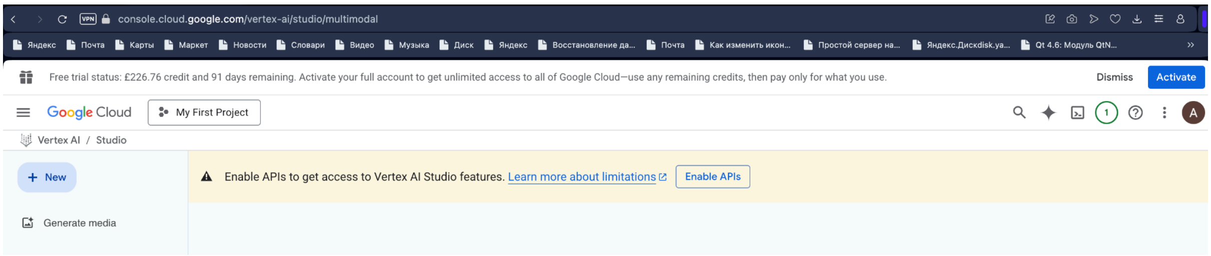
1. Create new google account or or use your account
2. Go to https://console.cloud.google.com/welcome/new and login
3. Start free trial and link your credit card (its free for 90 days)
4. Click Vertex AI tab in the left side
5. Enable APIs

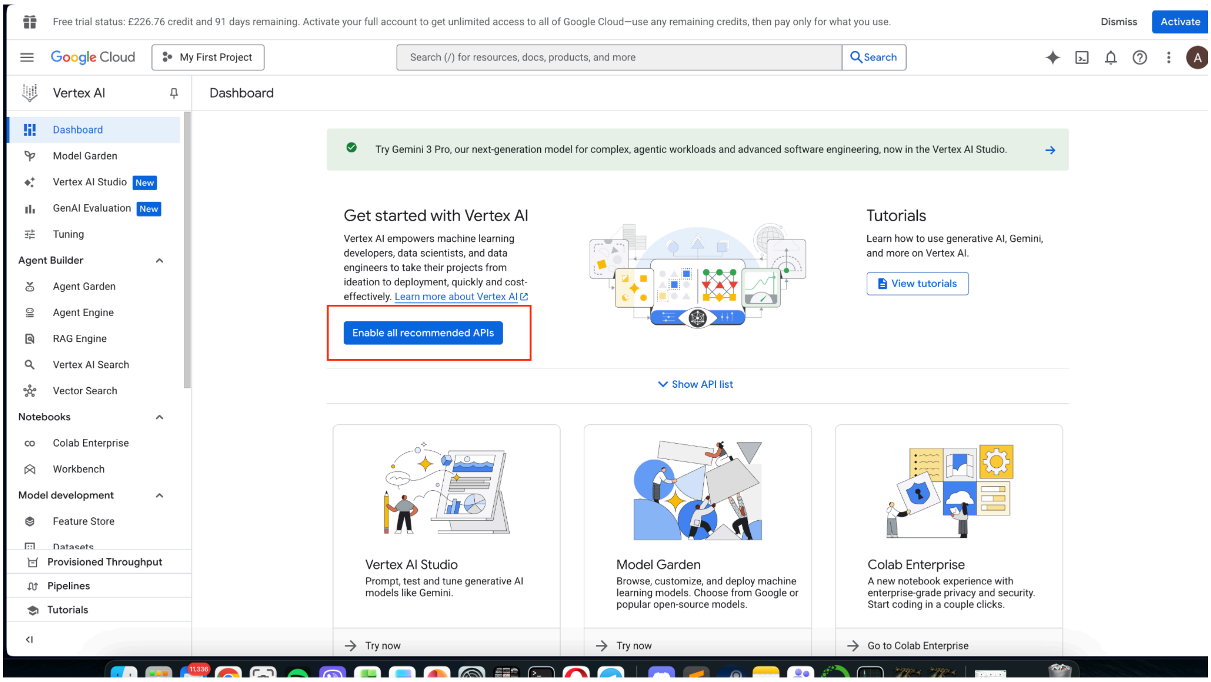


console.cloud.google.com/vertex-ai/studio/multimodal

Free trial status: £226.76 credit and 91 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

Enable APIs to get access to Vertex AI Studio features. Learn more about limitations

6. In Vertex AI tab click this button to enable other APIs. Wait few minutes



Free trial status: £226.76 credit and 91 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

Try Gemini 3 Pro, our next-generation model for complex, agentic workloads and advanced software engineering, now in the Vertex AI Studio.

