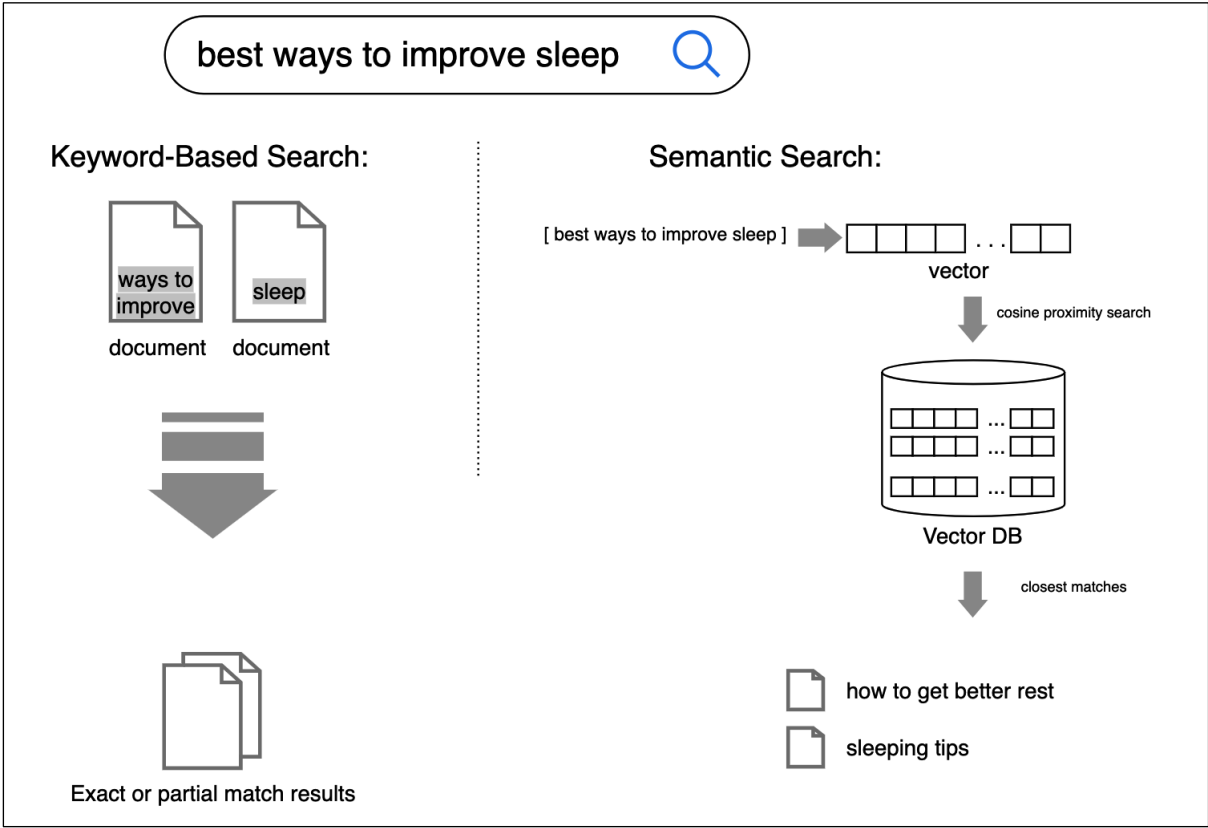
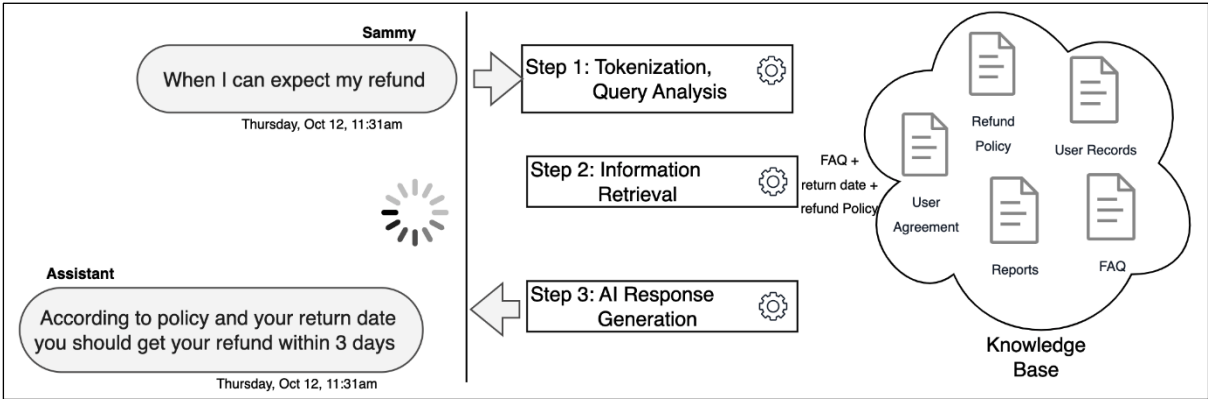
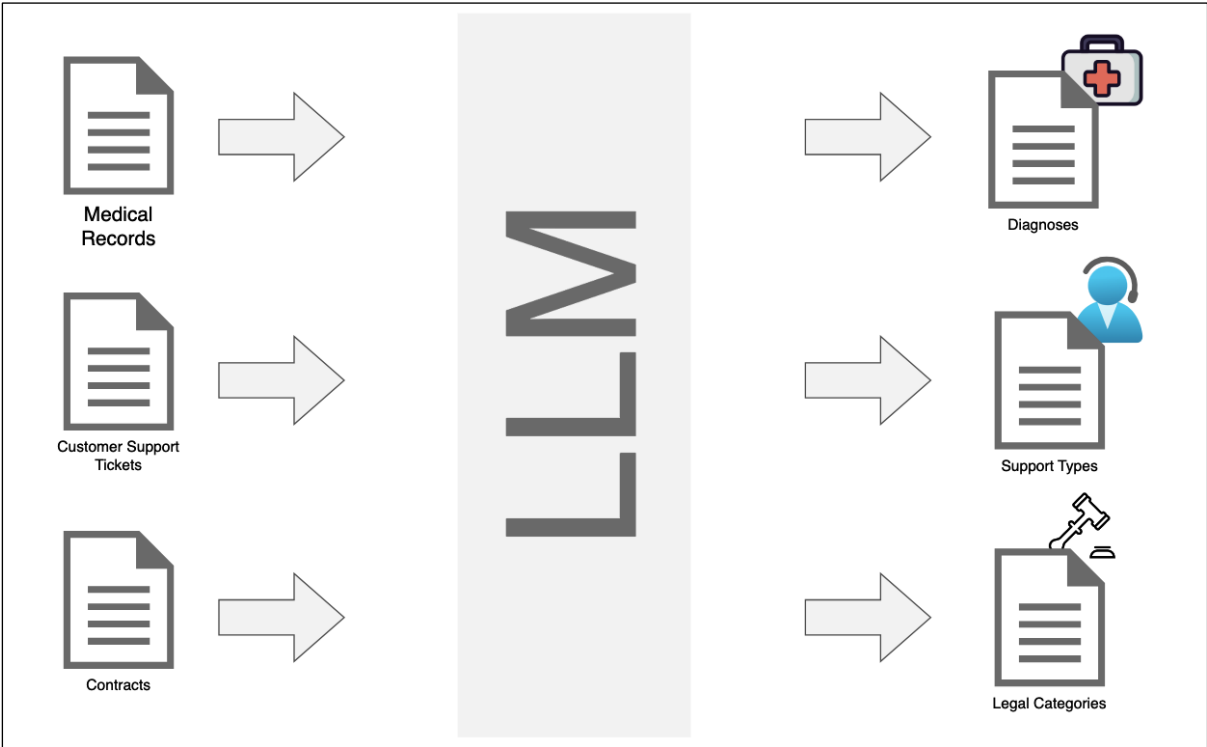
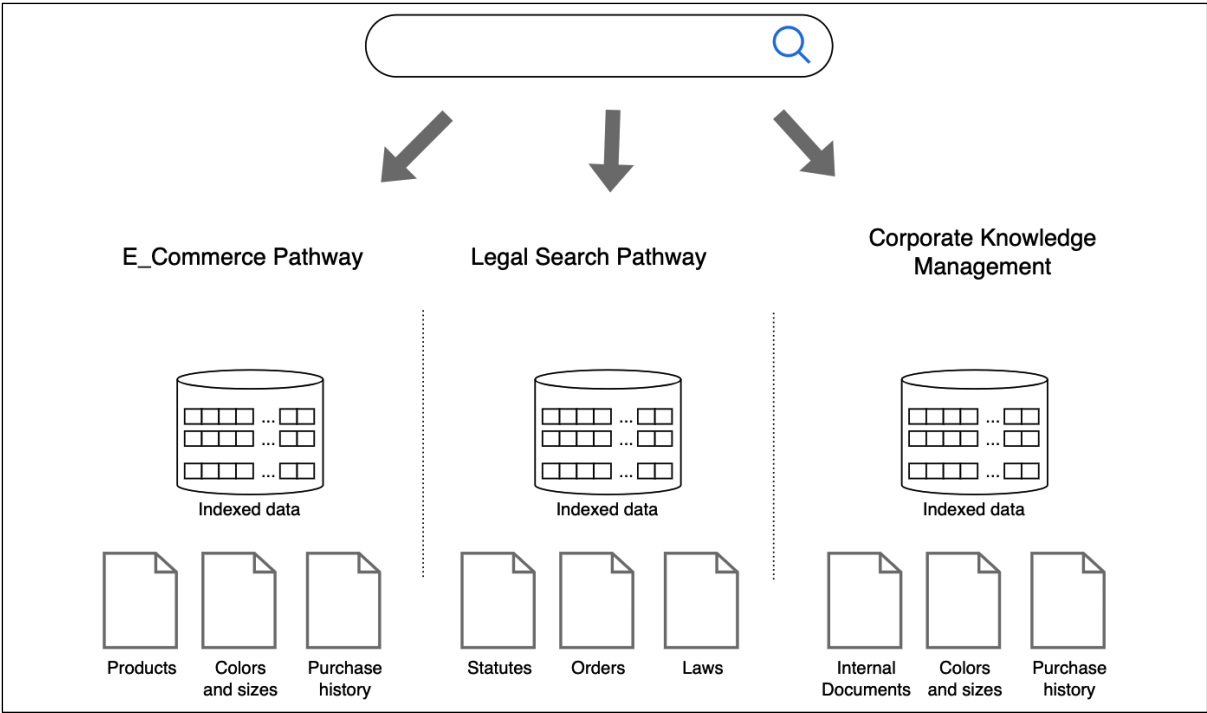


