RIS-Kernel

Sparse attention for CPU long-context inference
Method, results, and evidence boundaries

Paper author: Anderson R. Santos (UFU)

Independent paper-reading presentation

arXiv:2607.21927v1 | Submitted July 24, 2026

Reading date: October 1, 2026 | Approximately 25 minutes + backup

RIS-Kernel: A Model-Agnostic Architecture for Long-Context LLM Inference via Sparse Attention

The paper offers a CPU inference proof of concept

Core idea

Combine local connections and random sparse views, merge selected indices, and normalize them together without changing model weights.

- **32k:** best reported Stochastic score is 24/32; dense is 23/32.
- **64k:** best Structural score is 42/64; no context is 33/64.
- **Scope:** small models, four scientific documents, CPU hardware.

Reading note: Data inconsistencies and incomplete cost accounting limit the stronger claims.

Paper, Secs. 2.2, 4.4.1; Tables 2 and 5; PDF pp. 3–5, 9, 16–17, 20.

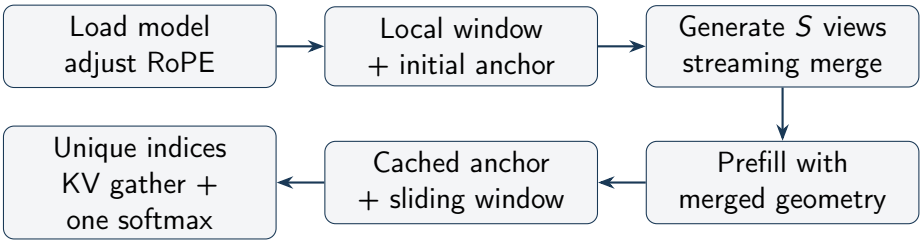
Three failure modes motivate the design

Missing facts	Merging attention	Position extrapolation
A sparse view can miss distant entities and their supporting context.	Separately normalized groups need a principled way to compete.	Selecting a useful token does not ensure usable positional features.
Multiple views + local structure	Unified softmax over an index union	Dynamic RoPE scaling

The target is inference on existing models; no weight update is reported.

Paper, Introduction and Methods; PDF pp. 1–2, 14–17.

The ensemble merges connections before attention



- T: sequence length; S: seed count. The paper overloads N.
- Methods describe an **index union**, not S answer-level votes.

Teaching schematic based on Secs. 4.1–4.4.2 and Figs. 8–9, PDF pp. 14–17.

Structural sampling spends budget on local blocks

Stochastic

Uniform global neighbors per pivot; independent views improve the chance of selecting distant content.

Structural

Build local cliques, then add random global edges from the remaining budget.

Trade-off: low-budget sampling can miss relevant neighborhoods.

$$B = \min(0.1T, B_{\max})$$

$$d_{\text{eff}} = d_{\text{clique}} + d_{\text{global}}$$

Both include base geometry: a 1,024-token local window and an initial anchor.

Reading note: A local clique preserves within-block links; it does not guarantee access to every answer.

Paper, Secs. 4.2, 4.4.2; PDF pp. 15, 17.

Streaming removes mask copies, not quadratic storage

One seed at a time

Sample indices → apply causality → merge into master structures → release the temporary seed object.

What it saves

- Avoids holding all seed masks at once.
- Reduces temporary memory as a function of S.

What remains

- The master boolean mask is $O(T^2)$.
- At 65,536 tokens: 4 GiB per full mask if one byte per entry.

Reading note: The stated length-independent initialization cost conflicts with the explicit quadratic matrix.

Paper, Sec. 4.1 and Fig. 8, PDF p. 14. Memory size is a reader calculation.