Get started with Vertex AI

Vertex AI empowers machine learning developers, data scientists, and data engineers to take their projects from ideation to deployment, quickly and cost-effectively. Learn more about Vertex AI

Tutorials

Learn how to use generative AI, Gemini, and more on Vertex AI.

Vertex AI Studio

Prompt, test and tune generative AI models like Gemini.

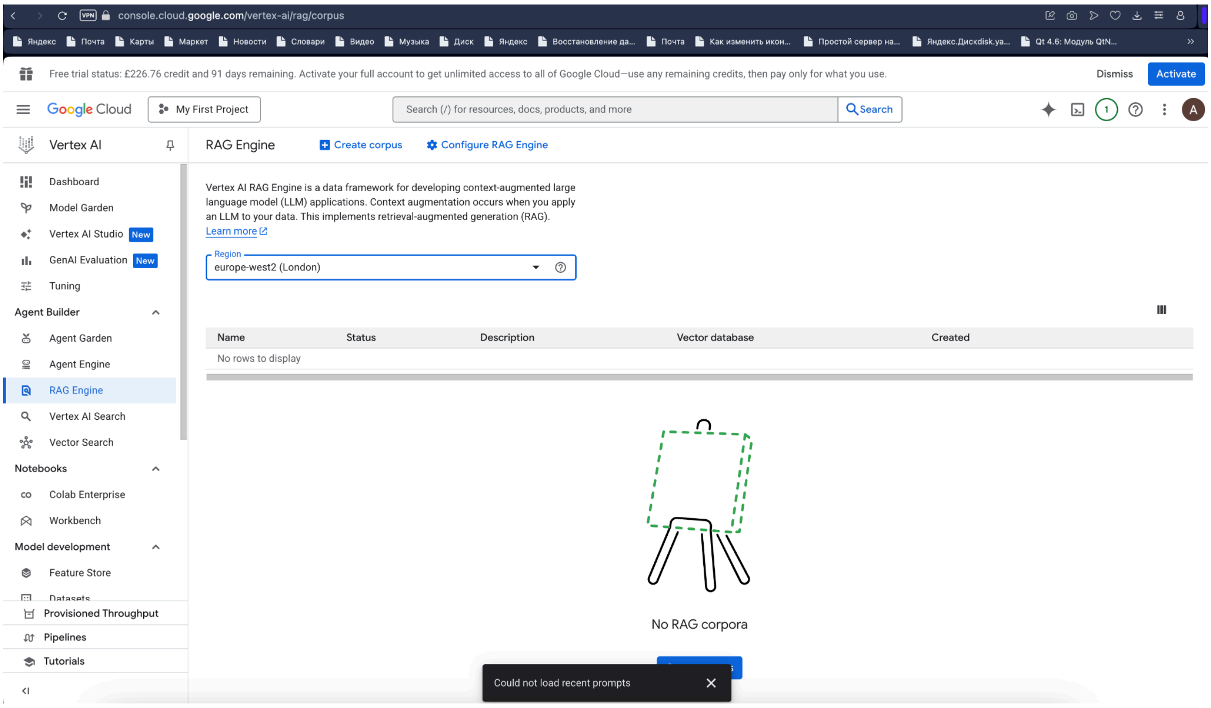
Model Garden

Browse, customize, and deploy machine learning models. Choose from Google or popular open-source models.

Colab Enterprise

A new notebook experience with enterprise-grade privacy and security. Start coding in a couple clicks.

