

## Machine Learning Failure Modes in Space Debris Tracking & Collision Avoidance

### 1. System Overview

This report documents Phase 2 of my Personal Research Portal, which implements a research-grade retrieval-augmented generation (RAG) pipeline for investigating machine learning failure modes in space debris tracking and collision avoidance. The system ingests a curated 20-source corpus, retrieves semantically relevant evidence for research queries, and generates citation-grounded answers where every factual claim is backed by a resolvable (source\_id, chunk\_id) reference.

The pipeline expands on my prompt engineering findings from Phase 1, particularly the importance of crafting structured citation formats and explicitly handling uncertainty, to present a fully automated retrieval and generation system.

### 2. Corpus and Manifest

The corpus comprises 20 sources spanning peer-reviewed journal articles (9), conference papers (6), and technical reports (5) published between 2008 and 2025. Sources were selected to address six research sub-questions covering machine learning (ML) approaches, data limitations, generalization failures, uncertainty quantification, operational integration, and validation challenges.

Each source is documented in `data_manifest.csv` with fields for source\_id, title, authors, year, document type, venue, DOI/URL, and relevance notes. Every citation produced by the system resolves through this manifest to a specific title, author list, and persistent identifier, providing citation traceability.

All 20 papers were chunked via a directed process using Claude Opus 4.6, following a highly structured and standardized protocol (documented in `CHUNKING_PROTOCOL.md` and later `CHUNKING_PLAYBOOK.md`). Automated extraction via OCR and PDF is something that was attempted but rejected due to poor accuracy, particularly for mathematical notation, table structures, and multi-column layouts. Claude-facilitated chunking produced 1,628 total chunks with section-aware paragraph IDs (e.g., `sec3.1_p2`), preserving the hierarchical document structure needed for meaningful citations.

The corpus was revised during the middle stages of Phase 2: `liou2008` (Liou & Johnson, "Instability of the Present LEO Satellite Populations") replaced `kessler1978` (Kessler & Cour-Palais, "Collision Frequency of Artificial Satellites") due to extended difficulty working with text extraction for such an old paper. It was originally included as a foundational source. However, due to the focus on machine learning in the research question, a more modern paper was selected to provide a more empirically grounded treatment of debris population dynamics with Monte Carlo simulation data directly relevant to ML validation challenges.

After finalizing the corpus selections, creating the data manifest, and completing chunking with Claude, the chunked content was ingested and parsed to create a vector store of embeddings using ChromaDB. While FAISS had been suggested, ChromaDB was selected to support downstream implementation of metadata filtering.

### 3. Pipeline Architecture

## Machine Learning Failure Modes in Space Debris Tracking & Collision Avoidance

The pipeline follows a four-stage architecture:

**Retrieve → Rerank → Generate → Log**

**Retrieval.** User queries are embedded using the `all-mpnet-base-v2` (768-dimensional, cosine similarity) sentence transformer model and matched against the 1,628 chunk embeddings stored in ChromaDB. The embedding model was selected over a smaller model (`all-MiniLM-L6-v2`) for its superior semantic matching on technical aerospace terminology. With this small, specialized, highly technical corpus, subtle distinctions (e.g., "collision probability" vs. "conjunction assessment") affect retrieval precision. The larger-sized model improves performance without introducing too much latency since the corpus is relatively small. The top 20 chunks by cosine distance are retrieved, along with full metadata (source\_id, chunk\_id, section\_title, year, document type, venue, authors). ChromaDB's metadata filtering capability supports optional constraints (e.g., year  $\geq$  2022, document type = peer-reviewed, tags), which will be exposed in the Phase 3 portal UI.

**Reranking.** The top 20 retrieved chunks are reranked using a cross-encoder model (`ms-marco-MiniLM-L-6-v2`) that scores each query-chunk pair jointly. Unlike the bi-encoder used in retrieval (which embeds queries and documents independently), the cross-encoder attends to both simultaneously, producing more accurate relevance judgments at the cost of higher latency. The top 10 chunks after reranking are passed to generation. This is the Phase 2 enhancement. Its measurable impact is evaluated in Section 6.

