

A Grounded Q&A Assistant for Intelligence Analysts

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

December 18, 2025

Abstract

We present a grounded question-answering system for intelligence analysis that retrieves relevant documents and generates answers with sentence-level citations. Starting from BM25 retrieval and LLM-only baselines, we develop six extensions: retrieval-augmented generation (RAG), cross-encoder reranking, LLM-based citation extraction, hybrid lexical-semantic retrieval, document chunking, and answer verification. On our 30-question benchmark (24 train, 6 dev/test) spanning CIA World Factbook entries, CNN articles, and Russia-Ukraine conflict reporting, our best system achieves 100% Recall@5, 0.63 citation F1, and 3.88/5 answer quality on the train set (vs. 2.58 for LLM-only). Each extension addresses a specific bottleneck, improving combined score from 0.30 to 0.71.

1 Introduction

Intelligence analysts often need to answer specific questions which require sifting through many documents such as government reports, court orders, policy briefs, and news articles. Today they face a trade-off: reading and extracting evidence manually is slow and error-prone, while querying large language models can be fast but may introduce hallucinated details or lack traceable sourcing. What’s missing is a middle ground: a system that produces concise answers while making it immediately clear which source text supports each claim. Our project tries to address this gap by building a grounded Q&A assistant that returns short, paraphrased answers with transparent citations to specific passages in the document collection.

This task involves several core tasks in NLP and computational linguistics: question answering, information retrieval, semantic matching, evidence grounding, and LLM evaluation. Unlike many Q&A settings that emphasize answer correctness alone, grounded Q&A emphasizes *attribution*. The system must be both correct and be able to provide verifiable evidence that supports each statement. This framing better matches real analyst workflows, where accountability and traceability are critical.

Question: According to the article, what specific political reforms did Sultan Qaboos enact in response to the 2011 demonstrations?

System Answer: In response to the 2011 demonstrations, Sultan QABOOS implemented economic and political reforms, including granting legislative powers to Oman’s bicameral legislature and authorizing direct elections for the lower house.

Citations: Sentences S8, S9 from CIA World Factbook – Oman (history section)

Figure 1: Example grounded Q&A output from our system. The answer paraphrases source content while citing specific sentences that support each claim.

Formal problem definition. Given a natural language question q about a specific intelligence topic and a corpus of documents D , the system must output (i) a short answer a consisting of 2-4 sentences and (ii) a set of citations C that map each claim in a to supporting evidence in D (down to the passage or sentence level). If the corpus does not contain sufficient supporting evidence, the system should return a refusal (e.g., “insufficient evidence”) or, when appropriate, surface multiple supported answers with notes indicating disagreement across sources.

We chose this task because it reflects a real reliability problem in applied NLP: users need answers that are both efficient and auditable. A grounded Q&A system addresses hallucination by requiring every claim to be traceable to source text.

2 Literature Review

Grounded question answering systems typically combine (i) a retriever that selects candidate evidence from a document collection and (ii) a generator that produces a natural-language response constrained by that evidence. Our project builds on three core research threads that motivate this design: retrieval-augmented generation, neural passage re-ranking, and strong sparse retrieval baselines.

Lewis et al. [2020] propose Retrieval-Augmented Generation (RAG), a hybrid model that couples a parametric generator with a non-parametric document memory accessed via dense retrieval. Their key motivation is that parametric-only models can be difficult to audit and update: retrieval provides an explicit, inspectable source of information that can be swapped or refreshed without retraining the entire generator. In their setup, a DPR-style bi-encoder retrieves top- k passages from a large Wikipedia-based corpus, and a BART generator conditions on retrieved evidence. They introduce two variants: RAG-Sequence, which conditions generation on a single selected document per output sequence, and RAG-Token, which can flexibly attend to different documents across tokens. Across multiple knowledge-intensive benchmarks (including open-domain QA and fact verification), RAG improves factuality and downstream performance relative to closed-book generation, supporting the view that retrieval can reduce hallucination by grounding outputs in external text.

While first-stage retrieval aims to ensure recall, high-precision ranking near the top is critical when downstream generation must cite a small number of passages. Nogueira and Cho [2019] adapt BERT as a cross-encoder passage re-ranker that jointly encodes the query and each candidate passage (as [CLS] query [SEP] passage [SEP]) and produces a single relevance score from the [CLS] representation. The model uses a simple pointwise training objective on MS MARCO candidates retrieved by BM25, with practical input constraints (e.g., truncating queries and fitting within BERT’s 512-token limit), making it well suited for reranking a modest top- k set. The reranker reports strong gains at top ranks and achieves state-of-the-art results on MS MARCO and TREC-CAR at the time, demonstrating that full cross-attention between query and passage is highly effective for identifying answer-bearing evidence. This directly motivates our use of reranking to reliably surface a small top- r evidence set for citation.

