RAG Readiness Audit

Sample Report

OVERALL STATUS

✓ RAG-READY (with noted limitations)

Audit Type: Pre-Deployment RAG Pipeline Assessment
Embedding Model: Configured per project (1024 dimensions)
Index Type: FAISS IndexFlatIP

This audit evaluates the semantic indexing pipeline and resulting vector stores for RAG readiness. Assessment covers data cleanliness, chunking quality, metadata completeness, and embedding suitability.

1. Data Cleanliness

Criterion	Status	Notes
Duplicate records removed	✓ PASS	Deduplication confirmed via hash comparison
Encoding issues resolved	✓ PASS	UTF-8 normalized; no orphan byte sequences
Empty/null entries filtered	✓ PASS	Zero-length chunks excluded from index
Whitespace normalized	✓ PASS	Leading/trailing whitespace stripped
Control characters removed	✓ PASS	Non-printable characters sanitized

Findings

Dataset A (Legal Corpus): Minor duplicates identified and removed. Final deduplication rate: <0.1%.

Dataset B (Mathematical Reasoning): Clean source data; no duplicates detected post-processing.

Dataset C (Mixed Prose Archive): Entries flagged for excessive whitespace; remediated during normalization pass.

2. Chunking Quality Assessment

Criterion	Status	Notes
Semantic boundary adherence	⚠ REVIEW	See findings below
Chunk size consistency	✓ PASS	Target: 512 tokens; variance within 15%
Overlap implementation	✓ PASS	64-token overlap between sequential chunks
Context preservation	✓ PASS	Source identifiers retained in chunk metadata

Findings & Remediation

Risk: Dataset A contains chunks where legal citations span boundaries — may degrade retrieval precision.
Remediation: Implement citation-aware chunking treating statutory references as atomic units.

Observation: Dataset B uses Q&A pairs as natural boundaries — optimal for QA retrieval.

Risk: Dataset C has chunks truncated mid-sentence.
Remediation: Add sentence-boundary detection before token limits.

3. Metadata Completeness

Criterion	Status	Notes
Unique chunk identifiers	✓ PASS	Sequential IDs assigned
Source document reference	✓ PASS	Original paths preserved
Chunk position tracking	⚠ PARTIAL	Page numbers missing
Timestamp metadata	✗ MISSING	No document timestamps
Category/domain labels	✓ PASS	Domain prefixes embedded

Current Schema

```
{
  "id": "integer",
  "text": "string (preview)",
  "source_path": "string",
  "domain_prefix": "string"
}
```

Gap: Missing timestamps prevent time-based filtering.
Gap: No PDF page mapping limits citation accuracy.

4. Embedding Model & 5. Index Integrity

Embedding Model Suitability

Criterion	Status
Model appropriate for domain	✓ PASS
Dimension alignment	✓ PASS
Context prefix application	✓ PASS
Query prefix documented	✓ PASS

Note: For domain-specific corpora, fine-tuned models may yield 5-15% precision improvement.

Index Integrity

Criterion	Status
Vector-chunk count match	✓ PASS
Index file integrity	✓ PASS
Shard alignment	✓ PASS
Backup artifacts	✓ PASS

Verification Results

Dataset	Chunks	Status
Legal Corpus	~14M	GREEN
Math Reasoning	~9K	GREEN
Mixed Prose	~1.2M	GREEN

6. Remediation Summary & 7. Certification

Remediation Priorities

Priority	Issue	Recommendation
HIGH	Citation boundaries	Entity-aware chunking
MEDIUM	Missing timestamps	Extend schema
MEDIUM	Mid-sentence truncation	Sentence-boundary detection
LOW	Page number mapping	Track PDF offsets

Certification

FINAL STATUS

✓ CONDITIONALLY APPROVED

Caveats:

→ Citation-heavy queries may have reduced precision until chunking remediation

→ Time-filtered retrieval not currently supported

This is a sanitized sample. Findings, thresholds, and schemas are adapted per dataset.