Data for RAG Systems

What Works, What Fails, and Why

Retrieval-Augmented Generation (RAG)

Why Data Matters More Than the Model

In RAG systems:

The model is useless without good data

- LLM does not “know” your documents
- Retrieval quality depends entirely on data quality
- Bad data → confident but wrong answers

What Is “Good Data” for RAG?

Good RAG data is:

- Text-heavy
- Knowledge-based
- Relatively stable
- Written for humans

What we need?
Information someone would normally read

Data Types That Work Best

Ideal Sources

- Policies & regulations
- Manuals & handbooks
- FAQs
- Legal documents
- Research papers
- Medical guidelines

These contain **explanations**, **definitions**, and **rules**.

Example: Good RAG Data

University Handbook

- Admission rules
- Grading policies
- Attendance requirements

Legal Policy Document

- Eligibility criteria
- Rights and responsibilities
- Procedures

Perfect Questions like:
“Who is eligible for X?”
“What happens if Y?”

What Data Is BAD for RAG

1. Tables Without Structure

- Raw tables
- Scanned tables
- CSV dumps with no explanation

Why?

- No semantic meaning
- Poor chunking
- Weak retrieval

Example: Bad Table Data

ID | Code | Value | Flag
12 | A32 | 0.87 | 1

LLM question:
“What does A32 mean?”

No explanation → no useful answer

What Data Is BAD for RAG

2. Constantly Changing Transactional Data

- Bank transactions
- Live sensor data
- Stock prices
- Attendance logs

Why?

- Needs real-time systems
- RAG is **static knowledge retrieval**
- Leads to outdated answers

Rule of Thumb

If data changes:

- every second
- every minute
- every hour

Don’t use RAG

Mixed Data: Be Careful

Some datasets contain:

- Text + tables
- Explanations + numbers

Keep:

- Explanatory text

Remove or summarize:

- Raw numeric dumps

The Role of Data Cleaning

Before RAG, documents often contain:

- Headers & footers
- Page numbers
- Repeated titles
- OCR artifacts
- Watermarks

These **pollute embeddings**.

Common Cleaning Issues

Example OCR Noise

G0VERNMENT 0F NEPAL
MINISTRY 0F H0ME AFFAIRS

Vectors become meaningless
Retrieval quality drops

What Should Be Removed?

Remove or clean:

- Page numbers
- Repeated headers/footers
- Table of contents
- References (optional)
- Scanned artifacts

Goal:
Only meaningful content remains

What Should Be Preserved?

Keep:

- Headings
- Section titles
- Definitions
- Lists and clauses
- Paragraph structure

These improve:

- Chunking
- Retrieval
- Answer quality

Data Cleaning Pipeline (Conceptual)

Raw Document
↓
Text Extraction
↓
Noise Removal
↓
Clean Text
↓
Chunking

Cleaning happens **before** embeddings.

Why Cleaning Affects Retrieval

Embeddings treat all text as important.

If noise exists:

- Noise gets embedded
- Noise gets retrieved
- LLM uses noise as context

Garbage in → garbage out

Practical Guidelines

Before using any dataset for RAG, ask:

- Is this text meant to be read by humans?
- Does it explain something?
- Will this still be valid next month?
- Can a paragraph answer a question?

If **yes** → good RAG data.

Key Takeaways

- RAG is **data-first**, not model-first
- Good documents matter more than fancy LLMs
- Cleaning improves retrieval more than tuning prompts
- Not all data belongs in a RAG system