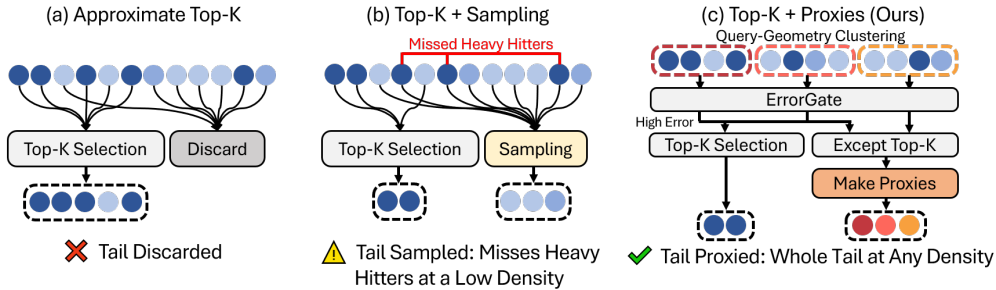
A 1% read retains 99% of dense accuracy

RULER-32K HARD, Llama-3.1-8B	Mean accuracy
Dense (100% read)	87.04
Best listed sparse baseline (1%)	48.22
OpenAttention (1%)	**86.30**

- Exact reads capture selected tokens.
- Count-aware proxies represent *every* unread token.
- A fused kernel runs the decode attention stage at 0.12 vs. 1.20 ms, batch 32.

Paper, Fig. 1, PDF p. 1; Table 11, p. 23; Table 9, p. 22. The 10× figure is attention-stage time.

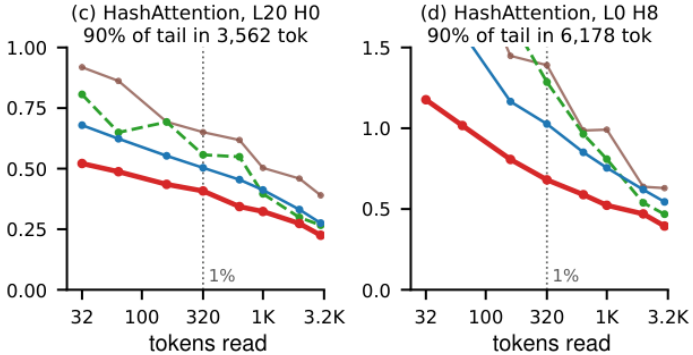
The tail remains after sparse top-k selection



- Discarding loses tail mass; sampling can miss residual heavy hitters.
- OpenAttention uses cluster proxies to carry the unread tail into softmax.

Paper, Fig. 2, PDF p. 2. Original figure.

One proxy wins in the low-read head diagnostic



- Red: top-k + one proxy; blue: top-k; green: top-k/sample; brown: sample only.
- The four panels in the paper test two heads and two selectors.

Paper, Fig. 3(c,d), PDF p. 3. Original figure cropped to the HashAttention selector. Relative attention-output error.

The proxy is a virtual token with a count

For exact set X and unread tail T, $N = |T|$:

$$z_i = d^{-1/2} q^\top k_i, \qquad \ell_T = \bar{z}_T + \log N, \qquad v_T^{\text{proxy}} = \bar{v}_T.$$

$$\hat{o} = \frac{\sum_{i \in X} e^{z_i} v_i + \underbrace{N e^{\bar{z}_T} \bar{v}_T}_{\text{one virtual token}}}{\sum_{i \in X} e^{z_i} + N e^{\bar{z}_T}}.$$

Main point

The $+\log N$ term makes one token stand for all N members. The mass $N e^{\bar{z}_T}$ usually underestimates $\sum_{i \in T} e^{z_i}$.

Paper, Eq. (1), PDF pp. 3–4; Appendix A, pp. 12–13.

Sampling nearly doubles a specific weight error

Define the *unnormalized* discrepancy

$$\mathcal{E} = \frac{1}{Z}\sum_i |\hat{w}_i - w_i|, \quad \tau = \frac{\sum_{i\in T} w_i}{Z}, \quad w_i = e^{z_i}.$$

For a *fixed* tail of N tokens and a uniform sample of size m:

$$\mathcal{E}_{\text{discard}} = \tau, \qquad \mathbb{E}[\mathcal{E}_{\text{sample}}] = 2\tau\left(1 - \frac{m}{N}\right).$$

Main point

As $m/N \to 0$, the sampling discrepancy approaches 2τ. This is not a twofold claim for normalized attention or output error.

