DoSA: Dots Sparse Attention

Exact token selection with conditional full scoring

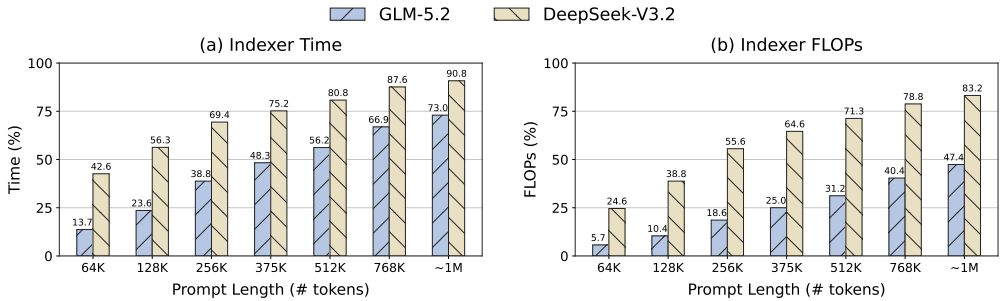
Paper authors: Fanxu Meng and Taosong Fang
Dots Studio, Xiaohongshu Inc.

Paper discussion | arXiv:2610.13990v1 | 19 October 2026

Prove that a key cannot win before computing its full score.

Independent reading presentation. arxiv.org/abs/2610.13990v1.

Choosing sparse attention can dominate prefill



At approximately 1M tokens

The Indexer accounts for **73.0%** of GLM-5.2 prefill GPU time and **90.8%** of DeepSeek-V3.2 prefill GPU time.

Paper, Fig. 1, PDF p. 2; profiling scope: App. D.1, p. 18. Time shares are GPU prefill shares, not TTFT shares.

The original scorer is a signed sum of ReLU heads

$$s_i = \sum_{h=1}^{H} w_h \left[\mathbf{q}_h^\top \mathbf{k}_i \right]_+, \qquad i \in \Omega.$$

- $\mathbf{q}_h, \mathbf{k}_i \in \mathbb{R}^D$; w_h may be positive or negative.
- Ω contains visible keys; return K distinct token IDs.
- SGLang scores all heads; LiteTopK reduces traffic through fusion.
- DoSA also avoids full scoring for certified losers.

The design problem

Build a useful upper bound more cheaply than evaluating all H heads.

Paper, Secs. 1–2.1, PDF pp. 1–2; Eq. (1). The model's original scoring rule is preserved.

A seed set gives a safe lower bound on the cutoff

Score distinct visible seeds $\mathcal{S}$, with $|\mathcal{S}| \geq K$:

$$\tau = \tau_K(\mathcal{S}) \leq \tau_K(\Omega) = \tau^*.$$

Certificate of rejection

$$s_i \leq U_i < \tau \quad \Longrightarrow \quad i \notin \mathrm{TopK}(\Omega).$$

Illustrative example: $K = 2$

Seed scores: $10, 7, 3$; hence $\tau = 7$.

$U_i = 6.5$: reject without full scoring.

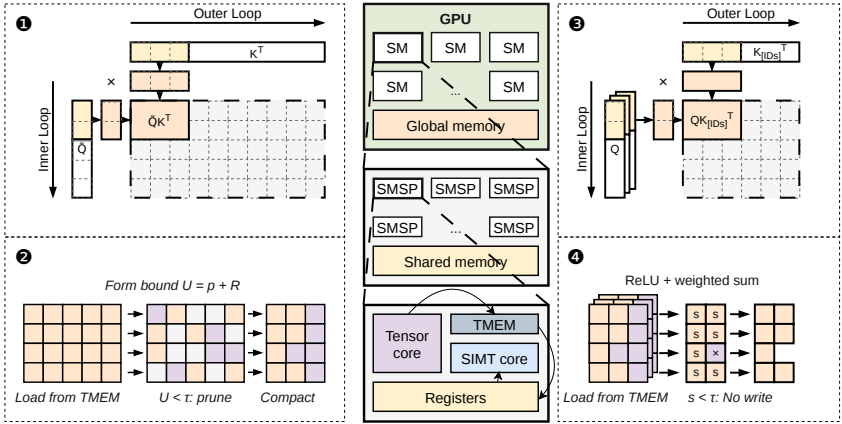
Equality must survive

$U_i = 7$: retain and compute s_i.

Weak seeds reduce efficiency, not safety.