Finally, sparse lexical retrieval remains a dominant baseline due to its robustness, interpretability, and efficiency. Robertson and Zaragoza [2009] summarize the Probabilistic Relevance Framework (PRF) underlying BM25, which ranks documents by estimating relevance from term evidence using saturated term-frequency contributions and length normalization. They also discuss practical considerations such as parameterization and optimization, and extensions including BM25F for fielded documents (e.g., title vs. body) where different fields can be weighted by importance. For structured policy-style text, BM25 provides a strong, transparent first-stage retriever that complements dense methods. Together, these works support our overall framing: use a strong sparse (and optionally dense) retriever for recall, apply a cross-encoder reranker for top- r precision,

and generate short answers that remain grounded in explicitly retrieved evidence.

3 Experimental Design

3.1 Data

We constructed a document-grounded QA dataset comprising 30 questions over a corpus of 787 documents. The corpus draws from three sources: CIA World Factbook entries (257 documents), CNN/DailyMail articles (300 documents), and Russia-Ukraine conflict reporting (217 documents).

Table 1: Dataset statistics by split.

Split	Questions	Purpose
Train	24	Development
Dev	3	Validation
Test	3	Final evaluation
Total	30	

Each question is paired with: (i) a source document ID, (ii) a reference answer, (iii) evidence sentence IDs (e.g., S8, S9) for citation evaluation, and (iv) a 5-point rubric for answer evaluation. Question types include specific fact retrieval, broad analytical overview, and unanswerable-from-source (requiring refusal).

Limitation: Our dev and test sets contain only 3 questions each, which is too small for reliable evaluation. We report primarily on train set results (24 questions) and acknowledge that conclusions may not generalize. Future work should expand evaluation data.

Example Question (from train set):

“According to the article, what controversies or issues arose during Griffin Bell’s confirmation process to become attorney general?”

Gold Answer: Griffin Bell’s confirmation hearings were difficult because of concerns about his past memberships in private segregated clubs and some of his judicial decisions. These issues became points of contention during the Senate review. Despite the controversy, he was ultimately confirmed in January 1977.

Evidence Sentences: S13, S14, S15

Source Document: cnn_dailymail_506e1ba...

Figure 2: Example question from the training set showing the question, gold answer, evidence sentence IDs, and source document.

3.2 Evaluation Metrics

We evaluate system performance along three dimensions: retrieval quality, citation accuracy, and answer quality.

Retrieval metrics. We measure Recall@ k , the fraction of questions for which the gold document appears in the top- k retrieved results:

$$\text{Recall}@k = \frac{1}{|Q|} \sum_{q \in Q} \mathbf{1}[\text{gold}_q \in \text{top-}k(q)] \quad (1)$$

We report Recall@1 and Recall@5, following standard IR evaluation practice [Robertson and Zaragoza, 2009].

Citation metrics. For questions with gold evidence sentences, we compute precision, recall, and F1 between predicted and gold sentence IDs:

$$\text{Precision} = \frac{|S_{\text{pred}} \cap S_{\text{gold}}|}{|S_{\text{pred}}|}, \quad \text{Recall} = \frac{|S_{\text{pred}} \cap S_{\text{gold}}|}{|S_{\text{gold}}|} \quad (2)$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Answer quality. We use LLM-as-judge evaluation, prompting Llama 3.1 8B to score each answer on a 1–5 scale using the question-specific rubric. This approach follows recent work on automated evaluation of open-ended generation [Zheng et al., 2023].

Combined score. We compute a combined metric that balances answer quality and evidence grounding:

$$\text{Combined} = 0.5 \times \frac{\text{AnswerScore}}{5} + 0.5 \times \text{EvidenceScore} \quad (4)$$

where EvidenceScore is the word recall between predicted and gold citation sentences:

$$\text{EvidenceScore} = \frac{|\text{words}(S_{\text{pred}}) \cap \text{words}(S_{\text{gold}})|}{|\text{words}(S_{\text{gold}})|} \quad (5)$$

This differs from Citation F1 (which matches sentence IDs) by measuring content overlap even when predicted citations differ from gold but cover similar information. For unanswerable questions (empty gold citations), EvidenceScore = 1.0 when the system correctly produces no citations.