**Generation.** The 10 reranked chunks are formatted into an evidence block and passed to Claude Sonnet (`claude-sonnet-4-5-20250929`) via the Anthropic API with a citation-enforcing system prompt (version `v1.0`). The prompt instructs the model to cite every factual claim using (`source_id, chunk_id`) format, refuse to invent citations, flag insufficient or conflicting evidence, and produce a references section listing each cited source. While Claude Opus 4.6 was the leading model selected in Phase 1, Claude Sonnet 4.5 was selected as the default model for retrieval as it provided cost efficiency across the 50+ evaluation runs. I also implemented an LLM-As-a-Judge evaluation scheme, utilizing Claude Opus 4.6 for that purpose. Having the stronger reasoning model available to judge, while still utilizing the same API key was valuable. While these defaults were used for my evaluation runs, the model is configurable via command-line flag.

**Logging.** Every pipeline run is logged as a JSON Lines entry containing the timestamp, query, all 20 retrieved chunks (with distances), the 10 reranked chunks (with scores and full text), the generated answer, model name, prompt version, and token usage. This structured logging facilitates reproducibility and post-hoc analysis of retrieval quality and generation behavior.

The entire pipeline executes via a single command:

None

```
python -m src.rag.query "What are the main failure modes of ML for collision avoidance?"
```

This produces retrieval results, a citation-backed answer, and a saved log entry.

## Machine Learning Failure Modes in Space Debris Tracking & Collision Avoidance

### 4. Query Set Design

The evaluation query set contains 25 queries in three categories designed to test different aspects of pipeline performance:

**Direct queries (12).** Each targets a specific finding from one or two papers with a known answer. These test retrieval precision (whether the system finds the right chunks) and citation accuracy (whether it cites the correct source). Examples include asking for specific statistics (e.g. class imbalance ratios in the ESA CDM dataset), named techniques (UQ methods in Licata 2022), and institutional conclusions (NASA CARA's assessment of ML viability). Direct queries cover all six research sub-questions.

**Synthesis queries (7).** These require reasoning across multiple papers to construct a coherent answer. They test whether the reranker can surface relevant chunks from diverse sources and whether the generator can integrate evidence with appropriate cross-citation. Examples include comparing UQ approaches across the corpus, tracing atmospheric drag uncertainty through the ML orbit prediction pipeline, and synthesizing institutional perspectives (NASA, ESA, RAND) on why ML has not achieved operational status in this sector.

**Edge-case queries (6).** These test the system's trust behavior (its ability to acknowledge the limits of its evidence). Two queries ask about topics completely outside the corpus (reinforcement learning for maneuver planning and quantum computing for debris tracking). These queries should elicit explicit "insufficient evidence" responses. Two are deliberately framed to invite overconfident answers (asking for "the single best ML approach" or whether deep learning outperforms traditional methods, when the corpus evidence points the opposite direction). Two are partially answerable, testing whether the system can provide what evidence exists while flagging gaps.

The query set was designed with known expected sources for each query, enabling automated verification of retrieval precision (whether the expected papers appear in the retrieved chunks).

### 5. Metrics and Rubric

Each query response is scored on five metrics: two judge-scored quality metrics from the initial evaluation, one judge-scored completeness metric added after initial results showed insufficient variance, and two mechanical metrics computed from log data. Claude Opus 4.6 served as the judge.

#### JUDGE-SCORED

**Groundedness:** Does every factual claim in the response have supporting evidence in the cited chunk?

Score	Definition
4	Every claim is supported by cited evidence; uncertainty is stated when evidence is weak
3	Mostly grounded; minor unsupported claims or slight overstatement of evidence
2	Partially grounded; some claims lack evidence or are loosely connected to citations

## Machine Learning Failure Modes in Space Debris Tracking & Collision Avoidance

1	Not grounded; hallucinated claims, fabricated evidence, or systematic misrepresentation
---	---

**Citation Correctness:** Do the cited (source\_id, chunk\_id) pairs resolve to real text that actually supports the claim?

