A Systematic Evaluation of Retrieval-Augmented Generation and Language Models for Space Operations

Anonymous CVPR submission

Paper ID *****

Abstract

The rapid expansion of space activities has led to an unprecedented accumulation of technical documentation, operational guidelines, and scientific literature, creating challenges for timely decision-making in space operations. Effective management in space operations requires tools capable of efficiently processing vast and heterogeneous information sources.

This paper systematically evaluates the performance of Retrieval Augmented Generation (RAG) pipelines, combining Large Language Models (LLMs) with information retrieval techniques for extracting and synthesizing actionable knowledge from domain-specific documents. We compare various retrieval strategies, embedding models, and LLM answers to assess their impact on information accuracy, relevance, and reliability.

Our results demonstrate that RAG pipelines can significantly enhance knowledge access, reduce uncertainty, and support decision-making in complex space operations.

1. Introduction

The growing complexity and cadence of modern space missions have led to a substantial increase in the amount of documentation that engineers and operators must handle. This material spans a wide range of sources, including engineering analyses, operational procedures, regulatory documentation, and scientific literature. Mission teams are therefore expected to make timely, high-impact decisions based on information that is both highly specialized and continuously evolving. Despite this, many operational workflows continue to depend on manual document inspection, legacy toolchains, and human expertise, resulting in processes that are inefficient, difficult to scale, and costly [20]. As a consequence, information handling in mission operations remains time-intensive, error-prone, and cognitively demanding for personnel [20, 23].

An example of such documentation is Concurrent Design Facility (CDF) reports, which typically span 200–300 pages and encompass a wide range of mission-related information. These documents commonly include detailed discussions of mission architecture, system integration, payload design, ground segment infrastructure, operational concepts, and risk assessments [12]. Another use case was identified at the German Space Operations Center (GSOC) of the German Aerospace Center (DLR), where operators must continuously analyze large volumes of technical data and react promptly to unanticipated events [23].

In this context, the integration of LLM-based tools offers the potential to reduce cognitive load by providing rapid, context-aware information support for engineers during both training activities and live mission operations [23].

The examples mentioned above underscore the increasing need for intelligent systems that can navigate domain-specific documentation and deliver accurate, context-aware information in real time. As space operations become more data-intensive, rapid retrieval and synthesis of relevant knowledge is now a critical requirement for mission success, operational safety, and long-term sustainability [12, 20, 23]. Using Natural Language Processing (NLP), LLMs, and artificial agents can enhance communication and decision-making for space operators, prevent potential collisions, and improve satellite trajectories [21]. Illustrating this potential, LLM-based agents have been explored for operational satellite tasks such as anomaly detection and real-time decision support [18]. Related work has also investigated the use of LLM and autonomous AI agents for spacecraft guidance and control, highlighting the expanding role of these systems across diverse space operations [3].

LLMs, when combined with RAG, offer a promising solution to the challenge of efficiently accessing, interpreting, and synthesizing large volumes of domain-specific and mission-critical documentation [1, 21, 23]. The effectiveness of RAG has also been demonstrated in other technical scientific domains, where grounding LLM outputs in authoritative sources improves factual accuracy and reliability and reduces hallucinations of the generated answers [17].

Finally, much of the data used in mission operations is

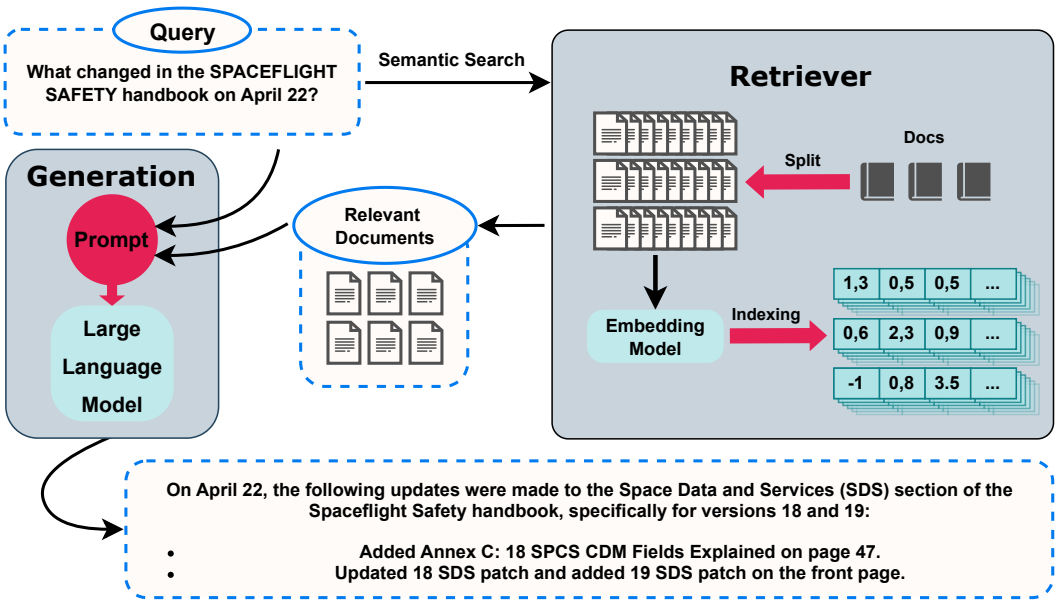

Figure 1. Overview of the information flow in a retrieval-augmented generation (RAG) pipeline for space-domain applications, illustrating document preprocessing and indexing, semantic retrieval of relevant passages, and their integration into a large language model for answer generation.

highly sensitive and unsuitable for processing via cloud-based services due to security and confidentiality constraints [18, 23]. Accordingly, this work emphasizes open-source models and frameworks that can be deployed within closed and secure environments, providing full control over data privacy and operational execution both on the ground and in space.

1.1. Related Work

The application of LLMs in space systems has attracted increasing attention as agencies face growing volumes of mission documentation. Early work adapted general-purpose models such as BERT [6] to domain-specific contexts [2] addressing the limitations of small, specialized corpora by further pretraining three models on curated space datasets relevant to space operations.

