NeurIPS 2026

Long-Context Language Models Require Extreme Sparsity in Context Dimension

Sahil Joshi* · Prithvi Dixit* · Agniva Chowdhury · Joseph E. Gonzalez · Anshumali Shrivastava · Ion Stoica · Kumar Krishna Agrawal · Aditya Desai

Rice University · UC Berkeley · IIT Bombay

github.com/skylight-org/skylight-research

*Equal contribution

MOTIVATION

Context Keeps Growing. Attention Remains Dense.

- **Workloads have evolved.** Agentic tool use, long dialogues, and multi-document reasoning push contexts towards 10^5-10^6 tokens.
- **Dense attention scales with context.** Prefill incurs quadratic attention computation, while each decoded token accesses the full KV cache.
- **Sparsity has not become standard in attention.** Mixture-of-Experts enables sparse FFN computation, yet attention remains predominantly dense.

33K
tokens in a 50-page PDF
Illustrative document

160K
tokens in one system prompt
reported for the OpenClaw

N^2
compute per prefill
Quadratic in context length

N
KV token reads per decoded token
Linear in context length

LIMITATION

Top-k Attention Is Accurate Only When Attention Is Peaked

DEFINITION · SPARSE TOP-k ATTENTION

$$a = softmax\left(\frac{qK^T}{\sqrt{d}}\right) \in \mathbb{R}^N, \qquad o = \sum_{i=1}^{N} a_i v_i$$

$$S = top_k(a), |S| = k \ll N, \qquad \tilde{o} = \sum_{i \in S} \frac{a_i}{\sum_{j \in S} a_j}$$

Keep the k keys with the largest attention scores: decoding reads $k \ll N$ cache entries, chosen by a cheap indexer.

Used in DeepSeek-V3.2 / V4, GLM-5.2.

PROBLEM · ATTENTION IS NOT ALWAYS PEAKED

Attention can be uniform: top-k sparse attention performs poorly during aggregation-like tasks.[1]

Aggregation-like: mass spread thin



[1] Z. Chen et al., "MagicPIG: LSH Sampling for Efficient LLM Generation," ICLR 2025.

LIMITATION

Linear Attention Cannot Retrieve Arbitrary Tokens

DEFINITION · LINEAR ATTENTION & SSMs

$$e^{\frac{q^T k_i}{\sqrt{d}}} \approx \phi(q)^T \phi(k_i), \qquad \phi\colon \mathbb{R}^d \to \mathbb{R}^r$$

$$\hat{o} = \frac{\phi(q)^T \sum_{i=1}^{N} \phi(k_i) v_i^T}{\phi(q)^T \sum_{i=1}^{N} \phi(k_i)}$$

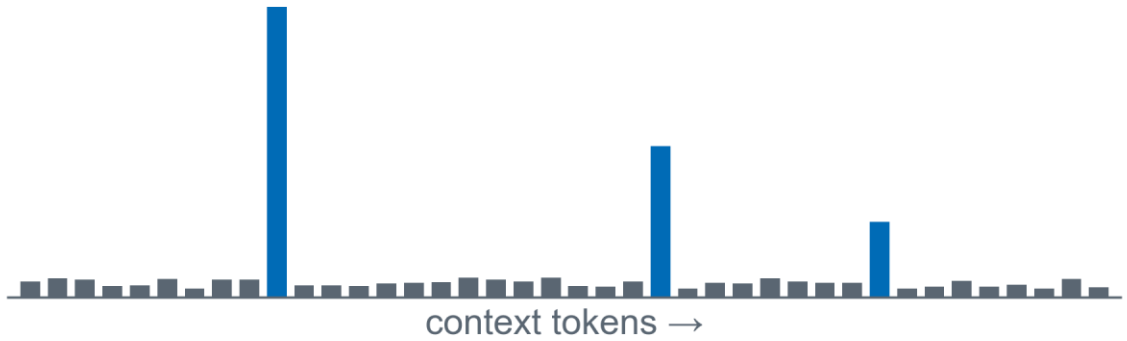
A feature map replaces the softmax kernel.

Variants: Mamba, Gated DeltaNet, KDA, lightning attention.

THE PROBLEM · RECALL IS BOUNDED BY THE STATE

Linear Attention is limited on retrieval tasks: Exact associated recall requires $\Omega(N)$ bits of state, motivating hybrid architectures.[1], [2]

Retrieval-like: a few tokens hold the mass



[1] S. Arora et al., "Simple linear attention language models balance the recall-throughput tradeoff," ICML 2024.
[2] S. Arora et al., "Zoology: Measuring and Improving Recall in Efficient Language Models," ICLR 2024.