Paper, Eq. (2), p. 2; App. A.2, p. 14. Example is instructional. The prefill scan uses a fixed threshold.

Screen, compact, fully score, and selectively write



Bounds decide eligibility; original scores decide the final ranking.

Paper, Fig. 2, PDF p. 3; Sec. 2.5, p. 4. Seed initialization and final Top-K are omitted in the original figure.

Spend direct evaluation on hard-to-bound heads

Retain r heads in R; omitted positives are P, omitted negatives are N.

If all positive heads fit

Retain them all. Fill the remaining slots with negative heads ranked by

$$|w_h| \, \|\mathbf{q}_h\|_2 \min(p_h, 1 - p_h).$$

- p_h: activation frequency on sampled seed keys.
- Strong heads with uncertain activation are valuable to evaluate.
- If positives exceed the budget, retain the r strongest by $w_h \, \|\mathbf{q}_h\|_2$.

Paper, Sec. 2.2, PDF p. 3; surrogate derivation: App. B.2, p. 16. Any retained set is safe; selection affects tightness.

Positive heads reuse a retained direction plus a residual

For an omitted positive head h, choose a retained anchor j:

$$\beta_h = \frac{\left[\mathbf{q}_h^\top \mathbf{q}_j\right]_+}{\|\mathbf{q}_j\|_2^2} \geq 0, \qquad \mathbf{d}_h = \mathbf{q}_h - \beta_h \mathbf{q}_j.$$

A reusable upper bound

$$\left[\mathbf{q}_h^\top \mathbf{k}_i\right]_+ \leq \beta_h \left[\mathbf{q}_j^\top \mathbf{k}_i\right]_+ + \|\mathbf{d}_h\|_2 \|\mathbf{k}_i\|_2 .$$

- Nonnegative projection respects ReLU positive homogeneity.
- Fold $w_h \beta_h$ into the anchor weight $\widetilde{w}_j$.
- Sum residual allowances into $E_P = \sum_{h \in P} w_h \|\mathbf{d}_h\|_2$.

Paper, Eqs. (4)–(6), PDF pp. 3–4. Choose the anchor minimizing the residual; zero anchors use coefficient zero.

Negative aggregation must avoid inactive-head cancellation

For any support $A \subseteq N$:

$$\mathbf{a}_A = \sum_{h \in A} |w_h| \mathbf{q}_h, \qquad \left[\mathbf{a}_A^\top \mathbf{k}_i\right]_+ \leq \sum_{h \in N} |w_h| \left[\mathbf{q}_h^\top \mathbf{k}_i\right]_+ .$$

Example: $w_1 = w_2 = -1$

True subtraction: $[5]_+ + [-4]_+ = 5$.
Aggregate both: $[5 - 4]_+ = 1$.
Aggregate the active head only: 5.

Activation-informed support

$A = \{h \in N : p_h \geq 1/2\}$.

Seed statistics improve tightness. Every support remains safe.

Paper, Sec. 2.4 and Eq. (8), PDF p. 4; App. B.2, pp. 15–16. Numerical example is instructional.

The certificate preserves every possible Top-K winner

$$U_i = \sum_{j \in R} \widetilde{w}_j \left[\mathbf{q}_j^\top \mathbf{k}_i\right]_+ + E_P \left\|\mathbf{k}_i\right\|_2 - \left[\mathbf{a}_A^\top \mathbf{k}_i\right]_+ \geq s_i$$

1. Screen non-seeds: $V = \{i : U_i \geq \tau\}$.
2. Fully score survivors; keep $C = \{i \in V : s_i \geq \tau\}$.
3. Select TopK$(\mathcal{S} \cup C)$ using original scores.

Why it works

Every discarded key has $s_i < \tau \leq \tau^*$. No winner is discarded, provided live candidates are never lost.

Paper, Eq. (9), p. 4; Algorithm 1 and Theorem 1, p. 14. Identical tied IDs require a common total tie order.