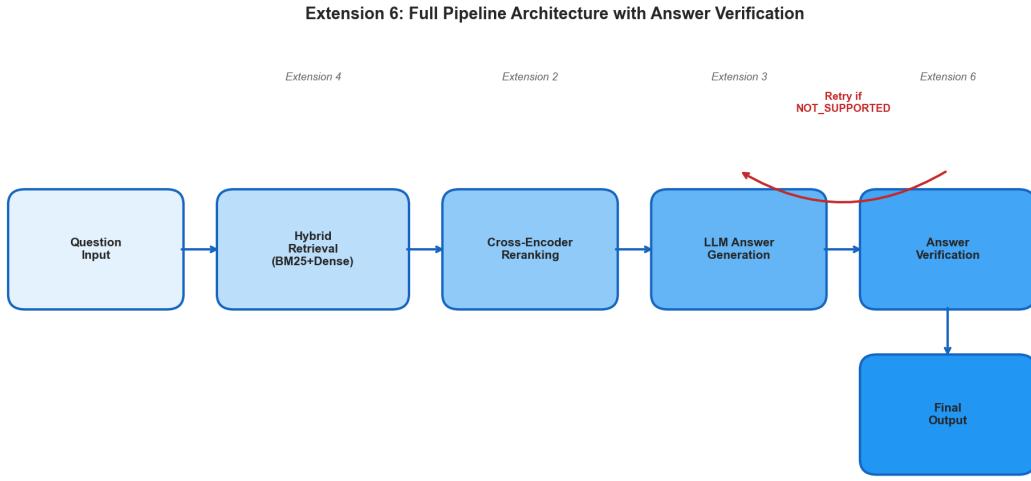


Figure 3: System architecture overview. Our final system (Extension 6) processes questions through hybrid retrieval (BM25 + dense embeddings), cross-encoder reranking, document chunking, LLM answer generation with verification, and LLM-based citation extraction.

3.3 Simple Baseline

Our simple baseline uses BM25 (Best Matching 25) for document retrieval [Robertson and Zaragoza, 2009]. Given a question, BM25 ranks all 787 corpus documents by lexical similarity using term frequency, inverse document frequency, and document length normalization. We retrieve the top-5 documents and extract the top-3 sentences from the highest-ranked document using word overlap with the question.

Table 2: Simple baseline (BM25) performance on train set.

Metric	Value
Recall@1	54.17%
Recall@5	62.50%
Citation F1	0.241
Answer Score	N/A (no generation)

The simple baseline achieves moderate retrieval performance but cannot generate natural language answers. Its citation F1 of 0.241 reflects the limitations of word-overlap matching for identifying evidence sentences.

4 Experimental Results

4.1 Strong Baseline: LLM-Only

Our strong baseline directly queries Llama 3.1 8B (via Groq API) without document retrieval. This tests whether the LLM’s parametric knowledge can answer domain-specific questions. We use temperature 0.7 and max tokens 128.

Table 3: Strong baseline (LLM-only) performance on train set.

Metric	Value
Recall@1	0%
Recall@5	0%
Citation F1	0%
Answer Score	2.58/5
Combined Score	0.30

The LLM-only baseline achieves 2.58/5 answer quality but cannot retrieve documents or cite evidence. Many answers contain plausible but unverifiable information, demonstrating the need for document grounding.

4.2 Extension 1: Retrieval-Augmented Generation

Our first extension combines BM25 retrieval with LLM generation, implementing the RAG paradigm [Lewis et al., 2020]. BM25Okapi (via the `rank-bm25` library) tokenizes documents with NLTK and retrieves the top-5 by lexical similarity. We pass the top-ranked document’s full text to Llama 3.1 8B (via Groq API),

prompting it to answer in 2–4 sentences based only on the provided text, or refuse if information is insufficient. We use temperature 0.7 for natural-sounding outputs. For citations, word-overlap scoring identifies the top-3 sentences sharing the most tokens with the question—a simple baseline we later replace with LLM extraction.

Results: RAG improves answer quality from 2.58 to 3.12 (+21%) and combined score from 0.30 to 0.48. Citation F1 reaches 0.241 with word-overlap extraction. Constraining the LLM to retrieved evidence reduces hallucination compared to the LLM-only baseline.

4.3 Extension 2: Cross-Encoder Reranking

Extension 2 adds neural reranking to improve retrieval precision [Nogueira and Cho, 2019]. BM25’s lexical matching misses semantically similar content when vocabulary differs—a cross-encoder can capture these relationships by jointly encoding the query and passage.