CONSENSUS

1M Context Models Keep Dense Attention Layers

Model	Release	Lab	Declared context	Full-attention layers	Context-sparse?
MiniMax-Text-01	Jan 2025	MiniMax	~1M	10 / 80	No
Qwen2.5-7B-Instruct-1M	Jan 2025	Alibaba Qwen	~1M	All layers	No
MiniMax-M1	Jun 2025	MiniMax	~1M	10 / 80	No
Kimi Linear 48B-A3B	Oct 2025	Moonshot AI	~1M	7 / 27	No
DeepSeek-V3.2	Dec 2025	DeepSeek	~128K	None · DSA	Yes
Mistral Small 4 119B	Mar 2026	Mistral AI	~250K	All layers	No
Nemotron 3 Super Base	Mar 2026	NVIDIA	~1M	8 / 8	No
GLM-5.1	Apr 2026	Z.ai	~200K	None · DSA	Yes
DeepSeek-V4	Apr 2026	DeepSeek	~1M	None · CSA, HCA	Yes
GLM-5.2	Jun 2026	Z.ai	~1M	None · DSA	Yes
Solar-Open2-250B	Jul 2026	Upstage	~1M	12 / 48	No
Kimi K3	Jul 2026	Moonshot AI	~1M	24 / 93	No
Inkling	Jul 2026	Thinking Machines Lab	~1M	11 / 66	No

PROPOSAL

Dense Attention Is Not Necessary.

THE POSITION

The future of LLM inference lies in extreme sparsity along the context dimension, but this sparsity cannot be naive.

SPARSE top−k ATTENTION

Retrieval-dominated attention

- Captures salient information

+

LINEAR ATTENTION & SSMs

Aggregation-dominated attention.

- Captures global information

COMPLETE CONTEXT SPARSITY

Complementary mechanisms. No dense attention layers.

IN PRACTICE · Qwen3.8-Flash-Next already does this: top-k attention + linear attention, no full attention.

THEORY

A Structural Motivation Behind Sparsity.

THEOREM 1

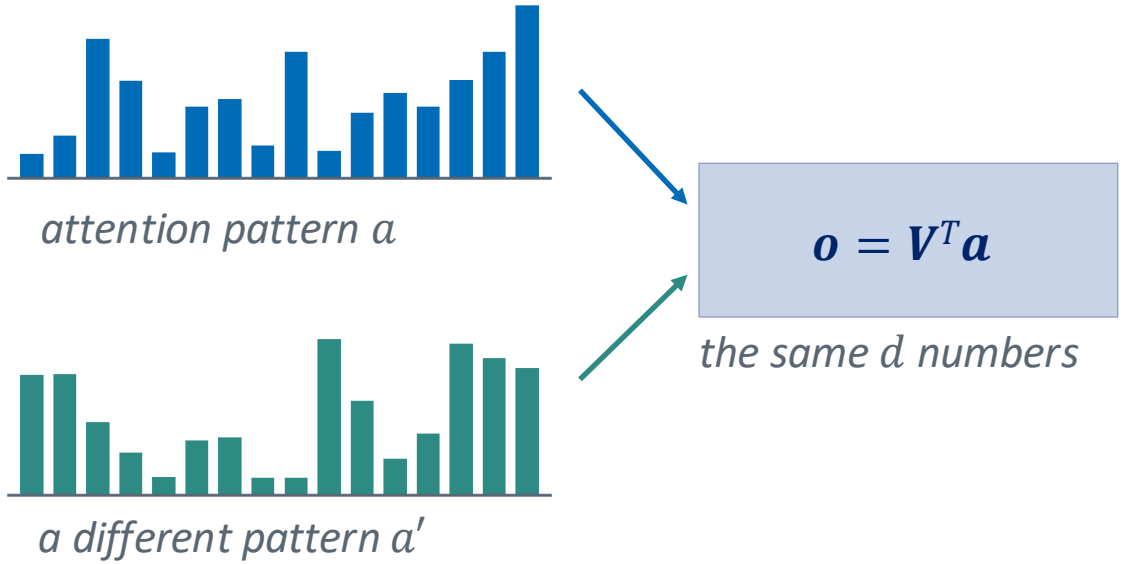
$$o = V^T a, \qquad a \in \Delta^{N-1}, \qquad V \in \mathbb{R}^{N \times d}$$

When $d < N - 1$, distinct attention distributions can produce identical outputs,

$$a \neq a', \qquad V^T a = V^T a'.$$

COROLLARY

Distinguishing every possible attention distribution requires $d \geq N - 1$.