FP8 exactness requires an upper bound on computed scores

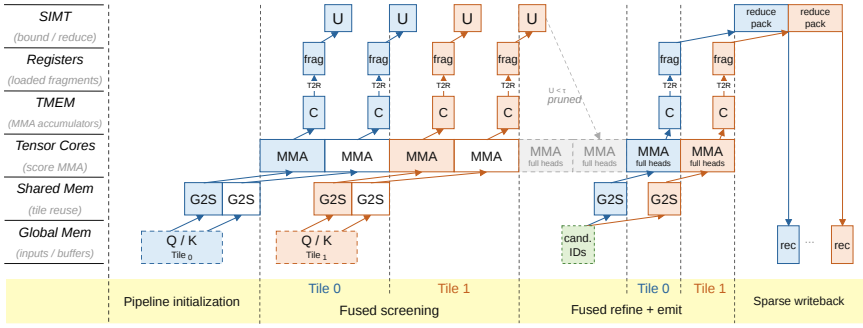
The implementation must establish $s_i^{\mathrm{ref}} \le U_i^{\mathrm{num}}$:

$$U_i^{\mathrm{num}} = \mathrm{Eval}_{\mathrm{up}}\left[\alpha_i\left(\widehat{U}_i + \rho_i C + \varepsilon_{\mathrm{raw}}\right) + \varepsilon_{\mathrm{out}}\right].$$

- s^{ref}: original scorer's computed score; use it for seeds and survivors.
- $\mathbf{k}_i = \alpha_i \bar{\mathbf{k}}_i$, $\alpha_i > 0$; $\rho_i \sqrt{D} \ge \left\|\bar{\mathbf{k}}_i\right\|_2$ must hold.
- $\widehat{U}_i$: screening result; $C, \varepsilon_{\mathrm{raw}}, \varepsilon_{\mathrm{out}}$ cover representation and arithmetic errors.
- Upward enclosure is essential; round-to-nearest alone is insufficient.
- Invalid certificates or lost storage capacity require safe recovery.

Paper, App. C, PDF pp. 17–18, especially Eq. (26). A real-arithmetic proof alone does not establish the FP8 guarantee.

GPU execution couples pruning with less data movement



- Group queries for regular Tensor Core screening.
- Four-query refinement groups share repeated key loads.
- Emit only threshold-qualified scores; resolve boundary codes exactly.

Paper, Sec. 3 and Fig. 3, PDF p. 5. Shared stream: at least two surviving queries and common visibility.

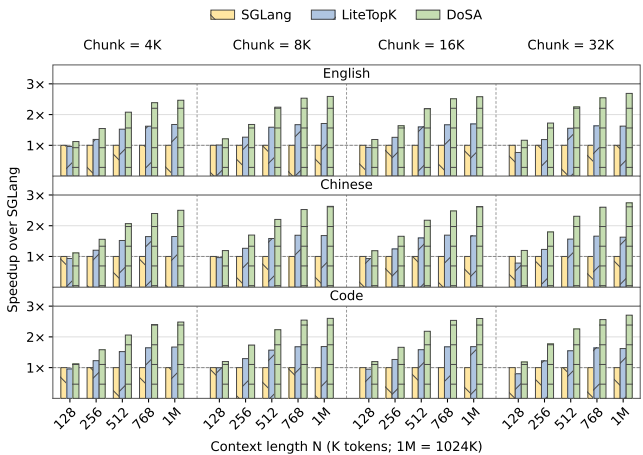
The evaluation separates Indexer cost from serving latency

	GLM-5.2	DeepSeek-V3.2
Indexer heads	32	64
Top-K	2,048	2,048
End-to-end parallelism	TP8	TP8
End-to-end prefill chunk	32K	16K

- English/Chinese books and code; 128K to approximately 1M tokens.
- Baselines: SGLang v0.5.18 and LiteTopK.
- Complete-Indexer: GLM, one GPU, all 21 Indexer-producing layers, including query/key preparation.
- TTFT: individual requests; prefix cache cleared between requests.

Paper, Sec. 4.1, PDF pp. 6–7; App. D.1, p. 18. GPU vendor is stated, but the specific GPU model is not reported.

DoSA wins all 60 complete-Indexer configurations



Overall speedup

SGLang: 1.12–2.74×
LiteTopK: 1.17–1.69×

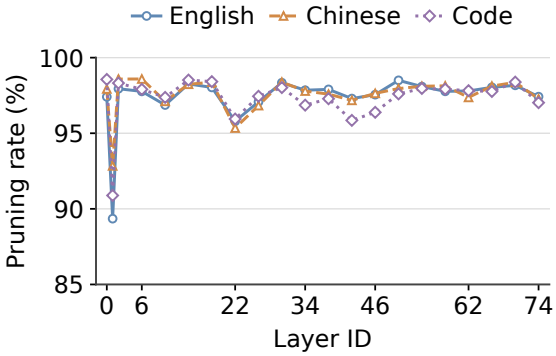
At approximately 1M

SGLang: 2.47–2.74×
LiteTopK: 1.47–1.69×

Gains grow mainly with context length.

