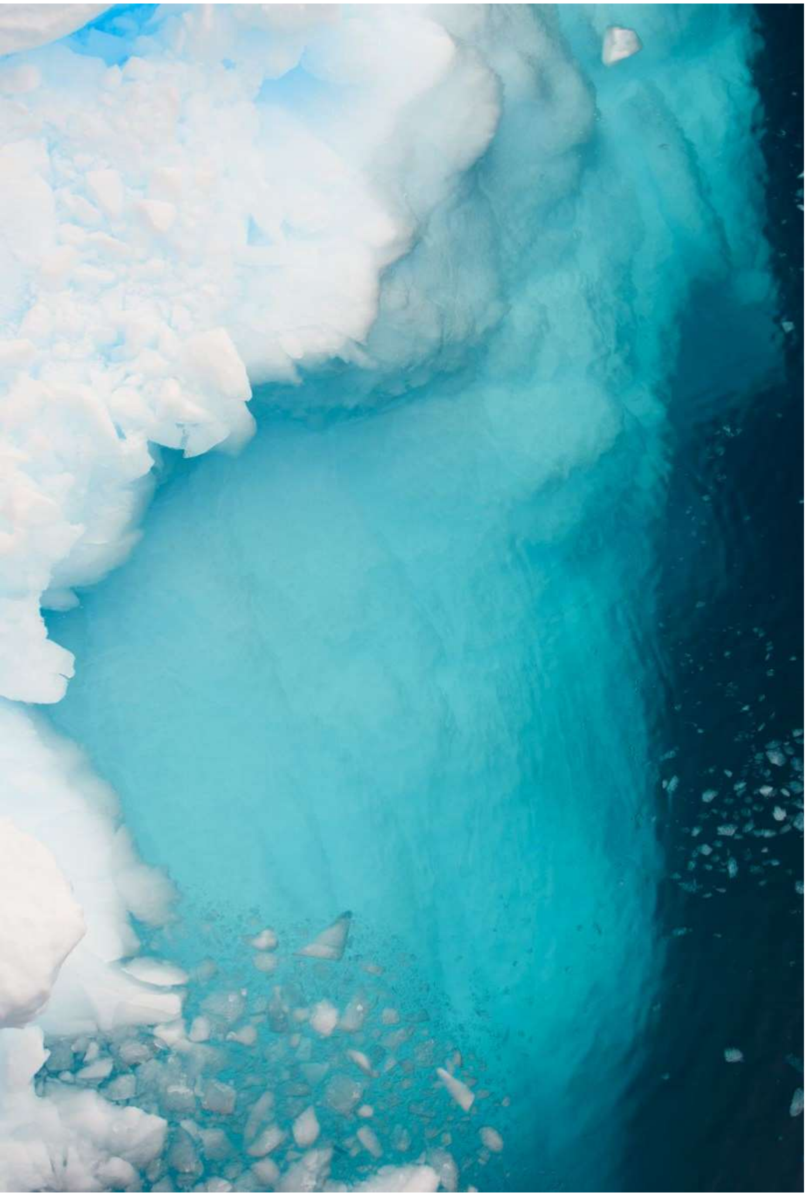


Space Science Information Retrieval System

BY RUCHIKA SANCHETI

Problem Statement

When you don't feel good, you may choose to visit your doctor. When you do a research, you can go to meet your supervisor. But when you realize that they are too busy engaged with so much small matters, you may want to take some of their work and try to do something by yourself? Then you need information. Can you read ten relevant medical books in a day? This is neither possible nor necessary. The information need is just a piece of pertinent information. Now, IR is such a technology that you cannot ignore!!

A model that can read and comprehend the meaning of language from the internet is a vital component of the semantic web. There are already models that can do this within a limited scope. However, there is a problem. These Language Models (LMs) need to learn before becoming these autonomous, language-comprehending IAs. It is difficult to train LMs because they require massive amounts of data. We can easily fine-tune a model in places where there is a large amount of relevant and labelled data. The issue with labelled data is that it must (almost always) be created by hand.

So we basically need a system that can answer queries specific to a domain without having to hand label data

Approach- GPL/SBERT/LFQA

To solve the issue of labelled data we use GPL. At its core, it allows us to take unstructured text data and use it to build models that can understand this text. These models can then intelligently respond to natural language queries regarding this same text data. Here GPL is used as a technique for domain adaptation of an already fine-tuned bi-encoder model (such as SBERT).

