

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

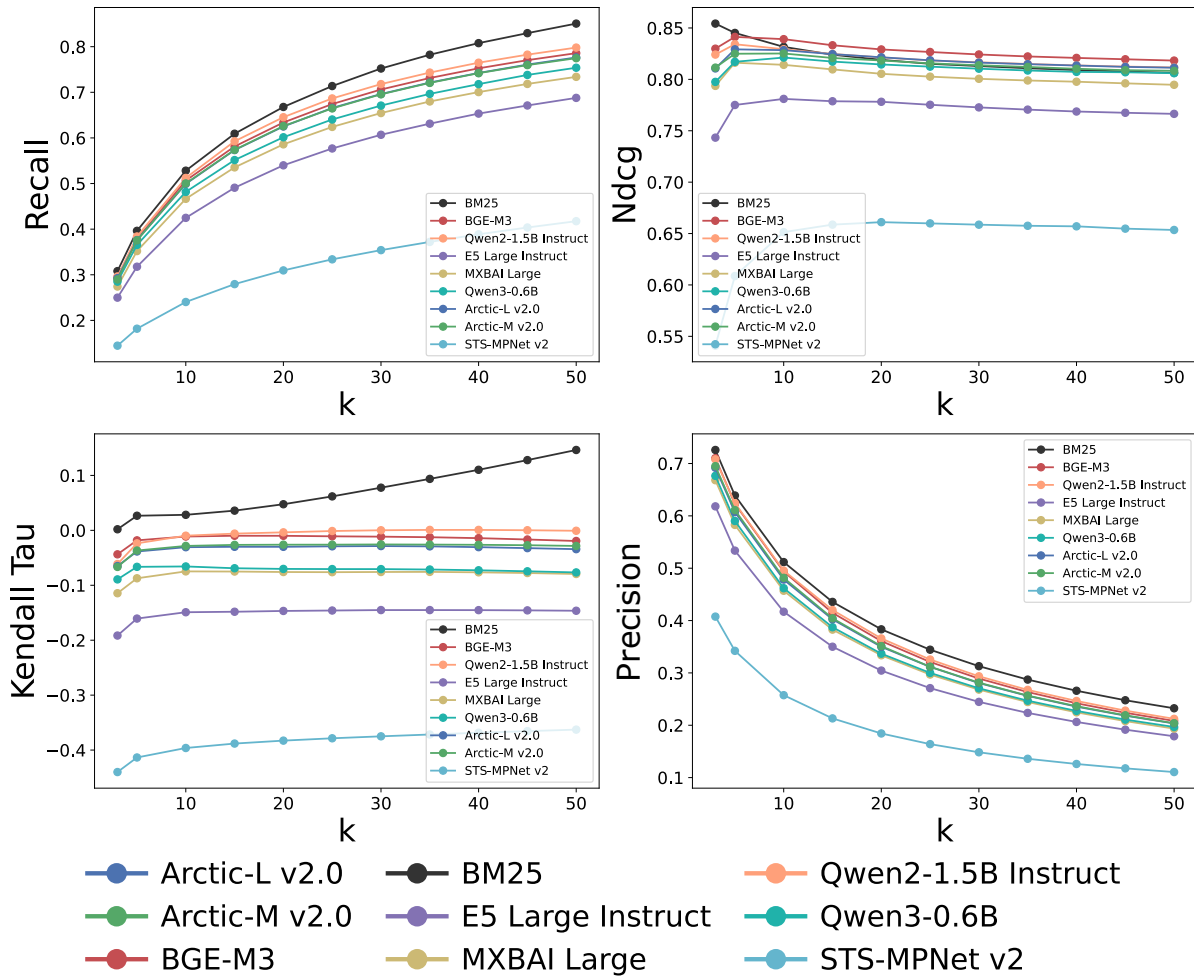


Figure 5. **Average Embedding model performance across retrieval metrics with 512-token document chunks:** BM25 and Qwen2-1.5B Instruct show the most consistent performance across all metrics and k values. STS-MPNet v2 and E5 Large perform worst, with the lowest scores across all metrics. The remaining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