Score	Definition
4	All citations resolve correctly and directly support their associated claims
3	Most citations are correct; minor issues (e.g., cites the right paper but a nearby chunk)
2	Some citations are incorrect, fabricated, or point to irrelevant text
1	Citations are systematically wrong or fabricated

**Answer Completeness:** Did the answer address the full scope of the question using the available retrieved evidence? Unlike groundedness (which asks "Are the stated claims supported?"), completeness asks "did the answer cover what it should have?" A response can be perfectly grounded but incomplete.

Score	Definition
4	Covers all aspects of the question using the full range of relevant retrieved evidence
3	Mostly complete; minor gaps or over-reliance on a single source when multiple were available
2	Partial; misses a major aspect or ignores clearly relevant retrieved chunks
1	Superficial or off-target despite relevant evidence being available

## CALCULATED

**Retrieval Recall (0.0–1.0):** What fraction of expected sources appeared in the top-20 retrieved chunks? This was computed for each of the 23 queries that have defined expected sources (excluding 2 out-of-scope edge cases with intentionally empty expected sources). This metric is identical across reranked and baseline modes because both share the same initial retrieval. Reranking only reorders the top 10.

**Context Utilization (0.0–1.0):** Of the 10 chunks sent to the generator, how many were actually cited in the answer? This was measured at exact (source\_id, chunk\_id) pair level. This metric distinguishes retrieval noise (chunks sent but ignored) from effective evidence use.

All judge-scored metrics used Claude Opus ([claude-opus-4-6](#)) as the evaluator, which is more capable than the generation model ([claude-sonnet-4-5-20250929](#)), reducing the risk of the evaluator missing errors the generator introduced. Initial evaluation with only groundedness and citation correctness produced

## Machine Learning Failure Modes in Space Debris Tracking & Collision Avoidance

near-perfect scores (4.00/4.00). This was suspicious and prompted the addition of completeness, retrieval recall, and context utilization as additional metrics to capture failure modes invisible to per-claim scoring.

## 6. Results

All 50 runs (25 queries × 2 modes) completed without errors.

### 6.1 Overall Scores

With Reranking:

Category	n	Groundedness	Citation	Completeness	Retrieval Recall	Context Utilization
Direct	12	4.00	4.00	3.75	0.88	0.38
Synthesis	7	4.00	4.00	3.57	0.89	0.64
Edge-case	6	4.00	3.83	3.83	0.67	0.38
Overall	25	4.00	3.96	3.72	0.84	0.46

Baseline (No Reranking):

Category	n	Groundedness	Citation	Completeness	Retrieval Recall	Context Utilization
Direct	12	4.00	4.00	3.75	0.88	0.30
Synthesis	7	4.00	4.00	3.57	0.89	0.57
Edge-case	6	4.00	4.00	3.67	0.67	0.40
Overall	25	4.00	4.00	3.68	0.84	0.40

Completeness proved to be the most discriminating metric, ranging 3–4 across queries and differentiating between categories, while groundedness and citation correctness were nearly uniformly perfect. This confirmed the hypothesis that per-claim scoring alone was insufficient to capture meaningful quality differences.

### 6.2 Enhancement Impact: Reranking

The five-metric view revealed some of the benefits of the reranking enhancement.

## Machine Learning Failure Modes in Space Debris Tracking & Collision Avoidance

**Context utilization improved measurably** from 0.40 (baseline) to 0.46 (reranked), a 15% relative improvement. The effect was strongest on synthesis queries ( $0.57 \rightarrow 0.64$ ), where the cross-encoder's joint query-chunk attention better identified relevant evidence across multiple papers. This means the reranker is promoting chunks that the generator actually uses, reducing retrieval noise in the evidence window.

**Completeness improved slightly** from 3.68 to 3.72 overall, with the gain concentrated in edge-case queries ( $3.67 \rightarrow 3.83$ ). The D11 case illustrates this well—the baseline produced a single-source answer from *mashiku2025*, while the reranked version additionally surfaced *nasa\_ca\_handbook2023*, contributing novel context about conjunction assessment workload scaling and data integration complexity. The baseline was the only run tagged MISSED\_EVIDENCE by the judge.

**Groundedness and retrieval recall were unchanged**, as expected. Groundedness reflects generation quality (unchanged by reranking), and retrieval recall measures the initial embedding search (identical for both modes).

