SemKV

SemKV

A Context Optimization Layer for LLMs

Reducing unnecessary context by **~80%** while preserving meaning and maintaining near-base latency



LLMs don’t need more context — they need **better context**

SemKV

02 THE PROBLEM

LLMs are inefficient in how they consume context

Excessive, noisy context

Retrieval systems return more text than needed, including irrelevant and redundant content.

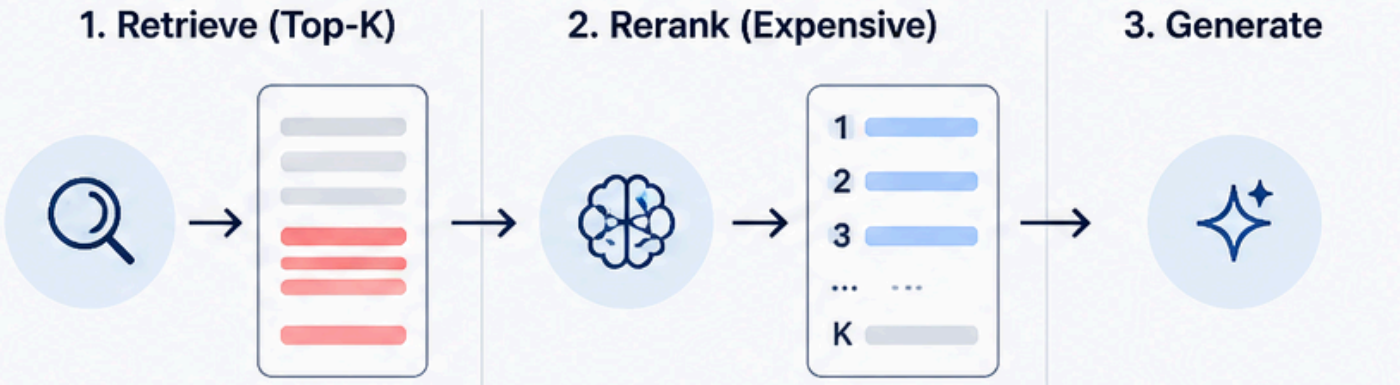
Higher cost and latency

More tokens mean higher compute costs and slower responses.

Reranking is expensive

Cross-encoder reranking improves relevance but adds an additional model pass, increasing latency significantly.

Current Pipeline: Retrieval → (Rerank) → Generation



We are optimizing models — but not what they read.

More tokens.
More cost.
More latency.

The problem is not the model. It’s the context we feed it.

SemKV

03 THE SOLUTION

SemKV: Dynamic Context Optimization Layer

A lightweight, data-driven layer that intelligently shapes context to deliver only what the LLM truly needs.