PFUS puts selected tokens in one normalization domain

PFUS: Pre-Fusion Unified Softmax.

$$\mathcal{I}_t = \mathrm{unique}(\mathcal{A}_{\mathrm{stoch}} \cup \mathcal{W}_t)$$

$$o_t = \sum_{j \in \mathcal{I}_t} \frac{\exp(q_t^\top k_j / \sqrt{D_h})}{\sum_{\ell \in \mathcal{I}_t} \exp(q_t^\top k_\ell / \sqrt{D_h})} v_j$$

- Cache the random anchor after the first generated token.
- Update the local window, deduplicate, then gather the selected KV.
- Every selected token competes using the same softmax denominator.

Reading note: No isolated PFUS ablation is reported. Exact denominator-weighted merging of separate groups can also recover unified softmax.

Equation is a teaching reconstruction of Sec. 4.3 / Fig. 9, PDF pp. 15–16. $q, k \in \mathbb{R}^{D_h}$; $\mathcal{W}_t$ is the dynamic local window.

One-percent views can produce a half-dense union

$$U = 1 - (1 - d)^S$$

- Independent views, equal per-target inclusion probability d.
- Probability of selection, not probability of a correct answer.
- Shared cliques violate simple independence assumptions.

d	S	Expected U
1%	10	9.56%
1%	70	**50.52%**
1%	100	63.40%
5%	10	40.13%
5%	40	87.15%
5%	70	**97.24%**

Single-view density is not final attention density or end-to-end cost.

Paper, Eq. (1), Sec. 2.2.3, PDF pp. 5–6. Table recalculated from the equation; the paper uses N for seed count here.

The stated implementation does not establish $O(T \log T)$

If each query keeps $r(T)$ keys, ideal sparse attention costs

$$O(T\, r(T)\, D_h)\,.$$

- $O(T \log T)$ requires $r(T) = O(\log T)$ at fixed head dimension.
- Under the paper's fixed-density union model, $r(T) \approx T[1 - (1-d)^S]$: still quadratic for fixed d, S.
- A sparse mask alone does not prove that prefill skips dense arithmetic.
- Construction, KV storage, projections, MLPs, and decoding need separate accounting.

Reading note: The paper gives no complete scaling curve that verifies the headline asymptotic claim for this implementation.

Reader analysis of Secs. 3.4–3.5 and 4.1–4.3, PDF pp. 13–16.

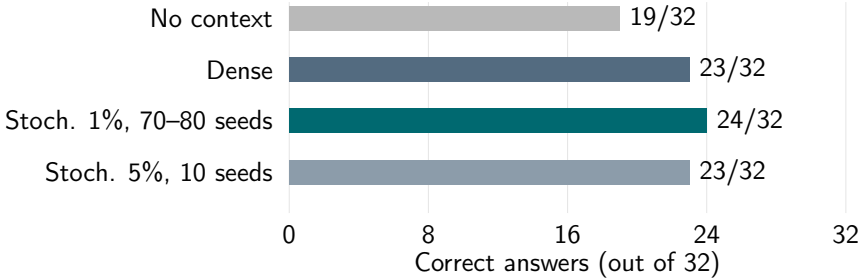
The main benchmark is small multiple-choice QA

	Native-window study	Extrapolation study
Model	Qwen2-1.5B-Instruct	Qwen2-1.5B-Instruct
Context	32,768 tokens	65,536 tokens
Questions	32, five choices	64, five choices
No-context score	19/32 (59.38%)	33/64 (51.56%)
Dense score	23/32 (71.88%)	OOM in tested setup
Main hardware	16 GB desktop CPU	128 GB Xeon CPU

- Context: four cleaned biological/scientific manuscripts.
- Choose the option with the highest log-probability; no free-text QA.
- Float32, CPU only; some 32k configurations move to the larger server.

Paper, Secs. 2.1–2.2, 4.4.1; PDF pp. 2–5, 15–17.

At 32k, the best reported result adds one correct answer



$$\text{Gap recovery} = \frac{24 - 19}{23 - 19} = 125\%.$$

Reading note: 125% is a ratio of net accuracy gains. It does not mean recovering 125% of document information. Dense is an empirical reference.