**Citation correctness dropped marginally** ( $4.00 \rightarrow 3.96$ ) due to a single edge-case query (E04) where the reranker narrowed the source pool, reducing citation diversity. This illustrates a design tension. The reranker optimizes for relevance, which can conflict with the diversity needed for out-of-scope queries that require breadth of corpus characterization.

### 6.3 Retrieval Precision

Retrieval recall averaged 0.84 across the 23 queries with defined expected sources. Direct queries achieved 0.88 (10/12 with perfect recall), while edge cases scored lowest at 0.67. Notable retrieval misses include:

- D03 ("What did NASA CARA conclude about ML viability?") had 0.00 retrieval recall. The *mashiku2025* paper, the sole expected source, directly relevant to the query, was entirely absent from the top 20. The system answered from *nasa\_ca\_handbook2023* instead.
- S04 missed both *mashiku2025* and *rand\_ai\_ssa2024* (0.50 recall).

Both failures stem from vocabulary mismatch. The queries reference organizational names and abstract concepts while the relevant chunks use domain-specific technical language.

### 6.4 Judge Calibration

The initial near-perfect groundedness and citation scores raised concerns about judge leniency. Adding completeness as a third judge-scored metric provided meaningful differentiation. Manual inspection of six records confirmed that the judge scores groundedness strictly on whether stated claims are supported, not on whether the answer is comprehensive. S04 received 4/4 on groundedness while explicitly stating "I cannot provide a comprehensive answer." The completeness metric captures this gap, scoring the same response 3/4. The mechanical metrics (retrieval recall, context utilization) provide an objective cross-check independent of judge behavior.

## Machine Learning Failure Modes in Space Debris Tracking & Collision Avoidance

# 7. Failure Cases

## Failure Case 1: Retrieval Miss on Multi-Source Synthesis Query (S04)

**Query:** "What are the common reasons cited across NASA, ESA, and RAND sources for why ML has not achieved operational status in space debris tracking?"

**Category:** synthesis

### Scores

Reranked | Groundedness: 4, Citation: 4, Completeness: 3, Retrieval Recall: 0.50, Context Utilization: 0.40

Baseline | Groundedness: 4, Citation: 4, Completeness: 3, Retrieval Recall: 0.50, Context Utilization: 0.20

**What happened:** Two of the four expected sources [mashiku2025 (NASA CARA's comprehensive ML assessment) and rand\_ai\_ssa2024 (RAND's AI for space domain awareness case studies)] were absent from all 20 retrieved chunks. The system correctly flagged this gap, stating "no RAND sources are included in the evidence," and provided a partial answer from the sources it did retrieve (uriot2022, acciarini2021, nasa\_ca\_handbook2023, catulo2023). The original judge awarded 4/4 on groundedness because the claims that were made were well-supported, but the completeness metric correctly identified the gap at 3/4, and retrieval recall at 0.50 confirms the root cause was in the embedding search.

**Why it failed:** The query uses organizational names ("NASA, ESA, and RAND") and an abstract concept ("operational status"), while the relevant chunks use domain-specific vocabulary ("stochasticity of orbital mechanics," "explainability paradox," "operator-machine teaming"). The bi-encoder embeds query and document independently. It cannot resolve this vocabulary mismatch without fine-tuning or query expansion. The cross-encoder reranking could not help because the relevant chunks were never retrieved in the initial top-20.

**Potential fix:** Query decomposition. Splitting this into sub-queries per organization would likely retrieve the missing sources. Alternatively, hybrid retrieval combining vector search with BM25 keyword matching on source metadata (authors, venue) could surface organization-affiliated sources directly.

