CompreSeed: A Patent-Pending Architecture for Semantic Compression, High-Speed Reasoning, and Zero-Leakage Knowledge Retrieval

Patent Application: JP 2025-201662 — "Compressed Semantic Search Inference System"

Inventor: Yoshikazu Nakamura

Executive Summary

Modern AI systems—particularly Large Language Models (LLMs)—face several structural limitations:
(1) unstable reasoning,
(2) high hallucination rates,
(3) an inability to retain and query large external knowledge,
(4) extreme GPU cost, and
(5) significant security and leakage risks when using vector databases or RAG pipelines.

CompreSeed introduces a *patent-pending* architecture that overcomes these limitations by enabling
semantic compression, search, and reasoning without decompression.

This new architecture has been prototyped successfully on a commercial laptop (NEC NS150K), storing **3 million Wikipedia documents** in a compressed semantic-core format and supporting **real-time retrieval and reasoning on CPU only**.

CompreSeed is not a variant of RAG, vector search, or embedding-based retrieval.

It is a **new knowledge-processing paradigm** that merges:

- irreversible semantic compression,
- CPU-native high-speed retrieval,
- integrated reasoning logic, and
- LLM interoperability as an external symbolic memory.

This technical foundation enables enterprise-scale AI with dramatically reduced computational costs and exceptional security.

1. Patent Overview

- **Title**: *Compressed Semantic Search Inference System*
- **Application Number: JP 2025-201662**
- **Filing Date**: Nov 21, 2025
- **International Classification: G06F 17/50**
- **Inventor**: Yoshikazu Nakamura
- **Additional Options**:
    - Divisional applications possible
    - One-year grace period for related filings
    - PCT international filing right retained

The patent covers:

- a semantic-core compression method,
- a compressed inference structure,
- a reasoning model integrated inside the compressed graph,
- and a unified query-to-answer pipeline requiring no decompression.

This creates strong **IP defensibility** and **exclusive implementation rights**.

2. Technical Background and Current Limitations

Traditional AI knowledge systems rely on:

- vector databases,
- RAG pipelines,
- full-text search engines,
- or monolithic LLM parametric memory.

These approaches suffer from well-known bottlenecks:

(1) High GPU cost and latency

Embedding generation, vector search, and LLM reasoning all require expensive compute.

(2) Information leakage and security risk

Vector databases store sensitive content in reversible forms.

(3) No integrated memory

RAG inserts knowledge temporarily; LLMs forget everything outside training.

(4) Zero compression reuse

Large knowledge bases scale linearly with size, with no lossless or semantic compression path.

These limitations collectively prevent stable, low-cost enterprise deployment.

3. The CompreSeed Architecture

CompreSeed introduces the world's first architecture enabling:

**✓ Semantic compression →

✓ Search →

✓ Reasoning

without decompression.**

The system constructs **irreversible semantic cores** that represent meaning, relations, and context with extreme compactness.

Core Components

1. **Semantic-Core Compression Engine**
    - Reduces millions of documents to a compact meaning graph
    - Irreversible by design → strong security
2. **Compressed Reasoning Graph**
    - Encodes conceptual relationships and inference paths
    - Enables "knowledge reasoning" at compression-level storage
3. **CPU-Only High-Speed Retrieval**
    - Achieves real-time responses for multi-million-document corpora
    - No GPUs required
4. **LLM Interoperability Layer**
    - Works as an external symbolic memory
    - Dramatically reduces hallucinations
    - Provides stable logical grounding for generative models

4. Prototype Demonstration

A working prototype demonstrates:

- Hardware: **Consumer laptop NEC NS150K**
- Dataset: **3,000,000 Wikipedia documents**
- Storage: compressed semantic cores
- Retrieval speed: real-time
- Language tasks: stable QA, summary, topic extraction

- Integration: usable as an external memory for GPT-class LLMs

This verifies the architecture is practical, lightweight, and production-ready.

5. Enterprise Advantages

1) 90%+ Cost Reduction

No GPU clusters, no embeddings, no expensive vector search.

2) Zero Leakage, High Security

Irreversible compression means raw data cannot be reconstructed.

3) Hallucination-Resistant Reasoning

Compressed knowledge graph stabilizes LLM outputs.

4) On-Premise or Offline Operation

Ideal for government, medical, legal, and industrial applications.

5) Massive Scalability

Millions of documents on low-end hardware.

6. Strategic Value for Large Technology Companies

CompreSeed provides **exclusive advantages** for companies seeking:

- superior AI reasoning foundations,
- low-cost knowledge integration at global scale,
- secure enterprise AI infrastructure,
- or next-generation memory systems for LLMs.

This is a **platform architecture**, not an application-level feature.

Exclusive rights to CompreSeed would provide:

- defensible technological differentiation,
- cost and performance advantages,
- and long-term competitive leadership.

7. Partnership and Acquisition Opportunities

The inventor is open to:

- **Technology acquisition**,
- **Exclusive license negotiations**,
- **Joint development**,
- **Enterprise deployment partnerships**,

- **Integration with existing LLM ecosystems**.

All associated IP — patent application, divisional rights, PCT rights, and future filings —
may be included in negotiated agreements.

Contact

Yoshikazu Nakamura
Email: **info@xinse.jp**
LinkedIn: https://www.linkedin.com/in/y-nakamura-ai/