Phase 2 Evaluation Report: Mars Life Research RAG System

1. Introduction

This report evaluates a RAG system built to answer research questions about potential life on Mars. The system addresses limitations from Phase 1, where LLMs generated plausible claims but rarely included citations and couldn't refuse questions when evidence was missing. Phase 2 implements retrieval-grounded answering with structured citations and evidence-gating to make sure all claims trace back to source text. The corpus has 20 peer-reviewed papers on Mars habitability, biosignatures, atmospheric measurements, and extremophile analog studies. The system uses semantic search with FAISS indexing over 421 text chunks to retrieve evidence and generate extractive answers with explicit source and chunk IDs.

2. Query Set Design and System Architecture

Twenty evaluation queries were created and distributed across three categories to test different system capabilities and map to the six research sub-questions from Phase 1. The distribution was chosen to emphasize basic retrieval, while ensuring the system can handle cross-source reasoning and refuse appropriately. Direct queries cover methane detection, geological habitability indicators, and biosignature interpretation. Examples include: "What evidence did Curiosity find in Gale Crater?" and "What minerals indicate long-term water-rock interaction?" Synthesis queries require comparison across sources, such as "Compare methane-based life arguments with hydrothermal vent habitability arguments." Edge cases ask for non-existent evidence using language like "confirmed" and "definitive proof" to test refusal behavior, addressing the Phase 1 failure mode where models couldn't distinguish between discussing a claim and confirming it.

The system follows a standard RAG pipeline. PDFs were downloaded then, converted to plain text and divided into approximately 4000-character chunks with 400-character overlap to preserve scientific context while staying within embedding model limits. Chunks were embedded using e5-base-v2 (768-dimensional embeddings) and indexed with FAISS exact search, prioritizing precision over speed. Answers are extractive. The system returns text snippets directly from retrieved chunks with citations in the format [source_id:chunk_id], mirroring the structured citation behavior from Phase 1 Prompt B.

3. Enhancement: Confidence Scoring

The baseline RAG system retrieves chunks and generates answers, but doesn't assess whether retrieved evidence is actually sufficient. This limitation also showed up in Phase 1, where LLMs generated answers even when evidence was weak. To address this, I implemented confidence scoring that combines three factors. First, retrieval confidence, which measures the average semantic similarity of the top-3 chunks. Second, lexical overlap, which measures the fraction of query keywords that appear in retrieved text and catches technical terms that semantic search might miss. Finally third, confirmation strength, for queries asking for "confirmed evidence" or "definitive proof," the system checks whether retrieved text actually uses confirmation language ("confirmed," "proven") in the same sentence as the claim. If the confirmation language isn't there, this score drops to zero.

The overall confidence score is a weighted combination. For normal queries, weights are 50% retrieval, 30% lexical, 20% confirmation, with threshold 0.5. For queries containing strong-claim language (such as "confirmed," "definitive," "proof," "does the corpus contain"), the system automatically detects these patterns and shifts weights to 20% retrieval, 20% lexical, 60% confirmation, with a higher threshold of 0.7. Normal queries consistently scored 0.7 or higher, while queries asking for confirmation scored 0.2-0.3, so the 0.5 and 0.7 thresholds create clear separation between answerable and unanswerable

queries. I chose confidence scoring over other enhancements because it directly addresses the Phase 1 limitations where models couldn't distinguish between discussing a claim and confirming it. Reranking would improve retrieval quality but would be more difficult to solve the refusal problem.

4. Evaluation Metrics and Results

I evaluated the system on three metrics to assess correctness and quality. Faithfulness measures whether answers are grounded in retrieved text. Answer relevance checks if answers address the query intent. Citation accuracy verifies all citations resolve to actual corpus chunks.

Faithfulness: 20/20 queries (100%). All 37 citations resolved to actual chunks in the corpus. Answered queries produced extractive text matching retrieved content. Refused queries stated insufficient evidence without speculation. This 100% rate is by design since the system only outputs retrieved text or refusal messages.