7. In Vertex AI tab find a RAG engine tab and click it. You should see: Select region europe-west2 **and copy it.** You will need it later



Free trial status: £226.76 credit and 91 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

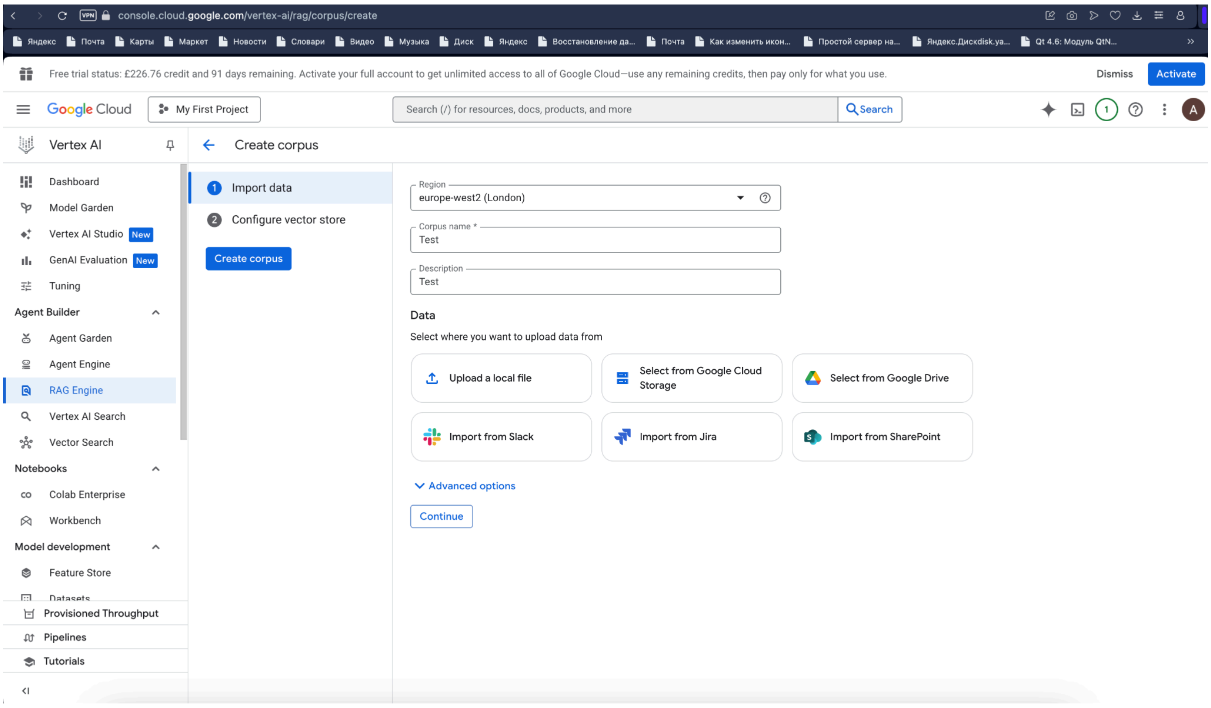
Vertex AI RAG Engine is a data framework for developing context-augmented large language model (LLM) applications. Context augmentation occurs when you apply an LLM to your data. This implements retrieval-augmented generation (RAG).
Learn more

Region: europe-west2 (London)

Name	Status	Description	Vector database	Created
No rows to display				

Could not load recent prompts

8. Click create corpus button



Free trial status: £226.76 credit and 91 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

Region: europe-west2 (London)