Redrawn from Table 5, PDF p. 20; Sec. 2.2.1, p. 3. Best-vs-dense gain: 3.125 percentage points.

The 32k Structural evidence is internally inconsistent

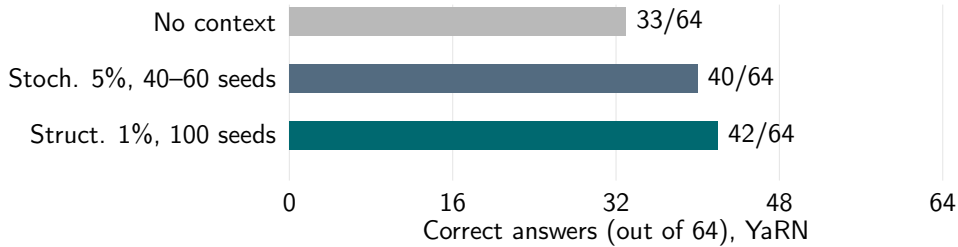
Configuration	Text / Table 5	Figure 1
Structural 1%, 10 seeds	68.75%	≈65.6%
Structural 1%, 40 seeds	71.88%	≈68.8%
Structural 5%, 10 seeds	65.62%	≈71.9%

- Figure 3 shows Structural peaks of 68.75% at both 1% and 2%.
- These conflict with the text's claim of reaching dense at those densities.
- The headline "75% gap recovery with 10 seeds" is therefore disputed within v1.

Reading note: The reading preserves both sources; raw run records are needed to resolve the discrepancy.

Paper, Sec. 2.2.1 / Fig. 1, p. 3; Fig. 3, p. 4; Table 5, p. 20.

At 64k, selected configurations improve on no context



- Structural gain: $9/64 = 14.0625$ percentage points.
- Stochastic gain: $7/64 = 10.9375$ percentage points.
- Dense failed in this setup; dense-relative quality is unknown at 64k.

Redrawn from Fig. 4 and Tables 2–3, PDF pp. 5, 9–10; paired results, p. 17. 42/64 is 65.625%, rounded as 65.62% or 65.63% in the paper.

The two modes favor different parts of the tested grid

	Stochastic		Structural	
Density	Mean	Peak	Mean	Peak
1%	52.13%	57.81%	**56.53%**	**65.62%**
2%	54.97%	60.94%	56.25%	62.50%
5%	**58.66%**	**62.50%**	55.26%	60.94%

- Local structure helps the low-density region in this benchmark.
- Stochastic global sampling benefits more from increasing density.
- Increasing seeds does not guarantee monotonically improving accuracy.

Reading note: Means summarize different grid configurations, not repeated estimates at a fixed configuration. No uncertainty intervals are reported.

Paper, Table 2 / Sec. 2.6, PDF pp. 8–9; Fig. 5, p. 6.

YaRN beats linear scaling in the reported peak comparison

Density	Linear peak (seeds)	YaRN peak (seeds)
1%	45.3% (50)	57.8% (40)
2%	54.7% (50)	60.9% (40)
5%	59.4% (40–50)	62.50% (40–60)

- 64k is a 2× extension of the model's stated 32k native window.
- Positional scaling and sparse selection jointly affect retrieval.
- These are separately selected peaks; seed budgets are not all matched.

Reading note: Do not compare linear single-seed scores (15.6–23.4%) against YaRN multi-seed peaks as a pure RoPE effect. Fig. 6 also mixes reporting conventions.

Paper, Table 1, PDF p. 7; Secs. 2.3–2.5, pp. 6–8.

TinyLlama shows a boundary, not a universal scaling law

Context	Native-window multiple	Questions	Correct
4,096	2×	4	2 (50.0%)
8,192	4×	8	1 (12.5%)
16,384	8×	16	1 (6.25%)
32,768	16×	32	6 (18.75%)

- TinyLlama-1.1B with YaRN; stated native window: 2,048.
- Paper reports no accuracy change over the tested densities and seeds.
- Adding views did not rescue these configurations.