Chapter 1: Strengths, Limitations, and Use Cases of Language Models





Ambiguity



Challenge:

Difficulty in understanding user intent when queries have multiple meanings, leading to inaccurate or mixed search results.

Example:

A query like "apple benefits" could refer to nutritional benefits of the fruit or advantages of Apple Inc. products.

Potential Solutions:

Disambiguation prompts, leveraging user history, and analyzing query patterns to better interpret user intent.

Context Loss



Challenge:

Loss of important context in multi-turn conversations, where models fail to maintain relevance across multiple user interactions.

Example:

If a user asks "What is the best way to manage diabetes?" followed by "What about diet?", the model must link the second query to the context of diabetes management.

Potential Solutions:

Using memory mechanisms, embedding-based context tracking, and extending context windows to retain information across interactions.

False Positives



Challenge:

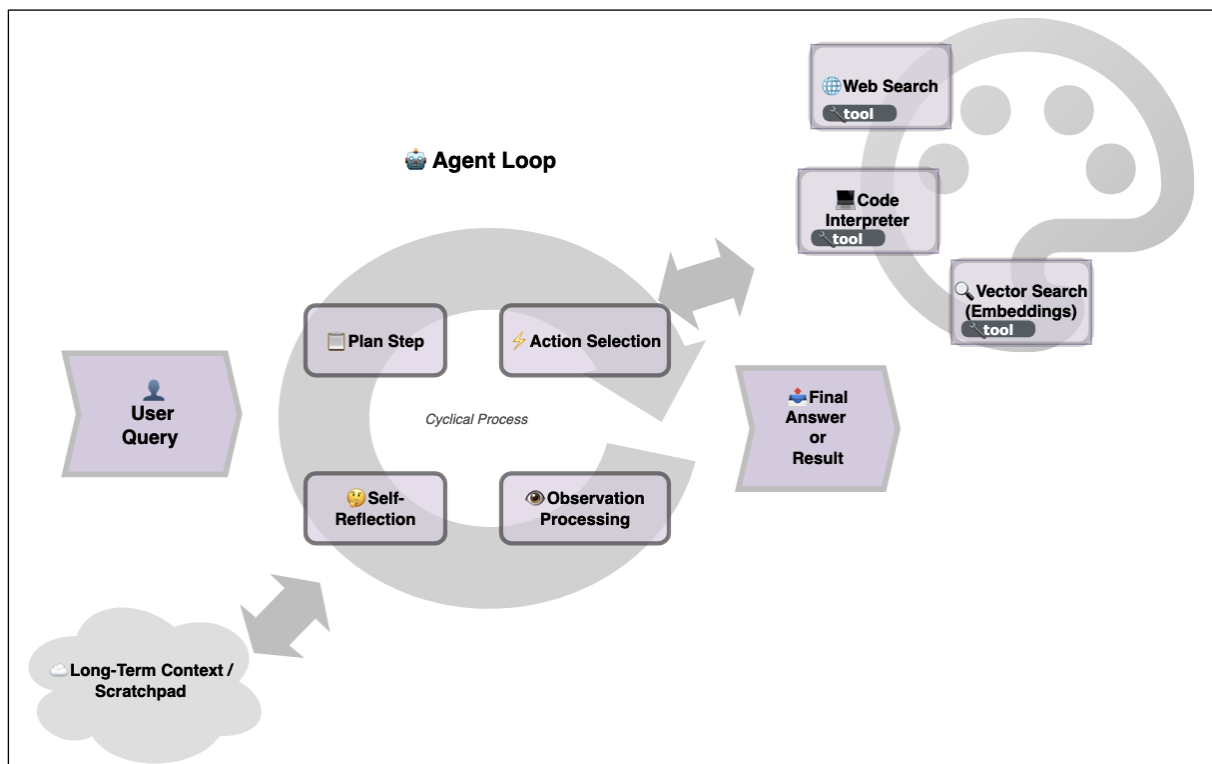
Models returning irrelevant results due to overgeneralization, where documents are matched based on superficial similarities rather than true relevance.