Open-domain question answering tailored to space missions has also been explored. SpaceQA [12] combined dense retrieval with neural reading to extract information from mission design documentation, highlighting the importance of domain-specific fine-tuning. Complementary efforts investigated Question and Answers (Q&A) over structured knowledge bases, translating natural language queries into formal database programs [5].

Beyond textual Q&A, multimodal approaches have been investigated, such as vision-language models fine-tuned on Martian terrain, demonstrating the potential of foundation models for scalable and autonomous space operations [10].

In operational settings, LLMs have been evaluated for rapid information retrieval and decision support at GSOC [23]. They have also been studied as expert agents for mission management, performing tasks such as data analysis, summarization, and planning under offline deployment constraints to ensure confidentiality [20].

Across these domains, a recurring challenge is the integration of external knowledge into LLMs. RAG has emerged as a widely adopted approach to improve factual accuracy and ensure adequate domain coverage, supporting both Q&A and operational decision-making [11, 20, 23]. In this context, Belo et al. [1] present an initial evaluation of RAG-based approaches for space operations, comparing two retrieval models and providing a brief qualitative analysis of the answers generated by an LLM using synthetically generated questions derived from space debris mitigation documents.

1.2. Our Work

In this paper, we extend the evaluation of key components in RAG pipelines applied to space-specific text, building on previous work by Belo et al. [1]. Consistent with recent studies [1, 23], the retrieval module (explained in Sec. 2) is of utmost importance, as the quality of the retrieved passages directly influences the quality of the generated answers. Thus, this paper analyzes retrieval models, providing insights into critical design decisions while ensuring strong retrieval effectiveness. Furthermore, we evaluate complete RAG pipelines not only to assess the factual accuracy of the generated answers, complementing prior work [1], but

also to analyze answer quality. Given the heterogeneity of subdomains in this field, our study focuses on space debris mitigation documentation and the European Space Agency (ESA) mission Q&A dataset as representative subdomains of space operations.

Our main contributions are:

1. We extend the previous dataset in [1] by adding negative passage samples per query, thereby creating a context relevance dataset (Sec. 4).
2. We evaluate three reranking models for context relevance classification, demonstrating their effectiveness in distinguishing relevant from irrelevant passages and validating their use for retrieval evaluation (Sec. 5.1).
3. We benchmark eight state-of-the-art embedding models and BM25 as retrievers within our evaluation framework using four metrics (Secs. 5.2). We further analyze relevance scores assigned by LLM over retrieved and reranked passages (Secs. **??**). Both scenarios are evaluated using different chunking strategies (512 vs. 2.000 tokens) to assess their impact.
4. We assess the factual accuracy of generated answers using a curated dataset from the ESA and further evaluate answer quality (Sec. 6).

2. How to Take Advantage of RAG to Improve LLMs Answers?

LLMs have demonstrated strong performance across a wide range of language tasks [9, 17, 19, 24], which justifies their increasing use across diverse applications. However, they continue to struggle with domain-specific or knowledge-intensive tasks [14]. RAG addresses this limitation by incorporating external knowledge relevant to the query as contextual grounding, thereby improving the reliability of LLM-generated answers [11].

The baseline RAG pipeline consists of two main components: the **Retriever** and the **Generator** [11]. The retriever identifies information that is likely to be useful for answering a given query. The generator uses this information to produce a coherent and informative response. The architecture and information flow of a typical RAG pipeline can be seen in Fig. 1.

To start, documents (e.g., guidelines, articles, papers, books, or reports) are first preprocessed by splitting them into smaller, semantically meaningful passages, called *chunks*. These chunks are then embedded and indexed. Here, embedding refers to transforming each text segment into a dense numerical vector that captures its semantic meaning, typically using a pretrained embedding model.

When a user asks a question, the retriever embeds the query in the same vector space, allowing it to select the most relevant passages based on cosine similarity, $\mathrm{sim}(\mathbf{q}, \mathbf{p}) = \frac{\mathbf{q} \cdot \mathbf{p}}{\|\mathbf{q}\| \, \|\mathbf{p}\|}$, where $\mathbf{q}$ and $\mathbf{p}$ denote the query and passage embeddings, respectively. $\mathrm{sim}(\mathbf{q}, \mathbf{p})$ yields a value in the range $[-1, 1]$, where -1 denotes the most dissimilar pair and 1 the most similar.

These retrieved passages are then provided as contextual input to an LLM, which combines them with the user query to generate a final answer. By leveraging this additional context, RAG systems combine the language understanding capabilities of LLMs with external, up-to-date information, without requiring model retraining. This makes RAG a valuable approach for improving both the accuracy and credibility of generated answers when compared to generation without retrieval [11] (Sec.6).

Beyond the core components of a RAG pipeline, additional modules may be incorporated (e.g., query rewriting, summarization, hybrid retrieval mechanisms, or prompt construction techniques) [11]. One such component is the **Reranker**, a deep learning language model that reorders retrieved passages by prioritizing the most relevant passage chunks earlier in the prompt and filtering out irrelevant ones, improving retrieval quality. However, this improvement comes at a higher computational cost [11], making rerankers unsuitable as direct replacements for retrievers in large-scale systems. In this work, we instead leverage rerankers as a tool to help us understand retriever performance.

3. Why and How We Evaluate the Retriever

We focus on the retriever, as it largely determines overall system performance due to its direct impact on downstream generation tasks [16]: without relevant and accurate context, faithful and robust responses are impossible.

In order to assess the retriever systematically without an in-domain dataset for this task, we rely on strong open-source context relevance classifiers (**Rerankers**), as a proxy relevance signal for ground truth. By doing so, we can approximate how well a retriever captures the passages that matter most for answering a query.