Answer Relevance: All 10 direct queries received on-topic answers as the system successfully retrieved specific details about organic molecules, methane mechanisms, and mineralogy. All synthesis queries were answered, but with mixed quality. While the system provided answers for all synthesis questions, it struggled with multi-document comparison. For example, Query 13 (Gale vs. Eridania) cited only a single source, failing to retrieve the second half of the comparison. All 5 edge cases correctly refused.

Citation Accuracy: 37/37 citations (100%). Every (source_id, chunk_id) pair exists in the corpus with no invented IDs.

Query breakdown: Direct queries: 10/10 answered with 2-5 citations (mean: 2.8). Two queries cited 5 sources each, while most queries cited 2 sources. Synthesis queries: 5/5 answered with 1-2 citations (mean: 1.8), but true multi-paper synthesis was rare. Only 2 queries (Q11, Q15) cited distinct papers. The other three (Q12, Q13, Q14) relied on a single author for their answer, and Q13 only had one chunk to cite. Edge cases: 5/5 refused with 0 citations. Confidence scores for answered queries ranged 0.66-0.93 (mean: 0.81). Refused queries scored 0.23-0.28 (mean: 0.26), verifying the 0.70 threshold is effective.

Baseline comparison: To measure the enhancement impact, I ran the same 20 queries without confidence scoring (system answers all queries regardless of score). Results: All 20 queries answered, including 5 edge cases that should have refused. This demonstrates the confidence scorer successfully prevents false answers when evidence is insufficient.

5. Failure Cases

Failure Case 1: Poor Synthesis with Single-Source Citation

Query 13	"Compare Gale Crater findings with Eridania basin models. What similarities exist in habitability potential?"
Expected Result	Should cite sources on both Gale Crater (e.g., Grotzinger2014, DesMarais2014) and Eridania (LaRowe2021) to construct a comparison.
Actual Result	Confidence = 0.80, cited only LaRowe2021_EridaniaLake_chunk_0002.

Retrieved Results	LaRowe2021 (0.860), LaRowe2021 (0.842), Kite2025 (0.837), Kite2025 (0.831), DesMarais2014 (0.830).
What went wrong	The retrieval system ranked Eridania-focused chunks (Score: 0.860) significantly higher than the first Gale Crater chunk (DesMarais2014 at 0.830). Additionally, the system fell into a "Bibliography Trap": the chunk from Kite2025 (Score: 0.837) was just a reference list containing the words "Gale Crater," which the embedding model mistakenly prioritized over the actual descriptive text in DesMarais2014. This pushed the relevant Gale Crater evidence out of the top context window, resulting in a one-sided answer.

Failure Case 2: Low Confidence Due to Sparse Corpus Coverage

Query 5	"How do carbonates serve as indicators of past CO_2-rich environments?"
Expected Result	Should cite Kite2025_Carbonates which directly addresses carbonate formation and CO_2 constraints.
Actual Result	Confidence = 0.66 (lowest among answered queries), cited Neveu2018 and Ehlmann2012.
Retrieved Results	Neveu2018 (0.828), Ehlmann2012 (0.820), Kite2025 (0.818), Hays2017 (0.818), Hays2017 (0.817).
What went wrong	The retriever prioritized high-level biosignature papers (e.g., Neveu2018) that mentioned "carbonates" in passing, rather than the core mineralogy paper Kite2025, which appeared 3rd in the ranking. “Carbonates” appears less frequently in the corpus than terms like “water” or “methane,” causing Kite2025 chunks to score lower. This suggests the semantic embedding model struggles to distinguish between a mention of a mineral and a definition of its formation process.

Failure Case 3: Garbled Text Due to PDF Parsing Issues

Queries	Queries 4, 5, and 6
Pattern	Answer text contains words concatenated without spaces, making text difficult to read.
Example from Query 4	"theregionalcontextcanhelpmakeacase" (instead of "the regional context can help make a case").
What went wrong	PDF-to-text conversion merged words together without spaces (e.g., "identifyfeaturesconsistentwith" instead of "identify features consistent with"), likely due to unusual formatting, multi-column layouts, or special characters. This affects readability across multiple queries and sources (Hays2017, Ehlmann2012) regardless of confidence score, appearing in both the lowest (Q5, 0.66) and highest (Q6, 0.93) confidence answers.