Paper, Eq. (2), PDF p. 4; derivation in Appendix A.2, PDF pp. 17–18.

Output error also depends on values and mass

$$\frac{\sqrt{\mathbb{E}[(\hat{Z}_s - Z)^2]}}{Z} = \tau \frac{\mathrm{CV}_T}{\sqrt{m}} \sqrt{\frac{N-m}{N-1}}.$$

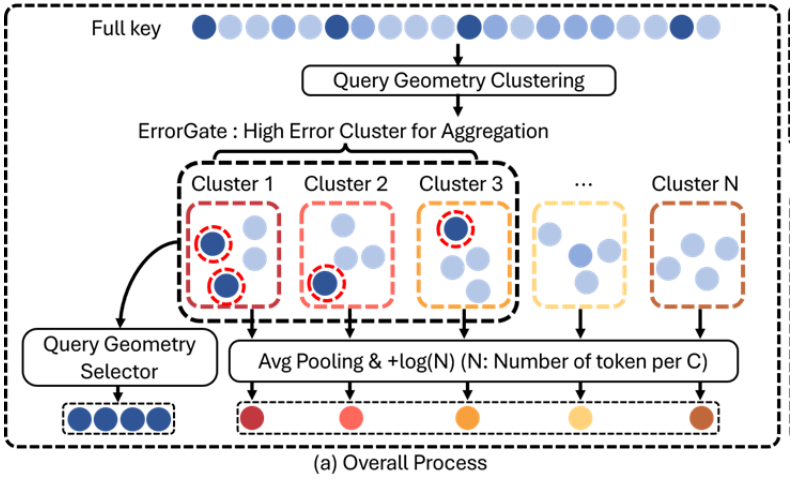
- Smaller m raises normalizer uncertainty.
- Concentrated tail weights raise CV_T.

Diagnostic at a 1% read

Behind HashAttention, the sample includes the heaviest missed token with only about **0.5%** probability.

Paper, Appendix A.2, PDF p. 17; Fig. 10, p. 20. The diagnostic uses 137,216 decode rows.

Four stages make a proxy-based read



Paper, Fig. 5(a), PDF p. 5. Original figure cropped to the pipeline.

Exact reads and residual proxies share softmax

After selected tokens are removed from each cluster:

$$\hat{o} = \frac{\sum_{i \in X} e^{z_i} v_i + \sum_{b \in \mathcal{P}} n'_b e^{\bar{z}'_b} \bar{v}'_b}{\sum_{i \in X} e^{z_i} + \sum_{b \in \mathcal{P}} n'_b e^{\bar{z}'_b}}.$$

- X: sink, local window, QGS top-k, and ErrorGate expansions.
- Primes denote statistics of *unread* cluster members.
- Subtracting exact tokens from pooled moments prevents double counting.

Paper, Appendix A, Eq. (8), PDF pp. 12–13; Appendix A.1, p. 16.

QGC clusters by expected logit distance

With query second moment $M = \mathbb{E}[qq^\top]$:

$$\mathbb{E}_q[(z_i - z_j)^2] = d^{-1}(k_i - k_j)^\top M(k_i - k_j).$$

Its leading eigenpairs (U_r, Λ_r) define

$$\Phi = KU_r\Lambda_r^{1/2} \in \mathbb{R}^{n\times r}, \qquad r = 16.$$

- Run k-means in this query-weighted key space, once after prefill.
- M mixes presampled and current question-query moments 50/50.

Paper, Sec. 3.2.1, Eq. (4), PDF p. 5.

QGC helps at the same 1% read

Clustering	Dimension	Mean accuracy
Token position	–	84.37
Key-only k-means	128	85.42
QGC	**16**	**86.30**

- Same $C = 64$ and total 1% read.
- Decode queries retain an absolute cosine of 0.82 (median) with M's first eigenvector.
- The clustering target is logit dispersion, not direct output error.

Paper, Table 1 and Fig. 6, PDF p. 5. RULER-32K HARD.

QGS finds high-logit tokens from int8 codes

$$c_i = Q_8(U_{32}^\top k_i), \qquad \tilde{z}_i = (U_{32}^\top q)^\top c_i.$$

Selector	Top-32 recall at 1/16	At 1/32	ms/layer
HashAttention	0.74	0.63	0.24
Double Sparsity	0.97	0.95	0.50
QGS	**0.97**	**0.94**	**0.22**

Exact attention recomputes the selected tokens' original logits.

Paper, Eq. (5) and Table 2, PDF p. 6. Same decode rows; selection time is separate from the attention kernel.