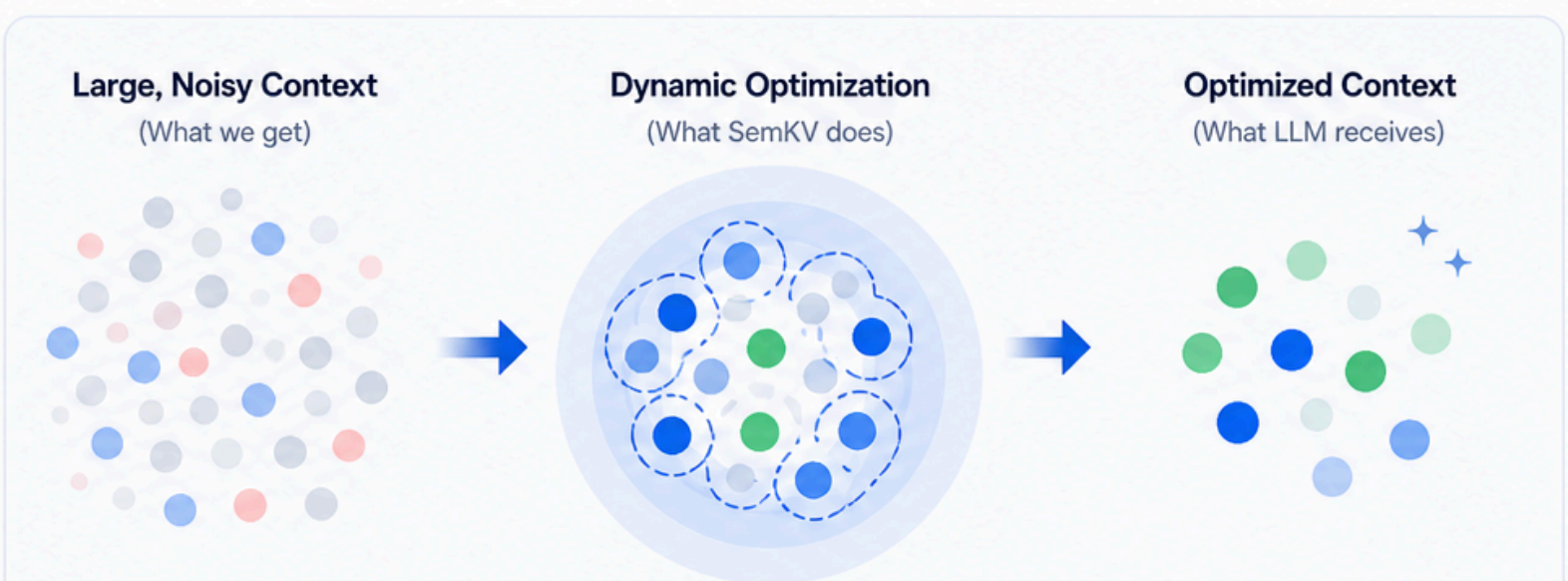
Data-Driven Selection
Identifies high-signal semantic regions from a large pool of candidate context.

Adaptive & Flexible
No fixed K or static thresholds. Selection adapts to each query and intent.

Expand When Needed
Intelligently broadens context only if additional information adds value.

Clean, Relevant, Compact
Removes redundancy and noise to keep context focused and highly relevant.

WHAT SEMKV DOES (AT A HIGH LEVEL)



SemKV doesn't just retrieve more. It delivers what matters.

- **More Relevant** — Higher signal, better answers
- **Fewer Tokens** — Lower cost, less compute
- **Faster Responses** — Reduced latency, better UX

SemKV is the missing layer between retrieval and generation. | Smarter Context | Higher Efficiency | Better Outcomes

SemKV

04 BENCHMARK RESULTS

Max Efficiency. Zero Compromise.

SemKV delivers rerank-level context quality with up to 85% fewer tokens and ~2× lower latency than rerank systems.

EVALUATION SETUP

- **Dataset:** Real-world queries across diverse domains
- **Pipelines Compared:**
  1. **Llama 3.1 Base** (RAG) (retrieval only)
  2. **Llama 3.1 + Rerank** (cross-encoder reranking)
  3. **SemKV (Ours)** (dynamic context optimization)
- **Metrics:** Token Usage, Token Reduction, End-to-End Latency, Semantic Similarity (embedding cosine)

Lower latency and higher token reduction are better. Quality within the comparable range (0.88 – 0.91).

SEMKV VS. BASELINES (AGGREGATE RESULTS)