6. Recommendations

Lower Citation Threshold for Multi-Source Queries: The current threshold (0.85) occasionally excludes highly relevant sources in synthesis queries. Query 13 retrieved DesMarais2014 at 0.830, highly relevant to the comparison but uncited due to missing the threshold by 0.02. While most comparison queries succeed, consider using a dynamic threshold (e.g., 0.80 for comparison queries) or citing all chunks within 0.05 of the top score to ensure comprehensive multi-source answers.

Filter Reference Sections to Prevent "Bibliography Traps": Analysis of Query 13 revealed that the system sometimes prioritizes bibliography lists over actual content. A reference list in Kite2025 scored 0.837 (higher than the actual descriptive text in DesMarais2014) because it contained a dense cluster of relevant keywords like "Gale Crater." It would be beneficial to implement a heuristic during chunking to detect and discard Reference sections to prevent metadata from displacing actual scientific prose.

Improve PDF Parsing Pipeline: Multiple queries (4, 5, 6) show garbled text where words merge without spaces. This affects both high-confidence (0.93) and low-confidence (0.66) answers across multiple sources (Hays2017, Neveu2018). It would be beneficial to upgrade PDF parsing tools (e.g., use PyMuPDF or Adobe Extract API instead of basic libraries), implement post-processing to detect concatenated words, and add quality checks to flag chunks with abnormal character patterns before indexing.

Enhance Retrieval for Specialized Terms: Query 5 on carbonates retrieved only generic biosignature papers (Neveu2018) instead of the more relevant Kite2025_Carbonates, resulting in the lowest confidence score (0.66). Specialized terms like "carbonates" appear less frequently than common terms like "water" or "methane," which may cause relevant chunks to rank lower. Consider implementing hybrid retrieval combining semantic search with keyword boosting for technical terms, or use query expansion to include synonyms and related concepts.

Add Source Diversity Scoring: Some queries may benefit from citing multiple sources when several relevant papers exist. Consider implementing a diversity penalty that slightly down-weights chunks from already-cited sources, encouraging the system to pull from multiple papers when appropriate. This could also help with comparison queries like Query 13.

Recommended Configuration: Based on this evaluation, the optimal system should combine hybrid retrieval (semantic + keyword boosting for technical terms) with an adjusted citation threshold of 0.80 for synthesis queries and upgraded PDF parsing (PyMuPDF with post-processing validation). This configuration addresses the key failure patterns: citation gaps in synthesis queries (Q13), retrieval traps in bibliographies, and readability problems (Q4-Q6). The 5 negative test cases (Q16-Q20) correctly returned low confidence, confirming the system appropriately flags insufficient evidence.

7. Conclusion

This evaluation demonstrates significant improvements from Phase 1, where the LLM systems struggled with consistent citations and confidence calibration. The current RAG system now performs more reliably on Mars habitability queries, answering questions with appropriate confidence scores while properly declining 5 negative test cases lacking sufficient evidence. The system's key advancement is its ability to assess when evidence is insufficient; all 5 unanswerable queries (Q16-Q20) were correctly flagged with low confidence scores below the 0.70 threshold, a marked improvement from Phase 1's over-confident responses.

However, there remain issues that impact answer quality. First, the fixed citation threshold (0.85) occasionally excludes highly relevant sources in synthesis queries. In Query 13, the system failed to cite the critical Gale Crater evidence (DesMarais2014, score 0.830) because it missed the cutoff, while mistakenly prioritizing a bibliography list in Kite2025 (score 0.837) due to keyword density. Second, PDF parsing errors produced garbled text in multiple queries, merging words without spaces and degrading readability regardless of confidence level. Third, specialized terminology queries like Query 5 on carbonates suffered from semantic drift, retrieving generic biosignature papers over specific mineralogy sources, resulting in the lowest confidence score (0.66).

The recommended configuration of Hybrid Retrieval (to capture technical terms), Bibliography Filtering (to remove metadata noise), and a Dynamic Citation Threshold (to capture multi-source evidence), addresses these remaining issues while preserving the system's improved confidence calibration. Phase 3 will build on these improvements by transforming the system into a personal research portal that supports the full workflow from question to evidence synthesis to exportable research artifacts.