DATA RESEARCH

1) Understanding Data

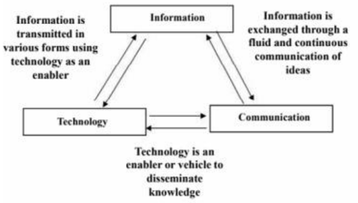
After going through the data , I have figured that the text in dataset has

➔ Long Paragraphs

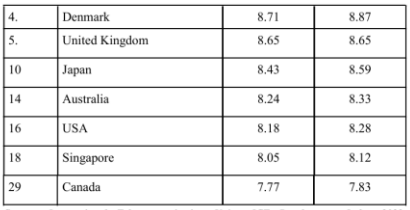
so we need embeddings that handle context and preserve semantic meaning.

➔ Content is Academic , not scientific (no maths , programming , etc)
Content includes Theories , communication models , logic etc
It is conceptual + theoretical
so , we need to use embeddings that are good fot conceptual semantic similarity.

➔ Technical Diagrams


Information is transmitted in various forms using technology as an enabler
Information
Information is exchanged through a fluid and continuous communication of ideas
Technology
Communication
Technology is an enabler or vehicle to disseminate knowledge

➔ Tables with Numeric Values



4.	Denmark	8.71	8.87
5.	United Kingdom	8.65	8.65
10	Japan	8.43	8.59
14	Australia	8.24	8.33
16	USA	8.18	8.28
18	Singapore	8.05	8.12
29	Canada	7.77	7.83

Source: International Telecommunication Union ICT Development Index 2023 (https://www.itu.int/itu-d/reports/statistics/IDI2023/)

require embeddings that interpret text + numbers

➔ Bullet Points, Lists & Structured Notes

➔ Learning outcomes (short sentences)
➔ Characteristics (medium paragraphs)
➔ Advantages & Disadvantages (bullet lists)
➔ Frameworks (structured lists + definitions)
➔ Check Your Progress (MCQs, exercises → long text)

Your dataset = mixed-length, multi-structure text.

So, the embedding model must:

1. Work well with short definitions
2. Work well with long educational content
3. Handle lists, tables, MCQs, structured text
4. Stay consistent across varied chunk sizes

In short, book text requires both short-text accuracy + long-text semantic accuracy.

EMBEDDING MODEL

bge-m3 (Best)

bge-m3 is explicitly designed for **multi-length (multi-granularity) embeddings**, so it preserves meaning equally well for:

- 1-line definitions
- Bullet points
- 1–2 page explanations

This is exactly what your educational data needs.

bge-m3 offers 3 modes:

Mode	**Use Case in Your Dataset**
Dense	Long paragraphs, conceptual explanations
Sparse (BM25-like)	Keywords, MCQs, lists, exact matches
Multi-vector	Frameworks, multi-part definitions, complex theories

Stable Across Chunk Sizes (Very Important)

Your PDFs will be chunked into:

- 100–250 token chunks for definitions
- 300–700 token chunks for theories
- 700–1500 token chunks for long explanations

Most embedding models degrade when chunk sizes vary.
Typical Embedding Models (SBERT, GTE, E5, Instructor-XL)

Short chunk and **long chunk** of the same concept drift apart.
In bge-m3 Short and long chunks are aligned in the same meaning cluster

bge-m3 stays consistent:

- Similar meaning is preserved
- Chunk-to-chunk similarity remains stable
- No loss in retrieval quality

This is essential for building a reliable retrieval pipeline.

It is also open source + Fast

SBERT / MiniLM

Strengths

Classical, reliable, widely used.

Good for tweets , emails , short general text

Very fast

Weakness

Extremely weak with:

- long text
- conceptual academic content

Severe chunk-size sensitivity.

Outdated compared to modern models

Cannot understand numerical tables properly

No multi-vector or sparse support

➔ Textbooks contain long, conceptual passages → SBERT will fail

BGE (older)

Industry use: general semantic search

Strengths

- Strong conceptual similarity
- Good dense embeddings

Weaknesses

- No multi-vector
- No sparse fusion
- Not multi-granular
- Cannot handle chunk mismatch well

VECTOR DATABASES

Qdrant → BEST for chunk variability

HNSW preserves distances even when embeddings come from different chunk sizes.

Open-Source LLM

Qwen2.5 (Open Source Local Model)

It has state-of-the-art accuracy in:

- reading comprehension
- educational QA
- conceptual reasoning
- long-text summarization
- Strong at MCQs, quizzes, and structured content