Cluster discrepancy is mass times spread

For residual cluster b, $\delta_i = z_i - \bar{z}_b$:

$$\varepsilon_b = n_b e^{\bar{z}_b} \underbrace{\frac{1}{n_b} \sum_{i \in b} |e^{\delta_i} - 1|}_{D_b}, \qquad Z\mathcal{E} = \sum_{b \in \mathcal{P}} \varepsilon_b.$$

- Equal logits give $D_b = 0$: the proxy is exact in weight space.
- Small mass gives small contribution at fixed spread.
- Reading a whole cluster exactly removes its nonnegative term.

Paper, Eq. (3), PDF p. 4; Appendix A, PDF p. 13. This is an exact finite-cache identity.

ErrorGate estimates that discrepancy cheaply

Gaussian logit-deviation model + diagonal key variance:

$$\hat{\sigma}_b^2 = d^{-1} \sum_j q_j^2 \mathrm{Var}_b(k_j),$$

$$\hat{\varepsilon}_b = n_b e^{\bar{z}_b} e^{\hat{\sigma}_b^2/2} \big(2\Phi_{\mathrm{N}}(\hat{\sigma}_b) - 1\big).$$

- Estimated error ranking: 86.30 mean accuracy at 1%.
- Centroid-mass ranking: 85.99; variance-only: 84.81.
- Diagonal vs. full model scores: Spearman 0.92 in the diagnostic.

Paper, Eq. (6), Tables 3 and Fig. 7, PDF p. 6. The score is neither an exact discrepancy nor a certified output bound.

The 1% read is split across three streams

$$q_b = \min\left(n_b, \lambda \frac{\hat{\varepsilon}_b}{\max_{b'} \hat{\varepsilon}_{b'}}\right), \qquad \sum_b q_b = E.$$

Deployed 1% configuration

$C = 64$ clusters; roughly 9 initial QGS reads; $E = 216$ added exact reads. Mean and variance statistics cost $1.5C = 96$ KV rows, about 29% of the read.

- Water filling spends the token budget exactly.
- Proportional partial-token allocation is a heuristic.

Paper, Eq. (7), PDF pp. 6–7; Appendix A.1, p. 17; hyperparameters, Appendix D, pp. 22–23.

One online softmax handles exact and proxy rows

1. Gather QGS and expanded-cluster exact KV rows into SRAM.
2. Subtract those rows from cached cluster moments.
3. Stream residual cluster means as virtual tokens with logits $\ell_b = d^{-1/2} q^\top \bar{k}'_b + \log n'_b$.
4. Update one running max, normalizer, and value accumulator.

State and representation

At $C = 64$, cached cluster statistics occupy about 12 MiB for the 32-layer model, versus a 4.0 GiB KV cache.

Paper, Sec. 3.4, PDF p. 7; Algorithm 1 and implementation, Appendix C, PDF pp. 20–21.

The main benchmark charges proxy reads

- Llama-3.1-8B-Instruct, bf16; seven RULER-32K HARD tasks, 200 samples each.
- Context prefill is dense. Sparse attention covers the question chunk and generation.
- Read budget: KV rows per query head, including 1.5 rows per proxy cluster.
- HashAttention/vAttention share signatures; MagicPIG uses LSH; OpenAttention uses QGS.
- Compressed-index scan is excluded from the read budget for all methods.

Paper, Sec. 4 Setup, PDF p. 7; baseline protocol, Appendix F, PDF pp. 23–25.

OpenAttention stays near dense down to 1%

Total read	OpenAttention	HashAttention	vAttention	MagicPIG
10%	**87.08**	81.23	83.08	52.52
5%	**86.82**	71.55	68.86	34.42
3%	**86.83**	63.48	55.74	24.01
1%	**86.30**	48.22	36.29	10.53

Reading the table

Dense mean = 87.04. At 1%, OpenAttention is 0.74 points below dense and 38.08 points above the best listed sparse baseline.

Paper, Fig. 8, PDF p. 7; Table 11, p. 23. MagicPIG realized reads: 10.0, 5.1, 2.9, 1.1%.

Proxies provide the largest ablation gain

5% read configuration	Mean accuracy	Gain vs. first row
Hash selector + residual sampling	67.35	–
Replace sampling with proxies	77.91	+10.56
Proxies + ErrorGate	85.90	+18.55
Proxies + QGS	86.73	+19.38
Proxies + ErrorGate + QGS	**86.82**	**+19.47**

Main point

A tail proxy already matters; clustering, allocation and selection recover most of the remaining loss.