Reading note: Very small, changing question sets and incomplete controls prevent attribution solely to RoPE or generalization to all small models.

Paper, Sec. 2.5, PDF pp. 7–8.

Sub-1% density retains less net retrieval gain than claimed

Structural setting	Best seeds	Best score	$B_{\max}$
0.3%	150–200	38/64 (59.38%)	197
0.5%	150–200	38/64 (59.38%)	328
0.8%	100	40/64 (62.50%)	524
1.0% baseline	100	42/64 (65.63%)	655

Absolute accuracy retained

$38/42 = 90.48\%$

Net gain over no context retained

$(38 - 33)/(42 - 33) = 55.56\%$

Reading note: Lower block density with more seeds does not establish a 70% reduction in total runtime.

Paper, Sec. 2.7 / Table 3, PDF p. 10. Retention ratios recalculated from question counts.

Cross-document generation is qualitative and expensive

Different experiment

- Qwen2.5-1.5B, linear RoPE.
- Six document pairs, 22.8k–32.4k tokens.
- Two phases change density, seeds, temperature, and repetition penalty.

Reported observations

- Phase 1: 90–113 min per pair.
- Phase 2: 16–63 min per pair.
- One semantic error in Phase 1; none observed in Phase 2.
- 36 GB peak across six active processes.

Reading note: Twelve runs with author assessment do not establish a general hallucination rate. Different output lengths and decoding parameters confound timing and quality attribution.

Paper, Sec. 2.8 / Table 4, PDF pp. 10–11. Times converted to minutes.

The strongest reported paired test is only marginal

Comparison	p	Interpretation
32k best Stochastic vs no context	0.227	Not significant
64k best Structural vs no context	0.078	Below 0.10 only
64k best Stochastic vs no context	0.189	Not significant
64k best Structural vs best Stochastic	0.6875	No detected difference

- These are the paper's McNemar results, not independently reproduced tests.
- No direct best-Stochastic-vs-dense significance result is reported.
- Reusing questions across a grid does not create independent evidence.

Reading note: No detected difference is not equivalence; post-selection uncertainty and multiplicity remain unresolved.

Paper, Sec. 4.4.1, PDF pp. 16–17. Full discordant-pair counts are not given there.

A stronger evaluation needs quality and actual cost together

1. **Resolve provenance:** align plots, Table 5, code snapshot, and per-question predictions.
2. **Separate selection and testing:** larger held-out QA sets, repeated random trials, uncertainty intervals.
3. **Measure real cost:** union density, mask construction, prefill, decode, peak memory, end-to-end latency.
4. **Isolate mechanisms:** PFUS, cliques, local window, cached anchors, and RoPE with matched controls.

CPU feasibility is useful evidence; it does not establish universal GPU speedups or architecture-independent quality.

Reader synthesis of the evidence gaps in Secs. 2–4, PDF pp. 2–17.

Keep the engineering idea and qualify the headline claims

What the paper supports

A runtime combination of sparse geometry, cached index selection, unified softmax, and RoPE scaling can deliver some long-context retrieval gains on the reported CPU benchmark.

- **Mechanism:** merge sparse views; do not confuse one-view density with total work.
- **Evidence:** promising selected scores, small samples, disputed Structural 32k data.
- **Boundary:** neither universal $O(T \log T)$ inference nor reliable dense superiority is established.

Synthesis of arXiv:2607.21927v1. Full evidence locators and discrepancies are in the companion Chinese notes.

Sources and reading scope

Anderson R. Santos. RIS-Kernel: A Model-Agnostic Architecture for Long-Context LLM Inference via Sparse Attention.
arXiv:2607.21927v1, July 24, 2026.