1,000,000 vs 128

context tokens vs. head dimension

≥ 999,871 directions of attention variation are invisible to the output

THE EVIDENCE SETUP

Three Design Choices for Sparse Decode

GRANULARITY

Per-query, per-head selection. Each query head independently selects its relevant tokens.

SELECTION RULE

Oracle-top-k is the quality ceiling. In practice, a learned or hashed indexer (HashAttention, Double Sparsity) can be used.

BASE SET

Always keep the anchors. Sink tokens and a local window stay in every query's set; the budget k is spent on the rest of the context.

What one decode step attends to, per query head (schematic)



$K\times$ sparsity means each query head attends to $1/K$ of the context: $5\times \approx 20\,\%$, $50\times \approx 2\,\%$, $100\times \approx 1\,\%$.

THE EVIDENCE SETUP

Experimental Setup: 4 Benchmarks, 5 Model Families

RULER-HARD · 32K context

Six hard tasks from RULER-32K: multi-key needle retrieval, variable tracking, frequent words, and qa. *Score: mean accuracy.*

LOFT · 32K and 128K context

Long-context retrieval / QA, five datasets: HotpotQA, Natural Questions, MuSiQue, QAMPARI, QUEST. *Score: mean subspan exact match.*

AIME 2025 · long reasoning

30 competition math problems; Qwen3.5-27B, generations capped at 65K tokens. *Score: accuracy; one problem = 3.33 points.*

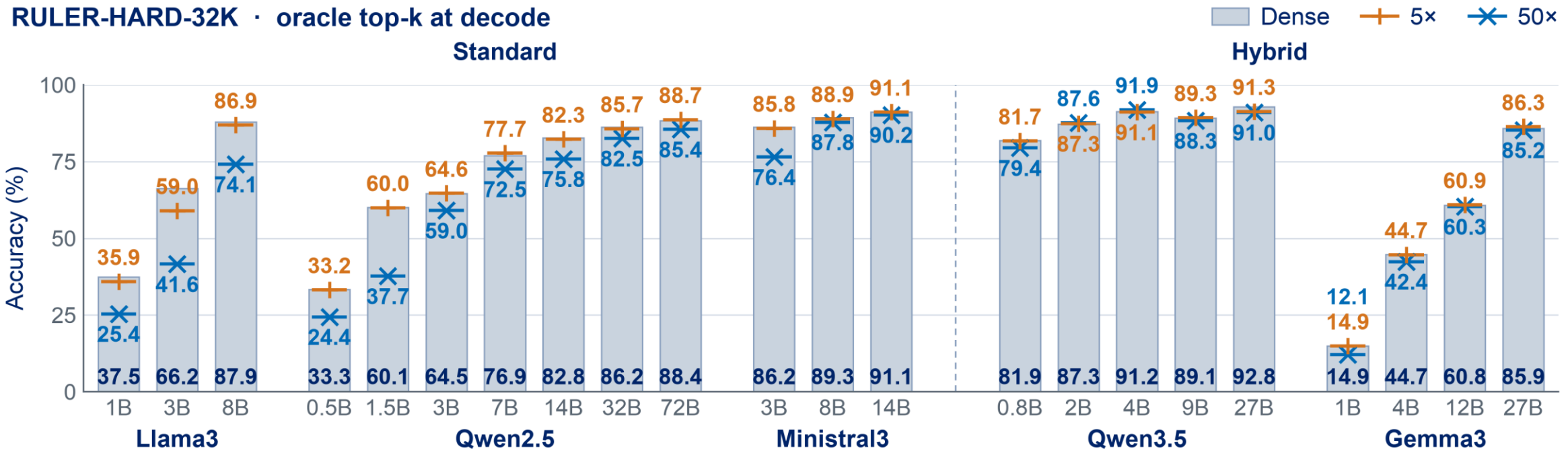
SWE-Bench Lite · Agentic Coding

300 real GitHub issues, Qwen3.5-27B as the agent, up to 100 turns per task. *Score: resolved tasks.*

MODELS · 5 families, 20 models from 0.5B to 72B, no sparse training. Standard transformers: Llama3, Qwen2.5, Ministral3. Hybrids with linear-attention/sliding-window layers: Qwen3.5, Gemma3.

