Document Retrieval System For GSI

BY TEAM: GeoManthan AI

Problem Context:

Objective Of RAG In GSI context:

The objective of Retrieval-Augmented Generation (RAG) in the context of thousands of GSI (Geological Survey of India) documents is to enable fast, accurate, and context-aware access to critical geological information (maps, mineral occurrences, stratigraphy, reports etc.) — without requiring geoscientists to manually sift through massive amounts of unstructured data, by combining:

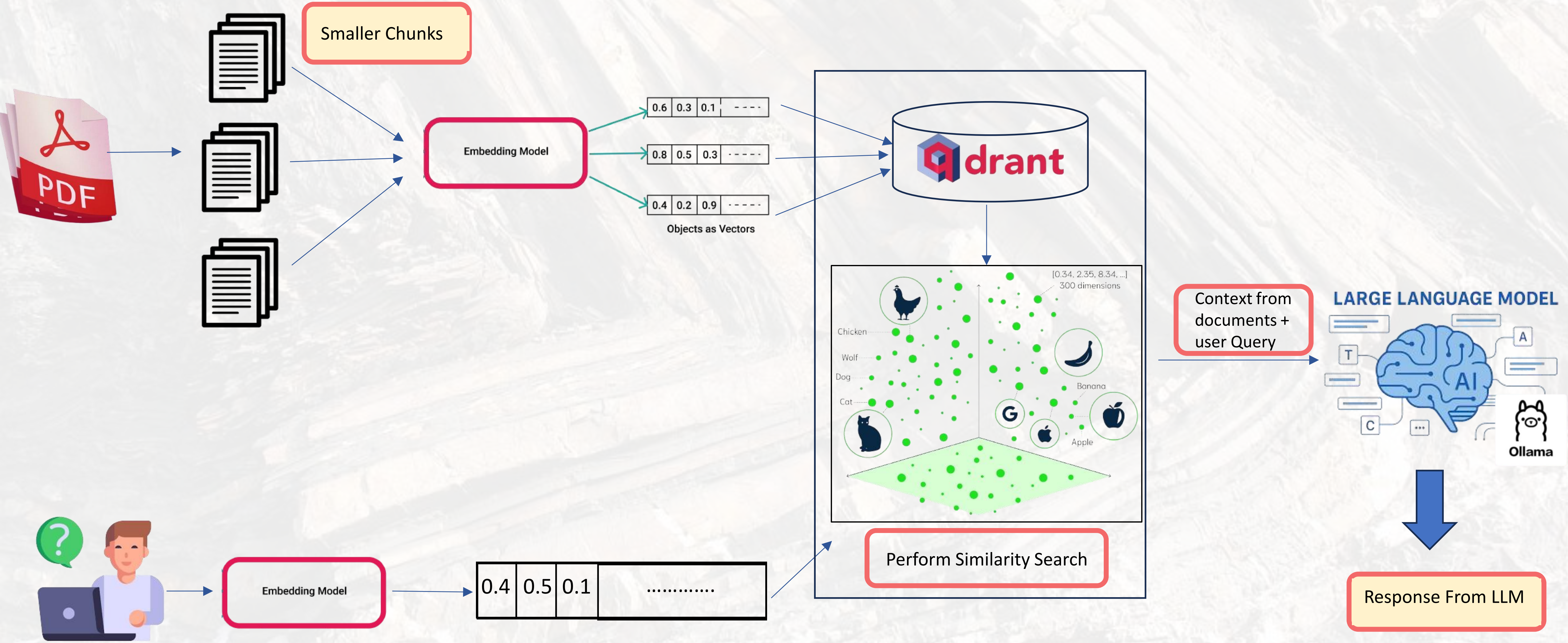
- **Document Retrieval** – Efficiently fetch the most relevant documents or sections from a large corpus.
- **Language Generation** – Use a language model to summarize, answer questions, or provide context-aware explanations from those documents.

3 Knowledge is often buried in **legacy reports**, **scanned documents**, or **fragmented data formats**.

2 Experts spend hours/days manually reading PDFs, field notes, maps, etc.

1 Thousands of reports: regional geology, mineral occurrences, magnetic/gravity surveys, exploration logs, etc.

High Level Design Of System



Smaller Chunks

Embedding Model

0.6 | 0.3 | 0.1 | - - - -

0.8 | 0.5 | 0.3 | - - - - -

0.4 | 0.2 | 0.9 | - - - - -

Objects as Vectors

Context from documents + user Query

Embedding Model

0.4 | 0.5 | 0.1 | ....

Perform Similarity Search

Response From LLM

Installation of Application

GeoManthanAIApp_IndiaAI_HACKATHON-2025.exe is uploaded on a Google Drive click on “Download Anyway” to install it on the system.

The link has been also provided in the Report. Download and install it.

LINK: https://drive.google.com/file/d/1MmUPyBJkxGVVLslo4wsL-6cYuoRMdM3A/view?usp=sharing

Note: Application is tested on Windows11

Prerequisite: 1

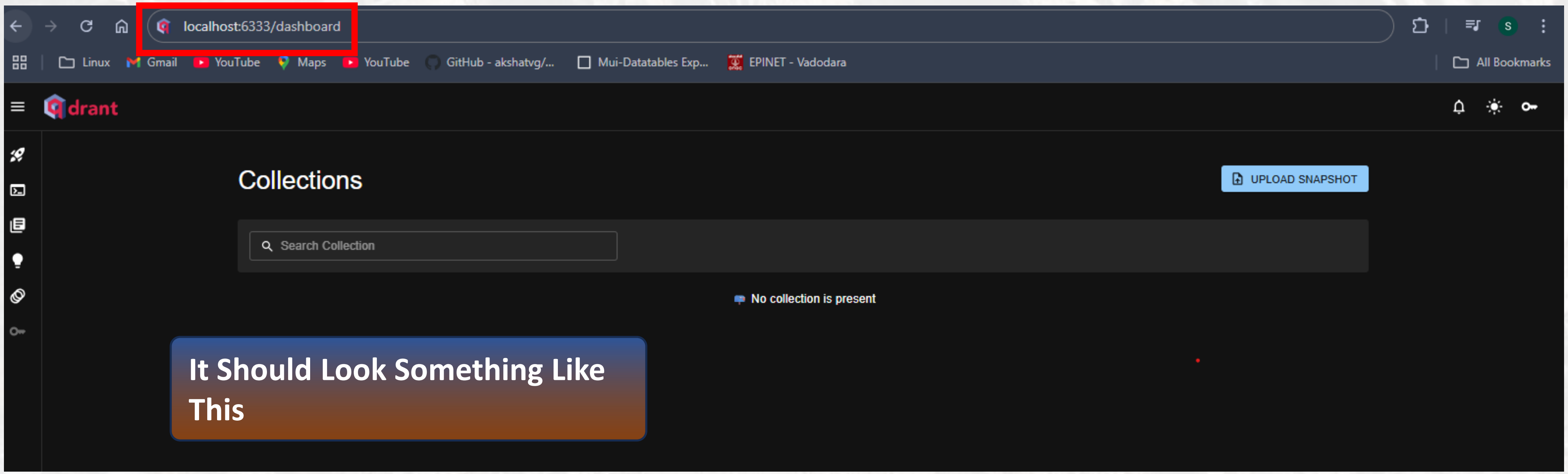
1

Install Vector Database (This Application only supports Qdrant Database)

Running via Docker:

```
docker run -d --name qdrant -p 6333:6333 -v C:\qdrant_data:/qdrant/storage qdrant/qdrant
```

(This will Start Qdrant database, check by going to browser and type: **http://localhost:6333/dashboard**)



Collections

No collection is present

It Should Look Something Like This

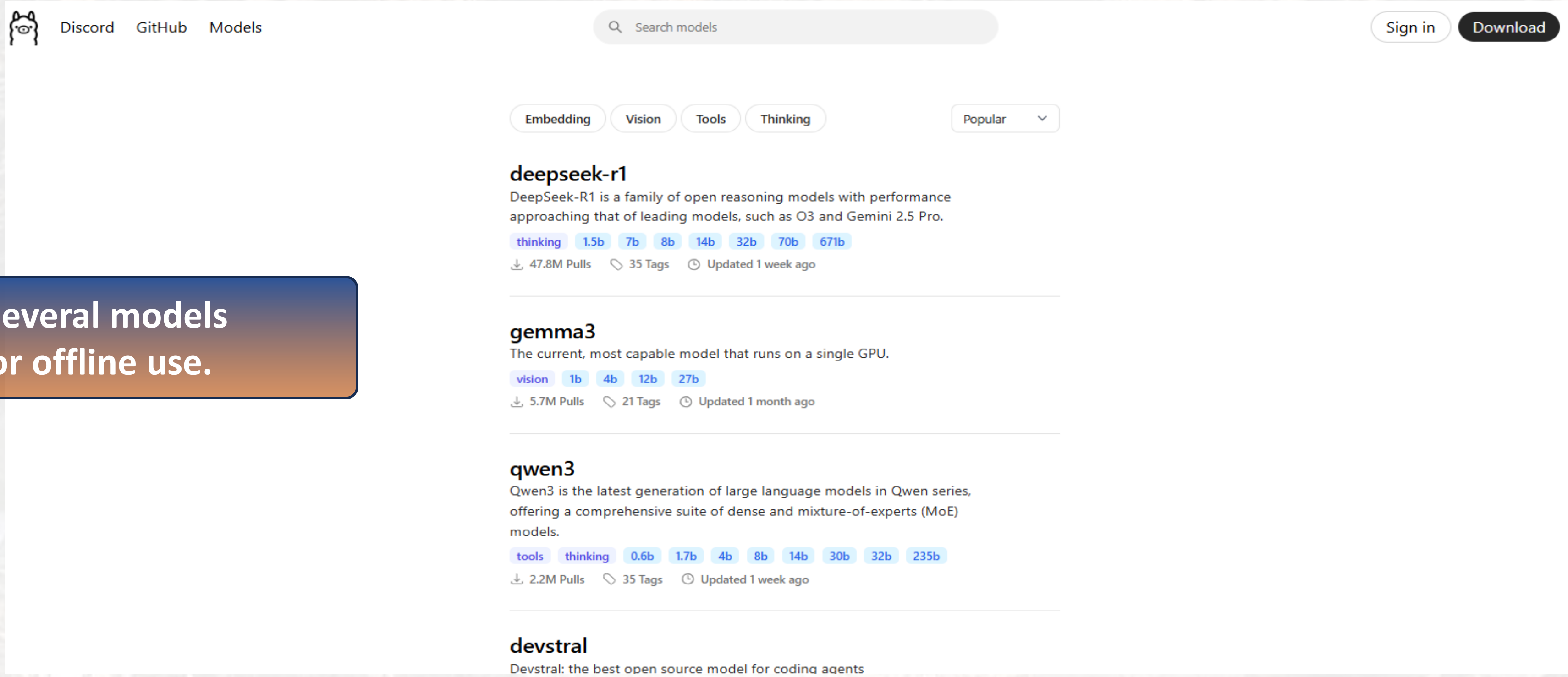
Prerequisite: 2

2

Local Setup: **Install Ollama and pull any LLM from https://ollama.com/search depending upon pc/workstation capacity**
Running via Ollama : ollama run phi4

Enterprise LLM Models: This application supports integration with all the major LLM providers (both opened or closed) through ChatCompletion API.

There are several models available for offline use.



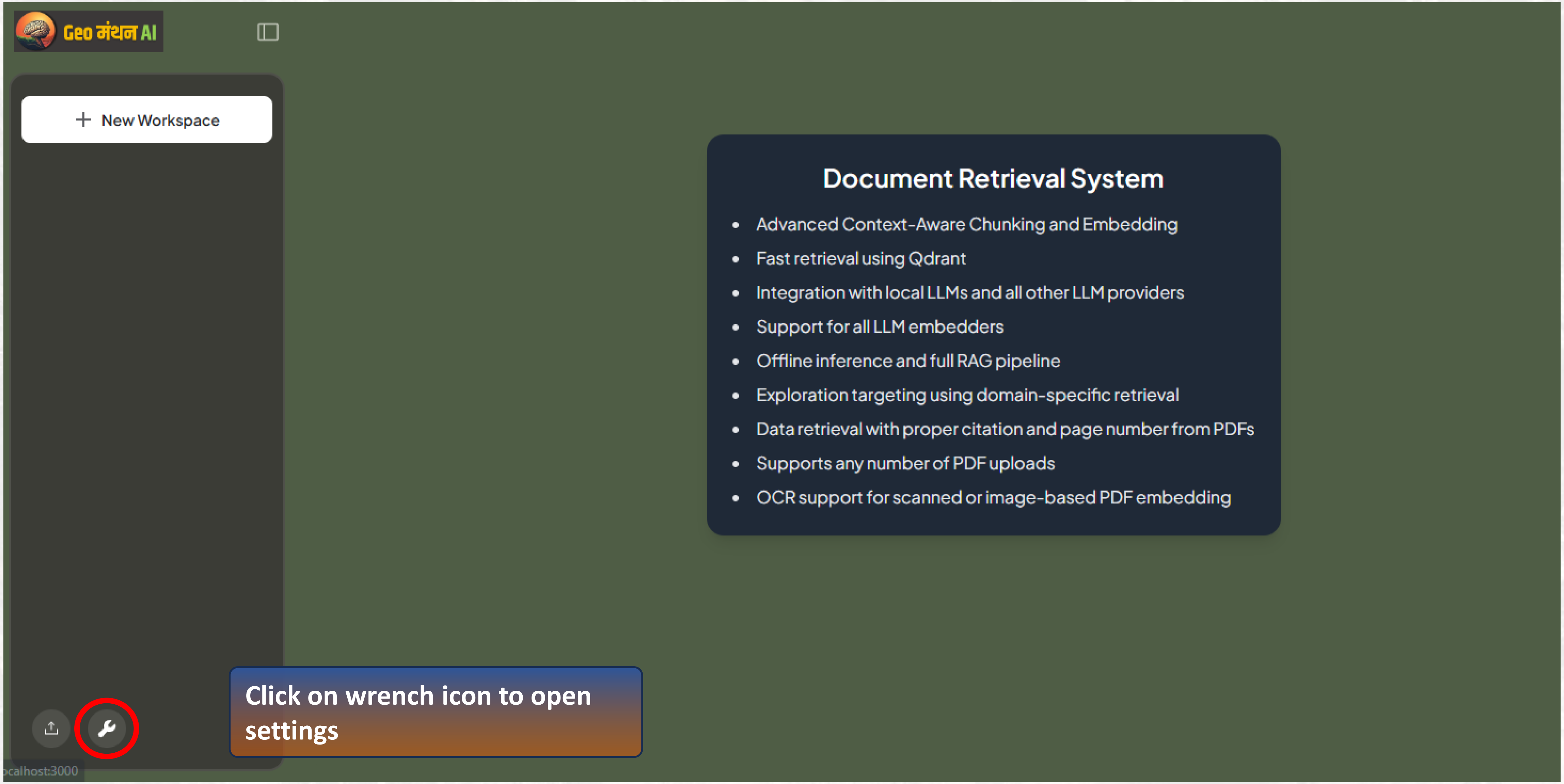
deepseek-r1
DeepSeek-R1 is a family of open reasoning models with performance approaching that of leading models, such as O3 and Gemini 2.5 Pro.
thinking 1.5b 7b 8b 14b 32b 70b 671b
47.8M Pulls 35 Tags Updated 1 week ago