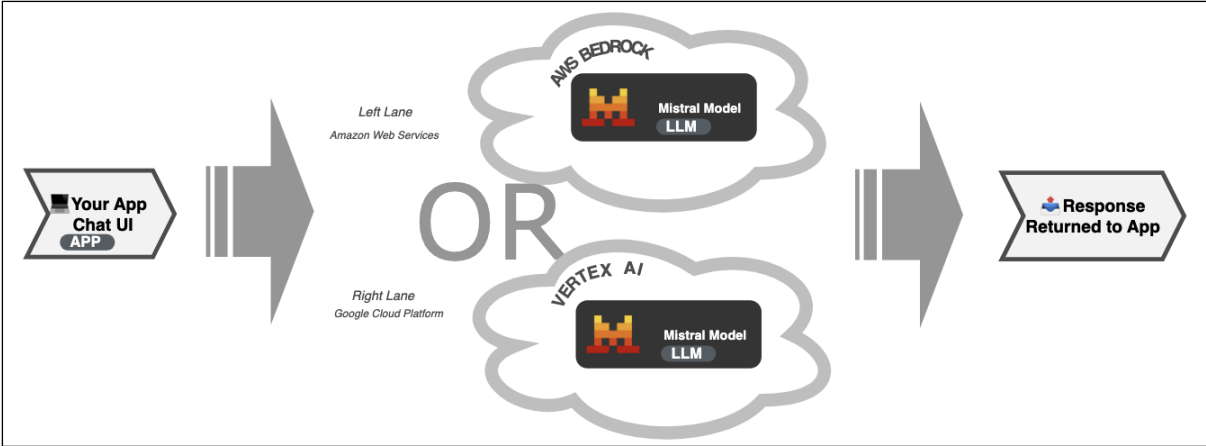
Example:

A search for "jaguar habitats" mistakenly retrieving documents about Jaguar cars.

Potential Solutions:

Implementing re-ranking mechanisms, confidence scoring for retrieved results, and fine-tuning with domain-specific datasets to reduce irrelevant outputs.





Chapter 2: Setting Up Your Own Chat



Get up and running with large language models.

Run [Llama 3](#), [Phi 3](#), [Mistral](#), [Gemma 2](#), and other models. Customize and create your own.

Download ↓

Available for macOS, Linux,
and Windows (preview)

Download Ollama



macOS



Linux



Windows

Download for macOS

Requires macOS 11 Big Sur or later

Install the command line

```
> ollama
```

Install

You will be prompted for
administrator access

🔍 mistral

Embedding

Vision

Tools

Popular



mistral-small3.1

Building upon Mistral Small 3, Mistral Small 3.1 (2503) adds state-of-the-art vision understanding and enhances long context capabilities up to 128k tokens without compromising text performance.

vision tools 24b

↓ 88.2K Pulls 🔖 5 Tags ⌚ Updated 5 weeks ago

mistral

The 7B model released by Mistral AI, updated to version 0.3.

tools 7b

↓ 13.4M Pulls 🔖 84 Tags ⌚ Updated 9 months ago

mistral-nemo

A state-of-the-art 12B model with 128k context length, built by Mistral AI in collaboration with NVIDIA.

tools 12b

↓ 1.7M Pulls 🔖 17 Tags ⌚ Updated 9 months ago

mistral-small

Mistral Small 3 sets a new benchmark in the "small" Large Language Models category below 70B.

tools 22b 24b

↓ 769.3K Pulls 🔖 21 Tags ⌚ Updated 3 months ago

mistral-openorca

Mistral OpenOrca is a 7 billion parameter model, fine-tuned on top of the Mistral 7B model using the OpenOrca dataset.

7b

↓ 170.3K Pulls 🔖 17 Tags ⌚ Updated 19 months ago

mistral-large

Mistral Large 2 is Mistral's new flagship model that is significantly more capable in code generation, mathematics, and reasoning with 128k context window and support for dozens of languages.

tools 123b

↓ 134.4K Pulls 🔖 32 Tags ⌚ Updated 5 months ago

BMO Chatbot

Show installed only

Showing 1 plugin:

BMO Chatbot

INSTALLED

By Longy2k

🔄 17,104

Updated 4 days ago

Generate and brainstorm ideas while creating your notes using Large Language Models (LLMs) such as OpenAI's "gpt-3.5-turbo" and "gpt-4".

Ollama Connection

OLLAMA REST API URL

Enter your REST API URL. Update [Ollama](#) to version >=0.1.42 to avoid CORS restriction.

http://localhost:11434

Enable Stream

Enable Ollama models to stream response.

API Connections

Anthropic



Google Gemini



Mistral AI



Mistral API Key

Insert Mistral API Key.

insert-api-key

Enable Stream

Enable stream for Mistral models.



OpenAI



OpenRouter



BMO Chatbot Settings

[Changelog](#) | [Wiki](#) | [Report a Bug](#) | [Support Me](#)

Type `/help` in chat for commands.

Profiles



General



Model

Choose a model.



✓ No Model

Ollama Models

emoji:latest

mistral-nemo:latest

mistral-small:latest

mistral:latest

llama3:latest

Max Tokens

The maximum number of tokens, or words, that the model is allowed to generate in its output. Some models require a minimum number of tokens to be set. The default value is empty.

BMO Chatbot Settings

[Changelog](#) | [Wiki](#) | [Report a Bug](#) | [Support Me](#)

Type ``/help`` in chat for commands.

Profiles >

General v

Model

Choose a model.



mistral-nemo:latest ▾

General v

Model

Choose a model.



No Model ▾

Max Tokens

The maximum number of tokens, or words, that the model is allowed to generate in its output. Some models require a minimum number of tokens to be set. The default value is empty.

4096

Temperature

Temperature controls how random the generated output is. Lower values make the text more predictable, while higher values make it more creative and unpredictable.

0.20

Enable Reference Current Note

Enable chatbot to reference current active note during conversation.

☐

Prompts v

Prompt

Select a prompt to provide additional context to the system role.

--EMPTY-- ▾

Prompt Folder Path

Select a prompt from a specified folder.

BMO/Prompts

Appearance >

Chat History >

Editor v

Editor System Role

System role for BMO Generate and 'Prompt Select Generate' command.

You are a helpful assistant. 



BMO

mistral:latest

YOU

Tell me about Adams Family

BMO

"The Addams Family" is a popular American comic strip created by Charles Addams, published in The New Yorker magazine from 1938 to 1988. The characters were later adapted into several television shows, films, and other media.

The main family consists of Gomez and Morticia Addams, their children Wednesday and Pugsley, Uncle Fester, Grandmama, and Thing (an animated hand). The Addams Family is known for their peculiar, macabre, and eccentric behavior, which often involves various forms of the supernatural.

