Milestone 2 Report

1 Task

For Milestone 2, we evaluate document-grounded QA systems that answer questions about news articles and CIA World Factbook entries. Each evaluation example consists of a question, a gold answer, a set of gold evidence sentences (sentence IDs from the source document), and a custom 1–5 rubric specifying what constitutes a good answer. The systems we evaluate must produce both an answer and a set of cited sentence IDs.

2 Simple Baseline: BM25 Retrieval

The simple baseline uses BM25Okapi for lexical document retrieval from a corpus of 1014 documents. Given a question, the system (1) ranks all documents by BM25 relevance score, (2) retrieves the top-ranked document, (3) extracts a text snippet (top-3 sentences by query word overlap) as the answer, and (4) identifies sentence IDs from the top document for citations.

This baseline demonstrates basic retrieval and citation capabilities without learned representations or language models.

2.1 Performance

Test set (3 questions): Recall@1: 33.33%, Recall@5: 66.67%, Citation Precision: 0.1111, Recall: 0.0256, F1: 0.0417.

Train set (24 questions): Recall@1: 58.33%, Recall@5: 66.67%, Citation Precision: 0.3030, Recall: 0.2307, F1: 0.2224.

The simple baseline achieves moderate retrieval performance but low citation precision due to the challenge of precisely identifying evidence sentences from retrieved documents.

3 Strong Baseline: LLM Query

The strong baseline uses llama-3.1-8b-instant via Groq API to generate answers directly from questions, without document retrieval. For each question, we send a simple instruction prompt to the model and treat its response as the answer. This baseline focuses solely on answer quality using a pre-trained language model, without attempting retrieval or citation.

This represents a standard LLM generation baseline from the literature and demonstrates the trade-off between answer quality and evidence grounding.

3.1 Performance

Test set (3 questions): Average Answer Score: 1.67/5, Evidence Score: 0.0000, Combined Score: 0.17.

Train set (24 questions): Average Answer Score: 2.58/5, Evidence Score: 0.0833, Combined Score: 0.30.

Since the strong baseline doesn't output evidence sentences, it scores 0.0 on evidence metrics (except for 2 train questions with no gold evidence, which score 1.0). The combined score reflects only answer quality (weighted by $\lambda = 0.5$), demonstrating that answer quality alone is insufficient without proper evidence grounding.

4 Evaluation Metrics

We evaluate three components: (1) retrieval accuracy (Recall@1, Recall@5), (2) citation correctness (Precision, Recall, F1 based on exact sentence ID matching), and (3) answer quality via LLM-as-judge plus evidence quality (word overlap) combined into a single score.

4.1 Retrieval Metrics

Recall@k measures the fraction of questions where the correct gold document appears in the top-k retrieved documents:

$$\text{Recall@k} = \frac{1}{|Q|} \sum_{q \in Q} \mathbb{1}[\text{gold\_doc}(q) \in \text{top\_k\_docs}(q)]$$

where Q is the set of questions and $\mathbb{1}[\cdot]$ is the indicator function. We report Recall@1 (correct document ranked first) and Recall@5 (correct document in top-5).

4.2 Citation Metrics

Citation metrics use exact sentence ID matching. For each question with gold evidence sentences:

$$\text{Precision} = \frac{|\text{gold} \cap \text{pred}|}{|\text{pred}|}, \quad \text{Recall} = \frac{|\text{gold} \cap \text{pred}|}{|\text{gold}|}, \quad \text{F1} = \frac{2PR}{P + R}$$

Dataset-level metrics average Precision, Recall, and F1 across all questions with evidence.

4.3 Answer Quality and Evidence

We use Groq `llama-3.1-8b-instant` as an automatic grader. For each example i, the grader sees the question, gold answer, rubric description, and system answer, then outputs a discrete score $A_i \in \{1, 2, 3, 4, 5\}$, normalized to $a_i = A_i/5$.

Evidence quality uses word overlap: $e_i = \frac{|\text{words}(E_i) \cap \text{words}(\hat{E}_i)|}{|\text{words}(E_i)|}$, where sentences are loaded from document JSON and words are extracted (lowercase, tokenized). Special case: if no gold evidence, score is 1.0 if no predictions, else 0.0.

The combined metric is: $c_i = \lambda \cdot a_i + (1 - \lambda) \cdot e_i$ with $\lambda = 0.5$ (equal weight). Dataset-level: we report mean raw rubric score $\overline{A}$, mean normalized answer $\overline{a}$, mean evidence $\overline{e}$, and mean combined $\overline{c}$.

5 Implementation

The evaluation script (`evaluation.py`) reads questions from a JSONL file, loads predictions from baseline output JSON files, computes all metrics (retrieval, citation, LLM-as-Judge), and saves results to `{split}_evaluation.json`. The simple baseline (`simple-baseline.py`) implements BM25 retrieval and sentence ID extraction. The strong baseline (`strong-baseline.py`) implements LLM querying via Groq API. Usage and file format details are documented in `evaluation.md`, `simple-baseline.md`, and `strong-baseline.md`.