gemma3
The current, most capable model that runs on a single GPU.
vision 1b 4b 12b 27b
5.7M Pulls 21 Tags Updated 1 month ago

qwen3
Qwen3 is the latest generation of large language models in Qwen series, offering a comprehensive suite of dense and mixture-of-experts (MoE) models.
tools thinking 0.6b 1.7b 4b 8b 14b 30b 32b 235b
2.2M Pulls 35 Tags Updated 1 week ago

devstral
Devstral: the best open source model for coding agents

Setting-Up Application before Use



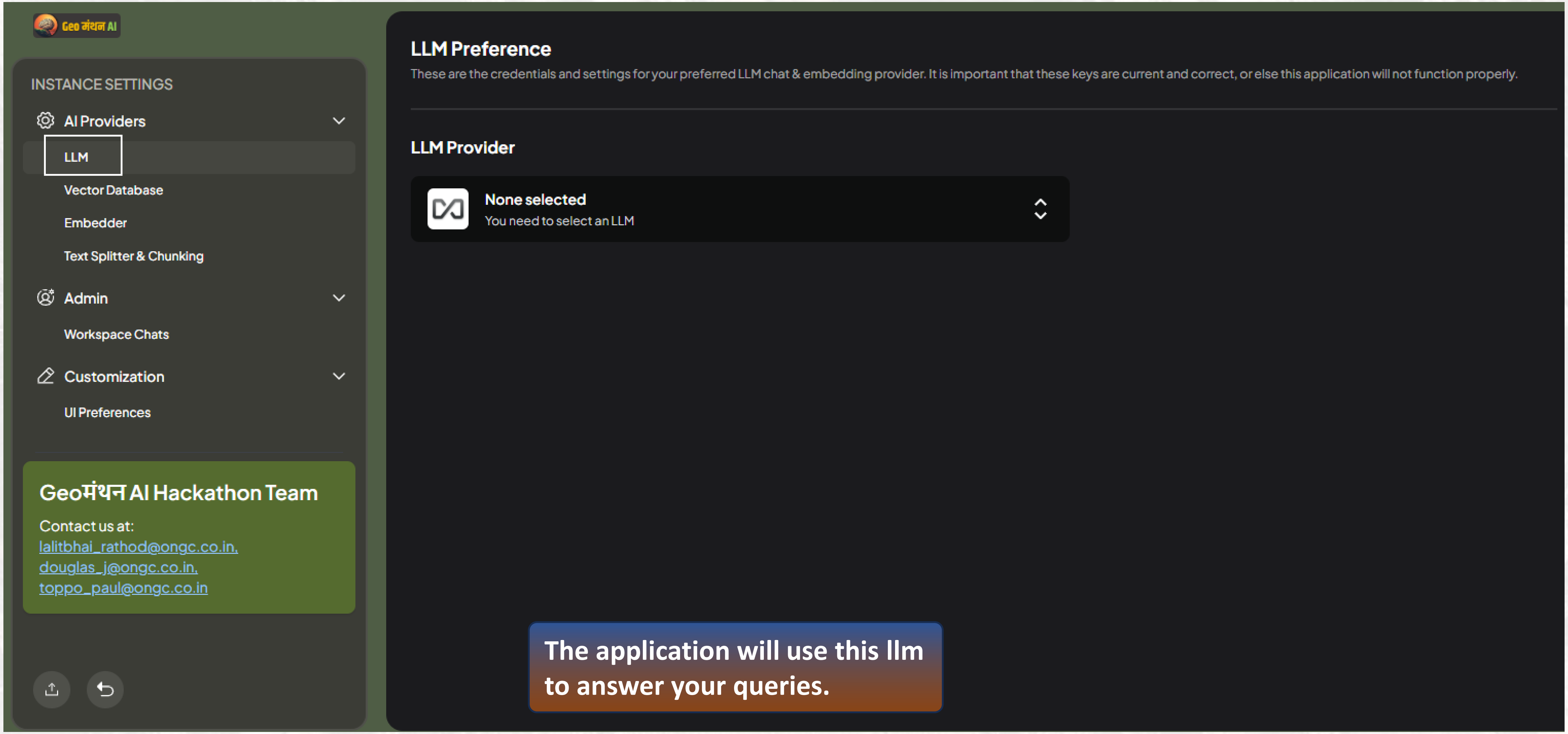
Document Retrieval System

- Advanced Context-Aware Chunking and Embedding
- Fast retrieval using Qdrant
- Integration with local LLMs and all other LLM providers
- Support for all LLM embedders
- Offline inference and full RAG pipeline
- Exploration targeting using domain-specific retrieval
- Data retrieval with proper citation and page number from PDFs
- Supports any number of PDF uploads
- OCR support for scanned or image-based PDF embedding

Click on wrench icon to open settings

SAGEFOX

Setting-Up LLM Provider



Geoमंथन AI Hackathon Team

Contact us at:
lalitbhai_rathod@ongc.co.in,
douglas_j@ongc.co.in,
toppo_paul@ongc.co.in

LLM Preference

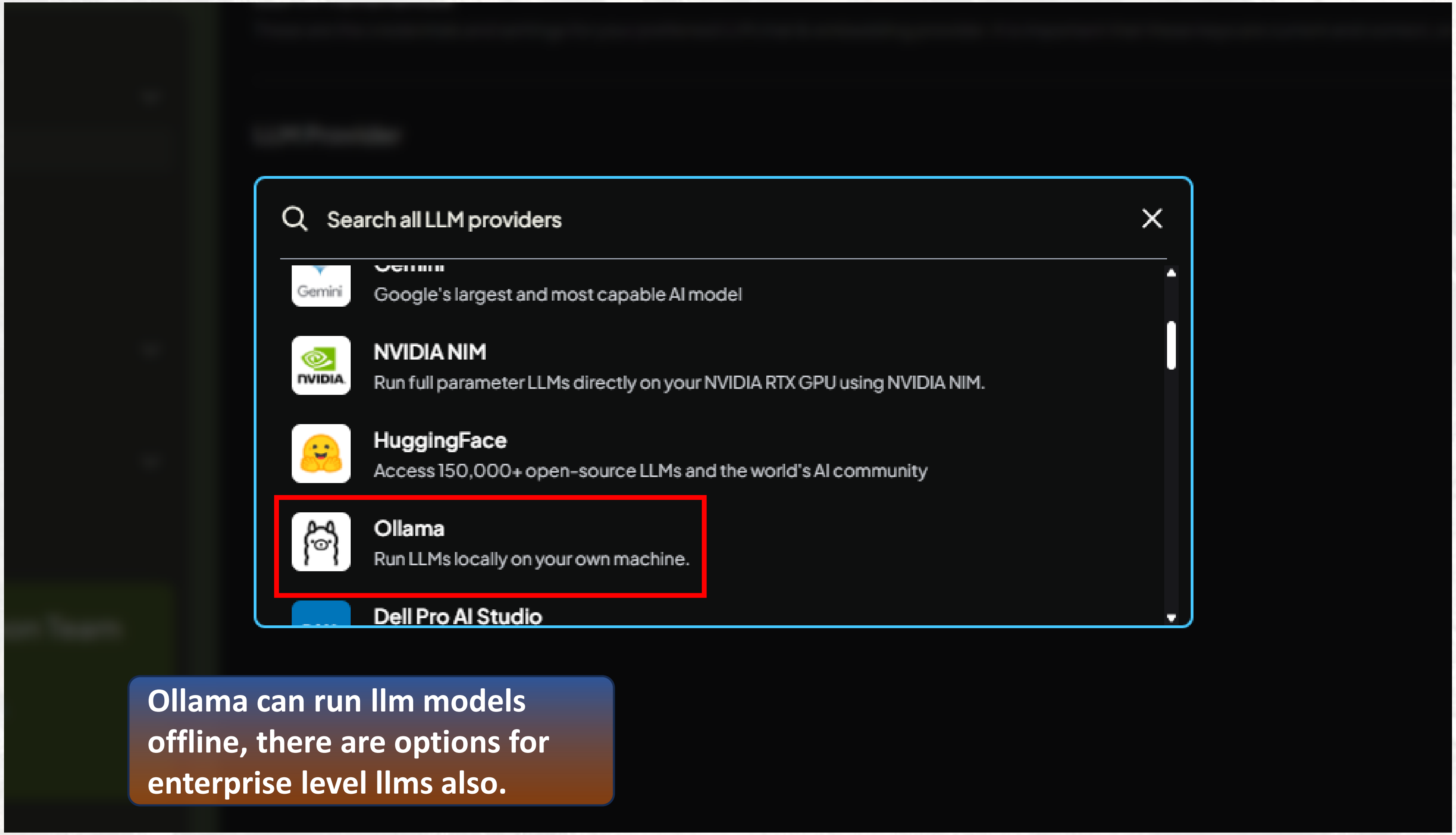
These are the credentials and settings for your preferred LLM chat & embedding provider. It is important that these keys are current and correct, or else this application will not function properly.

LLM Provider

None selected
You need to select an LLM

The application will use this llm to answer your queries.

Support for both local LLM and Enterprise LLMs



Search all LLM providers

Google's largest and most capable AI model

NVIDIA NIM
Run full parameter LLMs directly on your NVIDIA RTX GPU using NVIDIA NIM.

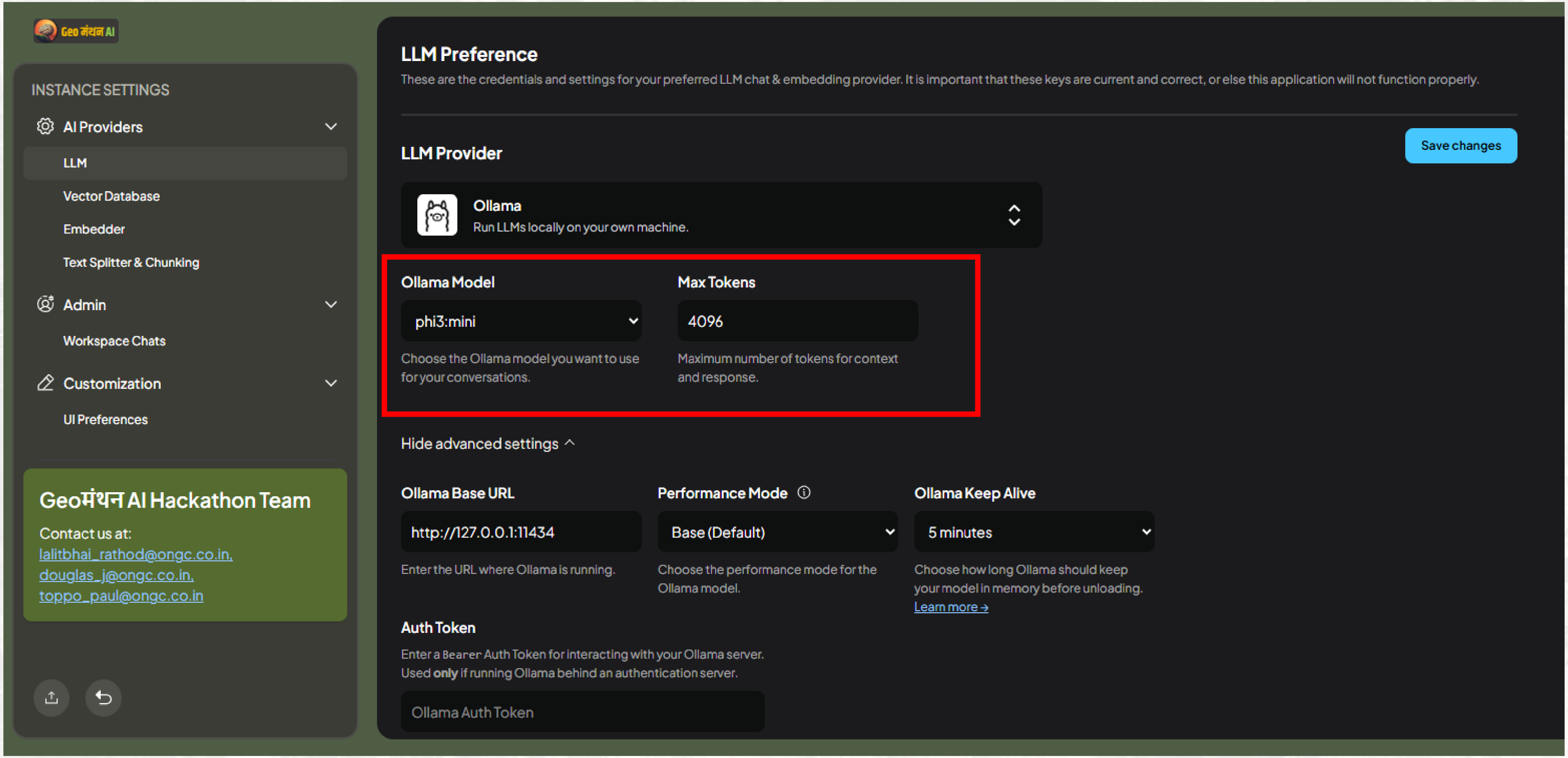
HuggingFace
Access 150,000+ open-source LLMs and the world's AI community

Ollama
Run LLMs locally on your own machine.

Dell Pro AI Studio

Ollama can run llm models offline, there are options for enterprise level llms also.

Setting Ollama as local LLM Provider



LLM Preference

These are the credentials and settings for your preferred LLM chat & embedding provider. It is important that these keys are current and correct, or else this application will not function properly.

LLM Provider

Ollama
Run LLMs locally on your own machine.

Ollama Model: phi3:mini
Choose the Ollama model you want to use for your conversations.

Max Tokens: 4096
Maximum number of tokens for context and response.

Ollama Base URL: http://127.0.0.1:11434
Enter the URL where Ollama is running.

Performance Mode: Base (Default)
Choose the performance mode for the Ollama model.

Ollama Keep Alive: 5 minutes
Choose how long Ollama should keep your model in memory before unloading. Learn more →

Auth Token
Enter a `Bearer` Auth Token for interacting with your Ollama server.
Used **only** if running Ollama behind an authentication server.