## Failure Case 2: Reranking Degrades Citation on Out-of-Scope Query (E04)

**Query:** "What role does quantum computing play in improving space debris tracking?"

**Category:** edge-case

### Scores

Reranked | Groundedness: 4, Citation: 3, Completeness: 4, Context Utilization: 0.40

Baseline | Groundedness: 4, Citation: 4, Completeness: 4, Context Utilization: 0.30

## Machine Learning Failure Modes in Space Debris Tracking & Collision Avoidance

*Note: Retrieval Recall is not applicable to this query, which is out of scope by design. There is no expected retrieval.*

**What happened:** Both versions correctly identified that quantum computing is absent from the corpus and refused to fabricate an answer. However, the reranked version scored 3/4 in Citation—the only imperfect citation score in the entire evaluation. The baseline retrieved chunks from 5 diverse sources (massimi2024, survey\_ml\_rso2024, precise\_orbit\_ml2024, uriot2022, vanslette2024) and cited them cleanly. The reranked version concentrated on only 2 sources (massimi2024, survey\_ml\_rso2024), and the judge noted that "not every chunk-content mapping is individually cited."

**Why it failed:** For out-of-scope queries, the reranker's relevance scoring actively narrows the evidence window. It demotes "off-topic" chunks rather than preserving diversity. With fewer sources to describe what the corpus does contain, the answer had less material for structured citation. This is a case where the reranker's design assumption that increased focus signals greater value conflicts with the task requirement (characterizing what evidence exists).

**Potential fix:** For queries where the system detects an out-of-scope condition (no highly relevant chunks), it could bypass reranking and use the full retrieval set to describe corpus coverage.

## Failure Case 3: Baseline Retrieves Narrower Source Set (D11)

**Query:** "What operational challenges does NASA identify for integrating ML into conjunction assessment workflows?"

**Category:** direct

**Scores:**

Reranked | Groundedness: 4, Citation: 4, Completeness: 4, Retrieval Recall: 0.50, Context Utilization: 0.80

Baseline | Groundedness: 4, Citation: 4, Completeness: 3, Retrieval Recall: 0.50, Context Utilization: 0.40  
Baseline tagged MISSED\_EVIDENCE.

**What happened:** The baseline answer drew almost exclusively from mashiku2025, producing a thorough but single-source response about data limitations, stochasticity, and explainability. The reranked version additionally surfaced nasa\_ca\_handbook2023, which contributed two novel aspects: conjunction assessment workload scaling as a function of CDM volume and the data integration complexity of receiving products from multiple commercial vendors. The new metrics quantify this difference precisely. Context utilization doubled from 0.40 (baseline) to 0.80 (reranked), meaning the reranker promoted chunks the generator actually used. Completeness improved from 3 to 4, and the baseline was the only run tagged MISSED\_EVIDENCE by the judge.

**Why it matters:** This is the strongest evidence that reranking provides value in this pipeline, despite the aggregate scores showing no improvement. The baseline retrieved nasa\_ca\_handbook2023 in positions 11–20 but the top-10 cutoff excluded it. The cross-encoder promoted it into the top 10 based on joint query-chunk relevance scoring, resulting in a more complete answer. Neither version retrieved newman2022 (an expected

Kaitlin Moore (kmoore2)

AI Model Development

### **Machine Learning Failure Modes in Space Debris Tracking & Collision Avoidance**

source), suggesting that newman2022's chunks may use vocabulary too distant from this query's framing to appear even in the top 20.

## Machine Learning Failure Modes in Space Debris Tracking & Collision Avoidance

## 8. Limitations and Next Steps

The evaluation revealed several limitations beyond what was anticipated from the architecture alone.

The most significant limitation is retrieval vocabulary mismatch. Queries framed around organizational names or abstract themes fail to retrieve chunks using technical domain vocabulary. Retrieval recall of 0.84 overall (dropping to 0.67 for edge cases) confirms this is systematic, not isolated. Query decomposition or hybrid retrieval (BM25 + vector) would address this directly.

The LLM-as-judge exhibited leniency bias on groundedness and citation correctness (near-uniform 4.00), providing no differentiation between queries. Adding completeness as a third judge-scored metric was essential for surfacing meaningful quality differences. Future evaluations should include completeness from the start.

Another limitation is that the embedding model and reranker are trained on general English rather than aerospace terminology. Semantic relationships could be strengthened with fine-tuning in this area.

Phase 3 will address user-facing gaps, implementing a UX that includes metadata filtering, research artifact generation, thread saving and export, and an integrated evaluation dashboard.