SemKV
Semantic Selection Infrastructure

CPU Deployment & Edge Compatibility

Evaluated on Real Hardware. Ready for the Edge.

SemKV's S3 architecture was evaluated for retrieval-quality optimization under CPU deployment using a 2,000-token dataset on a standard laptop processor. The results demonstrate efficient execution, low latency overhead, and high semantic retention—establishing a strong foundation for deployment across a wide range of edge and CPU-based systems.

Tested. Validated. Scalable.

Proven on a real laptop CPU with 2,000-token dataset.

- Retrieval-quality optimization
- Low latency overhead
- CPU-efficient execution

EVALUATION SETUP

Dataset	2,000-token dataset
Hardware	Standard Laptop Processor
Cores / Threads	4 Cores / 8 Threads
Base / Boost Clock	2.1 GHz / 4.1 GHz
Memory	16 GB RAM
Environment	Windows 11, Python Runtime

KEY OUTCOMES

Retrieval-quality optimization achieved on CPU
High semantic relevance with reduced token activation.

Low latency overhead
Efficient execution with minimal reranking dependency.

CPU-efficient operation
Optimized compute and memory footprint.

Scalable to edge environments
Architecture supports a wide range of CPU-based systems.

COMPATIBLE EDGE & CPU PROCESSORS (INDUSTRY WIDELY USED)

Intel® Core™ i5 (U-Series) — Mid-Range
4–6 Cores / 8–12 Threads
Base: ~1.2 – 2.4 GHz
Boost: ~3.5 – 4.2 GHz
TDP: 15W – 28W
Excellent compatibility

Intel® Core™ i7 (U-Series) — High-Performance
4–6 Cores / 8–16 Threads
Base: ~1.3 – 2.8 GHz
Boost: ~4.5 – 5.0 GHz
TDP: 15W – 28W
Excellent compatibility

AMD Ryzen™ 5 (U-Series) — Mid-Range
6 Cores / 12 Threads
Base: ~2.1 – 2.6 GHz
Boost: ~3.8 – 4.3 GHz
TDP: 15W – 28W
Excellent compatibility

AMD Ryzen™ 7 (U-Series) — High-Performance
8 Cores / 16 Threads
Base: ~2.0 – 2.7 GHz
Boost: ~3.8 – 4.3 GHz
TDP: 15W – 28W
Excellent compatibility

Qualcomm® Snapdragon 8 Gen Series — Mobile / Edge
8 Cores
Base: ~1.8 – 2.3 GHz
Boost: ~3.0 – 3.3 GHz
TDP: 5W – 10W
Good compatibility

Apple M1 / M2 (ARM) — Mobile / Edge
8 Cores
Base: ~2.0 – 3.2 GHz
Boost: ~3.0 – 3.5 GHz
TDP: 5W – 10W
Excellent compatibility

SemKV is designed to run efficiently across modern CPU architectures with 4+ cores and 8+ threads.

What This Means

SemKV's performance on a standard laptop CPU with a 2,000-token dataset proves its efficiency and scalability. You can expect similar retrieval-quality optimization on the above edge and CPU platforms widely used in industry.

CPU Optimized | Low Latency | Edge Ready | Industry Compatible

Results based on internal evaluations on a standard laptop processor with a 2,000-token dataset. Performance may vary based on hardware configuration and workload.

SemKV
Semantic Selection Infrastructure

Potential Deployment Environments

Extending Edge Compatibility Beyond Laptop Systems

SemKV's selective semantic activation architecture reduces retrieval overhead and compute requirements, enabling efficient execution across a wide spectrum of CPU-based and edge-oriented devices and systems.

Proven on Laptop. Ready for the Edge.

Validated on a standard laptop CPU with a 2,000-token dataset.

The same efficiency principles and reduced infrastructure dependency extend naturally to modern edge and embedded platforms.

WIDE RANGE OF COMPATIBLE DEPLOYMENT ENVIRONMENTS

Mobile AI Devices



Smartphones and mobile devices with on-device AI

- ARM-based CPUs
- Low power
- On-device privacy

Edge Gateways



Edge gateways and micro servers for local inference

- x86 / ARM CPUs
- Network efficient
- Local processing

Embedded AI Systems



Embedded boards and compact compute modules

- ARM / RISC-V CPUs
- Low power
- Compact footprint

Smart Cameras



AI-enabled cameras for real-time video analytics

- ISP + CPU SoCs
- Real-time capable
- Edge analytics

Robotics Controllers



Robots, drones and autonomous systems with on-board AI

- ARM / x86 CPUs
- Deterministic
- Reliable inference

On-device Assistants



Voice/vision assistants and personalized AI experiences

- Low latency
- Always-on ready
- Privacy focused

Industrial Edge Compute



Industrial PCs and edge servers for predictive analytics

- High reliability
- Extended lifetime
- Harsh environment

Consumer AI Devices



AR/VR, wearables, and next-gen consumer devices

- Power efficient
- Responsive AI
- Immersive apps

Why It Extends

SemKV minimizes broad retrieval, reduces reranking dependency, and decreases semantic operations—resulting in lower compute, memory, and latency requirements. These characteristics make it suitable for diverse CPU-based and edge-oriented platforms.

- Lower Compute Requirements
- Reduced Latency Overhead
- Energy Efficient
- Flexible Across Form Factors

Important Note

SemKV has been evaluated on a standard laptop CPU with a 2,000-token dataset. The compatibility with other devices and environments is based on architectural analysis and the efficiency characteristics of the system.

Future Work

Ongoing work includes validation and optimization on additional mobile, embedded, and robotics platforms to further demonstrate real-world performance across these environments.

Evaluation conducted on a laptop with 2,000-token dataset. | *Performance may vary based on hardware configuration and workload.*