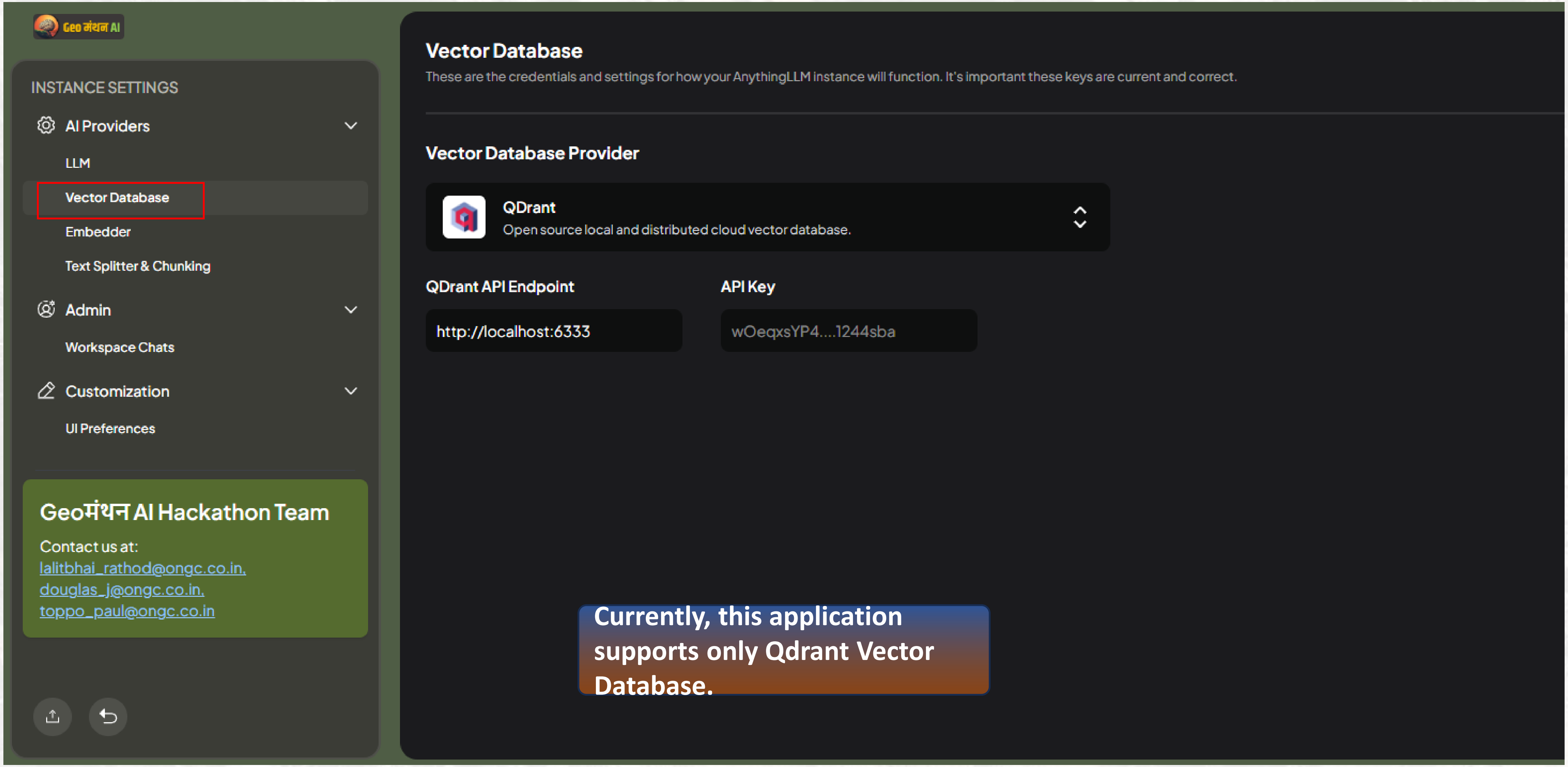
Ollama Auth Token

Geoमंथन AI Hackathon Team

Contact us at:
lalitbhai_rathod@ongc.co.in,
douglas_j@ongc.co.in,
toppo_paul@ongc.co.in

SAGEFOX

Setting-Up Vector Database



Geoमंथन AI Hackathon Team

Contact us at:
lalitbhai_rathod@ongc.co.in,
douglas_j@ongc.co.in,
toppo_paul@ongc.co.in

Vector Database

These are the credentials and settings for how your AnythingLLM instance will function. It's important these keys are current and correct.

Vector Database Provider

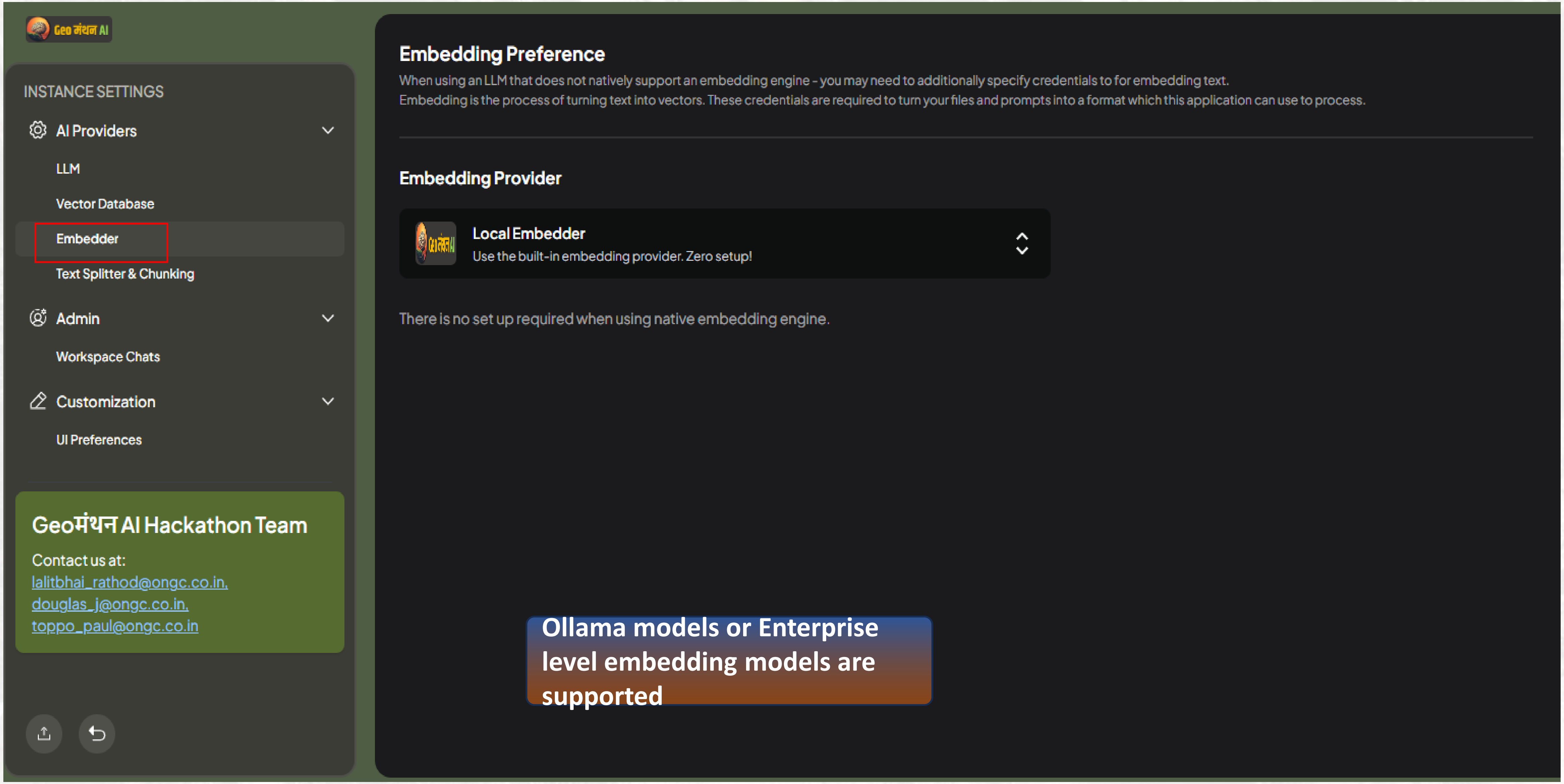
QDrant
Open source local and distributed cloud vector database.

QDrant API Endpoint: http://localhost:6333

API Key: wOeqxsYP4....1244sba

Currently, this application supports only Qdrant Vector Database.

Setting-Up Embedding Model



Embedding Preference

When using an LLM that does not natively support an embedding engine - you may need to additionally specify credentials to for embedding text.
Embedding is the process of turning text into vectors. These credentials are required to turn your files and prompts into a format which this application can use to process.

Embedding Provider

Local Embedder
Use the built-in embedding provider. Zero setup!

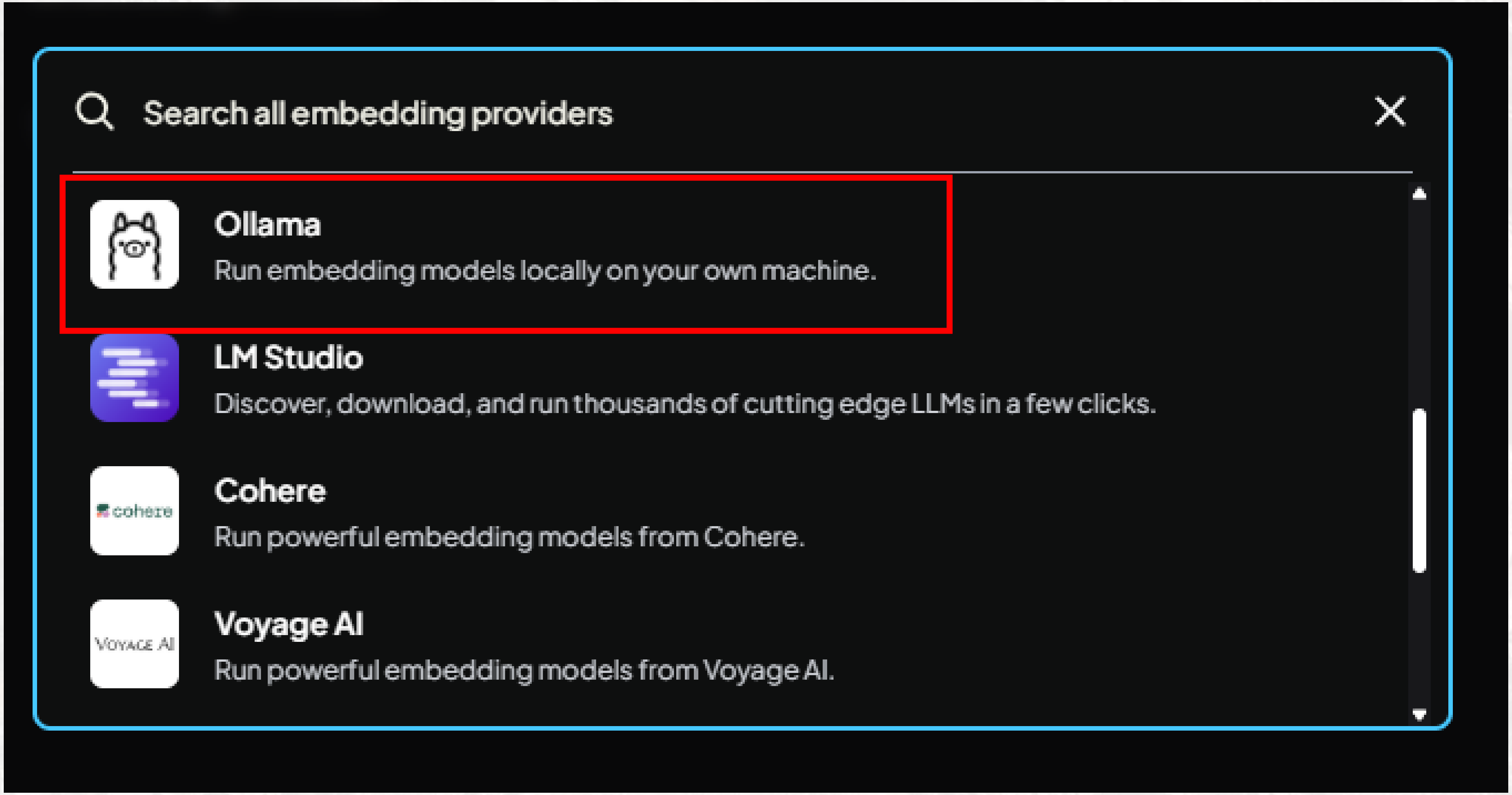
There is no set up required when using native embedding engine.