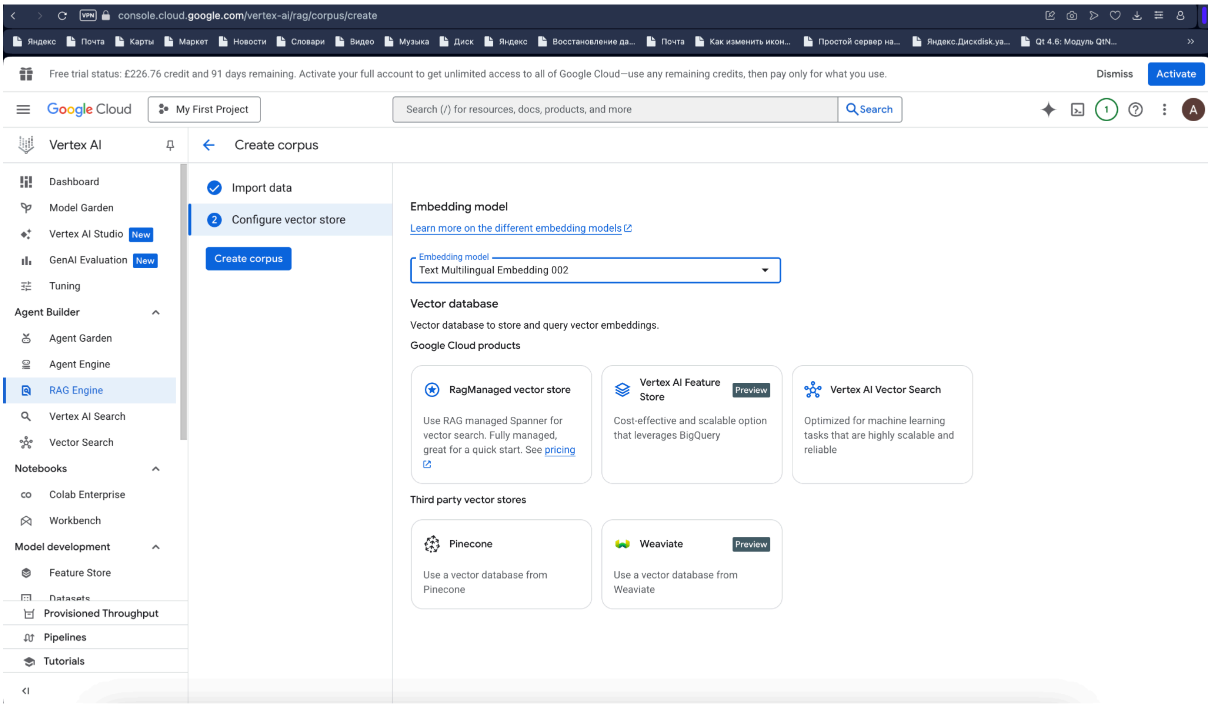
Corpus name *: Test

Description: Test

Data

Select where you want to upload data from

9.



Free trial status: £226.76 credit and 91 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

Embedding model

Learn more on the different embedding models

Embedding model: Text Multilingual Embedding 002

Vector database

Vector database to store and query vector embeddings.

