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 API

## Composable APIs.
## Full control over context.

Build, control and analyze optimized context before sending it to your LLM.

- **POST /compose** — Compose optimized context for LLM inference (core API)
- **POST /utils/prepare** — Prepare and normalize input data
- **POST /metrics** — Get compression and similarity metrics
- **POST /kv_cache** — Prepare and manage KV cache
- **POST /config/strategy** — Set optimization strategy and behavior
- **GET /config** — Get current configuration

Explore full documentation
→ **SemKV API Endpoint**

## POST /compose

Compose optimized context for LLM inference

Adaptive context selection (no fixed top-k). Semantic understanding decides what matters.

Mode controls behavior. Strategy controls trade-offs.

### Request

```
{
  "input": "your long context...",
  "query": "user question or intent",
  "mode": "extract | reason",
  "strategy": "coverage | precision | tuneable",
  "prompt": "default | custom"
}
```

Modes define how the model should treat the context.
Strategies define how aggressively to compress.

### Response (200 OK)

```
{
  "answer": "...",
  "context_reduction": 0.31,
  "similarity": 0.84,
  "strategy_used": "precision",
  "mode_used": "extract",
  "compression_applied": true,
  "full_context_tokens": 104,
  "filtered_context_tokens": 71
}
```

Similarity score indicates how close the optimized context is to the original meaning. Higher is better.

### Key Parameters

| | | |
|---|---|---|
| input | Full context to be optimized | |
| query | User question or intent | |
| mode | extract | → filter signal |
| | reason | → compose reasoning-ready context |
| strategy | coverage | → retain more context |
| | precision | → aggressive pruning |
| | tuneable | → adjust trade-offs |
| prompt | default | → built-in prompt |
| | custom | → your custom prompt/behavior |

### Common Use Cases

- **RAG Systems** — Retrieve less, send less, maintain answer quality
- **AI Agents** — Smarter decisions with leaner context
- **Enterprise Search** — Better results with optimized context

**Send less** — Fewer tokens
**Pay less** — Lower cost
**Same meaning** — High similarity
**Full control** — Configurable strategies

SemKV API



# Significant reduction.
# Same answer.

We remove noise. Not meaning.

SemKV keeps what matters and removes the rest using semantic understanding.

Similarity measures semantic closeness to original context.

## Full Context (Before)

Unoptimized · Noisy · Longer

```
GlobalTech Systems is a leading provider of industrial
solutions with a strong focus on innovation and customer
success. Founded in 2005, the company has grown to
serve clients in over 50 countries. Its product portfolio
includes advanced cooling systems, precision components,
and AI-driven monitoring platforms. Over the years,
GlobalTech has received multiple industry awards for
innovation and sustainability. The company invests
heavily in research and development, maintaining
state-of-the-art labs across North America, Europe,
and Asia. With a dedicated workforce of over 8,000
employees, GlobalTech is committed to delivering
high-quality, reliable, and efficient solutions that
help customers optimize operations and achieve long-term
growth in a rapidly evolving market landscape...
```

Tokens: 104

SemKV Compose →

## Optimized Context (After)

Optimized · Relevant · Shorter

GlobalTech Systems, founded in 2005, provides industrial solutions including advanced cooling systems, precision components, and AI-driven monitoring platforms. The company serves clients in 50+ countries and invests in R&D through labs across North America, Europe, and Asia to deliver reliable, high-quality solutions that optimize operations and drive long-term growth.

Tokens: 71

Key details and meaning preserved.

## Output & Metrics

Context Reduction
0.31 (example)

31% reduction
(this case)

Typical range: 30% – 70%
Depending on content and query

30% — 31% (this case) — 70%

Similarity (Semantic)
0.84

High similarity
Meaning preserved

Compression Applied
true

Optimization
success

Reduction varies with redundancy and query intent.

### Same Answer Quality

LLM gets the right context it needs to produce the same high-quality answer.

User Query → SemKV API → LLM → Accurate Answer

### Less is more

Less context. Lower cost. Better efficiency.

**Lower Cost**
Pay less

**Lower Usage**
Fewer tokens

**Faster Inference**
Less data to process

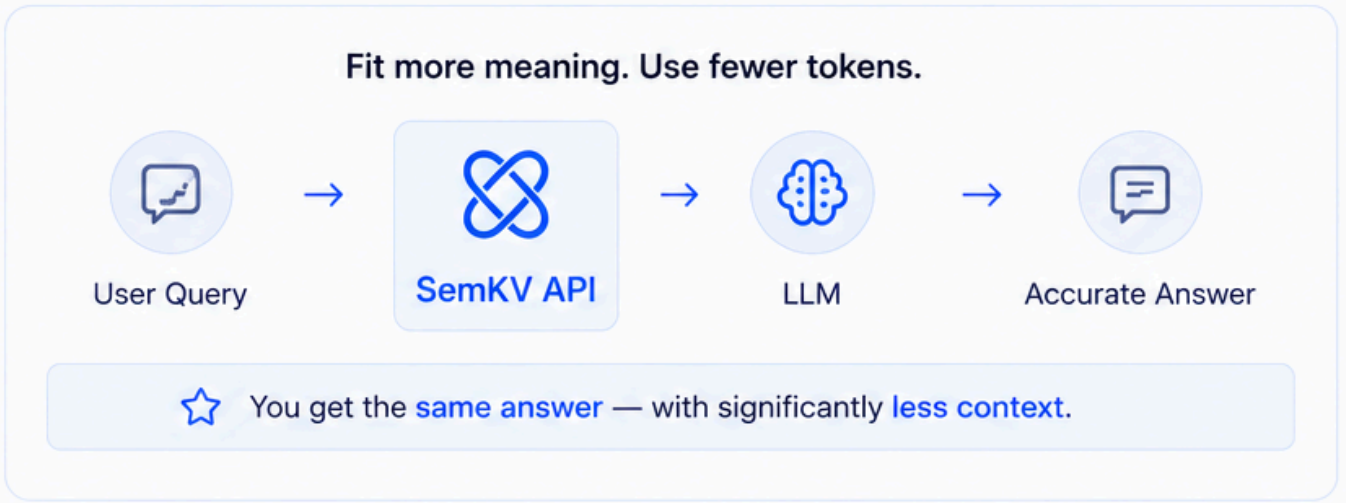
**Better Accuracy**
High similarity retained

SemKV API



# The missing layer in LLM pipelines.

Context optimization before inference for better cost, speed, and accuracy.




### Lower Cost
Send fewer tokens to the model and pay less.

**30–70%**
typical token reduction

### Faster Systems
Less data to process means lower latency.

**Lower latency**
faster responses

### Better Quality
Preserve meaning with high semantic similarity.

**High similarity**
meaning preserved

### More Context Room
Fit more relevant context within the same window.

**More room**
for what matters

**Adaptive context selection (no fixed top-k).**
Semantic understanding decides what matters for each query. No arbitrary limits.

- **No fixed top-k** — Not limited by K
- **Query-aware** — Adapts to intent
- **Only what matters** — Removes noise

## Send only what matters.

Drop SemKV into your pipeline and start optimizing context today.

## Explore the API

→ SemKV API Endpoint

Full documentation • API reference • Examples

Early access • Built for real-world evaluation • Continuously improving