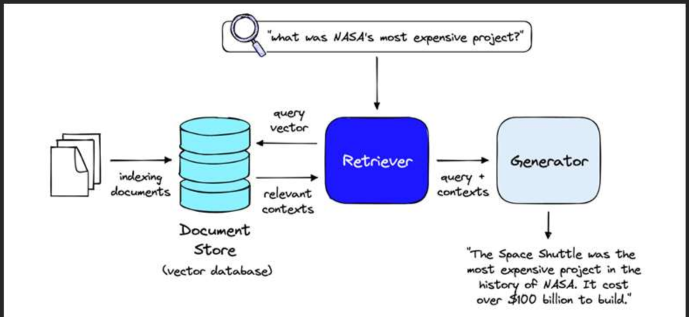
We then compare the answers returned when passing Astrophysics queries with the initial model (without GPL adaptation) as well the trained model.

Dataset Used: ArXiv Astrophysics
Initial Model: msmarco-distilbert-base-tas-b
Query Generation Model: doc2query/msmarco-t5-base-v1
Cross Encoder : ms-marco-MiniLM-L-6-v2
Loss Function : Margin MSE
Vector Database : Pinecone

LFQA- Long-Form Question Answering using BART

LFQA can produce multi-sentenced abstractive (generated) replies to open-ended questions, to provide extremely precise answers to inquiries about astrophysics from a given context. It operates by looking through vast document repositories for documents that contain relevant information, then synthesising an accurate multi-sentence response utilising this information. The relevant documents provide a wider background for developing unique, abstract long-form responses

Three main components are needed to build a LFQA pipeline in Haystack: DocumentStore, Retriever, and Generator.


"what was NASA's most expensive project?"
query vector
indexing documents
relevant contexts
Retriever
query + contexts
Generator
Document Store
(vector database)
"The Space Shuttle was the most expensive project in the history of NASA. It cost over $100 billion to build."

Superior Results with domain adapted model

Query	Model	Answer	Context 1	Context 2	Context 3
What guaranteed the presence of water in the moon?	Base Model	and nasa ' s lunar reconnaissance orbiter data	in the same issue of nature astronomy, scientists have published a paper using theoretical models and nasa ' s lunar reconnaissance orbiter data, pointing out that water could be trapped in small shadows, where temperatures stay below freezing, across more of the moon than currently expected. the results can be found here.	water is a precious resource in deep space and a key ingredient of life as we know it. whether the water sofia found is easily accessible for use as a resource	sofia has detected water molecules (h2o) in clavius crater, one of the largest craters visible from earth, located in the moon â€™ s southern hemisphere. previous observations of the moon
	Pretrained Model	The Moon is tidally locked to the Earth. This means that it orbits the Sun at the same time as the Earth, so it is always facing the same direction.	Under NASA's Artemis program, the agency is eager to learn all it can about the presence of water on the Moon in advance of sending the first woman and next man to the lunar surface in 2024 and establishing a sustainable human presence there by the end of the decade.	"We had indications that H2O – the familiar water we know – might be present on the sunlit side of the Moon," said Paul Hertz, director of the Astrophysics Division in the Science Mission Directorate at NASA Headquarters in Washington. "Now we know it is there.	Previous observations of the Moon's surface detected some form of hydrogen, but were unable to distinguish between water and its close chemical relative, hydroxyl (OH).
	Finetuned model	It's not that there's no water on the Moon, it's that the Moon's atmosphere is so thin that it doesn't reflect much of the Sun's light. This means that the only way for water to get to the surface of the Moon is for it to evaporate.	Under NASA's Artemis program, the agency is eager to learn all it can about the presence of water on the Moon in advance of sending the first woman and next man to the lunar surface in 2024 and establishing a sustainable human presence there by the end of the decade.	Another possibility is that the water could be hidden between grains of lunar soil and sheltered from the sunlight – potentially making it a bit more accessible than water trapped in beadlike structures.	Previous observations of the Moon's surface detected some form of hydrogen, but were unable to distinguish between water and its close chemical relative, hydroxyl (OH).

Model	Retriever	Generator/Reader
Base Model	BM25	BERT
Pretrained Model	Pretrained SBERT	BART
Finetuned model	Domain adapted SBERT	BART