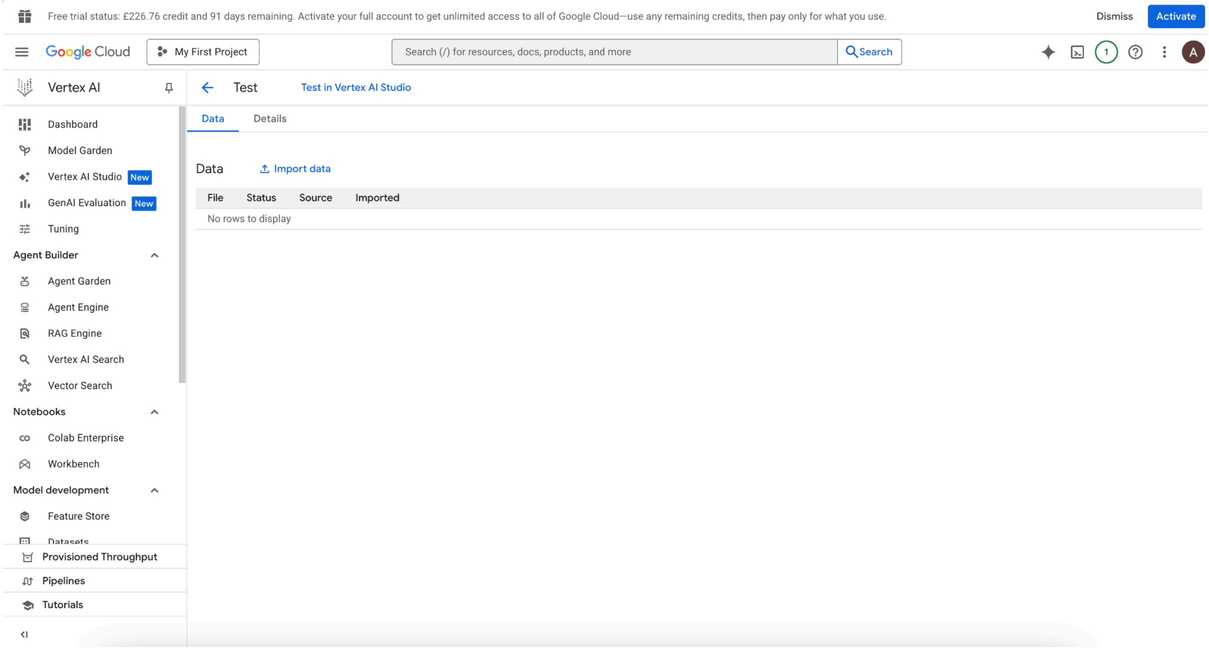
Google Cloud products

- RagManaged vector store — Use RAG managed Spanner for vector search. Fully managed, great for a quick start. See pricing
- Vertex AI Feature Store (Preview) — Cost-effective and scalable option that leverages BigQuery
- Vertex AI Vector Search — Optimized for machine learning tasks that are highly scalable and reliable

Third party vector stores

- Pinecone — Use a vector database from Pinecone
- Weaviate (Preview) — Use a vector database from Weaviate

10. Wait few minutes and you will see



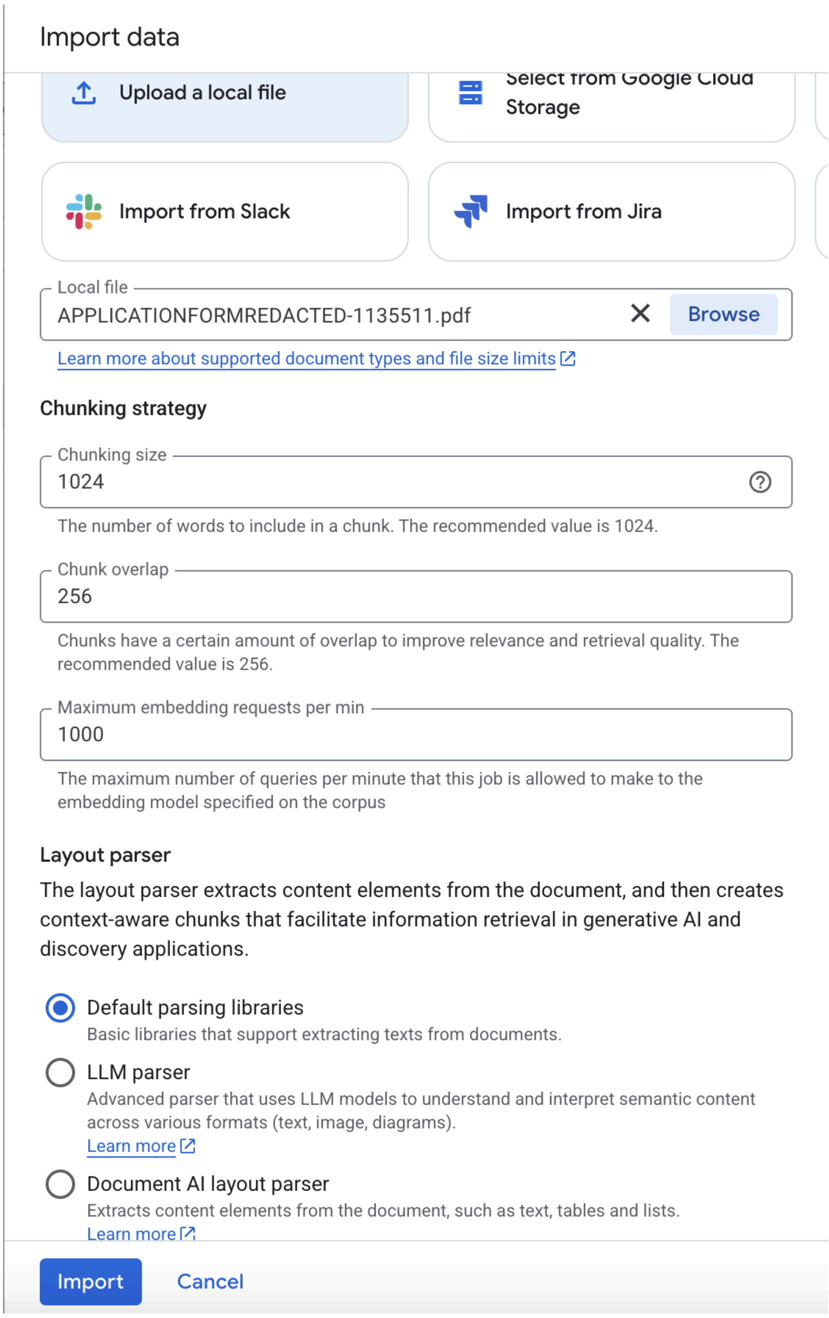
Free trial status: £226.76 credit and 91 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