We expand first-stage retrieval to top-30 candidates (from top-5) to give the reranker more to work with. The `cross-encoder/ms-marco-MiniLM-L-6-v2` model from HuggingFace scores each (question, document) pair using its `[CLS]` representation, and we re-sort by these scores to select the final top-5.

Results: Recall@1 jumps from 54% to 79% (+46%) and Recall@5 from 63% to 92%. Citation F1 edges up to 0.291, and answer quality reaches 3.75. The reranker promotes 6 additional correct documents to rank 1.

4.4 Extension 3: LLM-Based Citation Extraction

Extension 3 replaces word-overlap citation extraction with LLM-based semantic identification. Word overlap fails to match semantically equivalent phrases (e.g., “casualties” vs. “killed”, “enacted reforms” vs. “implemented changes”), so we ask the LLM to identify supporting sentences directly.

Gold documents contain pre-annotated sentence IDs (S_1, S_2, \dots); for plain-text documents, we segment using NLTK’s sentence tokenizer with abbreviation handling. We prompt the LLM with the question, answer, and all numbered sentences, asking it to identify 1–5 sentence IDs that support the answer’s claims. We parse sentence IDs via regex and use temperature 0.0 for deterministic extraction.

Results: Citation F1 nearly doubles, from 0.291 to 0.576. The LLM captures semantic relationships (“casualties” → “killed”) that lexical matching misses. Combined score reaches 0.70.

4.5 Extension 4: Hybrid Retrieval

Extension 4 combines lexical and semantic retrieval signals. While the cross-encoder reranks BM25’s top-30, it cannot promote documents that BM25 ranked below 30th. Hybrid retrieval incorporates dense semantic similarity from the start.

We encode all 787 documents and each query using `all-MiniLM-L6-v2` (384-dim embeddings) and compute cosine similarity. BM25 scores are min-max normalized to $[0, 1]$, then combined: $\text{score} = 0.5 \cdot \text{BM25}_{\text{norm}} + 0.5 \cdot \text{semantic}$. We weight both signals equally as a simple baseline; tuning α could yield further gains but we prioritized testing the core hybrid approach. The top-30 hybrid-scored documents are then reranked by cross-encoder, with LLM-based citation extraction downstream.

Results: Recall@1 improves to 83.33% and Recall@5 to 95.83%. Answer quality reaches 3.79/5. Citation F1

is slightly lower than Extension 3 (0.528 vs. 0.576) due to different retrieved documents affecting downstream citation, with combined score at 0.65.

4.6 Extension 5: Document Chunking

Extension 5 addresses a limitation of previous approaches: when documents are long, relevant sentences get diluted by irrelevant content. We implement document chunking to create smaller, more focused retrieval units.

Documents are split into overlapping windows of 8 sentences with 2-sentence overlap, ensuring evidence spanning boundaries appears in at least one chunk. We chose 8 sentences as a reasonable context window that fits comfortably within LLM limits while remaining focused. Each chunk is indexed separately for BM25 retrieval (expanding from 787 documents to \sim 3,500 chunks), then reranked by the cross-encoder at chunk level. The top-ranked chunk provides focused context for generation, reducing noise from irrelevant sections.

Results: Chunking achieves 100% Recall@5 (vs. 95.8% for Extension 4), ensuring the correct document always appears in the top-5. However, Recall@1 drops to 75% because chunk boundaries sometimes split the most relevant content across multiple chunks. Using LLM-based citation extraction (from Extension 3), citation F1 reaches 0.616, comparable to Extension 3’s performance. Answer quality is 3.67/5 with a combined score of 0.69.

4.7 Extension 6: Answer Verification

Extension 6 adds a verification step to detect hallucinated answers. Even with correct retrieval, the LLM may generate claims not supported by the retrieved text. We verify answers against the source and retry with alternative chunks when verification fails.

After generating an answer from the top-ranked chunk, we prompt the LLM to evaluate whether the answer is fully supported, returning “SUPPORTED”, “PARTIAL”, or “NOT_SUPPORTED”. If verification fails, we regenerate using the next-ranked chunk, allowing up to 3 attempts. We capped retries at 3 to balance answer quality against API costs and latency—enough to try alternatives without excessive overhead. The system returns the first verified answer, or the best attempt after exhausting retries.