The original television series, "The Addams Family," aired from 1964 to 1966 on ABC. It starred John Astin as

Images [Give feedback](#)

An image is a read-only template with instructions for creating a Docker container. [Learn more](#)

Local

Hub

Artifactory **EARLY ACCESS**

4.66 GB / 5.72 GB in use

5 images

Last refresh: about 7 hours ago

Q Search

III

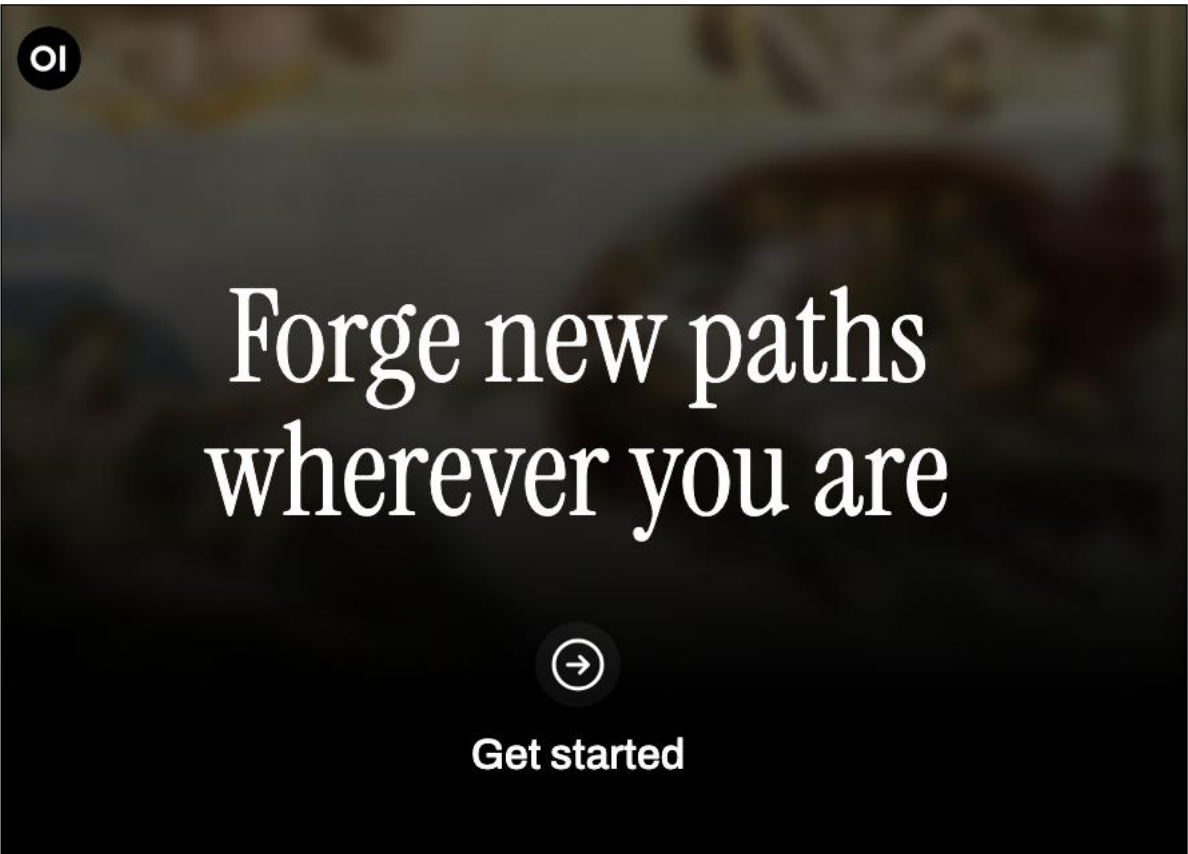
☐	Name	Tag	Status	Created	Size	Actions
☐	ghcr.io/open-webui/open-webui 4df5f45ff574	main	In use	2 months ago	2.67 GB	

Containers [Give feedback](#)

A container packages up code and its dependencies so the application runs quickly and reliably from one computing environment to another. [Learn more](#)

  Only show running containers

☐	Name	Image	Status	Port(s)	Last started	Actions
☐	 open-webui 24040fc24b31 	ghcr.io/open-webui/open-webui:main	Running	3000:8080 	10 days ago	  



OI

Get started with Open WebUI

① Open WebUI does not make any external connections, and your data stays securely on your locally hosted server.

Name

Enter Your Full Name

Email

Enter Your Email

Password

Enter Your Password

Create Admin Account

mistral-nemo:latest

Set as default

P

oi

mistral-nemo:latest

How can I help you today?

+

Code Interpreter

⚡ Suggested

Help me study

vocabulary for a college entrance exam

Give me ideas

for what to do with my kids' art

Overcome procrastination

give me tips

Admin Panel

Playground

Archived Chats

Settings

Sign Out

P

Pavlo

Settings

⚙️ General

☁️ **Connections**

📀 Models

💬 Interface

🔊 Audio

🖼️ Images

💬 Chats

👤 Account

ℹ️ About

OpenAI API

Show

Ollama Base URL

http://host.docker.internal:11434

+

↻

Trouble accessing Ollama? [Click here for help.](#)

Save

Admin Settings

⚙️ **General**

👥 Users

📀 Database

General Settings

Enable New Sign Ups

Default User Role

Webhook URL

JWT Expiration

Valid time units: 's', 'm', 'h', 'd', 'w' or '-1' for no expiration.