Test

Data

File	Status	Source	Imported
No rows to display			

11. Now you can import files to this storage. A model will see these files and will generate answers based on themes. Click import data to add file.



Import data

Upload a local file
Select from Google Cloud Storage
Import from Slack
Import from Jira

Local file: APPLICATIONFORMREDACTED-1135511.pdf

Browse
Learn more about supported document types and file size limits

Chunking strategy

Chunking size: 1024

The number of words to include in a chunk. The recommended value is 1024.

Chunk overlap: 256

Chunks have a certain amount of overlap to improve relevance and retrieval quality. The recommended value is 256.

Maximum embedding requests per min: 1000

The maximum number of queries per minute that this job is allowed to make to the embedding model specified on the corpus

Layout parser

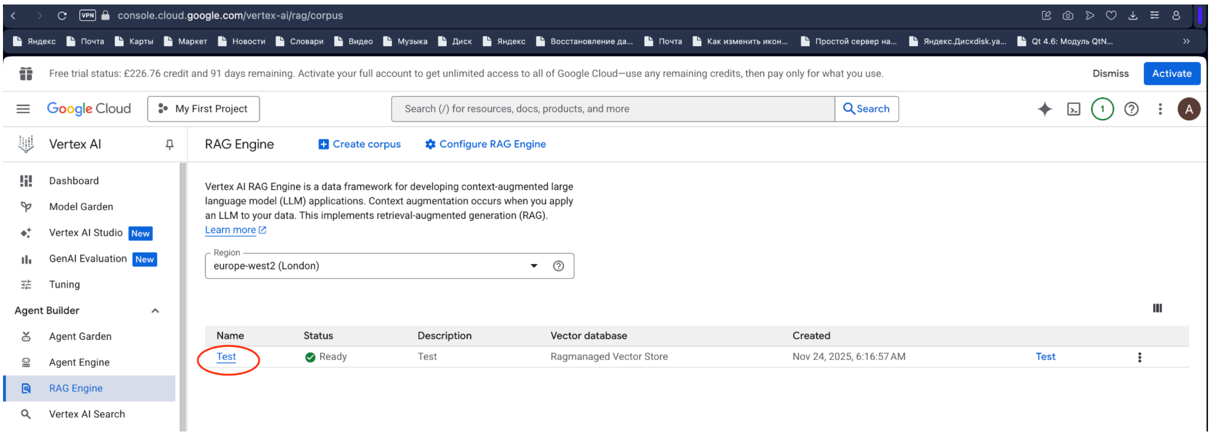
The layout parser extracts content elements from the document, and then creates context-aware chunks that facilitate information retrieval in generative AI and discovery applications.

● Default parsing libraries
Basic libraries that support extracting texts from documents.

