A Grounded Q&A Assistant for Intelligence Analysts

CIS 5300 Final Project

Vincent Lin, Risha Kumar, Chih Yu Tsai, Asher Ellis

December 18, 2025

The Problem

Intelligence analysts must sift through many documents to find information

Today’s options:

- Manual reading: Accurate but slow, tedious
- LLM summaries: Fast but may hallucinate

What's missing: Fast, trustworthy answers with clear provenance

Goal: Build system that retrieves docs, generates answers, cites sources

Sample Q&A

QUESTION:

According to the article, what specific impacts did Hurricane Maria have on Dominica in 2017, and which sectors or infrastructure were most heavily damaged?

SYSTEM ANSWER:

Hurricane Maria passed over the island in 2017, causing extensive damage to structures, roads, communications, and the power supply. The hurricane also largely destroyed critical agricultural areas.

SOURCE: CIA World Factbook - Dominica

CITED EVIDENCE:

[S7] *"Hurricane Maria passed over the island in September 2017 causing extensive damage to structures, roads, communications, and the electrical supply, and largely combating critical agricultural areas."*

✓

PERFECT MATCH

Successful example: correct document retrieved, accurate answer, correct sentence cited

Dataset

Document Corpus (787 docs)

CNN/DailyMail articles (300)

Russia-Ukraine conflict reporting (217)

CIA World Factbook entries (257)

Documents stored with sentence boundaries

Sentence IDs: S1, S2, S3...

QA Dataset (30 questions)

24 train / 3 dev / 3 test

Each entry includes:

- Question text
- Gold answer
- Source document ID
- Evidence sentence IDs

Edge cases: unanswerable questions, numeric precision queries

Evaluation Metrics

Retrieval & Citation

Recall@1: Correct doc ranked first?

Recall@5: Correct doc in top 5?

Citation Precision: Fraction of cited sentences that are correct

Citation Recall: Fraction of gold sentences that were cited

Citation F1: Harmonic mean

Answer Quality

LLM-as-Judge: Llama 3.1 8B scores answers 1-5 using rubric

Evidence Score: Word overlap between gold and predicted citations (less strict then SID matching)

Combined Score:

$0.5 \times (\text{Answer}/5) + 0.5 \times \text{Evidence}$

Rewards both answer quality and proper grounding

Simple Baseline: BM25 Retrieval

Classic lexical retrieval using term frequency and document length

Implementation:

- Tokenization: NLTK word_tokenize + lowercase
- Ranking: BM25Okapi scores all 787 documents
- Retrieve top-5 documents

Results: Recall@1 = 54.2%, Recall@5 = 62.5%, Citation F1 = 0.241

Strong Baseline: LLM-Only

Direct LLM generation without document retrieval

Implementation:

- Model: Llama 3.1 8B via Groq API
- No retrieval, no document access, no citations

Results: Answer Score = 2.58/5, Combined = 0.30 (no evidence!)

Extension 1: Retrieval-Augmented Generation

Combine BM25 retrieval with LLM generation

Pipeline:

- BM25 retrieves top-5 documents
- Pass top document to LLM as context
- LLM generates 2-4 sentence answer
- Word-overlap extracts top-3 evidence sentences

Results: Answer improves 2.58 → 3.12 (+21%), Combined 0.30 → 0.48

Extension 2: Cross-Encoder Reranking

Problem: BM25 misses semantically similar content

Solution: Reranking with cross-encoder

Implementation:

- BM25 retrieves top-30 candidates (expanded from top-5)
- Cross-encoder (ms-marco-MiniLM-L-6-v2) scores each pair
- Re-sort by scores, select top-5

Results: Recall@1 jumps 54% → 79% (+46%), 6 more correct docs at rank 1

Extension 3: LLM-Based Citation Extraction

Problem: Word-overlap misses semantically similar matches

E.g., "casualties" in question vs "killed" in document

Solution: LLM identifies supporting sentences

- Prompt LLM with question, answer, and numbered sentences
- LLM returns sentence IDs that support the answer
- Semantic matching instead of lexical matching

Results: Citation F1 jumps 0.29 → 0.58 (+98%)

Extension 4: Hybrid Retrieval

Combine lexical and semantic signals

BM25 scores (lexical matching) + dense embeddings (all-MiniLM-L6-v2) for semantic similarity

Combined score: $\alpha \times \text{BM25} + (1-\alpha) \times \text{Semantic}$, with $\alpha = 0.5$

Results: Recall@1 hits 83.3%, Recall@5 hits 95.8% (best retrieval)

Extension 5: Document Chunking

Problem: Long documents dilute relevant content

Solution: Break documents into overlapping chunks

Implementation:

- 8-sentence windows with 2-sentence overlap
- Each chunk scored independently -> best chunk used for answer generation