- Versioned arXiv record and 20-page paper PDF.
- Read the author TeX, bibliography, all sections, nine figures, five tables, and the supplementary table.
- Paper figures are reproduced with attribution under CC BY 4.0. Bars and arithmetic in this deck are explicitly labeled reconstructions.
- Author's GitHub project is a replication entry point. Current repository content is not treated as the v1 implementation.

Reading note: No model experiments or source-code audit were performed for this reading. Reported statistical tests are attributed to the paper.

Backup: why naive group averaging changes softmax

One global token has score 3; two local tokens have scores $0, 0$.

One shared denominator

$$p_{\text{global}} = \frac{e^3}{e^3 + 2} \approx 0.909.$$

Every token's score contributes to the same normalization.

Equal averaging of two groups

The global token gets probability 1 within its singleton group, then weight $1/2$ after averaging.

$$p_{\text{global}} = 0.5.$$

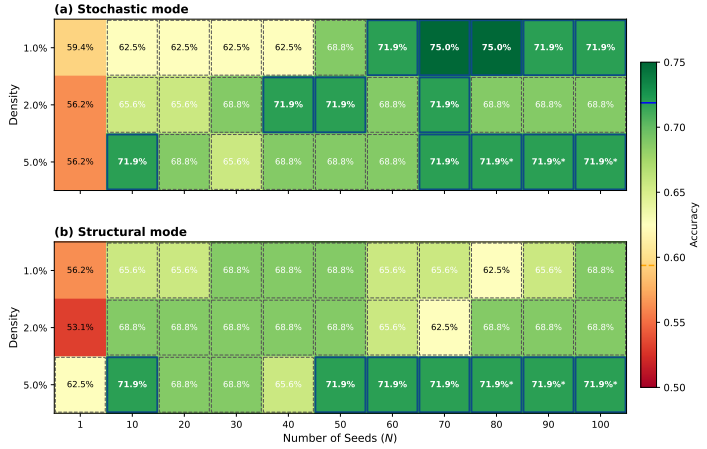
Exact merging remains possible with group denominator weights:

$$o = \frac{Z_G o_G + Z_L o_L}{Z_G + Z_L}.$$

Z_G, Z_L: softmax denominators of the disjoint global and local groups.

Teaching example, not an experiment in the paper. Motivation: Sec. 4.3, PDF pp. 15–16.

Backup: the original 32k figure exposes the conflict



Original Fig. 1, PDF p. 3, Anderson R. Santos, CC BY 4.0. Structural 1% at 10/40 seeds differs from Table 5 (p. 20).

Backup: further consistency checks change interpretation

- **Figure 2, p. 4:** right-axis gap recovery does not correctly align 0% with the floor and 100% with dense. Recalculate from counts.
- **Table 5, p. 20:** Stochastic 5% / one seed is 57.81%, which cannot be an integer-correct result on a single 32-question run.
- **Secs. 2.2.2 / 2.5:** 62.50% is called a 64k ceiling, but Tables 2–3 report Structural at $42/64 = 65.625\%$.
- **Sec. 2.7, p. 10:** 0.7% and 0.9% still improve from 150 to 200 seeds in Table 3, contrary to the saturation text.

Reader checks against the original PDF figures, tables, and TeX. These remain unresolved in v1.

Backup: sub-1% best scores are non-monotonic

Density	$B_{\max}$	Best accuracy	Best seeds	Grid mean
0.1%	66	51.56%	150, 200	48.44%
0.2%	131	54.69%	150	51.95%
0.3%	197	59.38%	150, 200	55.08%
0.4%	262	59.38%	150	56.25%
0.5%	328	59.38%	150, 200	58.20%
0.6%	393	57.81%	100, 150	56.64%
0.7%	459	60.94%	200	57.42%
0.8%	524	62.50%	100	60.16%
0.9%	590	57.81%	200	55.47%

Structural / Qwen2-1.5B / YaRN / 65,536 tokens.
Each density tests 50, 100, 150, 200 seeds; these are configuration means.

Selected columns and best-seed locations from Table 3, PDF p. 10. No-context score: 51.56%.