Save

🔒 Enabled

pending ▾

Admin Settings

General

Users

Database

User Permissions

Allow Chat Deletion

Allow

Manage Models

Model Whitelisting

On

mistral:latest

▼

+

1 Model(s) Whitelisted

Save

Add User

Form

CSV Import

Role

User▼

Name

Enter Your Full Name

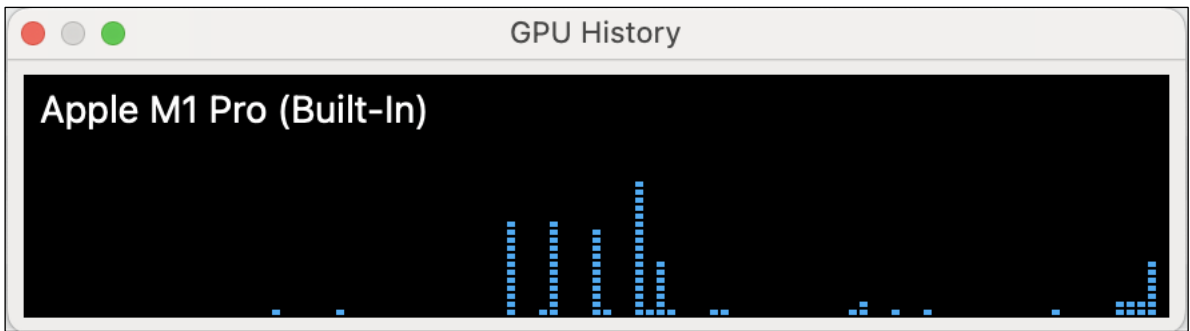
Email

Enter Your Email

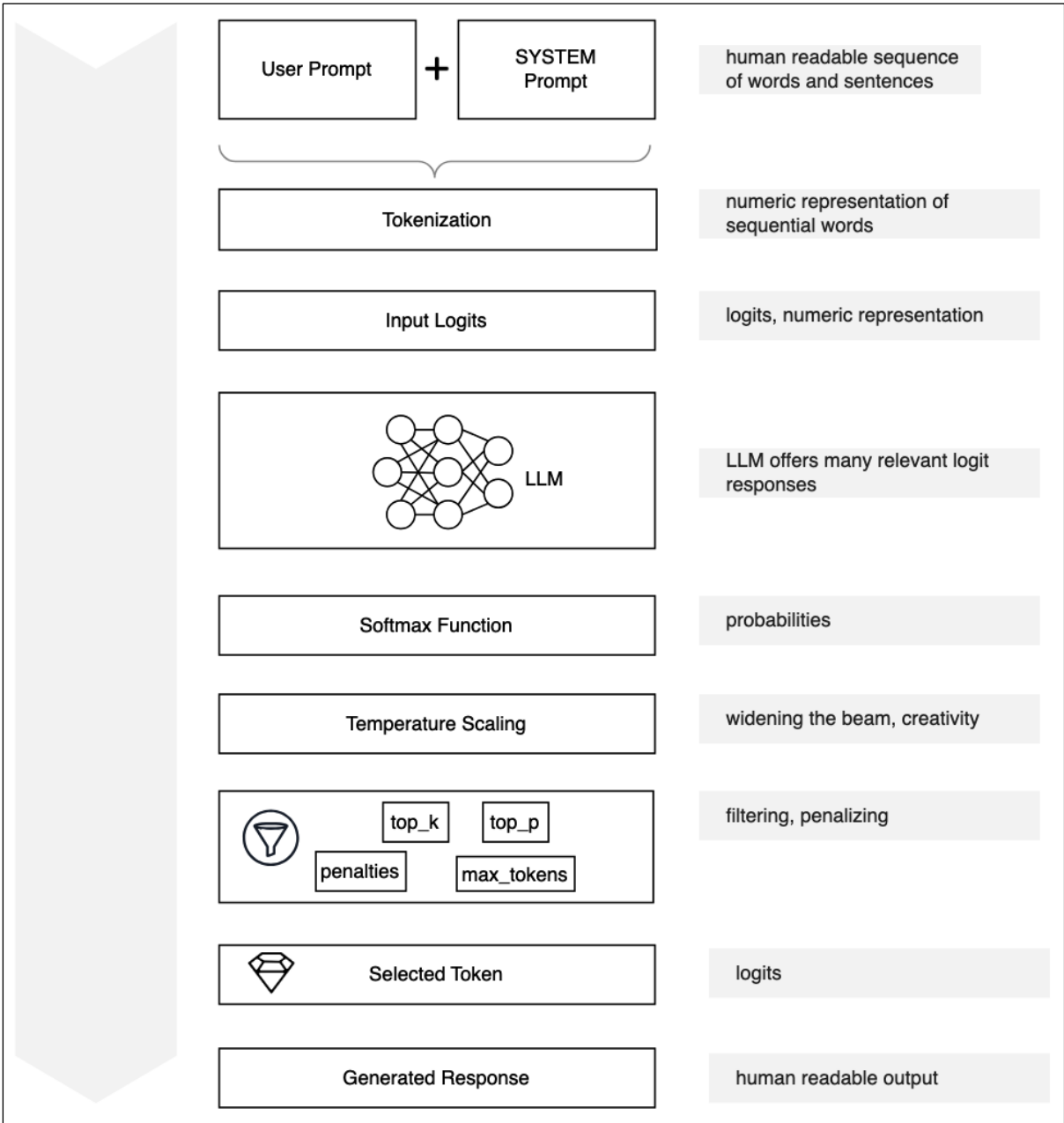
Password

Enter Your Password

Submit



Chapter 3: Managing the Model

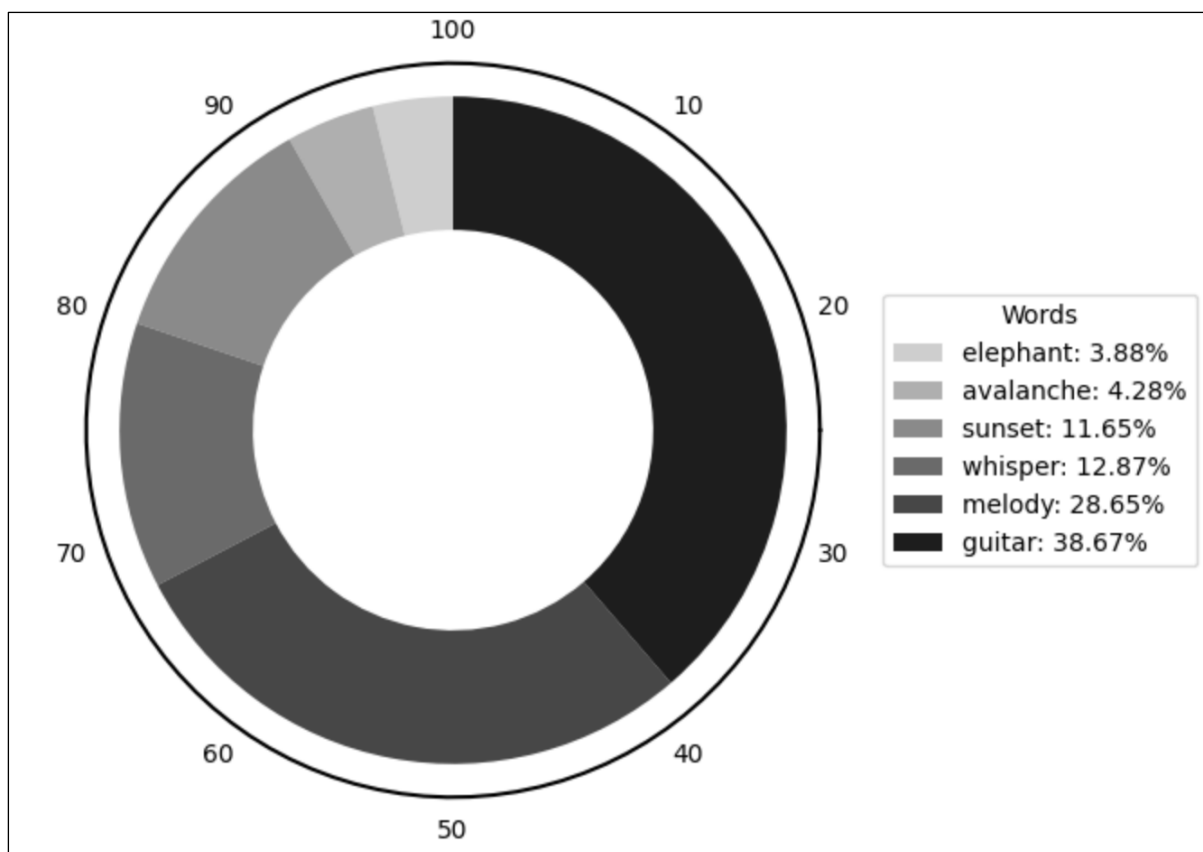


Prompt text:

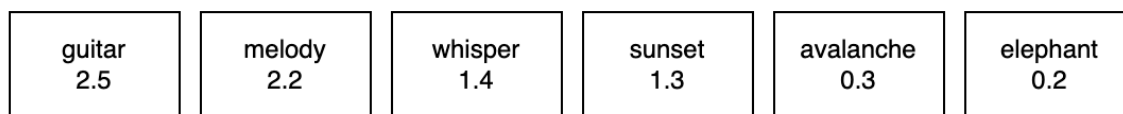
The concert started as the sun dipped below the horizon, casting a warm glow over the stage.The ..