Results: Recall@5 reaches 100% - correct doc always in top 5

Extension 6: Answer Verification

Problem: Generated answers may not be supported by retrieved doc

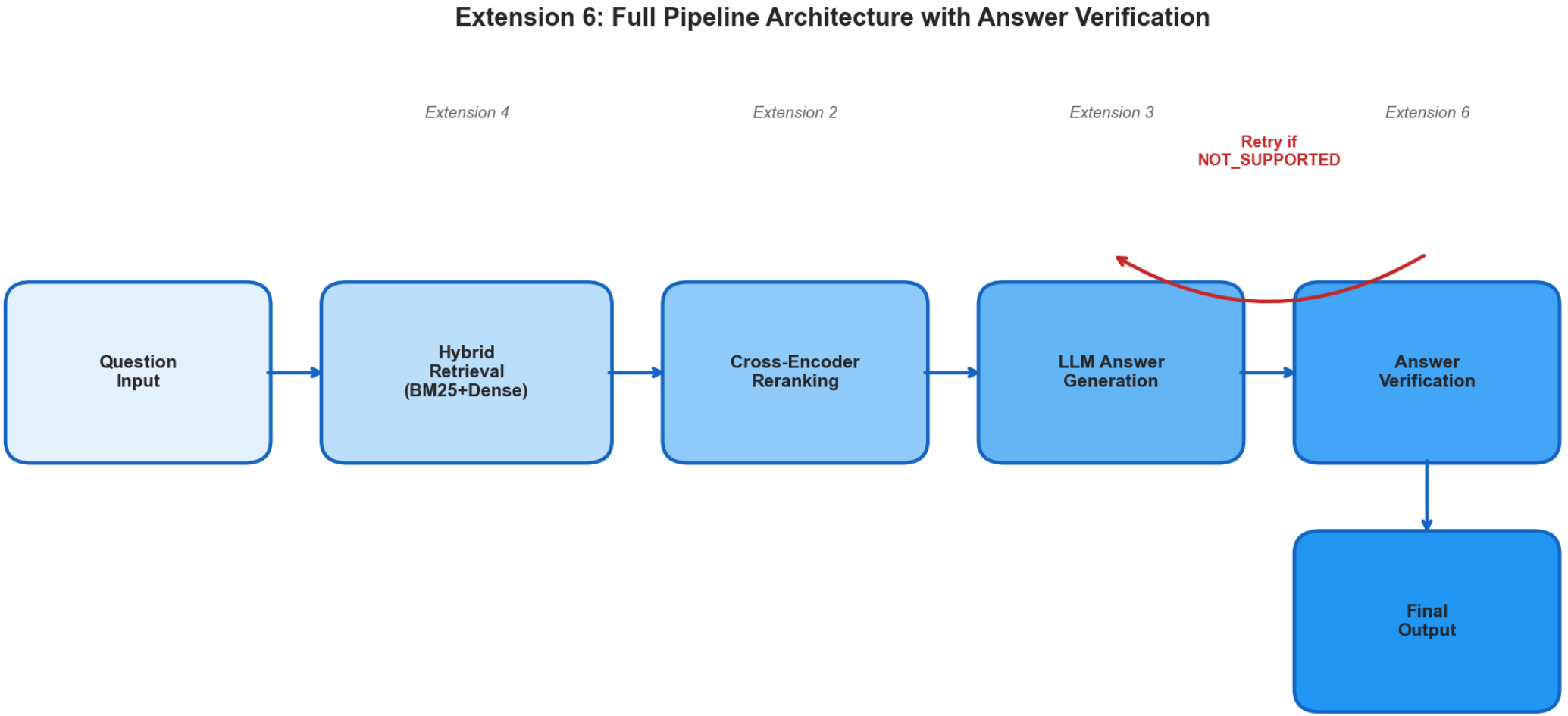
Solution: Verify answer against source, retry if needed

Implementation:

- Generate answer from top chunk
- LLM verifies: Is this answer supported by the document?
- If NOT_SUPPORTED: try next chunk (up to 3 attempts)
- Return best verified answer

Results: Best overall - Citation F1 = 0.63, Answer = 3.88/5, Combined = 0.71

Final Architecture



Extension 6 pipeline: Hybrid retrieval → Reranking → Generation → Verification with retry

Results: Train Set (24 questions)

Model	R@1	R@5	Cite F1	Answer	Combined
Simple (BM25)	54.2%	62.5%	0.241	--	--
Strong (LLM)	0%	0%	0%	2.58	0.30
Ext 1 (RAG)	54.2%	62.5%	0.241	3.12	0.48
Ext 2 (Rerank)	79.2%	91.7%	0.291	3.75	0.58
Ext 3 (LLM Cite)	79.2%	91.7%	0.576	3.83	0.70
Ext 4 (Hybrid)	83.3%	95.8%	0.528	3.79	0.65
Ext 5 (Chunk)	75.0%	100%	0.616	3.67	0.69
Ext 6 (Verify)	**75.0%**	**100%**	**0.632**	**3.88**	**0.71**

Key Improvements

Retrieval: 54% → 79% Recall@1 (+46% from reranking)

Citation: 0.24 → 0.63 F1 (+163% from LLM extraction)

Answer Quality: 2.58 → 3.88 (+50% from grounding)

Combined Score: 0.30 → 0.71 (+137%)

Grounding Analysis: Extension 6 achieves 100% answer grounding

- Every answer verified against source document
- Retry mechanism catches initially unsupported claims

Error Analysis

1. Retrieval Failures (6/24 questions)

   Similar documents confuse the retriever

   Example: Multiple Russia-Ukraine articles from same time period

2. Citation Misalignment
   - LLM cites semantically related but not gold sentences

3. Unanswerable Questions
   - System sometimes attempts answer when should refuse
   - Verification helps but doesn't fully solve this

Conclusions

What we built:

End-to-end document-grounded QA with sentence-level citations

Six extensions progressively improving each pipeline stage

Results (train set):

- Recall@1: 54% → 79% | Citation F1: 0.24 → 0.63
- Combined Score: 0.30 → 0.71

What we learned:

- Neural methods outperform lexical but are less interpretable
- "Hallucination-free RAG" is harder than it looks
- Each pipeline stage has failure modes that cascade

Questions?

github.com/asherellis/cis5300_project