4. Context Relevance Dataset and Negative Mining

To assess the suitability of open-source rerankers for our domain, we employ a dataset consisting of query-relevant passage pairs, with each query paired with two irrelevant passages. We build on previous work which created a dataset containing query-passage pairs from literature related to space debris mitigation [1]. We complement this by generating negative samples, passages not relevant to the query. To generate negative samples for our tasks, similar to the approach in [22], we employ two strategies:

- **Random Negative Sampling:** We randomly sample in-domain passages that are unrelated to the source document of the synthetic query. Passage irrelevance is as-

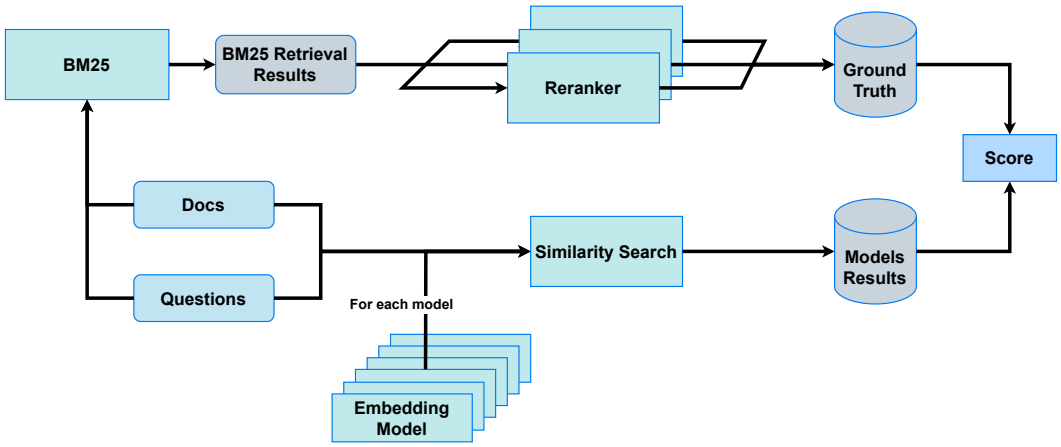

Figure 2. Embedding Models Evaluation Framework, where for each question we compute multiple passages rankings using the selected embedding models. These rankings are then compared with the reranker's ranking, obtained after an initial retrieval step with BM25. In addition, we evaluate the retrieval capabilities of BM25 against the reranking.

sessed using `BGE-M3` [4], from which we sample the least k relevant passages according to the model's scoring.

- **In-Document Negative Sampling:** This strategy follows a similar procedure to random negative sampling, but instead, we sample in-domain passages from the same document as the original passage. These passages are typically more topically similar, making them harder negatives for the model to distinguish.

4.1. Why it is not recommended to use this dataset directly for Retrieval Evaluation

There are several reasons why this dataset should not be used directly to evaluate retrieval methods. In synthetic datasets, the passage most relevant to a given query is not always the same as the passage from which the query was originally generated [15]. This misalignment makes it difficult to rely on origin passages as absolute ground truth.

Second, the dataset limits the types of evaluation metrics that can be applied. For example, metrics such as Normalize Discounted Cumulative Gain (NDCG), which depend on ranking multiple relevant passages and their scores, cannot be used reliably. This issue also affects other retrieval metrics, since retrievers may correctly prioritize different but still relevant passages.

To address this, we argue that the dataset is better suited for evaluating rerankers. Rerankers provide relevance scores over candidate passage sets, enabling the construction of ranked relevance judgments that can be used as a more reliable ground truth for retriever evaluation. A similar strategy was applied in Lee et al. [15], where an LLM was used to analyze the top-k retrieved passages and select the most relevant passage for a synthetic query to further train an embedding model. In our case, we instead use rerankers, as they provide intrinsic relevance scores for each query-passage pair that can be directly exploited for this purpose, while typically being more computationally efficient, with models often having one to two orders of magnitude fewer parameters.

5. Evaluating Retrieval Models

5.1. Validating Rerankers

Table 1. F1-Score and Accuracy comparison of BGE-M3, GTE-reranker-base, and Jina-reranker-v2 on the Context Relevance dataset.

Subset	BGE-M3		GTE		Jina	
	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy
Golden-Offset	.9395	.9463	.9568	.9618	.9543	.9602
Golden-Aligned	.9310	.9390	.9480	.9540	.9300	.9399

We selected three rerankers: BGE-M3 reranker, GTE-reranker-base, and Jina-reranker-v2. Using multiple rerankers rather than a single model helps mitigate biases[1]. Additionally, some rerankers are provided by the same organizations as certain embedding models, so including multiple rerankers helps reduce any potential advantage having a more neutral assessment of retrieval performance.

Table 1 reports F1-Score and Accuracy for each reranker. **Accuracy** measures the proportion of correctly classified contexts (relevant or irrelevant) and can be calculated as $\text{Accuracy} = \frac{\#\text{correct predictions}}{\#\text{total predictions}}$. **F1-Score** is the harmonic mean of precision and recall, given by $\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$, where $\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$ and $\text{Recall} =$

[1] Here, bias refers to systematic preferences introduced by model architecture, training data, or objective functions that may favor specific embedding models or textual patterns, potentially skewing relevance judgments.

$\frac{\text{TP}}{\text{TP}+\text{FN}}$, with TP, FP, and FN denoting true positives, false positives, and false negatives.

As shown, all rerankers demonstrate strong classification performance, indicating reliable distinction between relevant and irrelevant contexts, supporting their use to evaluate retrieval outputs in downstream tasks.

5.2. Evaluation of Embedding Models

To start our evaluation, we selected eight state-of-the-art embedding models from the Massive Multilingual Text Embedding Benchmark (MMTEB) leaderboard [7]. These embedding models can be commonly used as retrievers, and the leaderboard benchmarks them across a diverse range of tasks, serving as a standard reference in the field for evaluating new state-of-the-art models. The selection criteria, along with an overview of the embedding models used, are provided in B. In addition, we evaluated BM25 as a widely adopted retrieval method and to build our ground truth.

