Team Member 1		Team Member 2	
Name:	[Shubham Raut]	**Name:**	[Saurav Deshmukh]
Roll Number:	[2024201019]	**Roll Number:**	[2024201070]

Team Name: [SignalOps]
Domain: K-12 Education - Geometry (Grades 6-10)
Date: 25th October 2025

Subject Matter Expert AI Agent

Mid-Evaluation Report

K-12 Geometry Education (Grades 6-10)

October 25, 2025

Domain: K-12 Education - Geometry (Shapes, Angles, Theorems)

1 Executive Summary

This report presents the mid-evaluation deliverables for the Geometry Subject Matter Expert (SME) AI system, covering data collection, preprocessing, chunking, embedding generation, and vector database indexing as required up to **Section C** of the project requirements.

Key Achievements

- **19 documents** processed from authoritative sources
- **∼81,500 words** covering Grades 6-10 curriculum
- **3-level hierarchical chunking** (2048, 512, 128 tokens)
- **Vector database indexed** with 768-dimensional embeddings
- **Bonus features**: Reranking, hybrid search, caching

2 Data Collection & Organization

2.1 Data Sources

Our dataset comprises three primary source categories:

1. **NCERT Textbooks** (17 documents)
   - Official Indian school curriculum (Classes 6-10)
   - Covers: geometry fundamentals, shapes, angles, theorems, mensuration
   - Examples: `class6_9.pdf`, `class10_11.pdf`
2. **Comprehensive Textbook** (1 document)
   - *Geometry for Enjoyment and Challenge* (McDougal Littell)
   - 770 pages covering grades 6-12
   - Provides alternative pedagogical approaches and detailed proofs
3. **Presentation Materials** (1 document)
   - Visual learning content with structured explanations
   - Complements text-based materials

Metric	Value
Total Documents	19
Total Words	81,564
Total Characters	417,472
Estimated Chunks (after processing)	2,500+
File Formats	PDF, PPTX

Table 1: Dataset Overview

2.2 Data Statistics

2.3 Grade & Difficulty Distribution

Grade	Docs	%
Grade 6	5	26.3%
Grade 7	2	10.5%
Grade 8	1	5.3%
Grade 9	6	31.6%
Grade 10	4	21.1%
General	1	5.3%

Difficulty	Docs	%
Beginner	10	52.6%
Intermediate	7	36.8%
Advanced	2	10.5%

Table 2: Distribution by Grade and Difficulty

2.4 Justification

Quality over Quantity: We prioritized authoritative, curriculum-aligned sources over volume. NCERT textbooks ensure relevance to Indian education standards, while the comprehensive textbook provides international perspectives and rigorous proofs.

Comprehensive Coverage: The dataset covers all major geometry topics for grades 6-10: basic shapes, angles, triangles, quadrilaterals, circles, mensuration, coordinate geometry, and theorems.

3 Preprocessing & Chunking Strategy

3.1 Chosen Strategy: Hierarchical Chunking

We implemented a **3-level hierarchical chunking** approach to balance precision and context:

Level	Size	Purpose
0	2048 tokens	Broad context, full sections
1	512 tokens	Medium detail, subsections
2	128 tokens	Fine-grained, specific concepts

Table 3: Hierarchical Chunk Configuration

3.2 Rationale

- **Precision vs. Context Trade-off:** Fine-grained chunks (L2) enable precise retrieval, while coarse chunks (L0) provide surrounding context

- **Parent-Child Relationships:** Maintains document structure for context expansion
- **Overlap Strategy:** 20-token overlap ensures semantic continuity across chunks

3.3 Preprocessing Steps

1. **Text Extraction:** Format-specific extraction (PDF, DOCX, PPTX)
2. **Cleaning:** Remove excessive whitespace, normalize Unicode
3. **Grade Detection:**
   - Filename parsing (`class6_9.pdf` → Grade 6)
   - Content-based classification using keyword matching
4. **Metadata Extraction:** Topics, difficulty, formulas, theorems
5. **Deduplication:** Hash-based duplicate detection

4 Embedding & Indexing

4.1 Embedding Model

- **Model:** `sentence-transformers/all-mpnet-base-v2`
- **Dimensions:** 768
- **Rationale:** Top performance on MTEB benchmark, balanced speed/quality
- **Generation Time:** ∼50ms per embedding (with caching)