Paper, Fig. 4, p. 6; Sec. 4.2, p. 7. 3 domains × 5 lengths × 4 chunk sizes; plot normalized to SGLang.

Most pairs are rejected, but screening still costs work



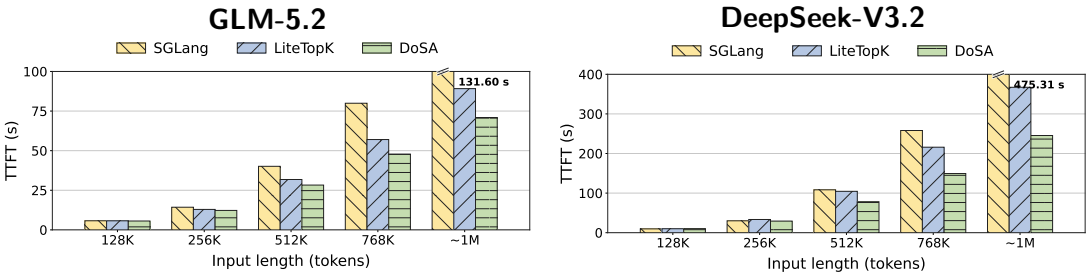
- Mean pruning: 97.2–97.5% by domain.
- 20 of 21 layers exceed 95% in each domain.
- Layer 1: approximately 89–93%.
- The rate counts query–key pairs avoiding full scoring.

Interpretation

High rejection does not eliminate the retained-head scan or key reads.

Paper, Fig. 6, p. 7; Sec. 4.2, pp. 7–8. Cost model: App. B.3, p. 16.

At approximately 1M tokens, TTFT is nearly halved



Model (approximately 1M tokens)	vs. SGLang	vs. LiteTopK
GLM-5.2	1.86×	1.26×
DeepSeek-V3.2	1.94×	1.50×

Paper, Sec. 4.3 and Fig. 8, PDF p. 8. TP8; chunks 32K / 16K. Panel y-scales differ; lower TTFT is better.

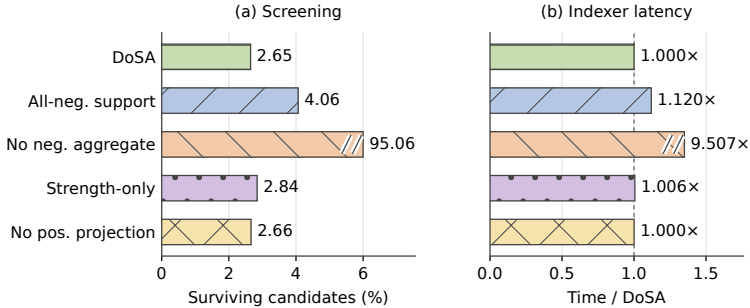
Screening and sparse writeback remove different costs

Variant	Screening	Sparse writes	Time (ms)	Reduction
Base	No	No	76.2	—
Full	No	Yes	58.4	23.4%
Dense	Yes	No	47.7	37.4%
DoSA	Yes	Yes	40.2	47.3%

- LOFT SciFact example: 16K queries, 896K keys; mean over 21 layers.
- Full still computes every head; Dense still pays for dense writeback.
- Together: 1.90× retrieval speedup over Base.

Paper, Fig. 7, p. 7; Sec. 4.4, p. 8. Times reproduced from figure labels; reductions from text.

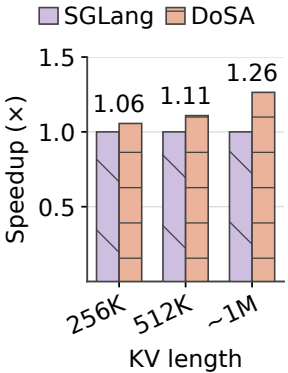
Negative aggregation is decisive in the measured workload



- Removing it raises survival from 2.65% to 95.06%.
- The 9.507× latency includes candidate-pool recovery.
- Positive projection is needed in only 0.42% of query-layer rows.

Paper, App. D.2 and Fig. 9, PDF pp. 18–19. Mean survival; geometric mean time ratio. Breaks mark truncated bars.

Decoding uses full scoring and a smaller selection problem



1. Re-score recent, distinct winner IDs for the current query.
2. Use them as witnesses for a conservative cutoff.
3. Fully score keys; compact qualifying records for exact selection.