Geoमंथन AI Hackathon Team

Contact us at:
lalitbhai_rathod@ongc.co.in,
douglas_j@ongc.co.in,
toppo_paul@ongc.co.in

Ollama models or Enterprise level embedding models are supported

Setting-Up Embedding Model



Search all embedding providers

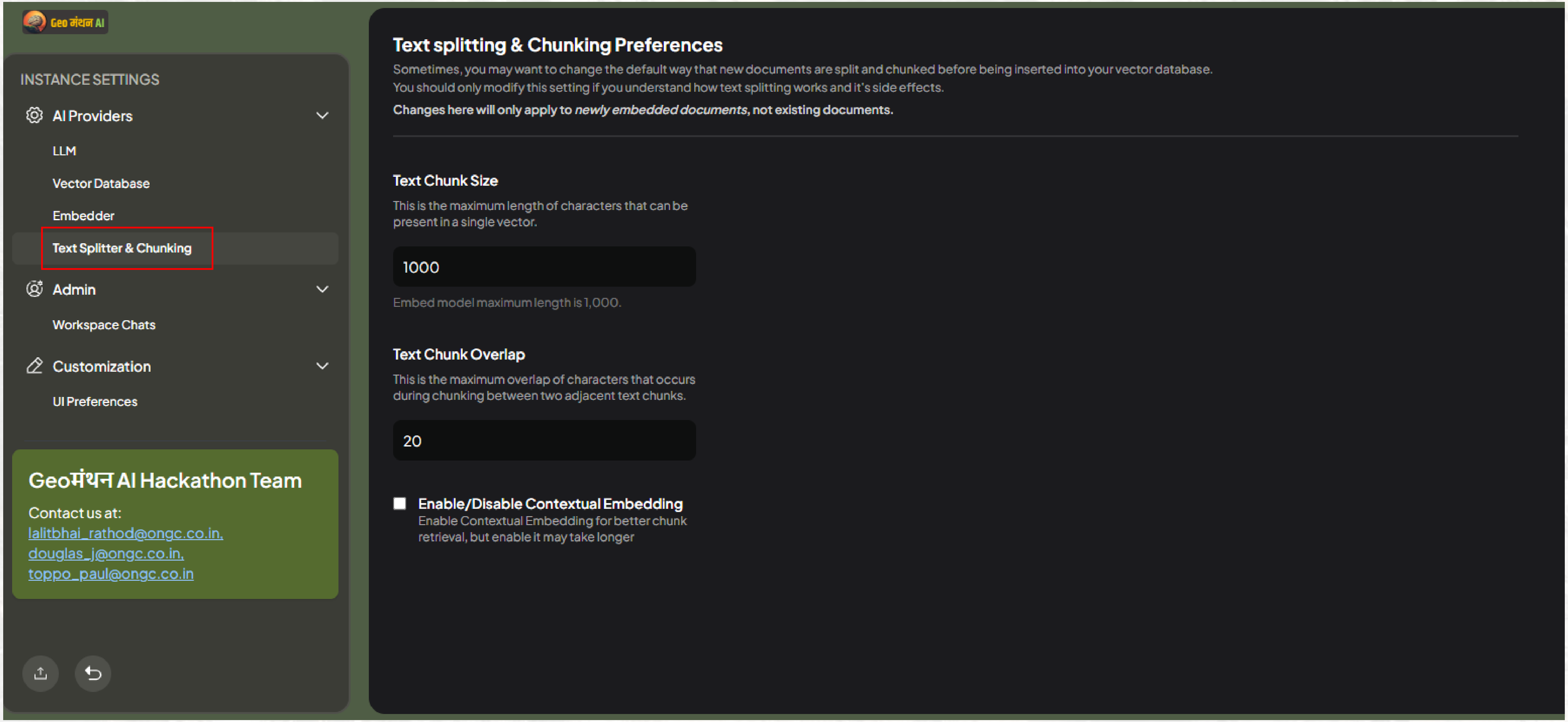
Ollama
Run embedding models locally on your own machine.

LM Studio
Discover, download, and run thousands of cutting edge LLMs in a few clicks.

Cohere
Run powerful embedding models from Cohere.

Voyage AI
Run powerful embedding models from Voyage AI.

Setting-Up Text splitting & Chunking Strategy



Geoमंथन AI Hackathon Team

Contact us at:
lalitbhai_rathod@ongc.co.in,
douglas_j@ongc.co.in,
toppo_paul@ongc.co.in

Text splitting & Chunking Preferences

Sometimes, you may want to change the default way that new documents are split and chunked before being inserted into your vector database. You should only modify this setting if you understand how text splitting works and it's side effects.

Changes here will only apply to *newly embedded documents*, not existing documents.

Text Chunk Size

This is the maximum length of characters that can be present in a single vector.

1000

Embed model maximum length is 1,000.

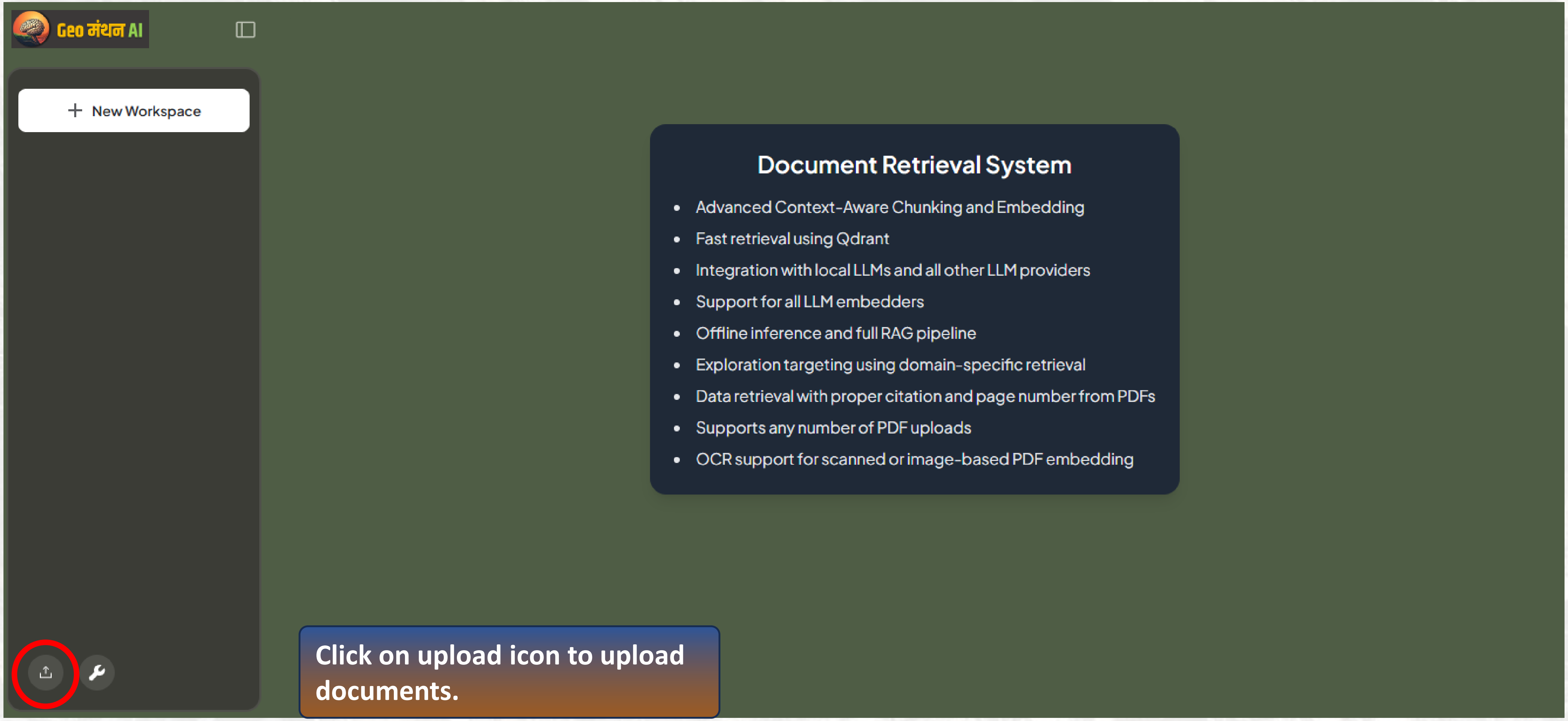
Text Chunk Overlap

This is the maximum overlap of characters that occurs during chunking between two adjacent text chunks.

20

☐ **Enable/Disable Contextual Embedding**
Enable Contextual Embedding for better chunk retrieval, but enable it may take longer

Upload All The Relevant Documents

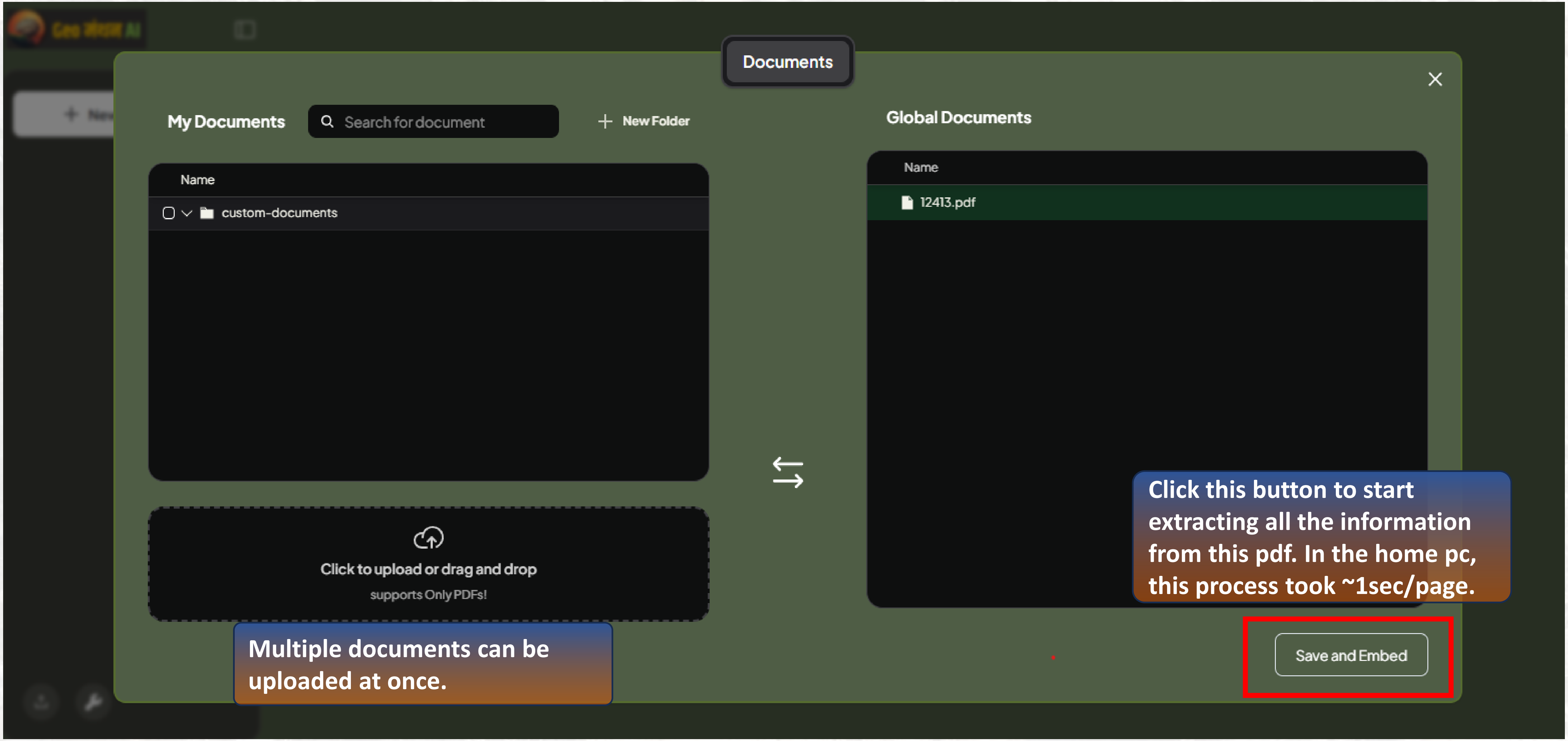


Document Retrieval System

- Advanced Context-Aware Chunking and Embedding
- Fast retrieval using Qdrant
- Integration with local LLMs and all other LLM providers
- Support for all LLM embedders
- Offline inference and full RAG pipeline
- Exploration targeting using domain-specific retrieval
- Data retrieval with proper citation and page number from PDFs
- Supports any number of PDF uploads
- OCR support for scanned or image-based PDF embedding

Click on upload icon to upload documents.

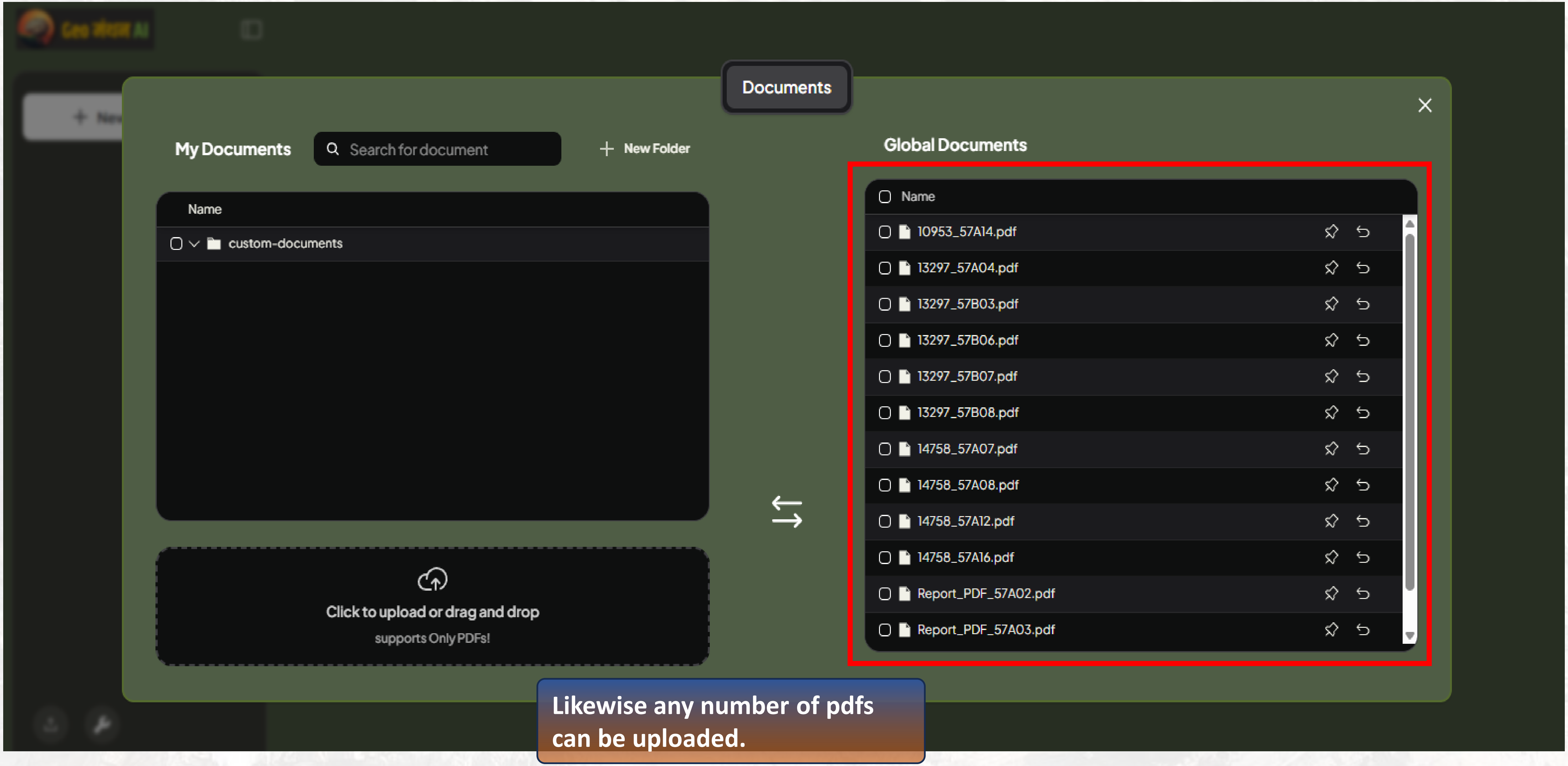
Upload All The Relevant Documents (13PDfs Uploaded)



Click this button to start extracting all the information from this pdf. In the home pc, this process took ~1sec/page.