○ LLM parser
Advanced parser that uses LLM models to understand and interpret semantic content across various formats (text, image, diagrams).
Learn more

○ Document AI layout parser
Extracts content elements from the document, such as text, tables and lists.
Learn more

12. Click to your corpus name



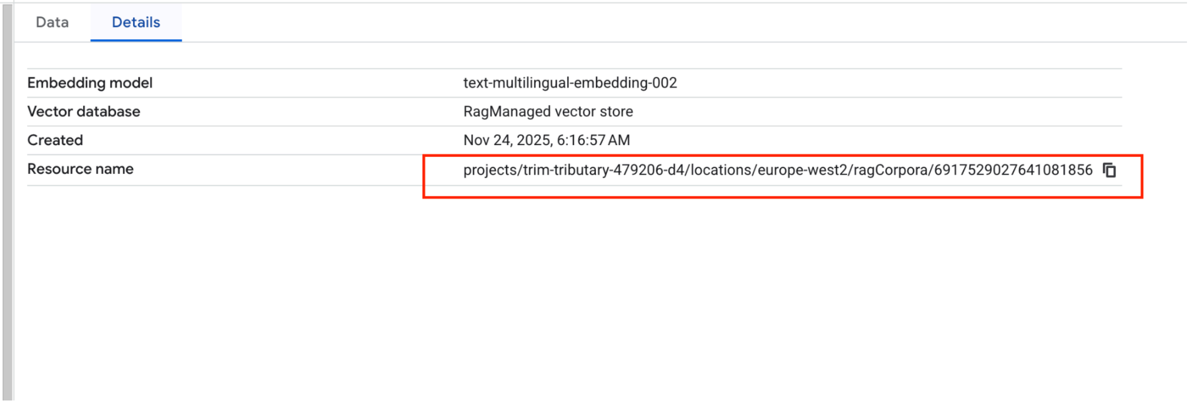
Free trial status: £226.76 credit and 91 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

Vertex AI RAG Engine is a data framework for developing context-augmented large language model (LLM) applications. Context augmentation occurs when you apply an LLM to your data. This implements retrieval-augmented generation (RAG).

Region: europe-west2 (London)

Name	Status	Description	Vector database	Created	
Test	Ready	Test	Ragmanaged Vector Store	Nov 24, 2025, 6:16:57 AM	Test

13. In details tab **copy** resource name.you will need it later. for example projects/**trim-tributary-479206-d4**/locations/**europe-west2**/ragCorpora/6917529027641081856 . **trim-tributary-479206-d4 - project id. europe-west2 - region**



Embedding model	text-multilingual-embedding-002
Vector database	RagManaged vector store
Created	Nov 24, 2025, 6:16:57 AM
Resource name	projects/trim-tributary-479206-d4/locations/europe-west2/ragCorpora/6917529027641081856

14. Cool. Now you should install gcloud CLI to your pc. This tool will allow your code to work with this api. Go to https://docs.cloud.google.com/sdk/docs/install and select your platform. Follow the instructions and install it.
15. When installed type ***./google-cloud-sdk/bin/gcloud auth login*** it will open browser and you should login using the account that you used before.
16. now type ***./google-cloud-sdk/bin/gcloud auth application-default login*** . you will be redirected. Don't forget to click select all before continuing. Then in the terminal select your project based on project id.
    a. In case of multiple accounts connected, these commands may be useful to select the project you need.
        i. ./google-cloud-sdk/bin/gcloud auth application-default set-quota-project ***[project_id]***
        ii. ./google-cloud-sdk/bin/gcloud config set project ***[project_id]***
        iii. ./google-cloud-sdk/bin/gcloud auth application-default set-quota-project ***[project_id]***
        iv. ./google-cloud-sdk/bin/gcloud auth list

17. If you don't see any errors or warnings, you've configured everything correctly. Now the code you're executing should work with the API.

18. In terminal write ***pip install google-cloud-aiplatform*** and ***pip install google-genai google-cloud-aiplatform*** to install libs
19. Execute code