Dataset Exploration

Dataset Overview

The RAG Mini Wikipedia dataset comprises 3,200 passages with the following distribution:

- Short passages (1–253 characters): 45.3%
- Medium passages (253–505 characters): 23.8%
- Long passages (505–757 characters): 16.5%
- Long tail: Passages extending beyond 2,500 characters

Sample Content Examples-

- **Example 1 (Short, 246 chars):** "Uruguay (official full name: Eastern Republic of Uruguay) is a country located in the southeastern part of South America. It is home to 3.3 million people, of which 1.7 million live in the capital Montevideo and its metropolitan area."
- **Example 2 (Medium, 312 chars):** "Montevideo was founded by the Spanish in the early 18th century as a military stronghold. Uruguay won its independence in 1828 following a three-way struggle between Spain, Argentina, and Brazil. It is a constitutional democracy, where the president fulfills the roles of both head of state and head of government."
- **Example 3 (Long, 592 chars):** "According to Freedom House, Uruguay ranked twenty-seventh in its 'Freedom in the World' index. According to the Economist Intelligence Unit, Uruguay scores 7.96 on the Democracy Index, the last position among the 28 Full Democracies. The report covers 60 indicators across five categories: free elections, civil liberties, functioning government, political participation, and political culture."

Data Quality Observations

- **Content accuracy:** Passages are factually reliable, including precise data points, dates, and statistics.
- **Information completeness:** Most passages provide comprehensive coverage. A few entries are split across consecutive passages; for example:
    - **ID 7:** “The name ‘Uruguay’ comes from Guaraní. It has many possible meanings. Some of the proposed meanings are:”
    - **ID 8:** “River of colorful or ‘painted’ chinchillas (birds)’: poetic interpretation attributed to Juan Zorrilla de San Martín.”

Infrastructure Components

Component / Development Environment	**Use Case**
Google Colab (Computational Platform)	Provides free cloud GPU/CPU resources to run notebooks and execute the RAG pipeline efficiently.
Google/flan-t5-small (LLM)	Answer generation in naive RAG system.
Google/FLAN-T5-base (LLM)	Answer generation in enhanced RAG system + query rewriting
Google Gemini 2.5 Pro	RAGAs evaluation judge LLM (assessing faithfulness, relevancy)
all-MiniLM-L6-v2	384-dimensional vector embeddings for semantic search
all-mpnet-base-v2	768-dimensional vector embeddings for semantic search
cross-encoder/ms-marco-MiniLM-L-6-v2	Reranking retrieved passages (enhanced system only)
Milvus Lite/FAISS	Vector storage and similarity search