4.2 Vector Database

- **Database:** Elasticsearch 8.11.0
- **Index Name:** `geometry_k12_rag`
- **Features:**
  - Dense vector search (cosine similarity)
  - BM25 keyword search
  - Parent-child relationship storage
  - Metadata filtering (grade, difficulty, topics)

4.3 Indexing Statistics

Metric	Value
Total Chunks Indexed	∼2,500
Level 0 Chunks	∼400
Level 1 Chunks	∼800
Level 2 Chunks	∼1,300
Chunks with Theorems	∼600
Chunks with Formulas	∼450
Index Size	∼150 MB

Table 4: Indexed Data Statistics

5 Bonus Features Implemented

Beyond the required components, we implemented:

1. **BGE Reranker:** CrossEncoder-based reranking for improved relevance (15-30% improvement)

2. **Hybrid Search:** Combines vector and BM25 search using Reciprocal Rank Fusion
3. **Redis Caching:** Embedding cache reduces query time from 50ms to ¡10ms
4. **Hierarchical Retrieval:** Multi-level retrieval with context expansion

6 System Architecture

Data Flow:

Raw Documents → Text Extraction → Preprocessing →
Hierarchical Chunking → Embedding Generation →
Elasticsearch Indexing → RAG Pipeline → Query Results

Figure 1: System Architecture Overview

Key Components:

- **Document Processor:** Multi-format support, grade detection, metadata extraction
- **Chunk Manager:** 3-level hierarchical chunking with overlap
- **Elasticsearch Client:** Vector + keyword indexing, parent-child relationships
- **RAG Pipeline:** Hybrid retrieval, reranking, context assembly

7 Sample Query Results

7.1 Example 1: Pythagorean Theorem

Query: "What is the Pythagorean theorem?"
Results: 5 relevant chunks retrieved
Top Score: 0.8734
Grade Level: Middle School (6-8), High School (9-10)
Sources: `class10_6.pdf`, `class9_7.pdf`

7.2 Example 2: Triangle Properties

Query: "Explain properties of isosceles triangles"
Results: 7 relevant chunks retrieved
Top Score: 0.8521
Grade Level: Grade 7, Grade 9
Sources: `class7_5.pdf`, `class9_6.pdf`

8 Design Decisions

8.1 Why Hierarchical Chunking?

Problem: Fixed-size chunks lose context (too small) or miss details (too large).
Solution: 3-level hierarchy captures both broad context and fine details.
Benefit: Retrieve precise answers at L2, expand to L1/L0 for context when needed.

8.2 Why Hybrid Search?

Problem: Vector search misses exact terms; keyword search misses semantics.
Solution: Combine both using Reciprocal Rank Fusion (RRF).
Benefit: Best of both worlds—semantic understanding + exact matching.

9 Challenges & Solutions

1. **Challenge:** Grade detection for non-NCERT documents
   - **Solution:** Content-based classification using grade-specific keywords
2. **Challenge:** Large textbook (770 pages) covers multiple grades
   - **Solution:** Chapter-wise grade mapping based on topic complexity
3. **Challenge:** Maintaining context in small chunks
   - **Solution:** Parent-child relationships + context expansion

10 Verification & Testing

Tests Completed:
- Document processing pipeline validation
- Chunk creation and hierarchy verification
- Elasticsearch indexing confirmation
- Sample query retrieval testing
- Grade-specific filtering validation

Test Results: All core functionality operational. System successfully retrieves grade-appropriate, relevant content for geometry queries.

11 Conclusion

We have successfully completed all required components for mid-evaluation:
Section A: Document collection from authoritative sources
Section B: Hierarchical preprocessing and chunking
Section C: Embedding generation and vector database indexing
Bonus: Reranking, hybrid search, and caching

The system provides a robust foundation for the Geometry SME agent, with comprehensive coverage of grades 6-10 curriculum, intelligent chunking, and efficient retrieval mechanisms.

Next Steps (Post Mid-Evaluation):
- Phase 3: LLM integration for answer generation
- Phase 4: API server and microservices
- Phase 5: Tool integration (PDF/email generation)
- Phase 6: Fine-tuning and evaluation

Repository & Code

GitHub Repository: [GitHub Repository Name]

Dataset (Raw + Processed): Google Drive Folder Link

Key Files:

- src/data_preparation/document_processor.py
- src/data_preparation/chunk_manager.py
- src/database/elasticsearch_client.py
- scripts/build_database.py