THE EVIDENCE

Extreme Sparsity Preserves Near-Dense Accuracy in Hybrids



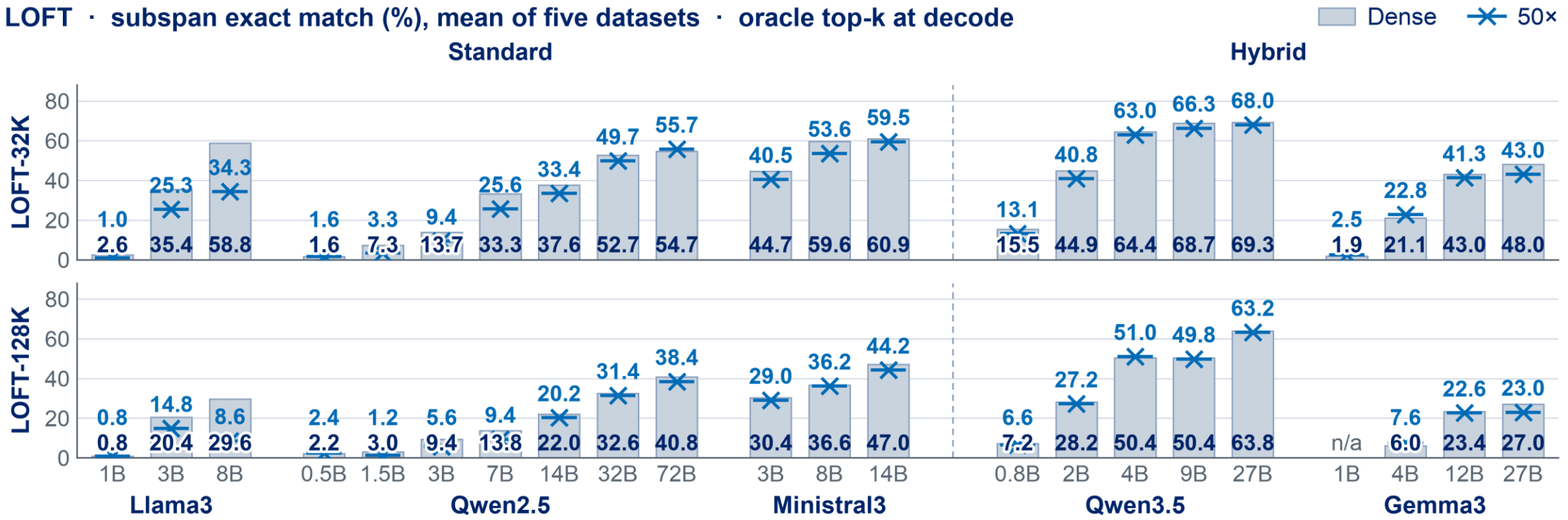
92.8 → 91.0
Qwen3.5-27B, dense → 50×

85.9 → 85.2
Gemma3-27B, dense → 50×

87.9 → 74.1
Llama3-8B, dense → 50×

THE EVIDENCE

Extreme Sparsity Remains Effective at Longer Contexts



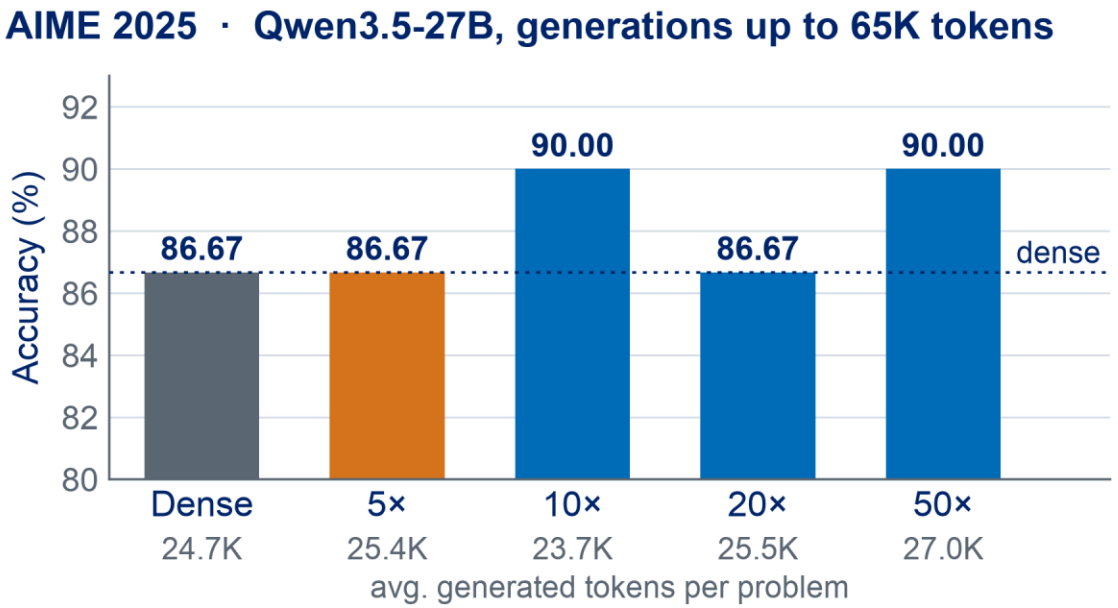
63.8 → 63.2
Qwen3.5-27B, 128K, dense → 50×