Multiple documents can be uploaded at once.

Upload All The Relevant Documents (13PDfs Uploaded)



Documents

My Documents

Search for document

+ New Folder

Name

☐ custom-documents

Click to upload or drag and drop

supports Only PDFs!

Global Documents

☐ Name

☐ 10953_57A14.pdf

☐ 13297_57A04.pdf

☐ 13297_57B03.pdf

☐ 13297_57B06.pdf

☐ 13297_57B07.pdf

☐ 13297_57B08.pdf

☐ 14758_57A07.pdf

☐ 14758_57A08.pdf

☐ 14758_57A12.pdf

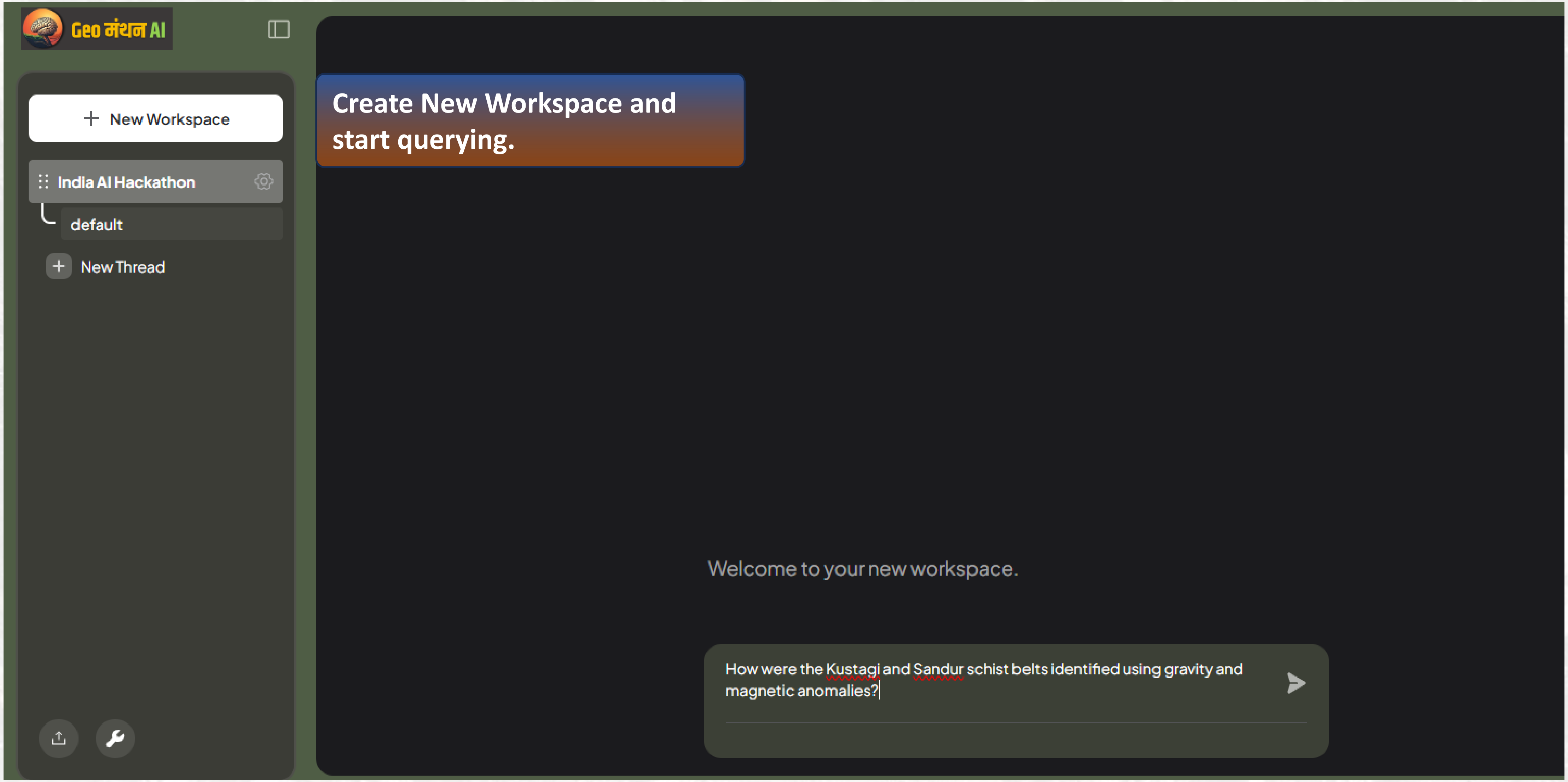
☐ 14758_57A16.pdf

☐ Report_PDF_57A02.pdf

☐ Report_PDF_57A03.pdf

Likewise any number of pdfs can be uploaded.

Query 1

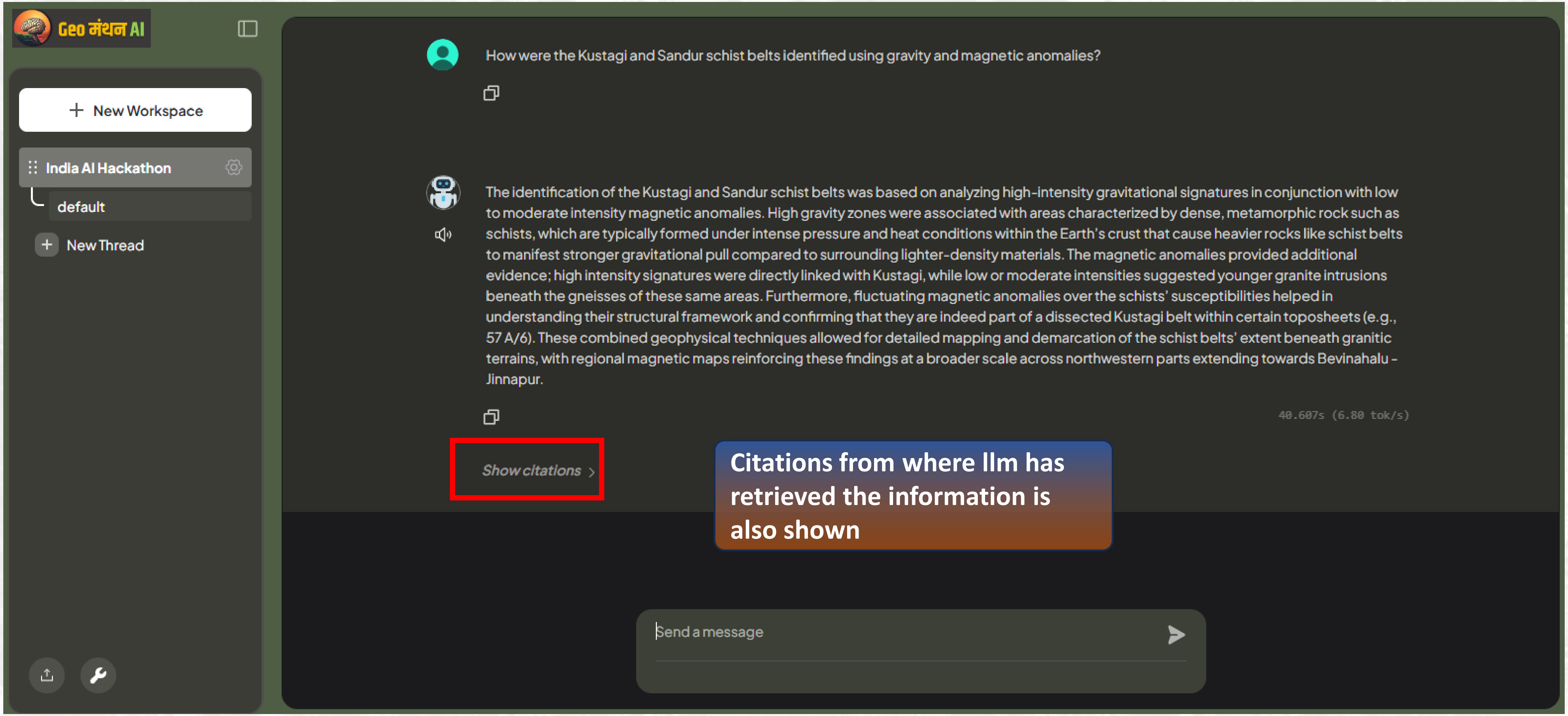


Create New Workspace and start querying.

Welcome to your new workspace.

How were the Kustagi and Sandur schist belts identified using gravity and magnetic anomalies?

Query 1 (LLM Response)



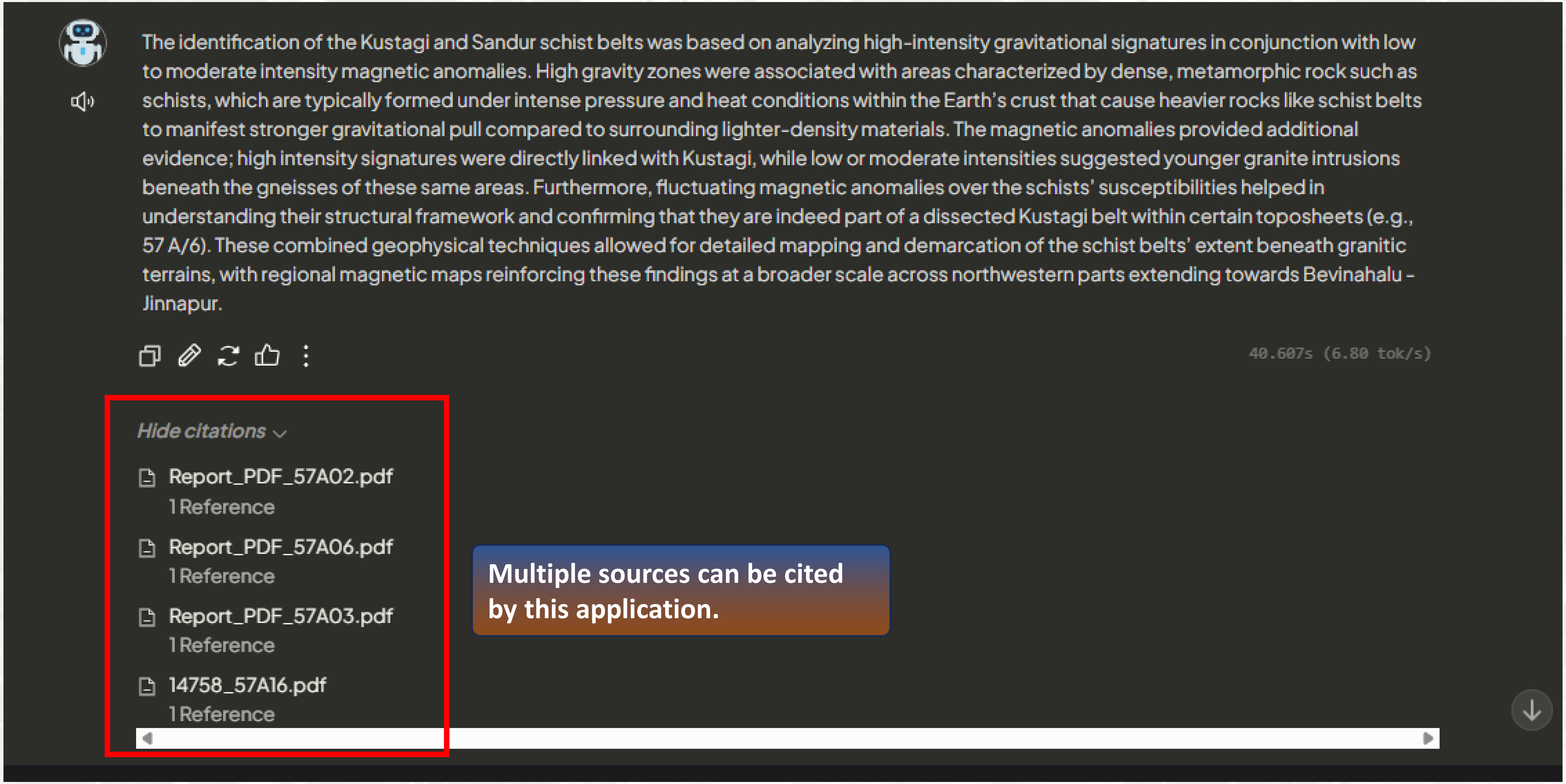
How were the Kustagi and Sandur schist belts identified using gravity and magnetic anomalies?

The identification of the Kustagi and Sandur schist belts was based on analyzing high-intensity gravitational signatures in conjunction with low to moderate intensity magnetic anomalies. High gravity zones were associated with areas characterized by dense, metamorphic rock such as schists, which are typically formed under intense pressure and heat conditions within the Earth's crust that cause heavier rocks like schist belts to manifest stronger gravitational pull compared to surrounding lighter-density materials. The magnetic anomalies provided additional evidence; high intensity signatures were directly linked with Kustagi, while low or moderate intensities suggested younger granite intrusions beneath the gneisses of these same areas. Furthermore, fluctuating magnetic anomalies over the schists' susceptibilities helped in understanding their structural framework and confirming that they are indeed part of a dissected Kustagi belt within certain toposheets (e.g., 57 A/6). These combined geophysical techniques allowed for detailed mapping and demarcation of the schist belts' extent beneath granitic terrains, with regional magnetic maps reinforcing these findings at a broader scale across northwestern parts extending towards Bevinahalu - Jinnapur.

40.607s (6.80 tok/s)

Citations from where llm has retrieved the information is also shown

Citations



The identification of the Kustagi and Sandur schist belts was based on analyzing high-intensity gravitational signatures in conjunction with low to moderate intensity magnetic anomalies. High gravity zones were associated with areas characterized by dense, metamorphic rock such as schists, which are typically formed under intense pressure and heat conditions within the Earth's crust that cause heavier rocks like schist belts to manifest stronger gravitational pull compared to surrounding lighter-density materials. The magnetic anomalies provided additional evidence; high intensity signatures were directly linked with Kustagi, while low or moderate intensities suggested younger granite intrusions beneath the gneisses of these same areas. Furthermore, fluctuating magnetic anomalies over the schists' susceptibilities helped in understanding their structural framework and confirming that they are indeed part of a dissected Kustagi belt within certain toposheets (e.g., 57 A/6). These combined geophysical techniques allowed for detailed mapping and demarcation of the schist belts' extent beneath granitic terrains, with regional magnetic maps reinforcing these findings at a broader scale across northwestern parts extending towards Bevinahalu - Jinnapur.

