# RAG Readiness Audit

Semantic Indexing Review

ENGAGEMENT OUTCOME

## ✓ Approved for Deployment

**Dataset:** Mathematical reasoning corpus — 8,792 vectors

**Scope:** Data quality, chunking integrity, metadata, embeddings

**Result:** Clear go/no-go assessment with documented rationale

Portfolio Validation Artifact — Pre-Deployment Assessment

# Overview & Challenge

This audit evaluated a mathematical reasoning dataset prior to RAG integration, assessing data quality, chunking integrity, metadata completeness, and embedding suitability.

## Client Challenge

The client maintained a curated dataset of mathematical word problems with step-by-step solutions. Before deployment, they required independent verification that the data would support reliable retrieval.

**Key Concerns:**

→ Unknown duplicate rates affecting retrieval diversity

→ Potential mid-solution chunk breaks disrupting answer completeness

→ Absence of structured metadata for filtering and attribution

→ No prior validation against retrieval-specific quality thresholds

# Audit Scope

The assessment covered five dimensions critical to RAG performance:

**1. Data Cleanliness**
Encoding consistency, whitespace normalization, duplicate detection

**2. Chunking Quality**
Chunk boundaries relative to semantic units — questions and solutions intact

**3. Metadata Completeness**
Schema review: identifiers, source attribution, domain labels, positioning

**4. Embedding Suitability**
Model alignment with retrieval requirements, prefix conventions, dimensionality

**5. Index Integrity**
Vector counts match chunk counts, index loads without corruption

# Execution Summary

## Data Inventory

Catalogued **8,792** complete question-answer pairs with step-by-step solutions.

## Cleanliness Assessment

| Check | Result |
|---|---|
| Encoding | UTF-8 normalized |
| Whitespace | Consistent formatting |
| Duplicates | Zero detected |

## Chunking Analysis

Natural question-answer boundaries used as chunk divisions. Each chunk is a complete, self-contained reasoning unit. **No mid-solution breaks identified.**

## Embedding Verification

| Parameter | Value |
|---|---|
| Dimensions | 1024 |
| Prefix | Applied per spec |
| Alignment | 100% match |

# Findings

## Strengths

**Zero duplicate contamination**
No redundant entries that would bias retrieval

**Intact semantic units**
Question-answer pairs remain whole within chunks

**Consistent formatting**
No encoding anomalies or whitespace issues

**Verified alignment**
1:1:1 correspondence: chunks, metadata, vectors

## Gaps Identified

| Priority | Issue | Impact |
|---|---|---|
| LOW | Missing timestamps | Cannot filter by recency |
| LOW | No source attribution | Limited provenance |

## Risk Assessment

No high-priority risks identified. Dataset exhibits characteristics well-suited for retrieval applications.

# Recommendation & Deliverables

FINAL STATUS

**APPROVED FOR DEPLOYMENT**

The mathematical reasoning corpus is ready for RAG integration without remediation. Validation confirms data integrity across all critical dimensions.

**Optional enhancements:**

→ Extend metadata schema with temporal fields

→ Add source attribution for audit trail

## Deliverables Provided

| # | Artifact |
|---|---|
| 1 | Audit Report |
| 2 | Validation Summary |
| 3 | Gap Register |
| 4 | Remediation Notes |

## Outcome

Client received clear go/no-go assessment with documented rationale, allowing integration to proceed with confidence.

Dataset characteristics and thresholds are engagement-specific.