RAG-Ready Semantic Indexing

FAISS & Embeddings

91,229

VECTORS INDEXED

Legal

DOMAIN CLASS

Zero

PROCESSING ERRORS

Delivered: FAISS index, aligned chunks, metadata, manifest

Integration: LangChain / LlamaIndex / Haystack compatible

Status: Production-ready — no preprocessing required

Portfolio Validation Artifact — End-to-End Indexing Delivery

Overview & Challenge

Transformed a legal benchmark corpus into a production-ready vector index with cleaned chunks, aligned metadata, dense embeddings, and FAISS index — structured for immediate RAG framework integration.

Client Requirements:

- ▸ Professional-grade chunking respecting legal text structure
- ▸ Dense vector embeddings for semantic similarity search
- ▸ FAISS index ready for framework integration
- ▸ Documented outputs enabling independent verification
- ▸ No ongoing infrastructure dependencies

Pipeline Execution

Stage 1: Data Preparation
Encoding normalization, control character removal, whitespace standardization

Stage 2: Deduplication
Hash-based duplicate detection to prevent retrieval bias

Stage 3: Semantic Chunking
Text segmentation with legal document structure awareness

Stage 4: Embedding Generation
Dense vectors via validated retrieval model with context prefixes

Stage 5: Index Construction
FAISS IndexFlatIP with inner product similarity

Stage 6: Validation
Post-processing verification of chunk-metadata-vector alignment

Execution Metrics & Quality

Processing Results

Metric	Value
Source Records	91,229
Final Vectors	91,229
Dimensions	1,024
Self-Contained Chunks	99.6%
Mid-Word Breaks	0.4%
Processing Errors	0

Quality Verification

Data

- UTF-8 ✓
- Clean ✓
- Normalized ✓

Embedding

- Prefix ✓
- L2 norm ✓
- No nulls ✓

Index

- Aligned ✓
- Loads ✓
- Search ✓

Deliverables & Integration

Delivered Artifacts

```
indexed_corpus/
├── chunks.jsonl      → 91,229 chunks
├── metadata.jsonl    → Aligned IDs
├── vectors.index     → FAISS index
└── summary.json      → Manifest
```

chunks.jsonl — Full text with domain prefixes
metadata.jsonl — Line-aligned identifiers
vectors.index — IndexFlatIP ready for Python
summary.json — VERIFIED status + config

Framework Compatibility

Framework	Integration
LangChain	`FAISS.load_local()`
LlamaIndex	Vector store API
Haystack	FAISS document store
Custom	Standard FAISS bindings

Query-time requirements documented in manifest.

Outcome

DELIVERY STATUS

✓ Production-Ready — Zero Errors

Complete, validated vector index ready for production integration. Legal benchmark corpus transformed from raw text to retrieval-ready format without client managing infrastructure.

Key Outcomes

- ▸ **Zero processing errors** — All records indexed
- ▸ **Verified alignment** — Chunks, metadata, vectors matched
- ▸ **Framework-ready** — No preprocessing required
- ▸ **Documented** — Manifest enables verification

Client Value

- ▸ No GPU infrastructure management
- ▸ No embedding pipeline maintenance
- ▸ Immediate RAG integration capability
- ▸ Independent verification possible

Output specifications and quality thresholds are engagement-specific.