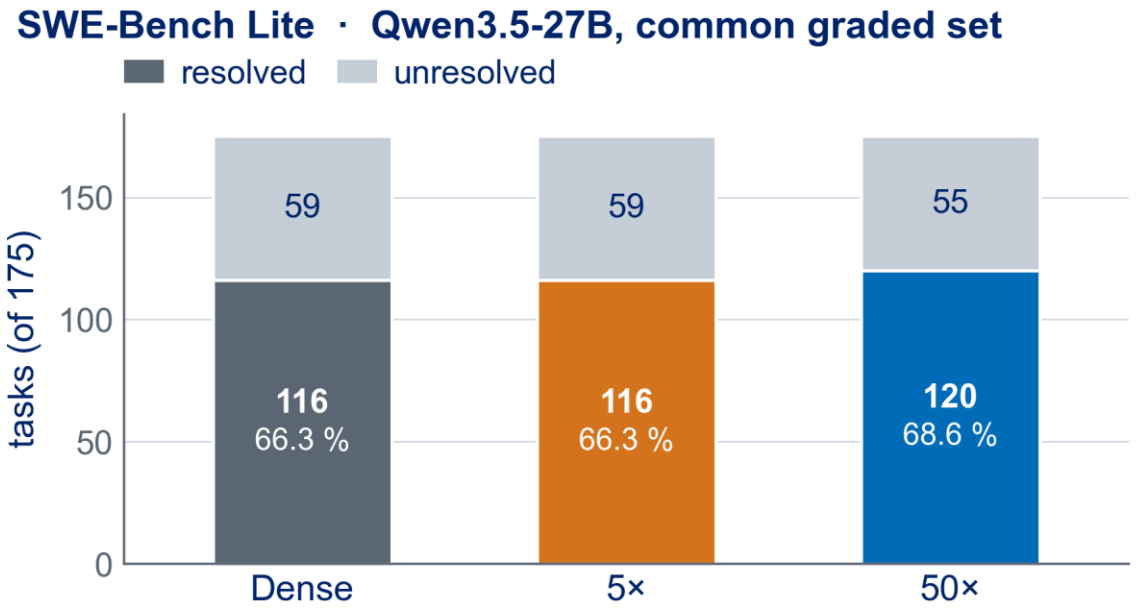
54.7 → 55.7
Qwen2.5-72B, 32K, dense → 50×