`40.607s (6.80 tok/s)`

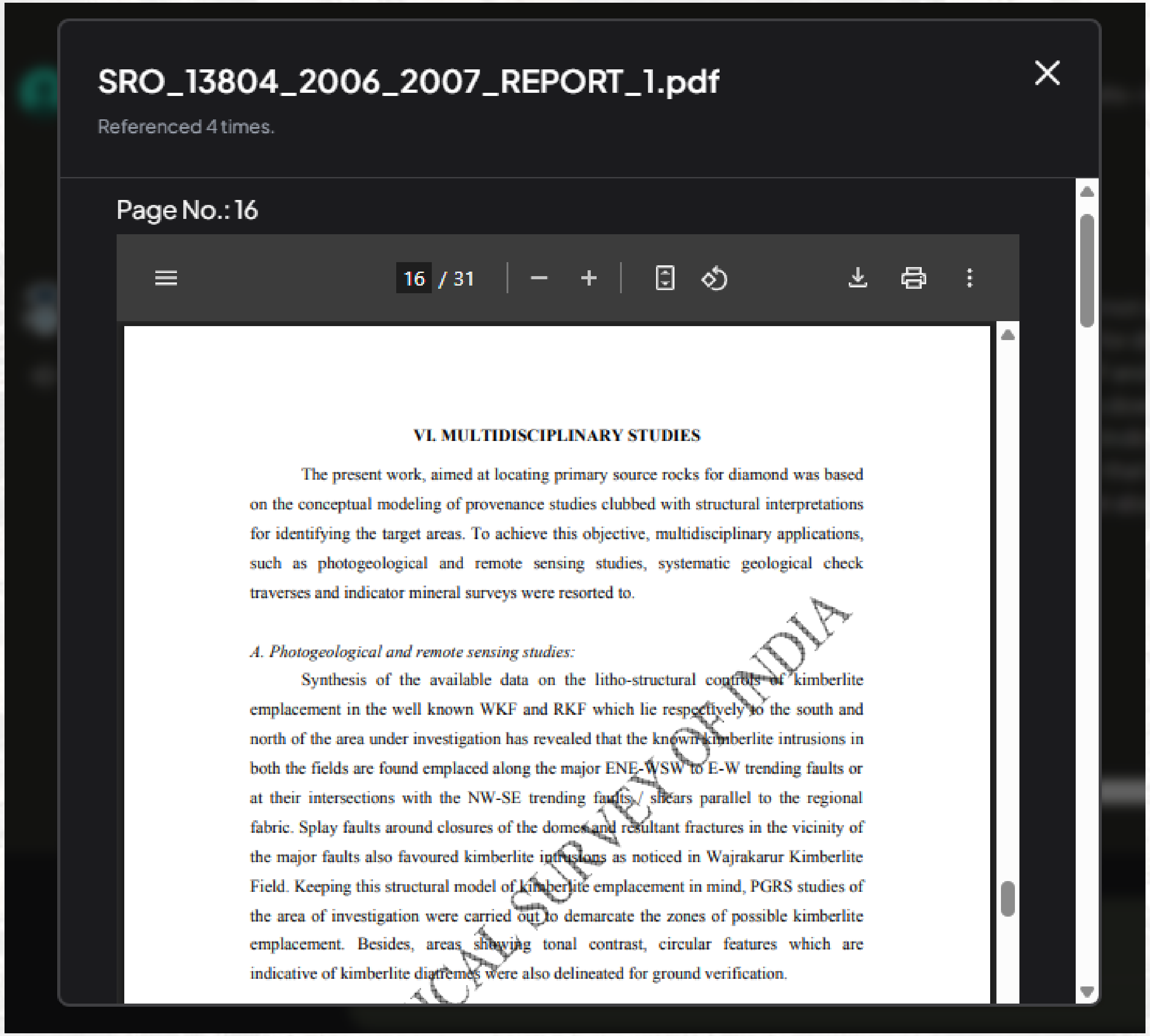
Hide citations ⌄

- Report_PDF_57A02.pdf
  1 Reference
- Report_PDF_57A06.pdf
  1 Reference
- Report_PDF_57A03.pdf
  1 Reference
- 14758_57A16.pdf
  1 Reference

Multiple sources can be cited by this application.

Query 1 (Citation)

On clicking references, we get pdf name along with page no. from where reference is being taken



SRO_13804_2006_2007_REPORT_1.pdf

Referenced 4 times.

Page No.: 16

VI. MULTIDISCIPLINARY STUDIES

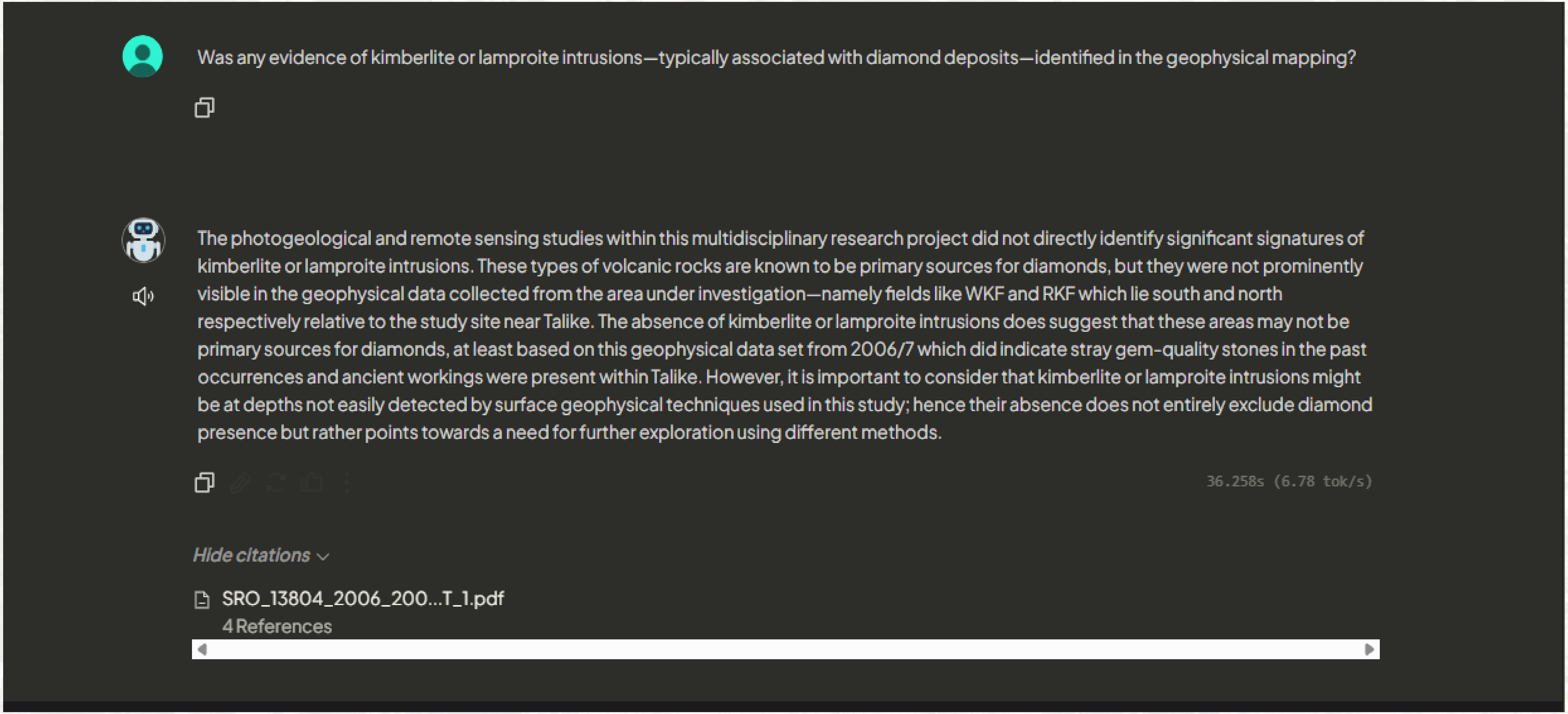
The present work, aimed at locating primary source rocks for diamond was based on the conceptual modeling of provenance studies clubbed with structural interpretations for identifying the target areas. To achieve this objective, multidisciplinary applications, such as photogeological and remote sensing studies, systematic geological check traverses and indicator mineral surveys were resorted to.

A. Photogeological and remote sensing studies:

Synthesis of the available data on the litho-structural controls of kimberlite emplacement in the well known WKF and RKF which lie respectively to the south and north of the area under investigation has revealed that the known kimberlite intrusions in both the fields are found emplaced along the major ENE-WSW to E-W trending faults or at their intersections with the NW-SE trending faults / shears parallel to the regional fabric. Splay faults around closures of the domes and resultant fractures in the vicinity of the major faults also favoured kimberlite intrusions as noticed in Wajrakarur Kimberlite Field. Keeping this structural model of kimberlite emplacement in mind, PGRS studies of the area of investigation were carried out to demarcate the zones of possible kimberlite emplacement. Besides, areas showing tonal contrast, circular features which are indicative of kimberlite diatremes were also delineated for ground verification.

ICAL SURVEY OF INDIA

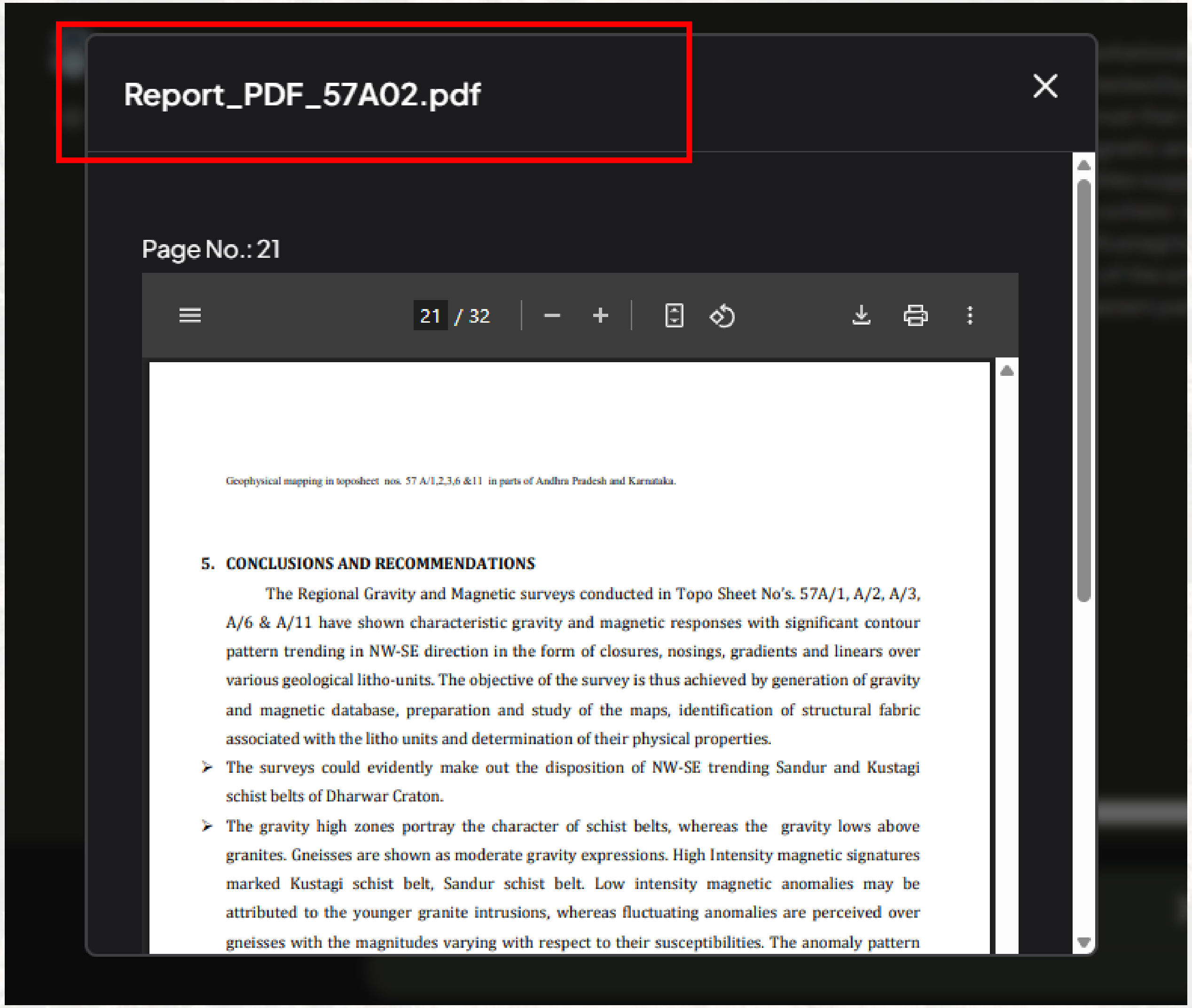
Query 2



Was any evidence of kimberlite or lamproite intrusions—typically associated with diamond deposits—identified in the geophysical mapping?

The photogeological and remote sensing studies within this multidisciplinary research project did not directly identify significant signatures of kimberlite or lamproite intrusions. These types of volcanic rocks are known to be primary sources for diamonds, but they were not prominently visible in the geophysical data collected from the area under investigation—namely fields like WKF and RKF which lie south and north respectively relative to the study site near Talike. The absence of kimberlite or lamproite intrusions does suggest that these areas may not be primary sources for diamonds, at least based on this geophysical data set from 2006/7 which did indicate stray gem-quality stones in the past occurrences and ancient workings were present within Talike. However, it is important to consider that kimberlite or lamproite intrusions might be at depths not easily detected by surface geophysical techniques used in this study; hence their absence does not entirely exclude diamond presence but rather points towards a need for further exploration using different methods.

Citations with PDF Name & Page No.



