1 Extension 1: Retrieval-Augmented Generation (RAG)

1.1 Overview

Extension 1 implements a Retrieval-Augmented Generation (RAG) system that combines our Milestone 2 baselines into a single, document-grounded question answering pipeline. The system retrieves relevant documents using BM25, generates answers using an LLM with document context, and extracts sentence-level citations.

1.2 Implementation

We combined components from both baselines: BM25 retrieval (simple baseline) and LLM answer generation (strong baseline). The key innovation is passing the top-ranked retrieved document directly to the LLM as context, creating a document-grounded prompt that instructs the model to answer only from the provided document and refuse when information is insufficient.

The system works as follows: (1) BM25 retrieves top-5 documents, (2) the top-ranked document is truncated and passed to Groq's `llama-3.1-8b-instant` with a document-grounded prompt, (3) the LLM generates a 2–4 sentence answer, and (4) word-overlap matching extracts top-3 evidence sentences for citations.

The implementation reuses document loading and BM25 retrieval from the simple baseline, and adapts the LLM generation approach from the strong baseline by including retrieved document context in the prompt. Citation extraction uses the same word-overlap method as the simple baseline.

1.3 Evaluation and Results

We evaluated Extension 1 on the train set (24 questions) using the same metrics as Milestone 2. Since Extension 1 combines both baselines, we compare it to each baseline on the metrics relevant to that baseline's purpose.

1.3.1 Comparison with Simple Baseline (Retrieval + Citations)

The simple baseline focused on retrieval and citation extraction, not LLM generation. Extension 1 maintains identical performance on these metrics:

- **Retrieval:** Recall@1: 58.33%, Recall@5: 66.67% — identical (same BM25 retrieval method)
- **Citations:** F1: 0.2224 — identical (same word-overlap extraction method)

This demonstrates that adding LLM generation does not degrade retrieval or citation performance.

1.3.2 Comparison with Strong Baseline (LLM Generation)

The strong baseline focused on answer quality via LLM generation without retrieval or citations. Extension 1 significantly improves on these metrics:

- **Answer Quality:** 3.25/5 — improved from 2.58/5 (+26%), showing the benefit of document grounding
- **Evidence Score:** 0.3230 — improved from 0.0833 (+288%), showing better grounding in sources
- **Combined Score ($\lambda = 0.5$):** 0.4865 — improved from 0.3000 (+62%)

These improvements show that document-grounded generation produces better answers than ungrounded LLM generation. Extension 1 successfully combines the strengths of both baselines: maintaining the retrieval and citation capabilities of the simple baseline while achieving better answer quality than the strong baseline through document grounding.

However, citation F1 (0.2224) remains low, indicating room for better sentence-level citation extraction methods (planned for Extension 2). Also, retrieval misses 41.67% of questions at Recall@1, suggesting potential to improve retrieval through semantic search or reranking.

1.4 Usage

Extension 1 is implemented in `extension1.py`. To generate predictions, run `python extension1.py [split]` where `split` is `train`, `dev`, or `test` (defaults to `train`). To evaluate, run `python evaluate_extension1.py [split]`. Requires `GROQ_API_KEY` environment variable.