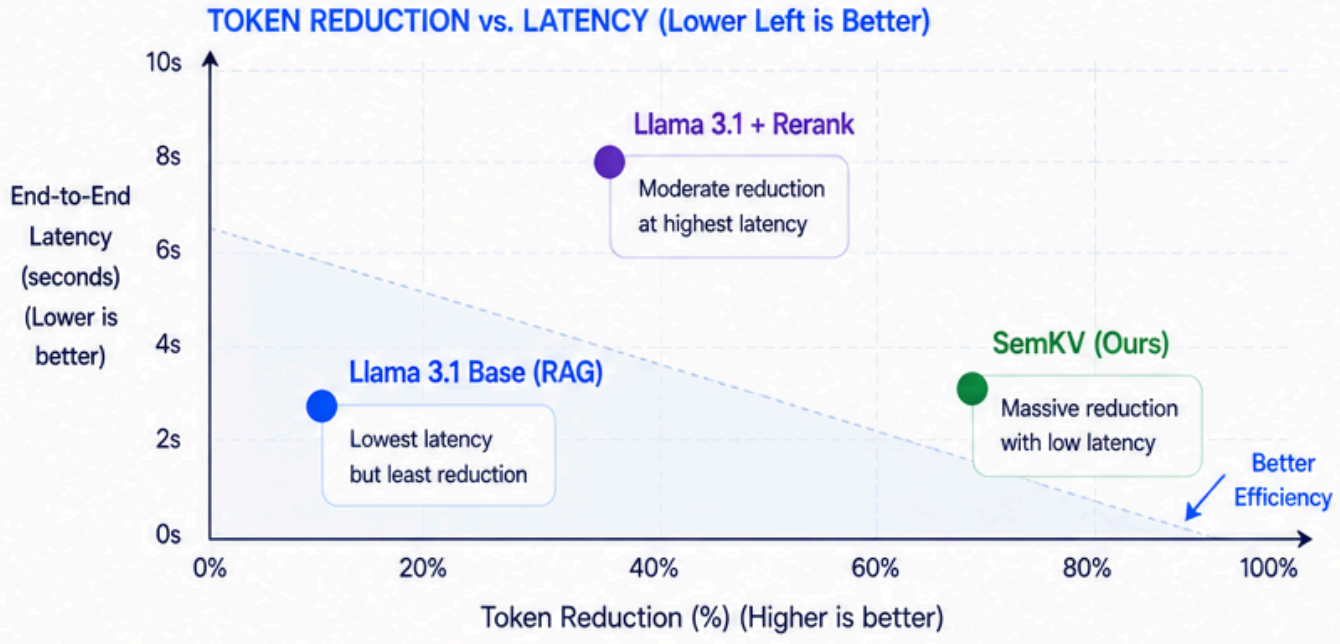
1 Llama 3.1 Base (RAG)
- TOKEN REDUCTION: 17% — vs. full corpus
- SEMANTIC SIMILARITY: 0.88 — (baseline quality)
- LATENCY (END-TO-END): 2.6s — Fastest but least efficient

2 Llama 3.1 + Rerank
- TOKEN REDUCTION: 41% — Moderate reduction
- SEMANTIC SIMILARITY: 0.90 — Highest quality
- LATENCY (END-TO-END): 7.8s — Highest latency

3 SemKV (Ours)
- TOKEN REDUCTION: 69 – 85% — Massive reduction
- SEMANTIC SIMILARITY: 0.88 – 0.90 — Comparable quality
- LATENCY (END-TO-END): 3.4s — ~2× faster than rerank (0.53× latency)

Results are averaged across all test queries. Similarity range shown where applicable.

TOKEN REDUCTION vs. LATENCY (Lower Left is Better)



THE TAKEAWAY

SemKV achieves rerank-level quality with **69–85% fewer tokens** and **~2× lower latency** than rerank systems.

- Dramatically lower token usage
- ~2× faster end-to-end latency
- Comparable quality of answers

SemKV redefines the efficiency frontier—more with less, without giving up quality.

- **Less Compute** — Lower inference cost
- **Less Latency** — Faster responses
- **Better Quality** — More relevant answers
- **Better Experience** — Higher satisfaction

Latency measured on identical hardware. SemKV vs. rerank latency multiplier = 0.53× (~2× faster).

SemKV 05 REAL-WORLD IMPACT

Real Impact. Real Use Cases.

SemKV is delivering measurable impact across real-world RAG applications where efficiency and quality matter most.