For our evaluation, we retrieved the top-50 passages from the entire corpus for each query using each method and model. As a proxy for ground truth, we applied a two-step procedure: first, retrieving the top-100 passages with BM25, and then classifying each passage as relevant and reranking using the rerankers. This reranked list served as our ground truth. A visualization of the procedure can be seen in Fig. 2.

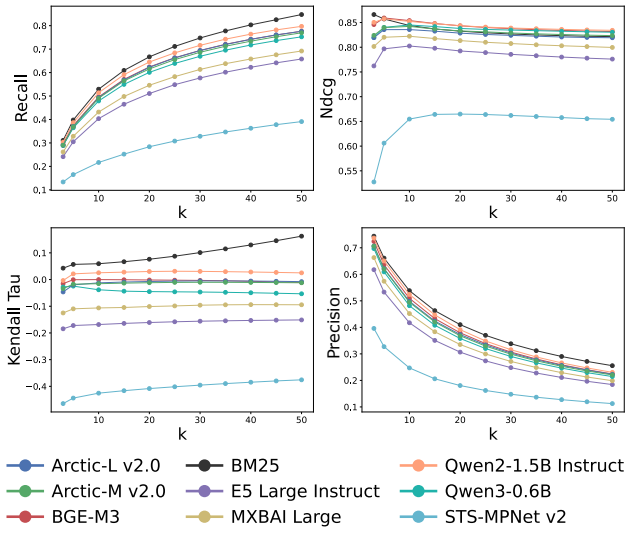

Figure 3. Average embedding model performance across retrieval metrics with 2.000-token passage chunks. BM25 and Qwen2-1.5B Instruct shows the most consistent performance across all metrics and k values. STS-MPNet v2 and E5 Large perform the worst, with the lowest scores across all metrics. The remaining models achieve comparable results, clustering closely across evaluation criteria.

We computed four distinct metrics: Recall and Precision, calculated using the relevant passages, and NDCG and Kendall Tau correlation, computed using the reranked list. Each metric was evaluated across multiple top-k thresholds. To assess the impact of document chunking on retrieval performance, we conducted the evaluation with two chunking strategies: one with a maximum chunk size of 2.000 tokens[2]. and another with 512 tokens. Results for the 2.000 token setting are shown in Fig. 3, while the results for the 512 token setting are provided in Fig. 5. The two figures show relatively similar behavior between models across the different metrics.

There are multiple ways to interpret these results. For instance, when designing a pipeline that includes a reranker, priority should be given to retrieval models with high recall and low latency (e.g., BM25), as the reranker is responsible for most of the relevance refinement. In such settings, maximizing the number of relevant passages passed to the reranker is critical, while efficiency remains important, as rerankers are significantly more computationally expensive than retrievers.

In contrast, when a pipeline relies solely on a retriever, greater emphasis should be placed on NDCG, which accounts not only for passage relevance but also for the position at which relevant passages appear within the top-k results, (e.g., BGE-M3 or Qwen 2).

5.3. What is the Impact of Raranking? Chunk Size Analysis

Previous work evaluated the relevance of each passage for each query but only with passages up to 2.000 chunk [1]. They used Llama 3.3 70B as the evaluator, since the use of LLMs for evaluation is already established in frameworks such as RAGAS [8] and is increasingly adopted in advanced retrieval setups [11], where the model acts as a judge. Recent methods further demon- strate that LLMs can reliably determine which documents to retain or discard, as well as when to actively retrieve additional information in an agentic manner [11]. We follow the same steps but also evaluate passages up to 512 tokens to see if reranking meaningfully improves the quality of retrieved content and the preference for the passage length.

We employ the same custom prompt (Fig. 6) asking the model to score each question-passage pair on a scale from 0 to 3, with 0 meaning completely irrelevant and 3 meaning highly relevant. These multi-level relevance labels have been shown to improve the performance of LLM-based classifiers [25] and we also evaluate four different top-k ranges: 3, 5, 7, and 10.

As shown in Table 2, several clear patterns emerge when comparing retriever and reranker outputs across both 2.000-token and 512-token settings.

[2]Tokens refer to the subword units produced by the tokenizer during the language model preprocessing step, rather than words or characters.

Table 2. Relative frequency (%) of relevance scores assigned by the Retriever and Reranker under 2.000-token and 512-token settings. Scores are on a 0-3 scale: 0 = completely irrelevant, 1 = slightly relevant, 2 = moderately relevant, and 3 = highly relevant. Results are reported for different Top-K retrieval ranges (3, 5, 7, 10).

Top-K	**Method**	**2.000 tokens**				**512 tokens**			
		0 ↓	**1** ↓	**2** ↑	**3** ↑	**0** ↓	**1** ↓	**2** ↑	**3** ↑
Top-3	Retriever	2.12	3.44	**54.78**	39.66	1.48	3.92	**52.06**	42.54
	Reranked	**1.29**	**1.78**	52.18	**44.76**	**0.73**	**1.78**	49.12	**48.37**
Top-5	Retriever	2.98	4.82	**59.96**	32.24	2.23	5.54	**57.64**	34.60
	Reranked	**2.13**	**2.76**	58.55	**36.56**	**1.17**	**2.71**	56.05	**40.07**
Top-7	Retriever	3.78	5.82	**62.52**	27.88	2.93	6.70	**60.39**	29.99
	Reranked	**2.65**	**3.70**	62.04	**31.61**	**1.57**	**3.62**	60.06	**34.76**
Top-10	Retriever	4.80	7.05	64.38	23.76	3.71	8.31	62.45	25.53
	Reranked	**3.51**	**4.80**	**64.87**	**26.82**	**2.13**	**4.78**	**63.46**	**29.63**

First, reranking consistently reduces the proportion of irrelevant documents (scores 0 and 1) across all top-k values, regardless of chunk size. For instance, in the 2.000-token setting at Top-3, the reranker decreases the share of completely irrelevant documents from 2.12% to 1.29% and increases the proportion of completely relevant documents (score 3) from 39.66% to 44.76%. A similar effect is observed in the 512-token setting, where top-3 reranked outputs reduce 0-score documents from 1.48% to 0.73% and increase 3-score documents from 42.54% to 48.37%. This demonstrates that reranking effectively filters out less relevant passages while boosting highly relevant ones, independent of chunk size.

