Multi-Domain Indexing Pipeline

Validated at Scale

155.7M Verified Vectors

244 DATASETS — 8 DOMAINS — SCALES FROM 8K TO 28M VECTORS PER CORPUS

244
DATASETS VERIFIED

8
KNOWLEDGE DOMAINS

Zero
SILENT FAILURES

Handled: Heterogeneous source data across legal, financial, scientific, cybersecurity, Q&A, and general knowledge domains

Verified: FAISS index integrity, vector-chunk alignment, and domain classification for every dataset

Delivered: Production-ready retrieval corpus with zero data loss and full validation artifacts

Challenge & Scale

This registry documents pipeline execution across 244 datasets to validate indexing reliability, domain versatility, and scale handling before client deployment. The pipeline was tested against legal filings, financial records, scientific papers, cybersecurity intelligence, developer Q&A archives, and general knowledge corpora — each with different structure, token density, and retrieval requirements.

Scale Challenges:

- **Volume:** 155.7 million vectors across 244 active datasets
- **Heterogeneity:** 8 distinct domains with different chunking strategies
- **Infrastructure:** 38 TB across two dedicated storage volumes
- **Verification:** Silent indexing failures are undetectable at scale
- **Classification:** Each corpus needs domain routing for retrieval
- **Range:** Individual datasets span 8,000 to 28.7 million vectors

Standard manual inspection does not work at this scale. Required automated verification with per-dataset integrity checks across every index, every alignment, and every domain classification.

Approach

Vector-Chunk-Metadata Alignment

Every dataset enforces 1:1:1 correspondence — each vector maps to exactly one text chunk and one metadata record. Any drift triggers automatic rebuild.

Domain Classification

Machine-readable domain tagging per dataset (financial, legal, cybersecurity, etc.) so the retrieval layer routes queries to the correct corpus automatically.

FAISS Index Verification

On-disk presence of FAISS index files confirmed per dataset. 1024-dimensional embeddings (intfloat/e5-large-v2). No valid index means no retrieval.

Multi-Phase Remediation

Automated repair passes applied domain declarations to legacy datasets, backfilled missing metadata, and purged 18 superseded duplicates.

Single-Source Registry

One canonical registry serves as the authoritative record for all datasets, produced through multi-phase reconciliation across both storage volumes.

Automated Compliance Checks

Eight verification checks per dataset: modality, alignment, embedding format, sharding, metadata completeness, domain classification, lifecycle status, and provenance.

Domain Coverage

Pipeline Tested Across 8 Domains

Domain	Datasets	Vectors	% of Total
General Knowledge	131	67,171,348	43.1%
Forum & Q/A	37	64,704,693	41.6%
Cybersecurity	19	737,895	0.5%
Legal & Regulatory	18	2,733,268	1.8%
Code & Development	12	3,820,942	2.5%
Financial & Economic	11	9,725,910	6.2%
Academic & Scientific	9	6,708,062	4.3%
Technical Documentation	7	104,386	0.1%
TOTAL	**244**	**155,706,504**	**100.0%**

226
RETRIEVAL-READY

244
DOMAIN-CLASSIFIED

8
KNOWLEDGE DOMAINS

Scale Demonstration

Largest Corpora Processed

#	Corpus	Vectors	Domain
1	harvard_cold_cases_v2	28,737,085	General Knowledge
2	stackoverflow_extracted	22,619,752	Forum & Q/A
3	pushshift_reddit_2013	17,361,608	Forum & Q/A
4	pushshift_reddit_2012	11,623,272	Forum & Q/A
5	wiki_v4236	11,142,472	General Knowledge
6	stackxchange	9,776,839	Forum & Q/A
7	open_orca_complete	6,933,169	General Knowledge
8	AMPS	5,369,384	General Knowledge
9	wikidata_5m_v4236	4,815,483	General Knowledge
10	yahoo_finance_v4235	3,685,870	Financial & Economic
11	bwzheng2010yahoo-finance-data_v4235	3,674,507	Financial & Economic
12	arxiv_physics	3,317,045	Academic & Scientific
13	cord19_v423	2,891,452	Academic & Scientific
14	codesearchnet	2,552,516	Code & Development
15	relbert_t-rex	1,616,065	General Knowledge

Outcome

PIPELINE STATUS

✓ Verified — Production-Ready

Pipeline Validation

- **155.7M vectors** — No data loss or silent failures
- **244 indices verified** — On-disk FAISS integrity confirmed
- **226 retrieval-ready** — Serving live inference queries
- **1:1:1 alignment** — Vector, chunk, and metadata in sync

Scale Parameters

- **38 TB** — Total across two dedicated storage volumes
- **8 domains** — Legal, financial, scientific, cybersecurity, and more
- **1024-dim** — intfloat/e5-large-v2 embeddings
- **FAISS IndexFlatIP** — Per-dataset index architecture

What This Means For Your Project

- ✓ **Your data** gets processed by the same pipeline validated across 155.7M vectors and 244 datasets with zero silent failures.
- ✓ **Domain flexibility** — legal, financial, technical, or custom corpora. The pipeline has handled comparable data at scale.
- ✓ **Verification artifacts** ship with every engagement — alignment reports, integrity checks, and domain classification.

Scale parameters demonstrated here. Engagement-specific thresholds calibrated to client requirements.