Vector of generated logits sorted by weight:

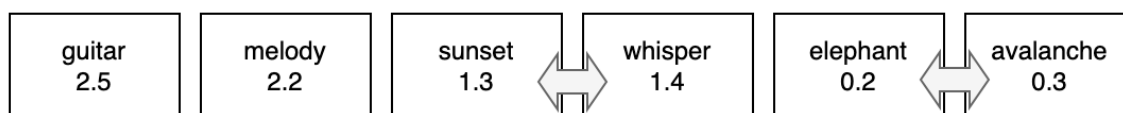
guitar 2.5	melody 2.2	whisper 1.4	sunset 1.3	avalanche 0.3	elephant 0.2
---------------	---------------	----------------	---------------	------------------	-----------------



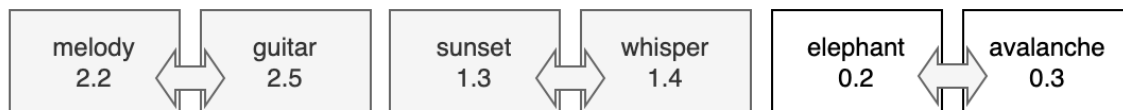
Original sorted vector of logits



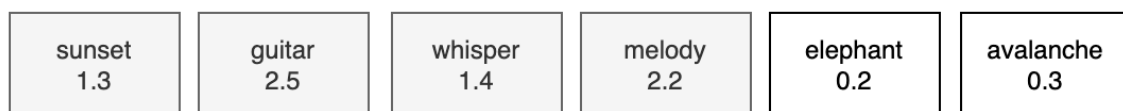
after setting temperature 0.1



after setting temperature 0.4



after setting temperature 0.7



After applying top_k=3

guitar P=38.67%	melody P=28.65%	whisper P=12.87%	sunset P=11.65%	avalanche P=4.28%	elephant P=3.88%
--------------------	--------------------	---------------------	--------------------	----------------------	---------------------

only top three remain

After applying top_p=0.6

guitar P=38.67%	melody P=28.65%	whisper P=12.87%	sunset P=11.65%	avalanche P=4.28%	elephant P=3.88%
--------------------	--------------------	---------------------	--------------------	----------------------	---------------------

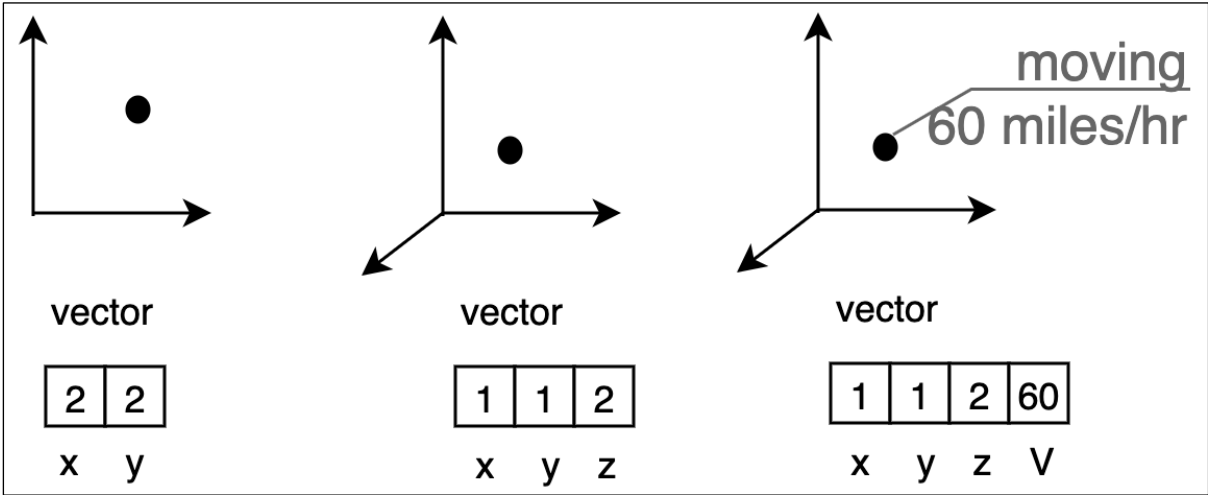
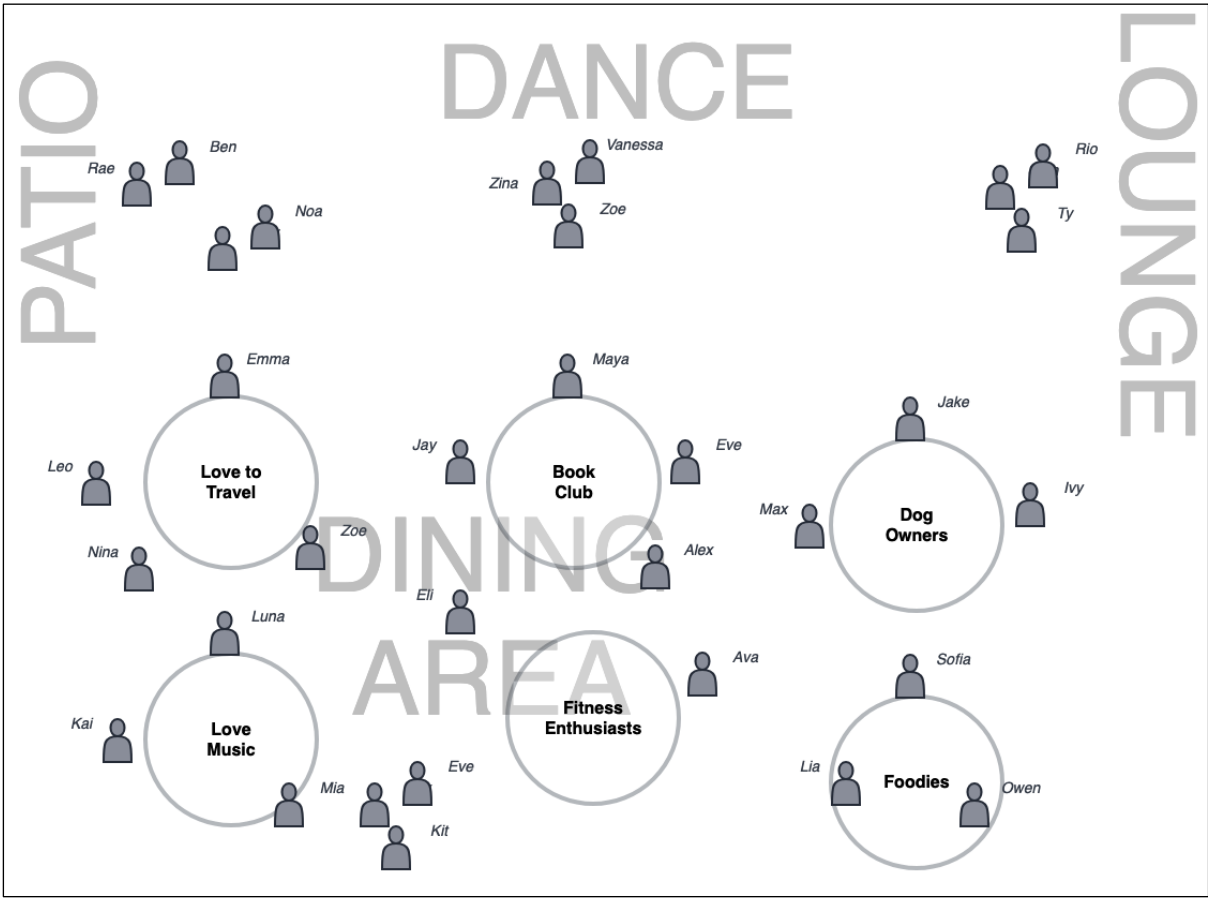
these two make
more than 60%