Second, for moderately relevant documents (score 2), the retriever generally achieves slightly higher proportions than the reranker, particularly at lower top-k values, indicating that the reranking step is most beneficial at lower top-k thresholds, where filtering out irrelevant content has a bigger impact on overall relevance. Third, comparing relevance scores across chunk sizes, the retriever with 2.000-token passages tends to have higher values for score 2, whereas the retriever with 512-token passages consistently shows higher values for score 3, this same trend also happens in the reranker. This aligns with the intuition that longer passages may contain more noise or irrelevant content, slightly diluting the relevance of top passages, and the LLM was a preference for this setup.

Overall the reranker with 512-token passages is the most consistently dominant setting. This suggests that, in practice, shorter, more focused passages combined with reranking provide the most effective retrieval setup, with the reranker leveraging the cleaner content of smaller chunks to improve the proportion of highly relevant documents.

6. Evaluating LLM Answer Accuracy and Quality On SpaceQA

Previous work by Belo et al. [1] focused on evaluating the quality of answers generated by LLMs in the space domain. In contrast, to assess answer accuracy, we use the SpaceQA dataset created by ESA, which enables evaluation of a model's ability to produce accurate answers. This dataset was previously used by Garcia-Silva et al. [12] to assess the performance of the first open-ended question answering systems in the space domain.

This curated dataset consists of 60 question–passage–answer triplets produced by ESA. While its subdomain differs slightly from our primary focus on space debris mitigation, it includes detailed information on mission design, system configuration, payloads, service modules, ground segments, operations, and risk assessment, thereby serving as comprehensive references of ESA missions [12].

We examine whether Llama 3 8B [13] can generate accurate answers and how it performs under noisy conditions, where the context is corrupted but relevant information is still present.

Setup To evaluate the answer-generation capabilities of our LLM, we provide the correct passage in the context for each question while also introducing noise by adding four random in-domain passages from the 2.000-token set. This setup maximizes Llama 3 8B [13] ratio of irrelevant to relevant tokens within the context. For evaluation, we measure **Answer Accuracy**, which quantifies the proportion of generated answers that match the available ground-truth answers in the dataset. In addition, we assess three metrics of answer quality that are commonly used in RAG evaluations: **Answer Faithfulness**, **Answer Relevance**, and **Noise Robustness** [1, 11]. Following prior work [1, 8], these metrics are evaluated using a LLM as a judge, with each metric

scored on a 1–5 scale [25], where higher scores indicate better quality.

6.1. Results

Answer faithfulness (4.02) and relevance (4.00) indicate that generated responses are generally accurate with respect to the provided evidence and align well with the questions. The relatively high robustness score (4.73) demonstrates that the model is able to filter out irrelevant information even when four noisy passages are included in the context. Importantly, the answer accuracy of $56/60$ shows that the model not only produces plausible text but also arrives at the correct ground-truth answers in most cases. In contrast, without RAG, accuracy drops drastically to $3/60$, highlighting the critical role of retrieved context in guiding the model toward correct answers.

Taken together, these results suggest that the LLM can leverage context effectively while remaining resilient to distractor passages. However, given the limited dataset of 60 examples, these findings should be interpreted as indicative rather than definitive.

Question:
What is the size of the solar array on ATHENA?
Passage:
This results in a total of 30m2 array area, 154 kg total mass. This is assumed to be implemented as 2 fixed deployable wings, each 15m2, 77 kg.
Answer:
The context provided does not mention ATHENA or its solar array size. The text only discusses CubeSats, collision risk, and space data exchange standards.

Figure 4. Example #1 of an incorrect answer generated by the model on the SpaceQA dataset. The model fails to extract the relevant information about ATHENA's solar array size from the provided passage, likely due to a weak semantic link between the question and passage content.

A closer examination of the incorrect cases when context is provided reveals that most failures arise when the connection between the question and the supporting passage is weak or only implicit. In these instances, the passage often contains the relevant information but does not explicitly reference the entities or terms used in the question. Correctly answering, therefore, requires the model to perform bridging inferences or make assumptions that go beyond the literal text. The model, however, generally avoids such speculation and defaults to conservative responses, claiming that the answer is not present in the context. This cautious behavior reduces the risk of hallucination but limits the model's ability to extract correct answers when indirect reasoning is required. To illustrate this limitation, we present the four incorrect examples in detail below.

6.1.1. Analysis of Incorrect Answers from SpaceQA

Example 1 In Fig. 4, the passage actually provides the solar array size (30 m² in total, split into two 15 m² wings), but the connection to ATHENA is implicit. Because ATHENA is not explicitly mentioned in the passage, the model treats the information as unrelated and refuses to attribute the array size to ATHENA. This shows the model's reluctance to bridge entity references when they are not directly aligned with the question, favoring caution over inference.

Example 2 In Fig. 7, the passage explicitly mentions a launch in 2028, which is the correct answer. However, because the passage frames the launch requirement indirectly and does not repeat the mission name "ATHENA," the model overemphasizes contextual uncertainty. Instead of extracting the straightforward launch year, it digresses into a lengthy discussion of TRL readiness and cost frameworks before ultimately disclaiming that ATHENA is not mentioned. This indicates a failure to prioritize direct numerical answers when they are embedded in a broader technical discussion.

Example 3 In Fig. 8, here the model focuses on the requesting and funding entities (ESA/SRE-FM and GSP) rather than the supporting contributors (JAXA and SRON experts). The correct supporting organizations are explicitly present in the passage, but the model fails to attend to the latter part of the text. This highlights a weakness in selective information retrieval: when multiple organizational roles are described in a passage, the model locks onto the first entities mentioned and disregards subsequent ones.