Results: Verification achieves the best combined score (0.71), with answer quality at 3.88/5 and citation F1 at 0.632. The retry mechanism adds latency (1.4 LLM calls on average) but catches unsupported claims before they reach the output. Grounding analysis confirms 100% of answers are supported by their cited documents.

4.8 Results Summary

Note on reproducibility: Results may vary slightly across runs because answer generation uses temperature 0.7 for more natural outputs. While our evaluation (LLM-as-judge) and citation extraction use temperature 0.0 for determinism, the underlying answers being evaluated differ between runs. The numbers reported here represent a single evaluation run; results from earlier milestones may differ by small margins.

Table 4: Full results comparison on train set (24 questions). R@1 = Recall@1, R@5 = Recall@5, Cite = Citation F1, Ans = Answer Score (out of 5), Comb = Combined Score.

Model	R@1	R@5	Cite	Ans	Comb
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

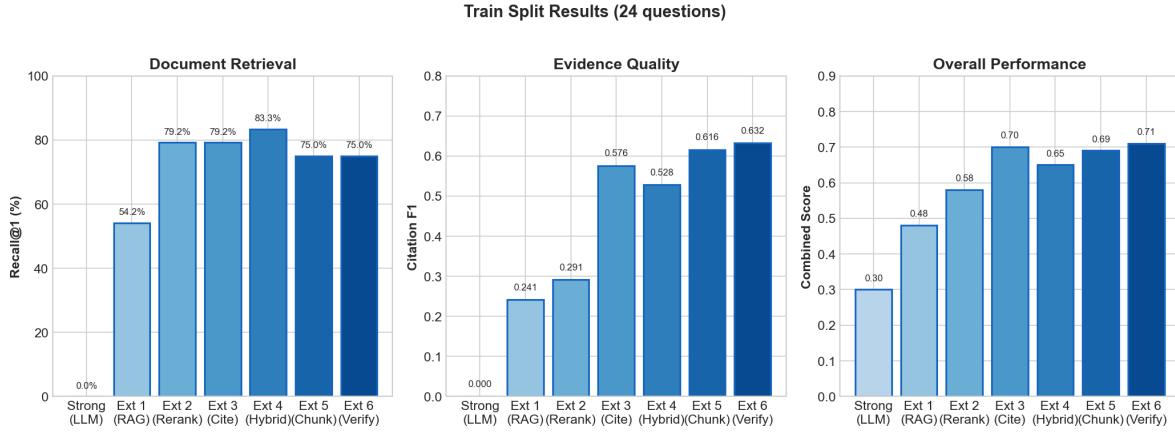


Figure 4: Performance comparison across all models. Extension 6 (Verification) achieves the best combined score (0.71) while maintaining 100% Recall@5.

4.9 Error Analysis

We analyzed errors from Extension 6 on the train set and found three main failure patterns.

Retrieval errors (25%, 6/24 at Recall@1). The most common issue is date confusion in our Russia-Ukraine corpus: articles from adjacent dates cover overlapping topics with similar vocabulary. For q013 (analyst predictions about Russia’s debt default), the system retrieved a July 4th article instead of the June 27th gold document—both discuss Russian default but only the earlier one contains the specific analyst quotes. Similarly, q023 about gas exports retrieved a July 27th article instead of July 21st. For CIA Factbook questions, retrieval sometimes grabbed the country overview instead of the history section. The 100% Recall@5 means the correct document is always available—the challenge is ranking it first.

Generation hallucinations (even with correct retrieval). Question q003 achieved Recall@1 but scored only 2/5 because the LLM hallucinated “250,000 extra postal ballots for soldiers and police”—a detail that appears nowhere in the source document. The evaluator noted: “introduces an incorrect detail and fails to mention the correct number.” Verification helps here (by catching unsupported claims), but doesn’t fully eliminate the problem.

Citation recall limits. Eight questions required 6+ gold evidence sentences, with two requiring 15–16 sentences. Our system predicts only 1–5 citations per answer, so recall is structurally capped. For q021, gold evidence spans 16 sentences but we predicted 4, yielding recall of 0.25 despite reasonable precision. This is a design challenge, since building a gold Q-A dataset often comes down to subjective human judgment and could affect the performance metrics.

Error Example (Retrieval Failure):

Q: “What broader economic and geopolitical consequences did analysts expect from Russia’s 2022 foreign-debt default?”

Gold document: russia-ukraine-conflict_2022-06-27

Retrieved: russia-ukraine-conflict_2022-07-04 (wrong date)

Consequence: Answer score 1/5, Citation F1 = 0.17

Figure 5: Example retrieval error where similar articles from different dates are confused. The retrieved article discusses related topics but lacks the specific analyst predictions from the gold document.