After applying min_p=0.1

guitar P=38.67%	melody P=28.65%	whisper P=12.87%	sunset P=11.65%	avalanche P=4.28%	elephant P=3.88%
--------------------	--------------------	---------------------	--------------------	----------------------	---------------------

these two make
less than 10%

Chapter 4: Mastering Embeddings



Vector 1

9	8	3	9
---	---	---	---

close

close

FAR

close

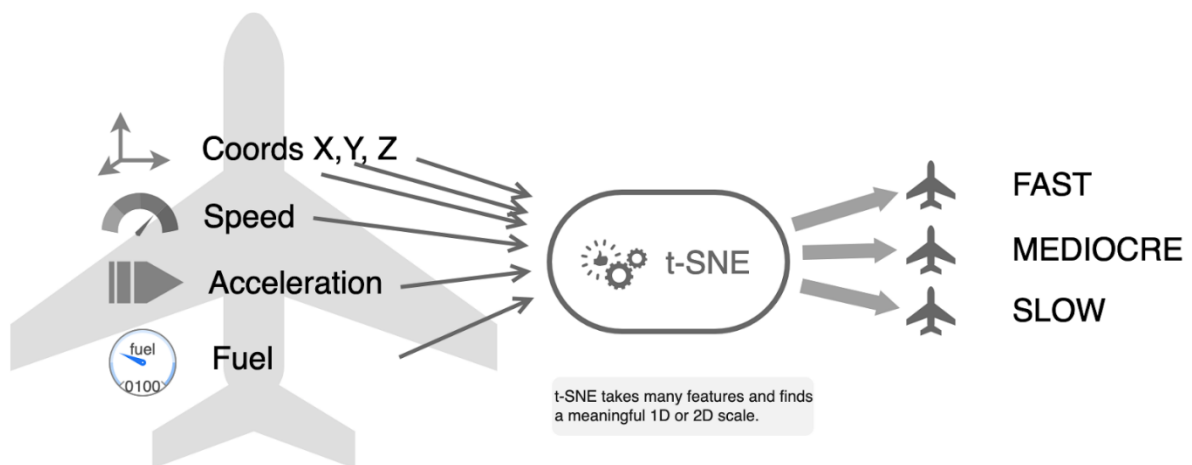
Vector 2

8	9	20	8
---	---	----	---

3 out of 4 close - 75% match

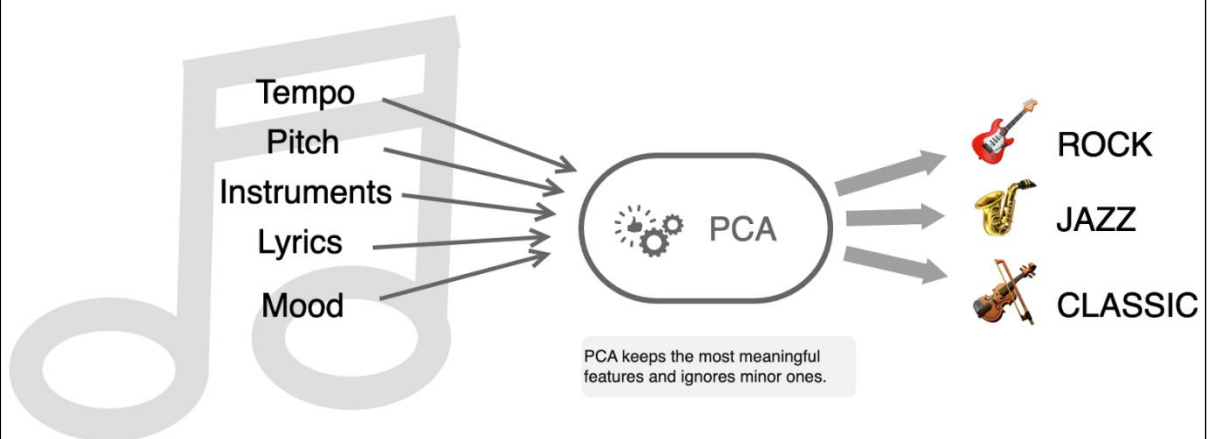
DIMENSIONS

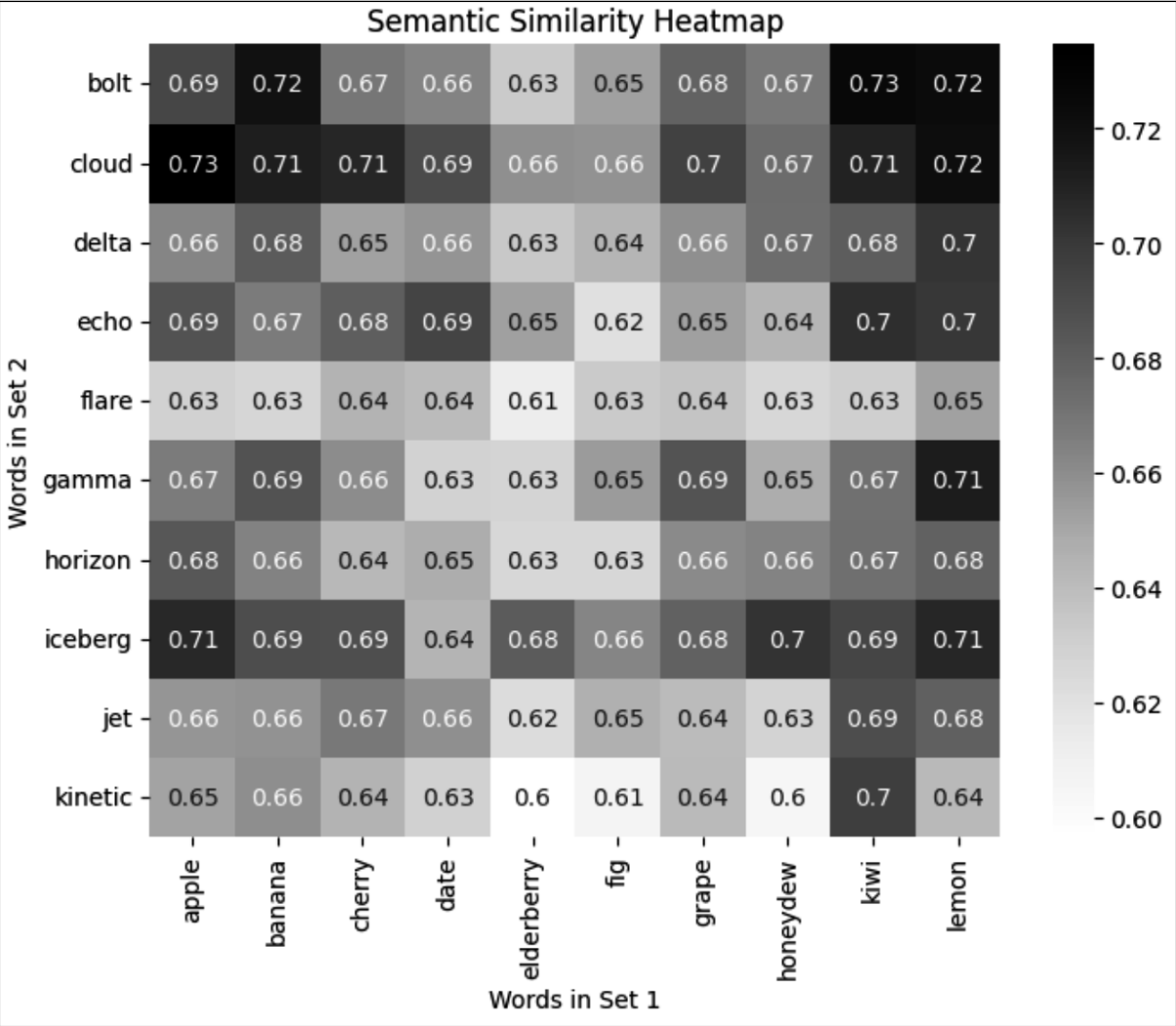
SCALE



DIMENSIONS

SCALE





Change runtime type

Runtime type

Python 3 ▼

Hardware accelerator

☐

CPU

☒

T4 GPU

☐

A100 GPU

☐

L4 GPU

☐

v2-8 TPU

☐

v5e-1 TPU

Want access to premium GPUs? [Purchase additional compute units](#)

Cancel

Save

 Secrets 

 Configure your code by storing environment variables, file paths, or keys. Values stored here are private, visible only to you and the notebooks that you select.