60–85%
Fewer Tokens
vs. baseline systems

~2× Faster
End-to-End Latency
vs. rerank systems (0.53× latency)

Comparable Quality
Semantic Similarity
0.88 – 0.90 (same as rerank)

Dynamic Selection
No fixed top-K
Adapts per query for optimal context

SEMKV IN ACTION: REAL-WORLD USE CASES

Use Case	Domain	Challenge	Impact with SemKV	Key Outcome
Enterprise Document Q&A	Enterprise Knowledge Management	Large, unstructured document corpora causing high token usage and latency.	↓ 72% Tokens ~2× Faster	✓ Faster answers for employees ✓ Lower infra cost per query
Finance Research Assistant	Financial Services	Extensive reports and filings increase context length and latency.	↓ 69% Tokens ~2× Faster	✓ Quick retrieval of critical information ✓ Reduced time-to-insight for analysts
Market Intelligence Search	Market Research	Large volumes of market reports slow retrieval and increase LLM load.	↓ 63% Tokens ~2× Faster	✓ Efficient handling of long documents ✓ Lower latency for complex queries
Customer Support Assistants	Customer Service	High token usage from verbose KB articles leading to slow responses.	↓ 70% Tokens ~2× Faster	✓ Faster resolutions ✓ Lower cost and higher customer satisfaction

Across diverse domains, SemKV consistently delivers up to **85% token reduction**, **~2× lower latency**, and **comparable answer quality.**

BUSINESS IMPACT

Lower Inference Cost
Up to 85% fewer tokens means significantly lower LLM usage cost.

Faster Time-to-Value
~2× faster responses drive better UX and higher productivity.

Scalable by Design
Handle more queries with the same infrastructure.

Quality You Can Trust
Rerank-level semantic quality with no compromise.

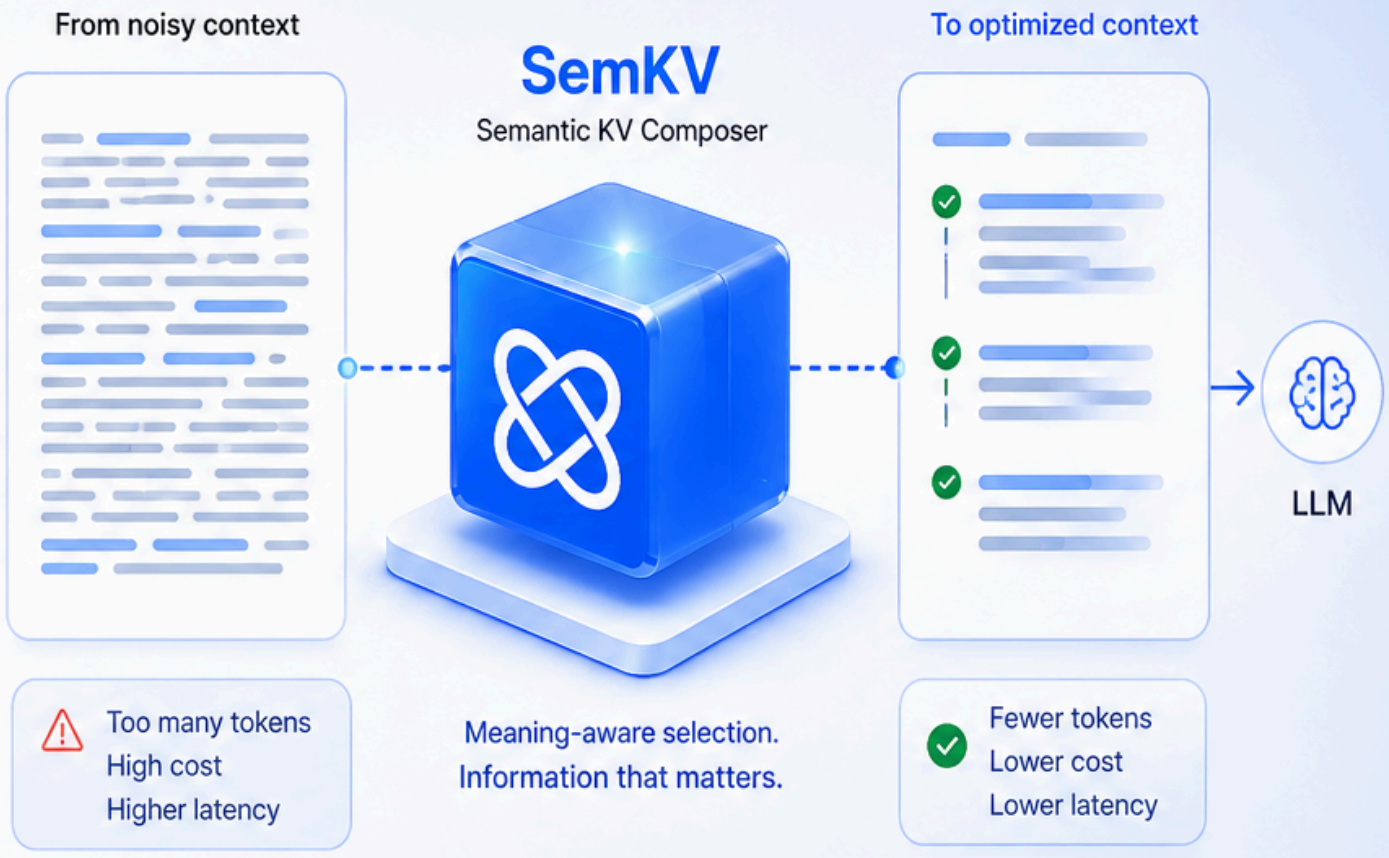
SemKV makes RAG systems faster, cheaper, and smarter — without compromising quality.