Equal-length requests fill approximately 1M total cached tokens.

Paper, Sec. 3, p. 5; Fig. 5 and Sec. 4.2, p. 7. Reported gains are decoding Indexer gains, not end-to-end token latency.

Margins and head budgets determine whether screening pays

Define $\Delta_i = \tau^* - s_i$, $\delta_i = U_i - s_i$, and $b = \tau^* - \tau$:

$$\boxed{U_i < \tau \quad \Longleftrightarrow \quad \Delta_i > \delta_i + b.}$$

- Better seeds reduce b; better geometry reduces δ_i.
- Many near-boundary keys can survive even a valid certificate.

$$\frac{\text{DoSA dot products}}{\text{baseline dot products}} \approx f + (1-f)\left(\frac{r + d_a}{H} + m\right).$$

f: initialization fraction; m: survivor fraction; d_a: auxiliary rows.

Paper, App. B.3, Eqs. (17)–(20), PDF p. 16. Logical-work model excludes planning, padding, selection, and recovery.

Layer-level gains come with a material storage qualification

Retrieval-only comparison	Faster cells	Peak speedup
vs. SGLang	251/252	3.74×
vs. LiteTopK	250/252	2.08×

- Prepared queries/keys; 32K queries; 12 replays per method and cell.
- Five cells remeasured after a candidate-capacity repair; 247 retained.
- Repair adds **9.94 GiB** of workspace at approximately 1M-key capacity.
- The full matrix was not rerun; three slower comparisons remain visible.

Paper, App. D.4, PDF p. 20; Figs. 11–12, p. 21. This scope differs from complete-Indexer benchmarks.

The contribution is exact conditional scoring inside an Indexer

What the evidence supports

A signed-head certificate can skip full scoring, and an aligned GPU pipeline turns that into substantial long-context prefill gains.

- Every key is still scanned; short contexts or loose bounds can erase gains.
- Numerical dominance, storage recovery, and tie handling are essential.
- Beyond-context measurements establish computational scaling, not task accuracy.
- Next steps: slack-aware training and certificates for whole key blocks.

Paper, Sec. 6, PDF p. 9; App. C and D.1, pp. 17–18. Future directions are proposals, not measured results.

Paper and discussion prompts

Fanxu Meng and Taosong Fang.
DoSA: Dots Sparse Attention. arXiv:2610.13990v1.

https://arxiv.org/abs/2610.13990v1
Author-linked repository: github.com/llm-sys/flash-indexer

- Can training improve both selection quality and certificate tightness?
- Can a block bound avoid key reads without becoming too loose?
- How should head budgets trade off arithmetic, memory, and fallback cost?

Figures are unmodified author assets. Instructional examples and discussion prompts are this presentation's analysis.

Backup: joint residuals exploit a shared key direction

Let $\mathbf{u}_h = w_h \mathbf{d}_h$ for omitted positive heads:

$$E_P^{\text{joint}} = \frac{1}{2}\left(\left\|\sum_h \mathbf{u}_h\right\|_2 + \sqrt{\sum_{h,\ell} |\mathbf{u}_h^\top \mathbf{u}_\ell|}\right) \le \sum_h \|\mathbf{u}_h\|_2 = E_P.$$

- Use $[x]_+ = (x + |x|)/2$, then bound signed sums via the Gram matrix.
- Opposite residuals $\mathbf{u}, -\mathbf{u}$: allowance falls from $2\|\mathbf{u}\|_2$ to $\|\mathbf{u}\|_2$.
- Coordinate groups combine as $\sqrt{\sum_g B_g^2}\,\|\mathbf{k}\|_2$.
- No extra key summary; query-side cost is quadratic in $|P|$.

Paper, Eq. (7), p. 4; App. A.1, pp. 13–14. Opposite-direction example is instructional.

Backup: negative slack explains majority support

With $x_h = \mathbf{q}_h^\top \mathbf{k}$, $c_h = |w_h|$, $Z_A^\pm = \sum_{h \in A} c_h [\pm x_h]_+$:

$$U - s = \underbrace{\delta_P}_{\text{positive allowance}} + \underbrace{\sum_{h \in N \setminus A} c_h [x_h]_+}_{\text{missed active negatives}} + \underbrace{\min(Z_A^+, Z_A^-)}_{\text{cancellation}}.$$

A bound on mean normalized slack (nonzero sample keys):

$$J(A) = \sum_{h \in A} a_h (1 - p_h) + \sum_{h \in N \setminus A} a_h p_h, \quad a_h = |w_h| \, \|\mathbf{q}_h\|_2 .$$

- Independent minimization gives $p_h \geq 1/2$.
- Retaining a negative head removes its $a_h \min(p_h, 1 - p_h)$ term.
- This minimizes a surrogate, not runtime or survivor count directly.

