Comparative Analysis of RAG Pipelines

In this exercise, you will compare two RAG-based systems or workflows on the same scientific information task.

The goal is not only to compare answer quality, but also to analyze:

- retrieval relevance
- evidence grounding
- citation fidelity
- failure modes
- reproducibility

Your report must be formal, structured, and evidence-based.

Task Scenario

Assume you are conducting a literature-based research task:

"Summarize the current state of methods for detecting hallucinations in large language models."

You must evaluate two different RAG pipelines.

Examples may include:

- two different AI tools / assistants
- two retrieval strategies
- two embedding models
- two chunking methods
- two prompting templates
- RAG vs non-RAG baseline

You may use any academically appropriate tools available to you.

Part A — System Design and Comparison Setup

Describe the two systems being compared.

For each pipeline, document:

- retrieval source(s)
- embedding / search method
- chunking strategy
- context window / number of retrieved chunks
- prompting template
- model used for generation

Represent each workflow as a simple pipeline:

query → retrieval → context injection → generation → output

A diagram is strongly encouraged.

Part B — Comparative Experiment (30%)

Run the same query on both systems.

Use at least 3 prompts / questions within the same topic.

For each prompt, collect:

- retrieved evidence
- generated answer

- citations / references
- latency (optional)
- observed failure modes

Present results in a comparative table.
Suggested structure:
Prompt System A System B Better System Why

Part C — Evaluation Framework
Evaluate both systems using the following criteria:
1. Retrieval relevance
Did the retrieved chunks actually match the query?

2. Grounding fidelity
Were claims directly supported by retrieved evidence?

3. Citation reliability
Were references real, accurate, and traceable?

4. Hallucination risk
Did the system introduce unsupported claims?

5. Scientific usefulness
Would the answer be suitable for research use?
Use either:

- qualitative evaluation
- rubric scoring (1–5 scale)
- hybrid evaluation

A rubric table is recommended.

Part D — Critical Error Analysis
Identify and discuss at least two failure modes.
Examples:

- irrelevant retrieval
- missing key papers
- context truncation
- fabricated citations
- overconfident synthesis
- contradictory sources

For each failure, explain whether the problem came from:

- retrieval
- chunking
- prompt design
- generation model

This section is particularly important.

Part E — Reflection and Recommendation (10%)
Write a short recommendation:
Which RAG design would you adopt for scientific research and why?
Your answer must explicitly discuss:

- reliability
- reproducibility
- scalability
- suitability for literature review workflows

Submission Requirements
Submit a report of 2–4 pages.
Include an appendix with:

- prompts used
- system outputs
- retrieved evidence snippets

RAG might be NotebookLM, non-RAG – ChatGPT-3. New versions of ChatGPT has web-search and access to different tools, so they are not vanilla LLM.