Phase 1: Dataset Exploration Report

This assignment employs the RAG Mini Wikipedia dataset from Hugging Face, comprising two splits: a text corpus for retrieval (3,200 passages) and a question-answer set for evaluation (918 pairs). Both splits loaded successfully in Google Colab using the datasets library.

Text Corpus Analysis: The corpus contains passages indexed by unique IDs with two fields: passage and id. Quality validation confirmed complete data integrity—no missing values or duplicates. Passage lengths exhibit substantial variability (1-2,515 characters; mean: 390, median: 299), indicating a balanced mix of concise factual statements and detailed descriptive content suitable for embedding-based retrieval.

QA Dataset Analysis: The evaluation set contains 918 question-answer with three fields: question, answer, and id. Data quality checks confirmed completeness and uniqueness. Questions average 53 characters while answers average 19 characters, with some extending beyond 400 characters. This distribution encompasses both simple binary queries and complex descriptive answers, enabling comprehensive evaluation of retrieval precision and generation quality.

Infrastructure: Development occurs in Google Colab with minimal dependencies (datasets, pandas, numpy). A local directory structure (assignment2-rag) organizes processed data and documentation, version-controlled via GitHub for reproducibility. Subsequent phases will implement embeddings using sentence-transformers' all-MiniLM-L6-v2 model with FAISS vector storage. The QA subset serves as the benchmark for systematic evaluation across retrieval configurations and prompting strategies.