Paper, App. B.2, Eqs. (14)–(16), PDF pp. 15–16. Displayed slack decomposition uses the triangle allowance.

Backup: the 32-head kernel also uses an absolute-value form

If all positive heads are retained, define

$$\mathbf{V} = \mathbf{a}_A - \tfrac{1}{2} \sum_{j \in R} w_j \mathbf{q}_j.$$

Then

$$\begin{aligned} U_{\mathrm{abs}}(\mathbf{k}) &= \tfrac{1}{2} \sum_{j \in R} w_j |\mathbf{q}_j^\top \mathbf{k}| - \mathbf{V}^\top \mathbf{k} \\ &= S_R(\mathbf{k}) - \mathbf{a}_A^\top \mathbf{k} \geq s(\mathbf{k}). \end{aligned}$$

- Additional slack versus the base certificate: $[-\mathbf{a}_A^\top \mathbf{k}]_+$.
- 32-head paths: $r = 14, 16$ use this form; $r = 20$ uses the base form.
- Two FP8 auxiliary rows encode $\mathbf{V}$; both forms cost $r + 2$ dots.

Paper, App. C.1, Eqs. (24)–(25), PDF p. 17. These budgets are specified for the 32-head implementation.

Backup: the RMS correction must enclose all relevant errors

A sufficient error contract is

$$s_i^{\mathrm{ref}} \le \alpha_i \left(\widehat{U}_i + \sum_g E_g^+ \left\| \bar{\mathbf{k}}_i^{(g)} \right\|_2 + \sum_d J_d |\bar{k}_{i,d}| + \varepsilon_{\mathrm{raw}} \right) + \varepsilon_{\mathrm{out}}.$$

It can be covered with a conservative coefficient

$$C \ge \sqrt{D} \left(\sum_g [E_g^+ + \left\| \mathbf{J}^{(g)} \right\|_2]^2 \right)^{1/2}.$$

- E_g^+ encloses positive residuals; $\mathbf{J}$ covers coordinatewise errors.
- Include screening and reference-scorer arithmetic, not just quantization.
- Conservative RMS, coefficients, scaling, and final evaluation are all needed.

Paper, App. C.2, Eqs. (27)–(28), PDF pp. 17–18. Absolute-value paths use directed FP64 intervals in planning.

Backup: candidate capacity and ties are correctness conditions

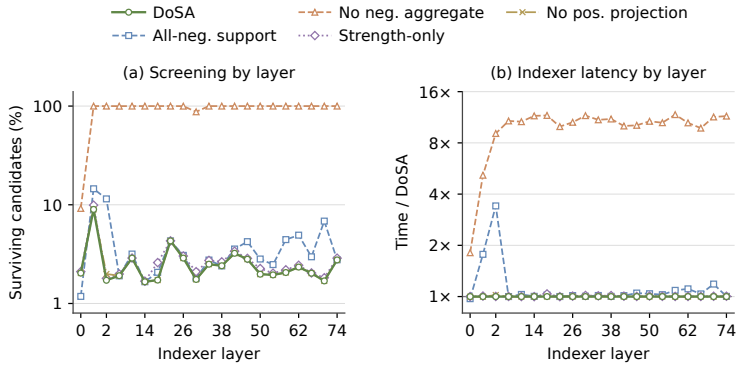
Top-K output size does not bound the live candidate set

If every score equals τ, every candidate survives, even with exact bounds.

- Full buffers: extend, process with exact Top-K merging, or recover globally.
- Never truncate live IDs or recover only from an incomplete buffer.
- Query-level numerical errors require a full visible-domain recovery, including previously discarded keys.
- Keep equality at both filters. Match total tie order for identical tied IDs.
- Different valid tied sets may produce different generated text.

Paper, Theorem 1, p. 14; App. C.3, p. 18; Sec. 6, p. 9. Finite-score theorem does not define NaN semantics.

Backup: all-negative support is especially costly in early layers



All-negative support reaches 3.41× relative Indexer latency at layer 2.

Paper, App. D.3, pp. 19–20; Fig. 10, p. 20. Both vertical axes are logarithmic; same samples as component ablation.