Maximum Efficiency | Real Impact | Built for Scale

SemKV API

Send less context. Get the same answer.

Semantic KV Composer API for
context optimization before inference



Reduce Tokens
Send less.
Pay less.

Reduce Latency
Faster inference.
Better UX.

Preserve Meaning
Same (or better)
answer quality.

Ready to try?
Explore the API

API Docs
→ SemKV API Endpoint
Interactive docs • Try endpoints • See examples

Secure by design | Fast by default | Built for builders

SemKV

Smarter Selection, Lower Latency, Higher Relevance.

BENCHMARK REPORT

RAG vs S3 (SELECTIVE STRUCTURAL SEGMENTATION) PERFORMANCE EVALUATION

S3 (Selective Structural Segmentation) delivers high-quality, semantically relevant context with ~67–83% less data and up to 5.41X lower latency compared to traditional RAG pipelines.

1. PARADIGM SHIFT

Moving beyond Retrieval-Augmented Generation (RAG) to Semantic Selection Architecture.

RAG Retrieval Architecture	S3 Selective Structural Segmentation
Chunk-first	Semantic Selection
Top-K Retrieval	Structural Formation
Re-ranking	Structured Segmentation
Context Expansion	Context Optimization
Static Boundaries	Dynamic Boundaries
More Data In	Only What Matters In

RAG retrieves more. S3 selects what matters.

2. DATASET

- **Domain:** Artificial Intelligence
- **Corpus:** Semantically dense knowledge corpus
- **Total Corpus Size:** ~2,000 tokens overall
- **Queries Evaluated:** 35

Query Types

- Factual: 40%
- Conceptual: 30%
- Analytical: 30%

3. EVALUATION SETUP

Retrieval / Selection Scope

RAG (Baseline): Top-10

S3 (~7xR content)

S3 follows the Selective Structural Segmentation (S3) architecture.

