

Benchmarking RAG vs Gemini for Document-Grounded Reasoning

Abhinav Pollepalli
November 2025

About

This paper evaluates whether a custom Retrieval-Augmented Generation (RAG) pipeline can match or exceed Gemini Pro model on document-grounded question answering. The goal was to measure reliability, citation accuracy, and hallucination resistance when both systems were restricted to using only a provided research paper.

What was tested?

To judge grounded reasoning performance, both systems were evaluated across:

- Document Alignment (DA): Answer must strictly reference the provided document
- Citation Precision (CP): Verbatim quotes with exact page and line references
- Factual Accuracy (FA): No incorrect claims
- Logical Completeness (LC): Fully answer the question
- External Dependence (ED): No hallucinations or outside knowledge unless clearly labeled

Setup

Data & Task

- Source: One academic PDF (Human-AI Complementarity: A Goal for Amplified Oversight)
- Task: Answer structured questions using only the PDF content, no outside knowledge

Rules for Both Systems

- Verbatim quotes required for supporting evidence
- Each quote must include page citations plus line citations
- If answer not found, respond with “Not found in document.”
- External info allowed only if explicitly labeled

Citation Format Evaluation Rule

- Answers were scored on fidelity to the document
- When a model located the correct evidence but failed to follow the strict citation format, a minor deduction was applied under “Citation Precision”
- Deductions reflected format compliance only, not reasoning, or factual correctness.

Table 1: Model Enforcement

Model	Grounding Method
RAG System	Retrieval + ranking + structured prompting
Gemini	Prompt-enforced grounding rules only

The RAG pipeline enforced grounding architecturally through:

- Sentence-level chunking
- Embedding-based retrieval
- Re-ranking
- Controlled answer formatting

Gemini followed the same rules, but only through prompt constraints (no retrieval).

Evaluation Results

The table below reports scores across five criteria for ten document-grounded questions.

Table 2: Benchmark Table

Question	System	DA	CP	FA	LC	ED	Total
Q1	Gemini	10	9	10	10	10	49
	RAG	10	9	10	9	10	48
Q2	Gemini	10	9	10	10	10	49
	RAG	10	9	10	10	10	49
Q3	Gemini	10	10	10	10	10	50
	RAG	10	10	10	10	10	50
Q4	Gemini	10	9	10	10	10	49
	RAG	10	10	10	10	10	50
Q5	Gemini	10	10	10	10	10	50
	RAG	10	10	10	10	10	50
Q6	Gemini	10	10	10	10	10	50
	RAG	10	10	10	10	10	50
Q7	Gemini	2	8	3	6	5	24
	RAG	10	9	10	10	10	49
Q8	Gemini	10	10	10	10	10	50
	RAG	10	9	10	10	10	49
Q9	Gemini	10	10	10	10	10	50
	RAG	10	9	10	10	10	49
Q10	Gemini	10	9	10	10	10	49
	RAG	10	9	10	10	10	49

Results

Table 3: Model Evaluation Scores

Model	Score (out of 500)
RAG	493
Gemini	470

The table summarizes evaluation scores across ten grounded question answering tasks. Both systems performed well, but differences appeared when the model was required to strictly verify claims using text from the source paper. The RAG pipeline showed consistently strong citation grounding and avoided unsupported statements, especially on questions that required precise evidence retrieval. A few responses cited the correct text but not the required page plus line format, resulting in a one-point Citation Precision deduction without affecting grounding or accuracy.

Conclusion

This project tested whether a custom RAG pipeline could stay accurate and grounded when answering questions about a research paper. Across ten questions, the RAG system scored 493 out of 500 (98.6%), while Gemini scored 470 out of 500 (94.0%).

Both models did well on straightforward questions that asked for definitions or process steps from the paper. The main difference showed up on Question 7, which required understanding of the paper’s core idea about human and AI complementarity. Gemini first responded, “Not found in document,” then contrasted the wrong concepts, both of which contradicted the paper. The RAG system located the relevant passages and supported its answer with direct evidence, resulting in the largest performance gap between the two models.

Overall, the results show that retrieval helps with reliability when answers must come directly from a source. The RAG system stayed grounded and avoided incorrect assumptions, while Gemini occasionally relied on its own reasoning instead of the document. This experiment suggests that a custom RAG setup can match or outperform a strong model on tasks that require evidence from a specific text.