Query Set Design:

I designed 20 queries to test different aspects of the RAG system's retrieval from its knowledge base and citation capabilities. We had 10 direct queries, 5, synthesis queries, and 5 edge case questions modeling after the example queries in the writeup. Half of the queries followed the two tasks I picked out in Phase 1 which were Claim-Evidence Extraction (CEE) and Cross-Source Synthesis (CSS), having 6 and 4 of each respectively. The other half were more general queries.

Here is how the queries mapped to our research question and sub-questions from Phase 1:

Main Research Question: How do different sleep factors (duration, quality, timing, consistency) affect physical health, mental well-being, and cognitive performance, and what interventions most effectively improve sleep outcomes?

Sub Questions:

1. What is the relationship between sleep duration and mental health outcomes (anxiety, depression, stress)? Queries direct_001_cee, direct_003_cee, direct_004_cee
2. How does sleep quality (measured by sleep stages, interruptions) affect physical health markers? Queries direct_002_cee, direct_006_cee
3. What role does sleep consistency (regular schedule) play vs total sleep hours? Queries direct_007, synth_001_css
4. How do lifestyle factors (exercise, caffeine, screens, diet) affect sleep quality and duration? Queries direct_008, synth_005
5. What sleep interventions (CBT-I, sleep hygiene, supplements) show the strongest evidence for improving well-being? Queries direct_009, direct_010, synth_002_css
6. How do individual differences (age, chronotype, health conditions) moderate sleep-wellbeing relationships? Queries edge_003

We utilize direct queries to establish a baseline performance for the system on straightforward tasks, have synthesis queries to test the system's aggregation, and edge cases to test that the system properly refuses fabricating citations when evidence is missing.

Metrics:

I used 3 metrics to access the RAG system performance:

1. Citation precision:
   - Measures whether citations in generated answers actually exist in the retrieved chunks.
   - Range: 0-1 (higher is better)
   - Why: In Phase 1 we saw a lot of citations being misplaced or missing, therefore it was a critical concern for Phase 2.
2. Groundedness score: measures whether claims in the returned answer are actually supported by the evidence further than just the correct citations.
   - Rubric: Adapted the provided scoring rubric from Phase 1 to a RAG context (can be seen in src/eval/metrics.py)
   - Method: GPT-4o-mini evaluated each answer against retrieved chunks

3. Answer relevance [one additional metric]: measures if the response actually answered the user's question
    - Rubric: Also followed a 1-4 scale
    - Method: GPT-4o-mini evaluated
    - Why: Even if an answer is grounded if it didn't answer the question then it is useless.

These 3 work together to properly evaluate high versus low quality answers

- High citation precision + low groundedness = valid citations but inaccurate claims
- High groundedness + low relevance = accurate information but doesn't answer the provided question
- High relevance + low citation precision = Answers the provided question but with fabricated sources

I utilized an LLM to evaluate groundedness and answer relevance, giving it the original query, retrieved chunks, the generated answer with citations, scoring criteria per task, and strict guidelines and output expectations with good reasoning.

Results:

Baseline System

Metric	Score	Intrepretation
Citation precision	77.6%	Over 3 out of 4 citations are valid
Groundedndess	3.4/4	Most answers are well-grounded with minor issues
Answer relevance	3.58/4	Answers are very relevant to queries
Average citations per query	4.3	Good citation coverage

The baseline system retrieves relevant evidence pretty well, the biggest weakness is the citation precision at 79.7% meaning about 1 in 5 citations are invalid.

Query Category

Category	Citation precision	Groundedness	Relevance	Average citation count
Direct (n=10)	85%	3.45/4	3.55/4	4.35
Synthesis (n=5)	89.5%	3.3/4	3.8/4	6.1
Edge Case (n=5)	51%	3.4/4	3.4/4	2.4

Some key observations are that synthesis has the best citation precision most likely because of how compare and contrast it is and edge cases correctly have lower citation precision because there should be nothing to cite if it's not within the existing system.

Phase 1 Task Performance

Task	Citation Precision	Groundedness	Relevance	Average citation count
General (n=10)	64.3%	3.5/4	3.6/4	3.75
CEE (n=6)	85%	3.5/4	3.25/4	3.75
CSS (n=4)	100%	3.38/4	4/4	6.5

Some key observations are that CEE has good citation precision because it explicitly requires supporting statistics, but may have slightly lower groundedness as it takes the direct jargon and turns it into more digestible, paraphrased explanations. CSS queries has the most precise and highest citation count because it is comparing and contrasting from multiple sources. General queries don't explicitly request citations, but still have good groundedndess and relevance scores.

Adding Enhancements

In Phase 1 there was a recurring issue with citations being fabricated and just plain wrong. The baseline RAG system would generate citations but had no verification if these citations actually existed in the retrieved chunks.

I decided to go with the structured citations enhancement which adds a post-processing validation layer after answer generation. It first parses the answer and extracts all citations using regex, then for each extracted citation verifies that the chunk_id exists in the retrieved chunks and marks it as valid or invalid. With the valid citations, it then looks up the source metadata in data_manifest.json and generates a formatted reference list so users can more easily access the original sources of data.

This enhancement is implemented as an optional flag (enhance=True).

Baseline vs Enhanced Comparison

Metric	Baseline	Enhanced	% Change
Citation precision	77.6%	78.5%	+1.1%
Groundedness	3.4/4	3.58/4	+5.1%
Answer relevance	3.58/4	3.73/4	+4.2%
Average citation count	4.3	3.95	-8.1%

The citation precision increased slightly by 1.1%, demonstrating that the enhanced system is encouraging more accurate citations as it explicitly flags invalid citations instead of sneaking them into the response unnoticed. Groundedness improved by 5.1% as the fact that the citations will be validated encourages the LLM to be more careful about grounding claims in actual evidence. Answer relevance increased by 4.2% because the reference list allows users to see which papers support each claim without having to search through the data manifest on their own. The average citation count went down by 8.1% because the enhanced system generates fewer, higher-quality citations and discourages overciting.

Query Baseline vs Enhanced

Category	Citation precision [baseline]	Citation precision [enhanced]	Groundedness [baseline]	Groundedness [enhanced]	Relevance [baseline]	Relevance [enhanced]
Direct (n=10)	85%	85%	3.45/4	3.45/4	3.55/4	3.6/4
Synthesis (n=5)	89.5%	94%	3.3/4	3.5/4	3.8/4	3.8/4
Edge Case (n=5)	51%	50%	3.4/4	3.9/4	3.4/4	3.9/4

Some key improvements are the increase in groundedness and relevance for edge cases, both going from 3.4/4 to 3.9/4. This demonstrates that the invalid citations being caught early improves the actual claims being generated.

This enhancement improves system performance and improves the overall user experience. It removes friction for users who want to manually verify their claims, allowing them to trust the system and its outputs more.

Failure Cases

1. direct_006_cee is a direct CEE query that was marked as a failure because only 3/4 of its citations were valid. I asked the system about how sleep duration affects physical health markets like inflammation, to which it cited a chunk that was in the knowledge base but not in the retrieved chunks.
2. edge_004 is a general edge case query that was marked as a failure because of missing citations, but this is an edge case where we don't have citations for caffeine consumption timing on sleep quality.
3. edge_005 is a general edge case query regarding how sleep affects athletic performance in professional athletes. Some chunks pulled didn't even contain much content at all, including the DOI link and paper titles. This will be explored further in Phase 3.