Example 4 In Fig. 9, the passage does in fact state both a technical goal (demonstrating long-distance driving within a short timeframe) and a scientific goal (investigating multiple sites under similar conditions). The model fails to connect "this mission" to "MarsFAST," perhaps due to a mismatch in naming: the mission name is not explicitly repeated in the passage. Consequently, the model defaults to claiming irrelevance. This shows that the model struggles with pronoun and co-reference resolution when mission names are not reiterated in the supporting text.

It is important to note that these limitations are not solely attributable to the LLM itself. Many of the observed incorrect answers arise from situations where critical information or entity mentions are not present in the necessary chunk, making it difficult for the model to answer correctly without making assumptions or adding additional context. Such issues could potentially be addressed with alternative

Table 3. Evaluation of answer generation on the SpaceQA dataset with noisy context. Reported are average scores ± standard deviation, as well as accuracy with and without (w/o) context over the 60 examples.

	Answer faithfulness	Answer relevance	Noise robustness	Answer accuracy	Answer accuracy w/o Context
SpaceQA (Top-5)	$4.02_{[3.17;4.87]}$	$4.00_{[3.12;4.88]}$	$4.73_{[3.72;5.00]}$	56/60	3/60

chunking strategies or the inclusion of richer metadata that provides mission-specific information (e.g., mission name, mission type, or equipment used). Since the dataset was predefined, we had no control over the presence or distribution of this information. These observations highlight that retrieval design plays a critical role in overall system performance, and that careful consideration of these factors is necessary when evaluating and deploying LLM-based Q&A systems for space-domain applications.

7. Conclusion

In this work, we systematically evaluated state-of-the-art models and their use in the principal components of a RAG pipeline with the specific focus in the space domain. More specifically, we evaluate 8 different embedding models as well as the BM25 algorithm as retrieval options. We also evaluate two different retrieval strategies with and without reranking and different passage lengths. With this we learn that some models are up to the task, the use of the reranker is strongly beneficial to maximize context relevance as well as the preference for passages up to 512 tokens. We evaluate the generator with four complementary metrics, simulating from standard performance up to extreme system failure. We also complement a previously made dataset with passage-irrelevant passage triplets.

Our results show that current models are promising for controlled deployment in space operations with some retrievers and rerankers showing impresive performance, and with the LLM exhibiting both strong answer generation capabilities and robustness in scenarios with highly corrupted retrieval.

Our results show that current models are promising for controlled deployment in space operations, with some retrievers and rerankers demonstrating impressive performance, and with the LLM exhibiting both strong answer generation capabilities and robustness in scenarios with highly corrupted retrieval.

References

[1] Ruben Belo, Claudia Soares, and Marta Guimaraes. Language Modeling and Retrieval-Augmented Generation for Integration in Space Operation Decision Support Tools . In *Proceedings of the 76th International Astronautical Congress (IAC 2025)* , Sydney, Australia, 2025. International Astronautical Federation. IAC Congress Proceedings. 1, 2, 3, 5, 6

[2] Audrey Berquand, Paul Darm, and Annalisa Riccardi. SpaceTransformers: Language Modeling for Space Systems . *IEEE Access*, 9:133111–133122, 2021. 2

[3] Alejandro Carrasco, Victor Rodriguez-Fernandez, and Richard Linares. Large language models as autonomous spacecraft operators in kerbal space program . *Advances in Space Research*, 76(6):3480–3497, 2025. 1

[4] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Annual Meeting of the Association for Computational Linguistics*, 2024. 4

[5] Paul Darm, Antonio Valerio Miceli Barone, Shay B. Cohen, and Annalisa Riccardi. DISCOSQA: A knowledge base question answering system for space debris based on program induction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 487–499, Toronto, Canada, 2023. Association for Computational Linguistics. 2

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding . In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1*, pages 4171–4186, 2019. 2

[7] Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzeminski, Genta Indra Winata, et al. MMTEB: Massive Multilingual Text Embedding Benchmark . In *International Conference on Learning Representations* . International Conference on Learning Representations, 2025. 5, 10

[8] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation . In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* , pages 150–158, 2024. 5, 6

[9] Lizhou Fan, Lingyao Li, Zihui Ma, Sanggyu Lee, Huizi Yu, and Libby Hemphill. A bibliometric review of large language models research from 2017 to 2023 . *ACM Transactions on Intelligent Systems and Technology*, 15(5):1–25, 2024. 3

[10] Matthew Foutter, Daniele Gammelli, Justin Kruger, Ethan Foss, Praneet Bhoj, Tommaso Guffanti, Simone D'Amico, and Marco Pavone. Space-LLaVA: a Vision-Language Model Adapted to Extraterrestrial Applications . 2024. 2

[11] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997, 2023. 2, 3, 5, 6

[12] Andres Garcia-Silva, Cristian Berrio, Jose Manuel Gomez-Perez, Jose Antonio Martínez-Heras, Alessandro Donati, and Ilaria Roma. Spaceqa: Answering questions about the design of space missions and space craft concepts . In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* , pages 3306–3311, 2022. 1, 2, 6

[13] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 6

[14] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, pages 15696–15707, 2023. 3

[15] Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, et al. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327*, 2024. 4

[16] Alexandria Leto, Cecilia Aguerrebere, Ishwar Bhati, Ted Willke, Mariano Tepper, and Vy Ai Vo. Toward Optimal Search and Retrieval for RAG . In *Second NeurIPS Workshop on Attributing Model Behavior at Scale* , 2025. 3

[17] Wenchen Li, Su Lu, Hongqi Zhu, Peijun Wu, and Wuhe Zou. Quantum Physics Intelligent Question Answering (QA) System Based on Retrieval-Augmented Generation . *Concurrency and Computation: Practice and Experience*, 38, 2025. 1, 3

[18] Zhenping Li. Developing AI Agents for Satellite Operations . *The Journal of Space Operations Communicator*, 21:9, 2025. 1, 2