29.6 → 8.6
Llama3-8B, 128K, dense → 50×

THE EVIDENCE

50× Sparsity Preserves Reasoning and Agentic Performance



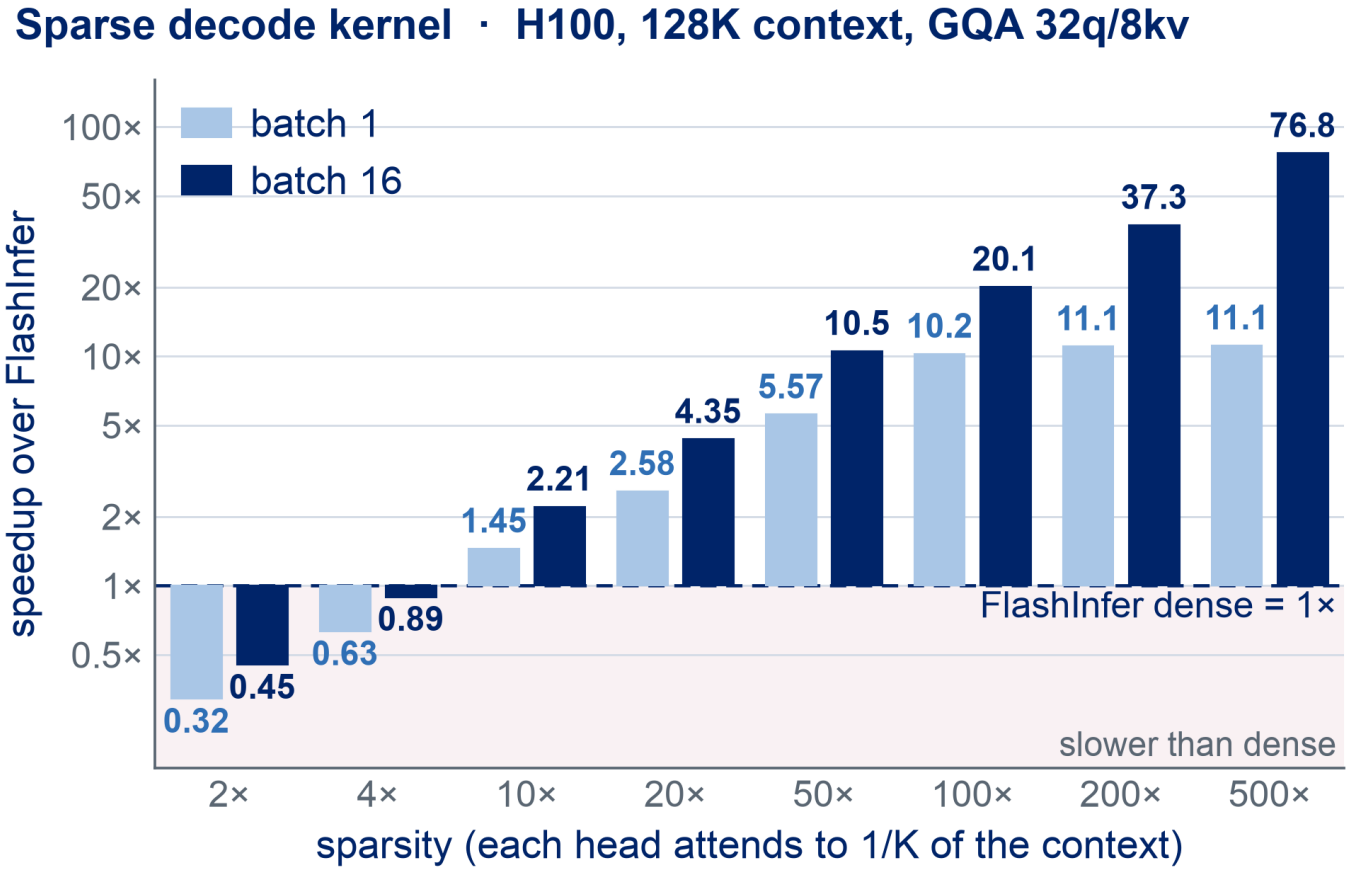


90 vs 86.67
AIME 2025, 50× vs dense

68.6 % vs 66.3 %
Resolved / 175, 50× vs dense

THE EVIDENCE

Irregular Sparsity Already Achieves Speedups on an H100



[1] S. Yang et al., "Post-Training Sparse Attention with Double Sparsity," arXiv:2408.07092, 2024.

≈10×

sparsity to break even with FlashInfer

slower at ≤ 4×, faster from 10×, at every batch size

5.6-10.5×

faster at 50× sparsity

batch 1 → batch 16, 128K context

76.8×

at 500× sparsity, batch 16

no block structure imposed on the sparsity

3.9× / 2.5×

MHA / GQA at 50×, indexer included, batch 16

8-channel Double Sparsity[1]; up to 4.2× / 2.8× at 100×

THE TAKEAWAY

Open Questions

01

KV page granularity

What size suits per-token selection, and is it the same for training, prefill and decode?

02

Top-*k* as a system primitive

Every sparse method keeps approximate-key state (hashes, PQ codes). Should it get its own memory tier?

03

Fused Gather + Matmul

Merge them in hardware, or is a software gather plus dense matmul enough?

04

Tiered KV cache

Sinks and recent tokens are read every step; most tokens rarely. Should the memory hierarchy mirror that?

THE TAKEAWAY

Treat context sparsity as a design principle for models, training systems and hardware - not as a heuristic.

Thank you!

Long-Context Language Models Require Extreme Sparsity in Context Dimension

NeurIPS 2026

github.com/skylight-org/skylight-research