Large-Scale Semantic Indexing

With Validation

4.2 Million Vectors

SEC REGULATORY FILINGS — DETERMINISTIC EXECUTION

9.3B	~10 hrs	Zero
TOKENS PROCESSED	EXECUTION TIME	DATA LOSS

Handled: Memory limits, checkpoint recovery, sharded storage

Avoided: Silent failures, index corruption, resource exhaustion

Delivered: Production-ready index with full validation artifacts

Challenge & Scale

Processed an extensive archive of SEC 10-K regulatory filings for semantic retrieval supporting compliance research, competitive analysis, and due diligence workflows.

Scale Challenges:

- **Volume:** Billions of tokens requiring batch control
- **Memory:** Single-pass would exhaust resources
- **Reliability:** Multi-hour execution needs recovery
- **Verification:** Silent failures undetectable at scale
- **Storage:** Index exceeds single-file limits

Standard indexing approaches insufficient. Required infrastructure-aware execution with operational safeguards.

Approach

Memory Discipline
Controlled batch sizes with explicit cleanup between stages. Memory monitoring throughout.

Multi-Stage Validation
Chunk-vector alignment, shard integrity, index load testing, null embedding detection.

Checkpoint Recovery
Intermediate state preservation. 10-hour job interrupted at hour 8 doesn't restart from zero.

Deterministic Outputs
Identical source data → byte-identical results. Reproducibility required at scale.

Sharded Storage
Vectors distributed across sequential shard files. No single file exceeds practical limits.

Document Classification
LEGAL_REGULATORY class, LONG token regime. Calibrated for complex regulatory text.

Execution & Quality Metrics

Processing Summary

Metric	Value
Source	SEC 10-K Filings
Tokens	9.3 billion
Vectors	4,229,632
Dimensions	1,024
Duration	~10 hours
Errors	0

Quality Thresholds

Metric	Result	Threshold
Self-Contained	**97.05%**	≥ 90%
Mid-Semantic	**2.95%**	≤ 10%
Mid-Word	**0.14%**	≤ 1%
Alignment	**100%**	100%
Null Vectors	**0**	0

Deliverables & Validation

Delivered Artifacts

```
indexed_corpus/
├── chunks.jsonl          → 4.2M chunks
├── metadata.jsonl        → Aligned IDs
├── vector_shards/
│   └── ... (43 shards)   → Distributed
├── summary.json          → VERIFIED
├── semantic_split_analysis.json
└── batch_calibration.json
```

Validation Artifacts

Artifact	Purpose
semantic_split_analysis	Chunking quality breakdown
batch_calibration	Reproducibility params
Integrity verification	Alignment + load test
Processing log	Audit trail

Outcome

VALIDATION STATUS

✓ Verified — Production-Ready

Validated, production-grade vector index for high-availability retrieval systems. Regulatory corpus now supports compliance research, competitive intelligence, and due diligence.

Key Outcomes

- **4.2M vectors** — No data loss or silent failures
- **Zero errors** — Across 10-hour execution
- **Verified integrity** — Chunk, vector, index levels
- **Production storage** — Memory-efficient shards

Operational Guarantees

- **Deterministic** — Identical source → identical output
- **Recoverable** — Checkpoint prevents full restart
- **Verifiable** — Validation artifacts enable audit
- **Scalable** — Sharded for large corpora

Scale parameters, quality thresholds, and validation criteria are engagement-specific.