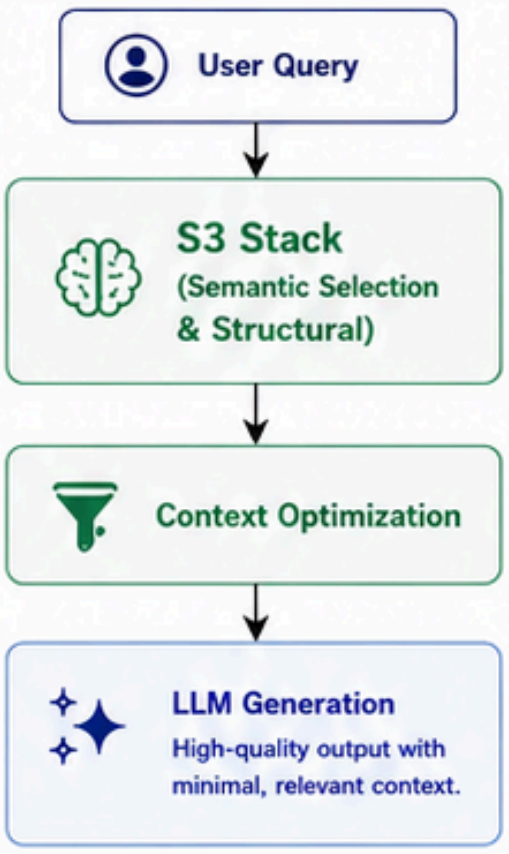
Baselines Compared (RAG)

- Llama 3.1 70B (Base)
- BGE Reranker
- MiniLM Reranker

Metrics

- Data Reduction Similarity (higher is better)
- Latency in seconds (lower is better)

4. S3 PIPELINE (HIGH LEVEL)



S3 intelligently selects, structures, and optimizes context—so the LLM only sees what matters.

5. RESULTS SUMMARY (Pathological Outliers Removed)

Comparison	RAG Pipeline (Baseline Top-60)		S3 Architecture (SemKV)		Token Reduction (vs RAG)	Avg Data Reduction Similarity (↑)	Latency Improvement (S3 vs RAG)	
	Avg Tokens Returned	Avg Latency (seconds)	Avg Tokens Returned	Avg Latency (seconds)			Faster by	Speedup (X)
vs Llama 3.1 70B (Base)	791.0	3.00	131.5	2.20	83.38%	0.888	26.7%	1.36X
vs BGE Reranker	395.1	11.91	131.5	2.20	66.72%	0.872	81.5%	5.41X
vs MiniLM Reranker	426.5	7.72	131.5	2.20	69.17%	0.895	71.5%	3.51X

S3 (Selective Structural Segmentation) returns ~67–83% less data than RAG baselines while maintaining high semantic similarity, achieving up to 5.41X lower latency.

AVERAGE TOKEN REDUCTION (ACROSS BASELINES) **73.09%** TOP VALUE

6. KEY TAKEAWAYS

- Up to ~83% reduction in data returned to the LLM.
- Up to 5.41X lower latency compared to RAG pipelines.
- Strong semantic similarity preservation across all baselines.
- Consistent performance across factual, conceptual, and analytical queries.
- Lower latency translates to better user experience and lower inference cost.

7. WHY S3 IS THE FUTURE

BUILT FOR SPEED
Optimized for low latency, high-throughput agentic interactions.

EDGE AI READY
Minimal data transfer makes S3 ideal for edge environments.

AGENTIC SYSTEMS
Efficient, selective context enables high-resolution agent-to-agent communication.

COST EFFICIENT
Less data, less compute, lower cost—without compromising quality.

RAG is overkill for edge AI and high-speed agentic systems.
S3 is built for the next generation of intelligent, efficient AI.

ABOUT SEMKV & S3

SemKV is powered by the S3 architecture (Selective Structural Segmentation)—a next-generation semantic selection infrastructure that identifies, structures, and optimizes only what truly matters—eliminating data bloat, latency, and cost while preserving the quality of downstream generation.

Lower Latency | Lower Cost | Higher Quality | Built for Agents

BENCHMARK DATE: MAY 2026 | QUERIES EVALUATED: 35 | DATASET: AI Semantically Dense (~2K TOKENS OVERALL)

LLM costs don’t scale with usage.

They scale with context.

Most systems keep
adding more context.
Few question
what actually goes in.