Report_PDF_57A02.pdf

Page No.: 21

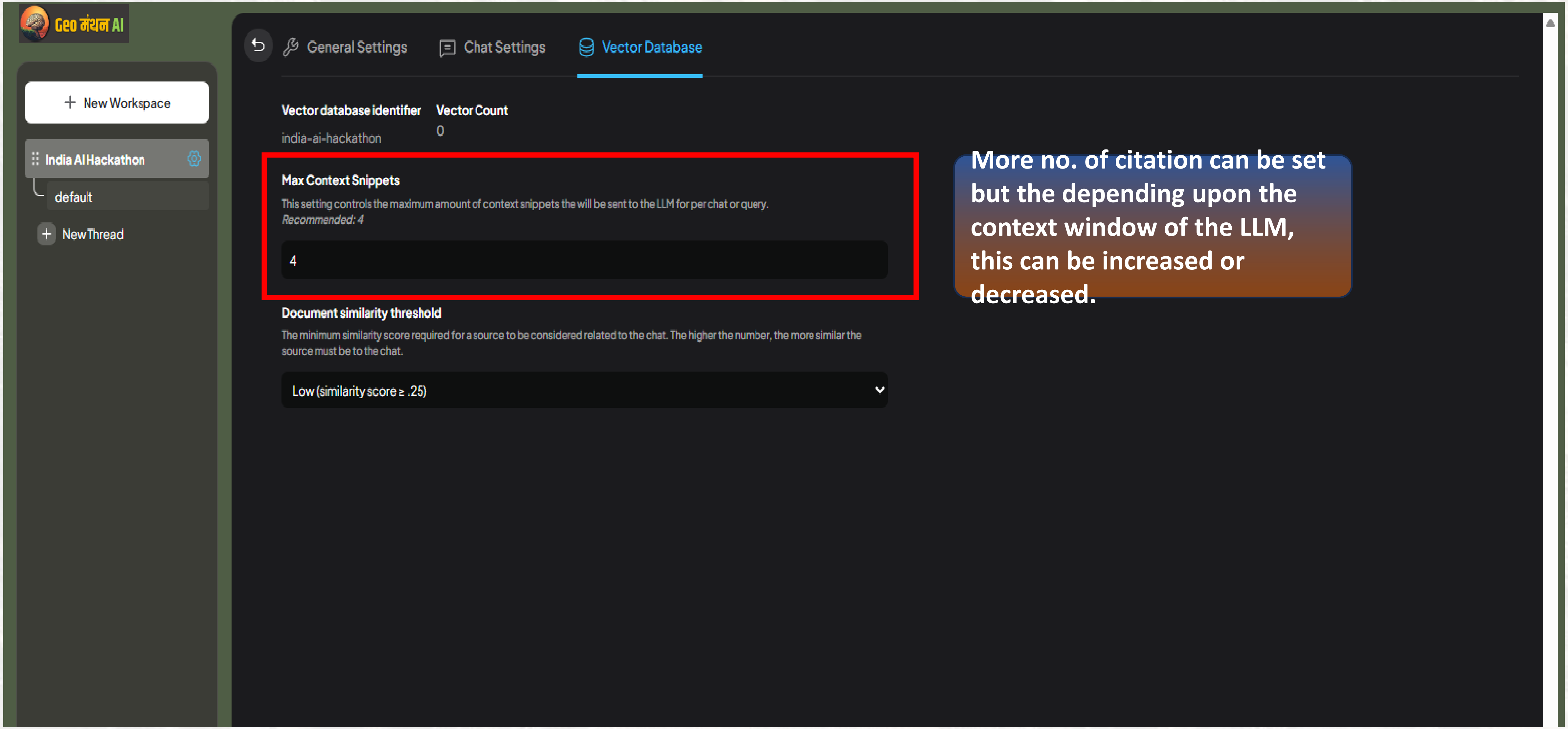
Geophysical mapping in toposheet nos. 57 A/1,2,3,6 &11 in parts of Andhra Pradesh and Karnataka.

5. **CONCLUSIONS AND RECOMMENDATIONS**

The Regional Gravity and Magnetic surveys conducted in Topo Sheet No's. 57A/1, A/2, A/3, A/6 & A/11 have shown characteristic gravity and magnetic responses with significant contour pattern trending in NW-SE direction in the form of closures, nosings, gradients and linears over various geological litho-units. The objective of the survey is thus achieved by generation of gravity and magnetic database, preparation and study of the maps, identification of structural fabric associated with the litho units and determination of their physical properties.

- The surveys could evidently make out the disposition of NW-SE trending Sandur and Kustagi schist belts of Dharwar Craton.
- The gravity high zones portray the character of schist belts, whereas the gravity lows above granites. Gneisses are shown as moderate gravity expressions. High Intensity magnetic signatures marked Kustagi schist belt, Sandur schist belt. Low intensity magnetic anomalies may be attributed to the younger granite intrusions, whereas fluctuating anomalies are perceived over gneisses with the magnitudes varying with respect to their susceptibilities. The anomaly pattern

Retrieving More Citations



Vector database identifier
india-ai-hackathon

Vector Count
0

Max Context Snippets

This setting controls the maximum amount of context snippets the will be sent to the LLM for per chat or query.
Recommended: 4

4

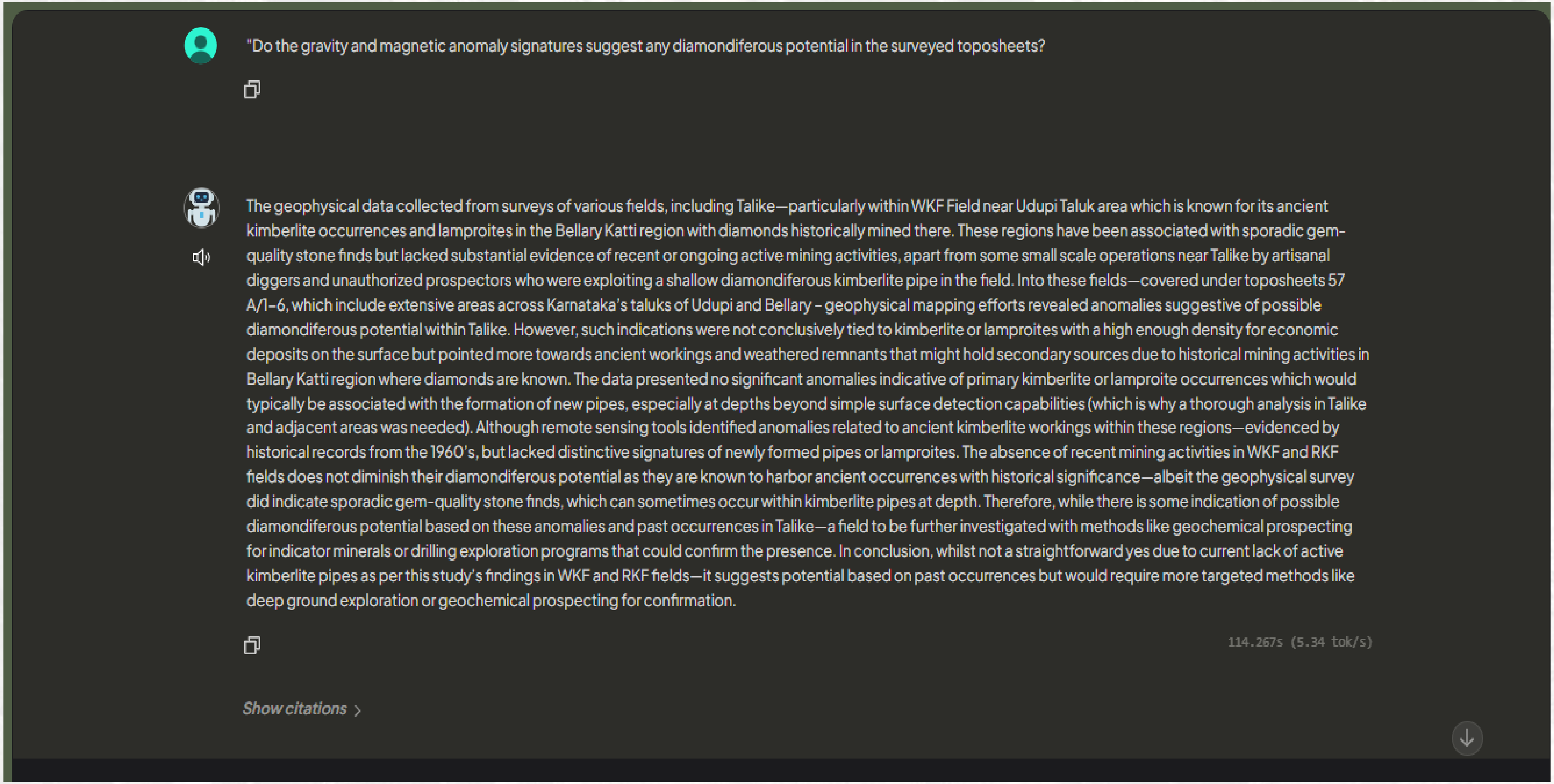
Document similarity threshold

The minimum similarity score required for a source to be considered related to the chat. The higher the number, the more similar the source must be to the chat.

Low (similarity score ≥ .25)

More no. of citation can be set but the depending upon the context window of the LLM, this can be increased or decreased.

Query 3

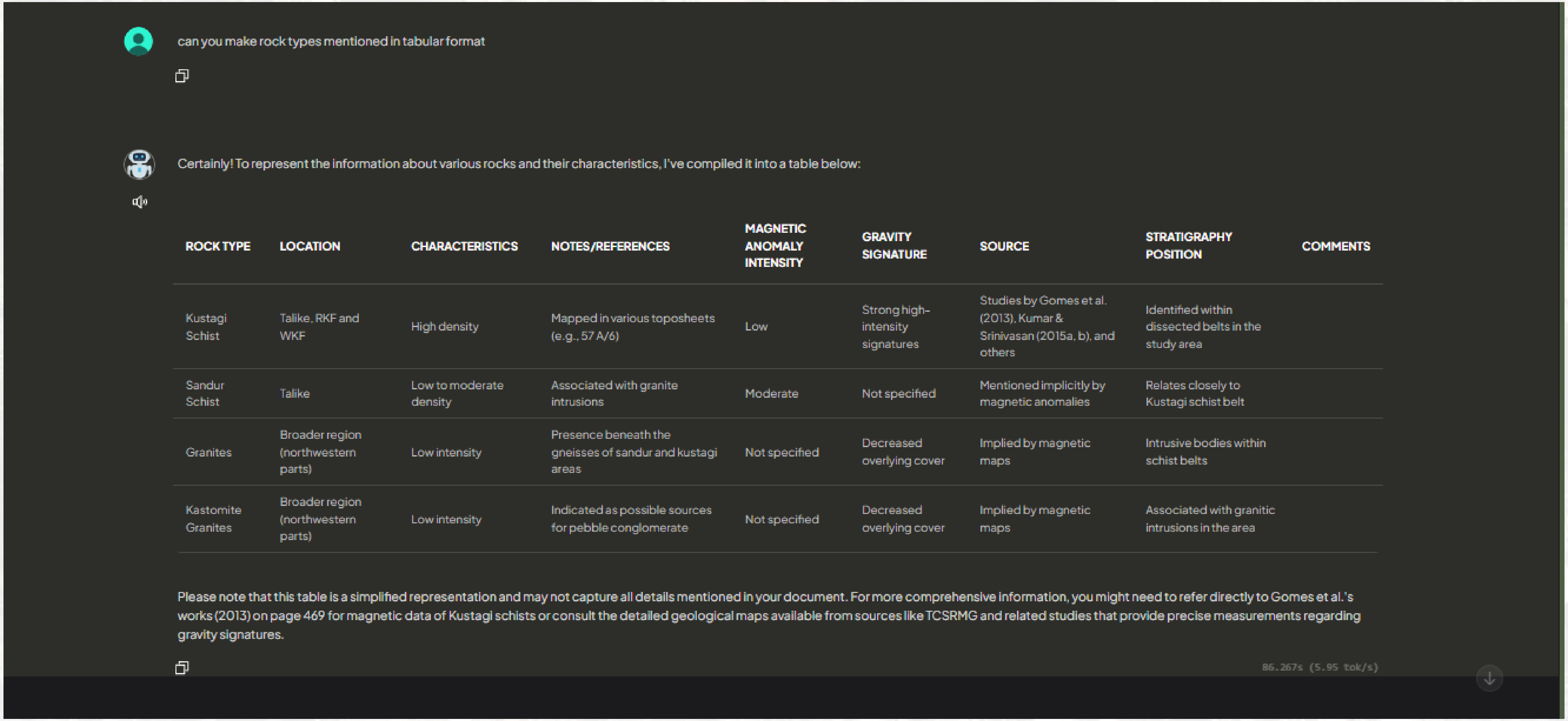


"Do the gravity and magnetic anomaly signatures suggest any diamondiferous potential in the surveyed toposheets?

The geophysical data collected from surveys of various fields, including Talike—particularly within WKF Field near Udupi Taluk area which is known for its ancient kimberlite occurrences and lamproites in the Bellary Katti region with diamonds historically mined there. These regions have been associated with sporadic gem-quality stone finds but lacked substantial evidence of recent or ongoing active mining activities, apart from some small scale operations near Talike by artisanal diggers and unauthorized prospectors who were exploiting a shallow diamondiferous kimberlite pipe in the field. Into these fields—covered under toposheets 57 A/1-6, which include extensive areas across Karnataka's taluks of Udupi and Bellary - geophysical mapping efforts revealed anomalies suggestive of possible diamondiferous potential within Talike. However, such indications were not conclusively tied to kimberlite or lamproites with a high enough density for economic deposits on the surface but pointed more towards ancient workings and weathered remnants that might hold secondary sources due to historical mining activities in Bellary Katti region where diamonds are known. The data presented no significant anomalies indicative of primary kimberlite or lamproite occurrences which would typically be associated with the formation of new pipes, especially at depths beyond simple surface detection capabilities (which is why a thorough analysis in Talike and adjacent areas was needed). Although remote sensing tools identified anomalies related to ancient kimberlite workings within these regions—evidenced by historical records from the 1960's, but lacked distinctive signatures of newly formed pipes or lamproites. The absence of recent mining activities in WKF and RKF fields does not diminish their diamondiferous potential as they are known to harbor ancient occurrences with historical significance—albeit the geophysical survey did indicate sporadic gem-quality stone finds, which can sometimes occur within kimberlite pipes at depth. Therefore, while there is some indication of possible diamondiferous potential based on these anomalies and past occurrences in Talike—a field to be further investigated with methods like geochemical prospecting for indicator minerals or drilling exploration programs that could confirm the presence. In conclusion, whilst not a straightforward yes due to current lack of active kimberlite pipes as per this study's findings in WKF and RKF fields—it suggests potential based on past occurrences but would require more targeted methods like deep ground exploration or geochemical prospecting for confirmation.

114.267s (5.34 tok/s)

Presenting Response in Tabular Format



can you make rock types mentioned in tabular format

Certainly! To represent the information about various rocks and their characteristics, I've compiled it into a table below:

ROCK TYPE	LOCATION	CHARACTERISTICS	NOTES/REFERENCES	MAGNETIC ANOMALY INTENSITY	GRAVITY SIGNATURE	SOURCE	STRATIGRAPHY POSITION	COMMENTS
Kustagi Schist	Talike, RKF and WKF	High density	Mapped in various toposheets (e.g., 57 A/6)	Low	Strong high-intensity signatures	Studies by Gomes et al. (2013), Kumar & Srinivasan (2015a, b), and others	Identified within dissected belts in the study area	
Sandur Schist	Talike	Low to moderate density	Associated with granite intrusions	Moderate	Not specified	Mentioned implicitly by magnetic anomalies	Relates closely to Kustagi schist belt	
Granites	Broader region (northwestern parts)	Low intensity	Presence beneath the gneisses of sandur and kustagi areas	Not specified	Decreased overlying cover	Implied by magnetic maps	Intrusive bodies within schist belts	
Kastomite Granites	Broader region (northwestern parts)	Low intensity	Indicated as possible sources for pebble conglomerate	Not specified	Decreased overlying cover	Implied by magnetic maps	Associated with granitic intrusions in the area	

Please note that this table is a simplified representation and may not capture all details mentioned in your document. For more comprehensive information, you might need to refer directly to Gomes et al.'s works (2013) on page 469 for magnetic data of Kustagi schists or consult the detailed geological maps available from sources like TCSRMG and related studies that provide precise measurements regarding gravity signatures.