In contrast, Figure 6 shows a successful case where the system correctly retrieves, answers, and cites evidence.

Sample Q&A: Hurricane Maria Impact

SOURCE: CIA World Factbook - Dominica

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.

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**

Correct document retrieved, accurate answer generated, exact evidence sentence cited

Figure 6: Example of successful grounded Q&A. The system correctly retrieves the CIA World Factbook entry for Dominica, generates an accurate answer about Hurricane Maria’s impact, and cites the exact supporting sentence (S7)—demonstrating the pipeline working end-to-end.

5 Discussion and Conclusions

We built an end-to-end grounded Q&A system that retrieves documents and generates answers with sentence-level citations, improving combined score from 0.30 (LLM-only) to 0.71 through six incremental extensions (Figure 7). The biggest wins came from cross-encoder reranking (+46% Recall@1), LLM-based citation extraction (+98% Citation F1), and answer verification (best overall quality at 3.88/5).

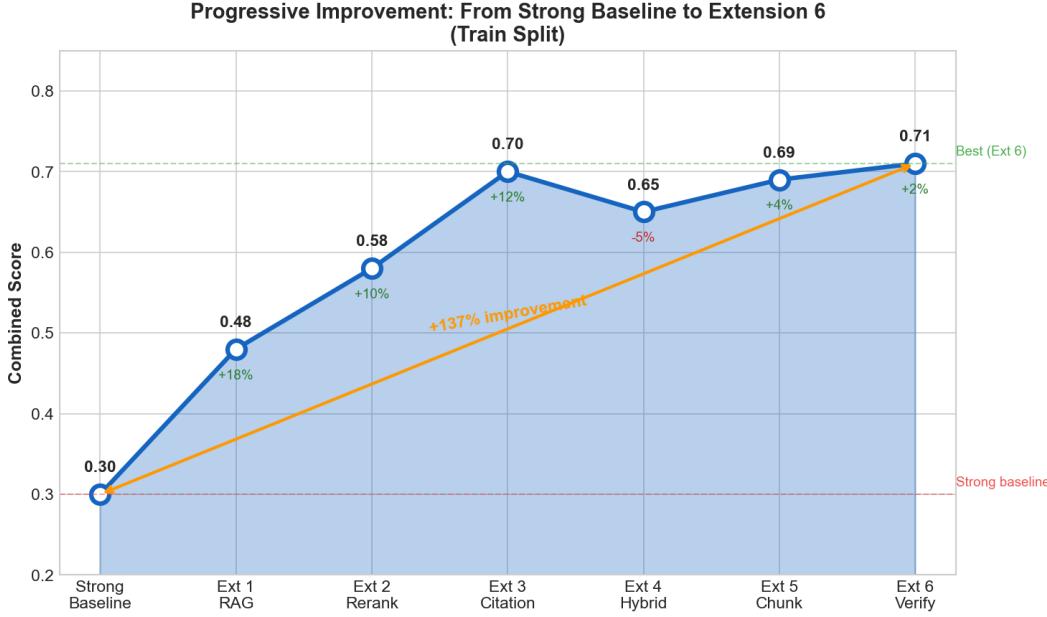


Figure 7: Metric progression across extensions. Each extension targets a specific bottleneck: reranking improves retrieval precision, LLM citations capture semantic matches, chunking ensures perfect Recall@5, and verification reduces hallucination.

Building a “foolproof” RAG system turned out to be harder than we expected. Each pipeline stage introduces its own failure modes, and fixing one doesn’t automatically fix others. Verification (Extension 6) addresses generation failures by catching unsupported answers and retrying, but no amount of verification fixes upstream retrieval errors when semantically similar documents exist. Our decomposed evaluation—separately measuring retrieval, citation, and answer quality—proved more valuable than a single end-to-end metric for diagnosing where things broke.

The main limitations are retrieval errors on semantically similar documents (particularly date-adjacent news articles) and our small evaluation set (30 questions total). Future work should expand the benchmark and explore date-aware retrieval, query expansion, or multi-document synthesis.

“Hallucination-free RAG” remains an open problem, but our incremental approach—adding reranking, then semantic citations, then chunking, then verification—demonstrates how modular improvements can systematically address different failure modes in the pipeline.

Acknowledgements

We thank Anirudh Bharadwaj for providing guidance throughout the project. We also thank Groq API and Huggingface for providing access to pretrained models.

References

- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.
- Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019.
- Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.