Paper, Table 4, PDF p. 8; per-task Table 12, p. 24. The ablation's first row is a separate configuration from the main vAttention row.

The 3% setting transfers across models

Method	Llama-8B	DeepSeek-8B	Mistral-7B
Dense	87.04	73.46	65.46
HashAttention	63.48	59.85	51.24
vAttention	57.15	49.26	46.14
OpenAttention	**86.59**	**73.49**	**64.83**

- Same seven hard RULER tasks at a 3% read.
- OpenAttention remains within 0.7 accuracy points of dense on these models.

Paper, Table 5, PDF p. 8.

Accuracy tracks dense from 16K to 96K

Context	16K	32K	64K	96K
Dense	87.14	87.75	85.49	78.01
OpenAttention (3%)	**87.49**	**87.38**	**84.82**	**78.10**
HashAttention (3%)	67.41	63.87	49.65	27.99

- 100 samples per task; 64K and 96K are RULER-generated extensions.
- Cluster count scales with length; this sweep uses $E = 0$.

Paper, Table 7, PDF p. 9; Appendix I, p. 27.

The 10× result is an attention-stage timing

B=32, 32K, H100 NVL	1% read	3% read
Dense fused SDPA	1.20 ms	1.22 ms
OpenAttention.flash	**0.12 ms**	**0.24 ms**
Stage speedup	10.0×	5.1×

- Kernel table uses $E = 0$, unlike the 1% accuracy deployment.
- Excludes index scan, selection, ErrorGate scoring, and residual updates.
- At $B = 1$, 1% times are 0.06 vs. 0.05 ms.

Paper, Table 9, PDF p. 22; discussion of excluded work, PDF p. 21.

What the evidence supports

1. At small reads, preserving the unread tail by a few count-aware proxies can retain much more task accuracy than discarding or sampling it.
2. Query geometry helps both clustering and top-k selection; estimated proxy error guides the exact-read budget.
3. The weight-error formula is exact for the defined metric, while the Gaussian ErrorGate score and partial-token allocation are approximations.
4. Full decode latency still requires accounting for index scanning, selection, cluster construction and updates.

Paper, Secs. 2–4 and 6, PDF pp. 3–9; Appendix C, PDF pp. 20–22.

Paper and source

Primary source

Hyunwoo Yu et al., *OpenAttention: One Percent Is Enough with Tail Proxies*, arXiv:2610.30951v1 (2026).

- Paper: arxiv.org/abs/2610.30951v1
- Project: hyunwoo137.github.io/OpenAttention
- Code: github.com/hyunwoo137/OpenAttention

All numerical results and original figures in this deck come from the cited arXiv v1 PDF.

Backup: output error has a value-dependent identity

For the residual-proxy estimator:

$$\hat{Z}(\hat{o} - o) = -\sum_{b \in \mathcal{P}} \sum_{i \in b} (e^{z_i} - e^{\bar{z}_b})(v_i - o).$$

Hence

$$\|\hat{o} - o\|_2 \leq \sum_b \hat{a}_b^{\mathrm{virt}} V_b D_b, \quad V_b = \max_{i \in b} \|v_i - o\|_2.$$

- Exact finite-cache identity and deterministic bound.
- Explains why the value-free $\mathcal{E}$ need not order output errors.

Paper, Appendix A, Lemma 1 and Proposition 1, PDF pp. 13–15.

Backup: comparisons use specific baselines

- HashAttention and vAttention share the public HashAttention signatures.
- vAttention is tuned per target read under the paper's matched-read accounting.
- MagicPIG hash tables are calibrated to realized reads near 1, 3, 5 and 10%.
- Main protocol sparsifies question plus generation; MagicPIG's original protocol keeps the question dense.
- The 1% timing uses $E = 0$; the 1% accuracy configuration uses $E = 216$.

Paper, Sec. 4 Setup, PDF p. 7; Appendix F, pp. 23–25; Table 9, p. 22.

Backup: other tasks and KV compression

Question-free tasks, 3% read

	Dense	Ours
gov_report ROUGE-L	36.08	35.53
multi_news ROUGE-L	26.61	26.80
LongBench-v2 accuracy	30.00	33.30

With 2048-entry KV compression

	Alone	+ Ours
StreamingLLM	28.00	28.28
SnapKV	69.98	69.49
PyramidKV	70.34	70.14

Paper, Tables 6 and 15–16, PDF pp. 9, 25–26. LongBench-v2 uses only 60 requests.