[19] Weixin Liang, Yaohui Zhang, Mihai Codreanu, Jiayu Wang, Hancheng Cao, and James Zou. The widespread adoption of large language model-assisted writing across society. *Patterns*, 6, 2025. 3

[20] Sobhana M., Syama Gudipati, and Satwik Panda. LLM based expert AI agent for mission operation management . *Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska*, 15:88–94, 2025. 1, 2

[21] Chiara Manfletti, Marta Guimarães, and Claudia Soares. AI for Space Traffic Management . *Journal of Space Safety Engineering*, 10(4):495–504, 2023. 1

[22] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems . In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* , pages 338–354, 2024. 3

[23] Clemens Schefels, Carsten Hartmann, Leonard Schlag, and Kathrin Helmsauer. Evaluating Large Language Models for Space Operations . 2024. 1, 2

[24] Ashok Urlana, Charaka Vinayak Kumar, Ajeet Kumar Singh, Bala Mallikarjunarao Garlapati, Srinivasa Rao Chalamala, and Rahul Mishra. LLMs with Industrial Lens: Deciphering the Challenges and Prospects - A Survey. *ArXiv*, abs/2402.14558, 2024. 3

[25] Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. Beyond Yes and No: Improving Zero-Shot Pointwise LLM Rankers via Scoring Fine-Grained Relevance Labels . In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)* , 2024. 5, 7

A. Rerank Evaluation extra Data

Table 4. Classification performance comparison of BGE-M3, GTE-reranker-base, and Jina-reranker-v2 on the Context Relevance dataset.

Setting	**Label**	**BGE-M3**			**GTE**			**Jina**		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Golden-Offset	Irrelevant Context	.9590	.9604	.9597	.9669	.9761	.9715	.9514	.9909	.9708
	Relevant Context	.9206	.9180	.9193	.9512	.9332	.9421	.9802	.8988	.9378
	Macro Avg	.9398	.9392	.9395	.9591	.9546	.9568	.9658	.9449	.9543
	Accuracy		.9463			.9618			.9602	
Golden-Aligned	Irrelevant Context	.9498	.9592	.9544	.9609	.9704	.9657	.9272	.9873	.9563
	Relevant Context	.9167	.8986	.9075	.9397	.9211	.9303	.9709	.8451	.9036
	Macro Avg	.9332	.9289	.9310	.9503	.9458	.9480	.9491	.9162	.9300
	Accuracy		.9390			.9540			.9399	

B. Model Information and Selection Criteria

Table 5. Comparison of selected embedding models. Columns show: **Organization** (model developer), **Model Name**, **Parameters** (trainable weights), **Context Length** (maximum tokens per input), and **Embedding Dimension** (vector size). Performance metrics from the MMTEB leaderboard [7] include: **Clustering** (grouping similar texts), **Reranking** (reordering retrieved results), **Retrieval** (finding relevant documents), and **STS** (Semantic Textual Similarity).

Org.	**Model**	**Params**	**Context**	**Dim**	**Clust.**	**Rerank**	**Retr.**	**STS**
Gameselo	STS-MPNet v2	278M	514	768	32.38	47.43	34.66	69.33
intfloat	E5 Large Instruct	560M	514	1024	50.75	**76.81**	62.61	57.12
Qwen	Qwen3-0.6B	595M	32768	1024	**52.33**	60.26	**64.65**	**74.94**
Alibaba-NLP	Qwen2-1.5B Instruct	1B	32768	8960	52.05	61.61	60.78	70.36
BAAI	BGE-M3	568M	8194	1024	40.88	61.98	54.60	72.99
Snowflake	Arctic-M v2.0	305M	8192	768	42.24	60.45	54.83	65.91
Snowflake	Arctic-L v2.0	568M	8192	1024	42.76	62.89	58.36	69.48
mixedbread-ai	MXBAI Large	335M	512	1024	42.49	44.51	40.30	61.20

Our selection criteria focused on open-source models with fewer than 2 billion parameters, ensuring diversity in model size, embedding dimensions, as well as differences in architecture and spanning different research groups and organizations. The selected embeddings can be seen in Table 5.