86.267s (5.95 tok/s)

Advanced Embedding Method Employed

Why Contextual Embedding

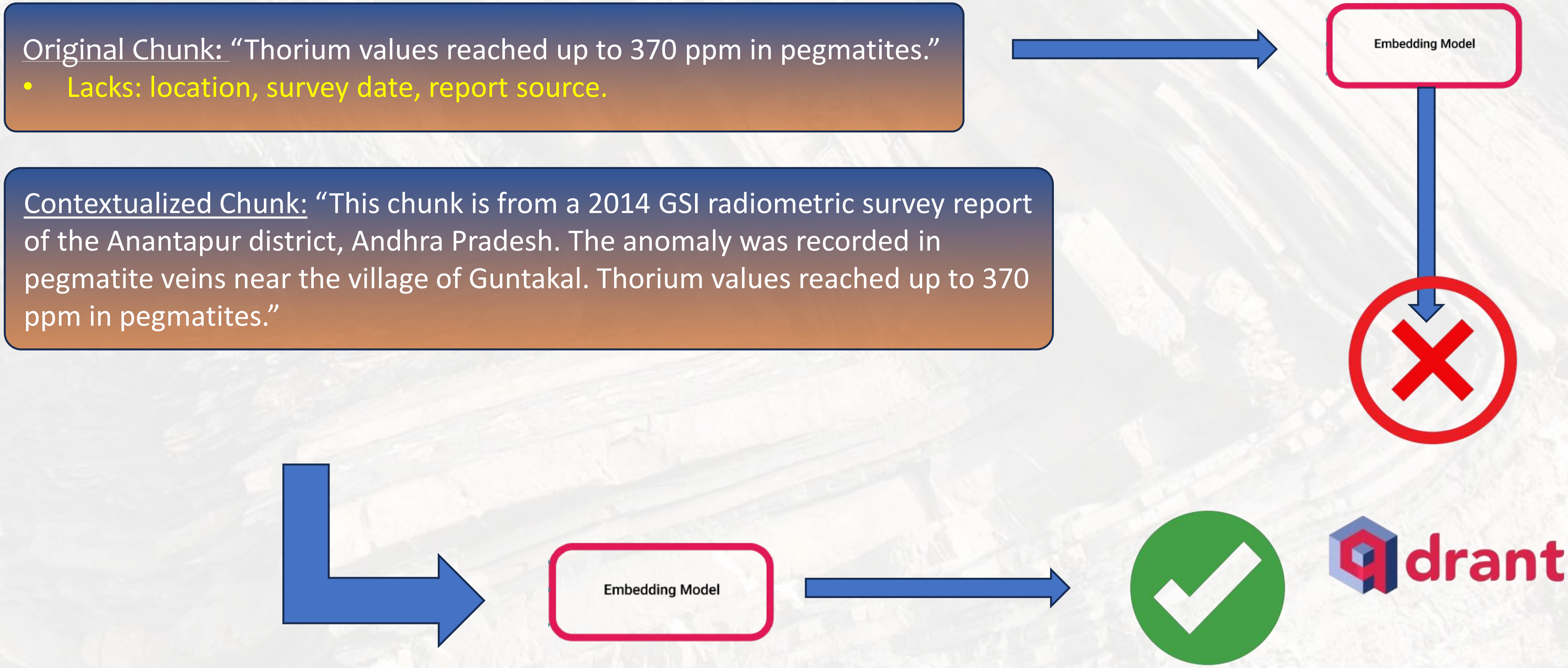
A **Improves retrieval accuracy by preserving section titles, block names, and geospatial context—crucial for distinguishing similar terms across different regions.**

B **Enhances semantic understanding by embedding not just the content but its surrounding meaning, reducing misinterpretation by the language model.**

C **Preserves document hierarchy, allowing low-context data (like assay tables or lithologs) to be correctly linked to their geological setting or purpose.**

D **Reduces hallucination risk** by anchoring responses to well-contextualized sources, making the RAG outputs more trustworthy and traceable.

Contextualized Embedding



Original Chunk: "Thorium values reached up to 370 ppm in pegmatites."

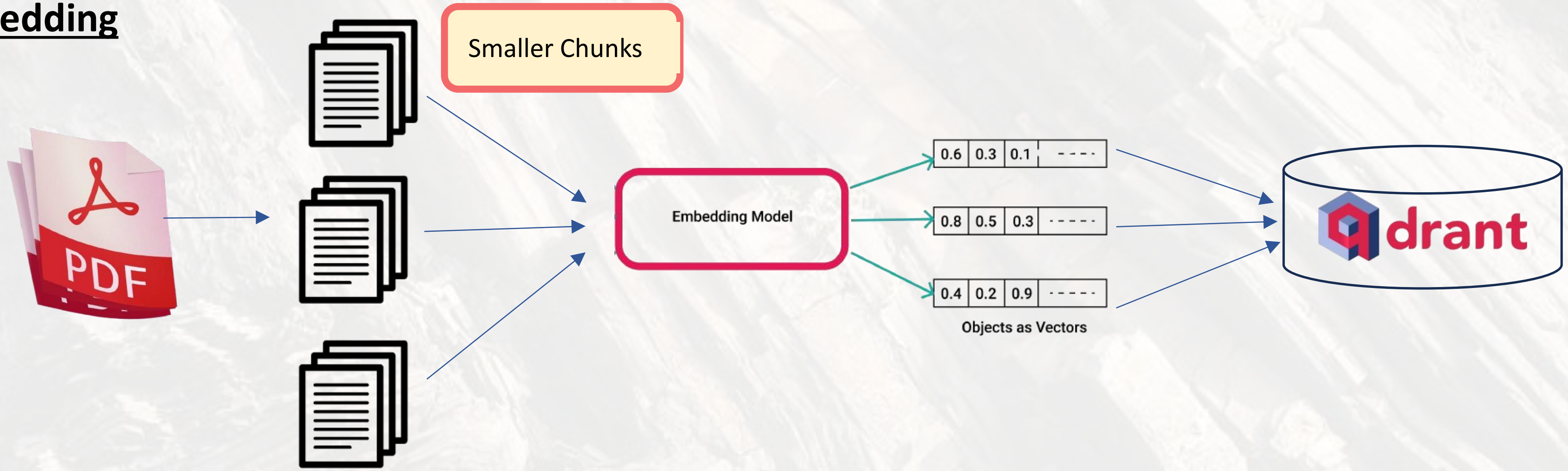
- Lacks: location, survey date, report source.

Contextualized Chunk: "This chunk is from a 2014 GSI radiometric survey report of the Anantapur district, Andhra Pradesh. The anomaly was recorded in pegmatite veins near the village of Guntakal. Thorium values reached up to 370 ppm in pegmatites."

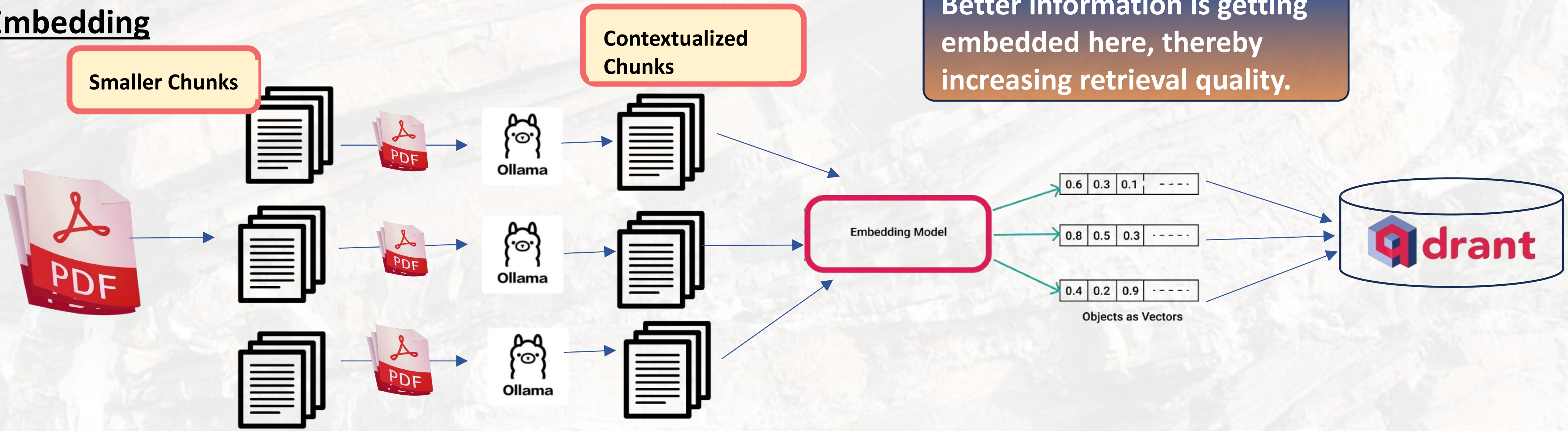
Now, More Contextualized Information is getting stored in Vector Database

Contextualized Embedding

Regular Embedding

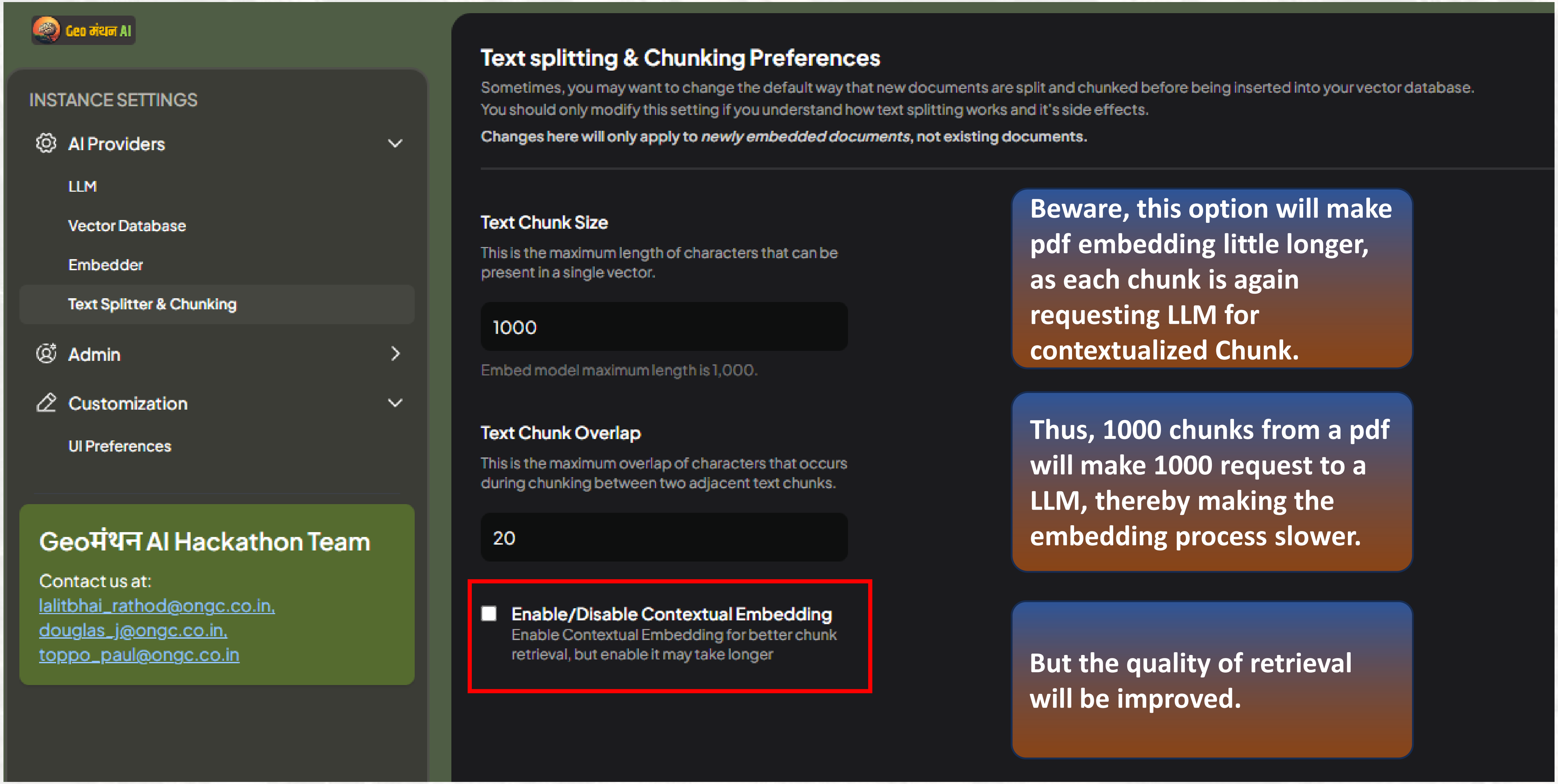


Contextual Embedding



Better information is getting embedded here, thereby increasing retrieval quality.

Enabling Contextual Embedding



Geoमंथन AI Hackathon Team

Contact us at:
lalitbhai_rathod@ongc.co.in,
douglas_j@ongc.co.in,
toppo_paul@ongc.co.in

Text splitting & Chunking Preferences

Sometimes, you may want to change the default way that new documents are split and chunked before being inserted into your vector database. You should only modify this setting if you understand how text splitting works and it's side effects.

Changes here will only apply to *newly embedded documents*, not existing documents.

Text Chunk Size

This is the maximum length of characters that can be present in a single vector.

1000

Embed model maximum length is 1,000.

Text Chunk Overlap

This is the maximum overlap of characters that occurs during chunking between two adjacent text chunks.

20

☐ **Enable/Disable Contextual Embedding**
Enable Contextual Embedding for better chunk retrieval, but enable it may take longer

Beware, this option will make pdf embedding little longer, as each chunk is again requesting LLM for contextualized Chunk.

Thus, 1000 chunks from a pdf will make 1000 request to a LLM, thereby making the embedding process slower.

But the quality of retrieval will be improved.