C. Evaluating Retrieval Models Extra

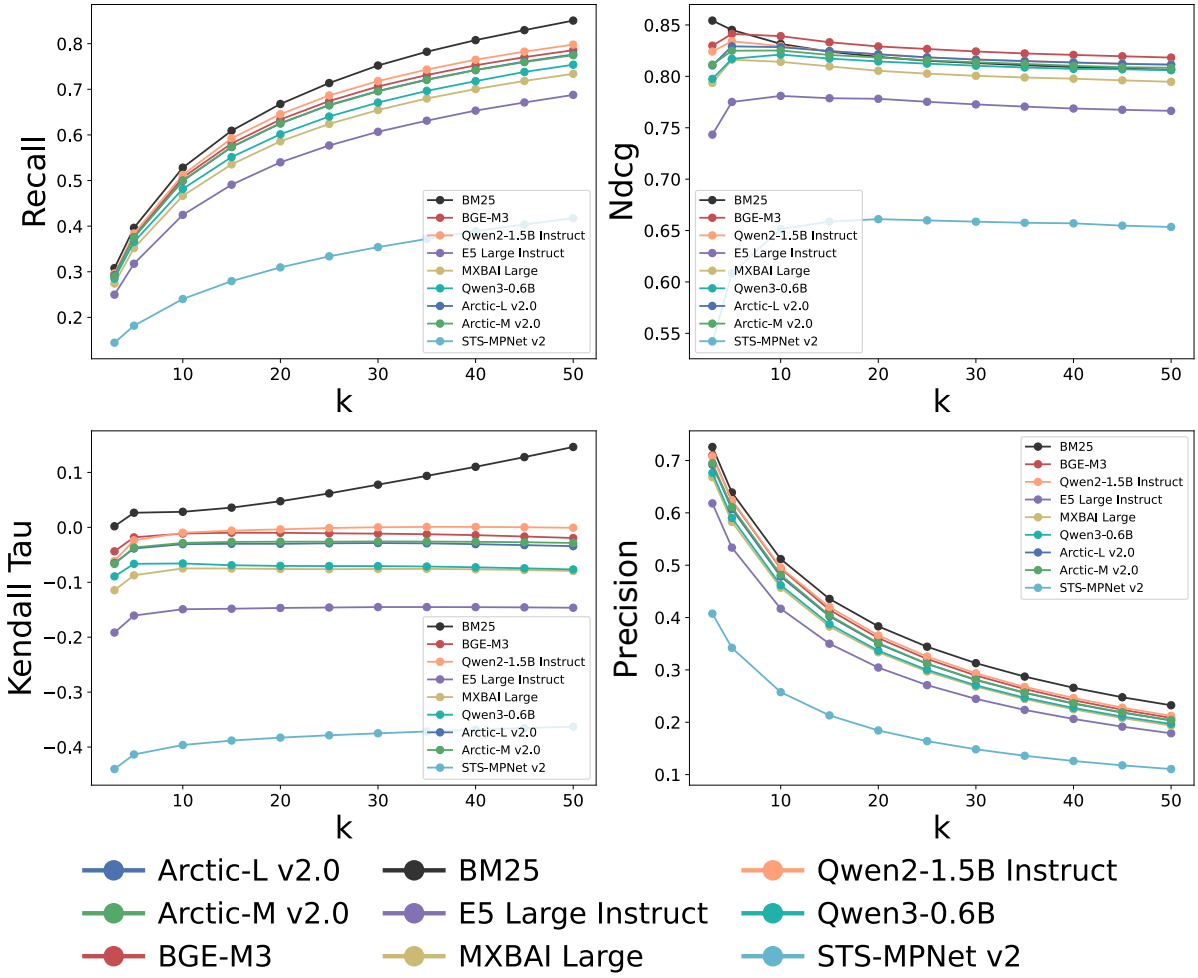

Figure 5. **Average Embedding model performance across retrieval metrics with 512-token document chunks:** BM25 and Qwen2-1.5B Instruct show the most consistent performance across all metrics and k values. STS-MPNet v2 and E5 Large perform worst, with the lowest scores across all metrics. The remaining models achieve comparable results, clustering closely across evaluation criteria.

D. Prompts used for evaluation

System prompt:
You are an AI assistant that judges the relevance of a document to a given question. Respond with a score from 0 to 3.
QUESTION: User question
DOCUMENT: doc
Rate how relevant the document is to answering the question using the following scale:
0 = Completely irrelevant
1 = Slightly irrelevant
2 = Slightly relevant
3 = Completely relevant
Respond with a single number between 0 and 3.

Figure 6. **Context Relevance Prompt:** Used to test both the Retriever and the Reranker.

E. Incorrect Answers in SpaceQA

Question:
When will ATHENA be launched?
Passage:
The main challenge for this project is to find technical solutions fitting to the cost frame work. Next to that all technologies with low TRL need to be advance such that they reach TRL 6 (RD[47]) before the start of the implementation phase. This could become difficult, although the launch is only required in 2028, because some parts of the X-Ray Field Unit are reported to be only at TRL 2 presently.
Answer:
Based on the provided context, it seems that the main challenge is to find technical solutions that fit within a cost framework and advance technologies with low TRL (Technology Readiness Level) to TRL 6 before the implementation phase. This is important because some parts of the X-Ray Field Unit are currently at TRL 2 and need to be advanced before the launch in 2028.
To address this challenge, it would be helpful to identify the specific technologies that need to be advanced and prioritize them based on their potential impact on the project's success. Additionally, it may be necessary to allocate additional resources or funding to support the development of these technologies.
It's also important to note that the project has a specific timeline and budget constraints, so it's crucial to manage the project's scope, schedule, and budget effectively to ensure its successful completion.
Regarding the question about ATHENA's launch, it's not mentioned in the provided context. ATHENA is not a project mentioned in the document, so it's not possible to provide an answer to this question based on the provided information.

Figure 7. Example #2 of an incorrect answer generated by the model on the SpaceQA dataset. The passage explicitly mentions a launch in 2028, but the model fails to extract this direct numerical information and instead focuses on contextual uncertainties.

Question:
Who supported ESA in the NG-CryoIRTel study?
Passage:
The study was requested by ESA Science Directorate SRE-FM and funded by the General Studies Programme GSP. The study was carried out in 8 sessions, starting with a kick off on the 13th November 2014 and ending with an Internal Final Presentation on the 18th December 2014 by an interdisciplinary team of specialists from ESTEC and ESOC and supported by SPICA/SAFARI experts from JAXA and SRON.
Answer:
The study was requested by ESA Science Directorate SRE-FM and funded by the General Studies Programme GSP.

Figure 8. Example #3 of an incorrect answer generated by the model on the SpaceQA dataset. The model fails to identify the entities supporting ESA in the NG-CryoIRTel study and instead reports who requested and funded the study, demonstrating confusion between different organizational roles described in the passage.

Question:
What is the goal of the MarsFAST mission?
Passage:
The technical demonstration goal of this mission, to cover a long distance drive within a few weeks only, supports the scientific goal to investigate multiple and different places at similar boundary (= weather) conditions. However, the possible requirement to enter "difficult" terrain like the edges of dune field or potentially soft old river beds must be carefully analysed and traded against the inherent risks.
Answer:
The context provided does not mention the MarsFAST mission. The text appears to be a technical paper discussing various topics such as statistical sampling, Monte Carlo simulations, and uncertainty quantification in space situational awareness (SSA) and space debris mitigation. There is no mention of a specific mission called MarsFAST. If you meant to ask a different question, please feel free to rephrase it, and I'll do my best to assist you.

Figure 9. Example #4 of an incorrect answer generated by the model on the SpaceQA dataset. Although the passage clearly describes both the technical and scientific goals of the mission, the model fails to associate "this mission" with MarsFAST and incorrectly concludes that the mission is not mentioned, highlighting a failure in co-reference resolution.