< > Secret name cannot contain spaces.

 Notebook access	Name	Value	Actions
	MISTRAL_API_KEY		  
	PINECONE_API_KEY		  

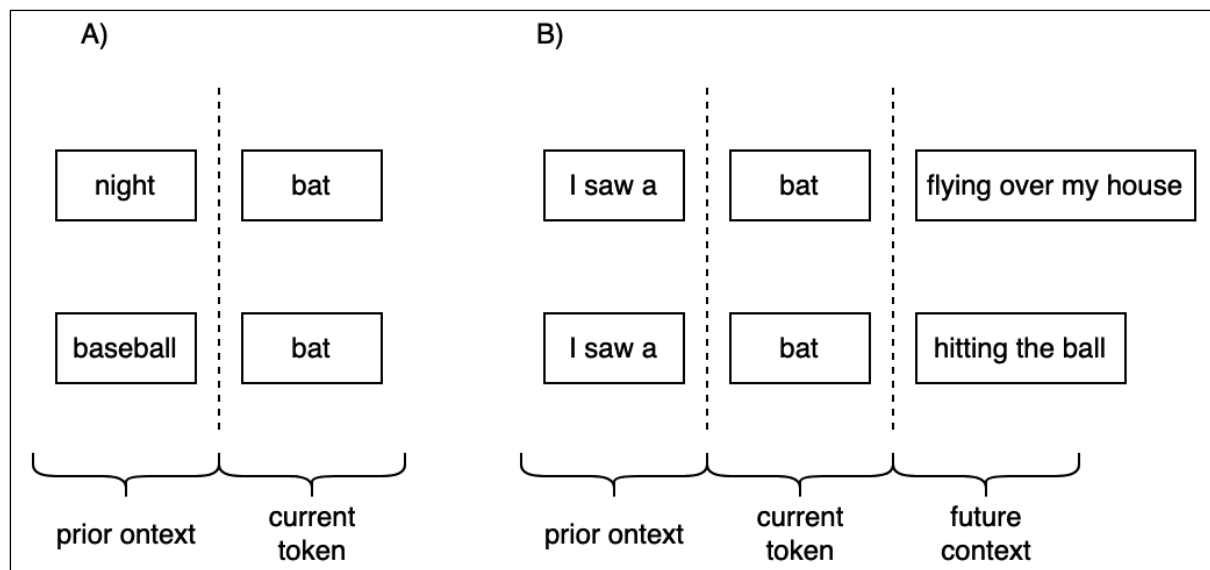
[+ Add new secret](#)

[Gemini API keys](#) ▼

Access your secret keys in Python via:

```
from google.colab import userdata
userdata.get('secretName')
```

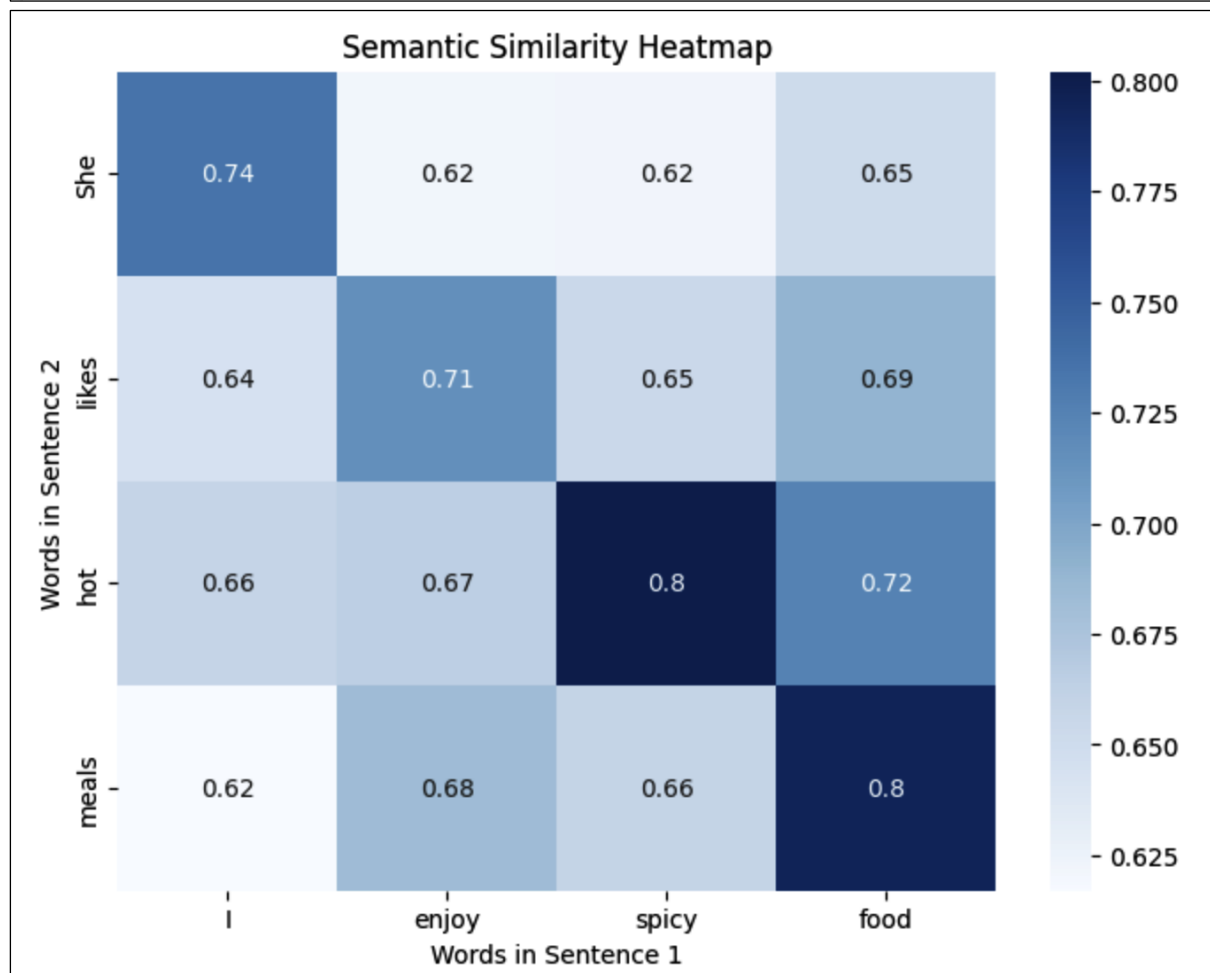


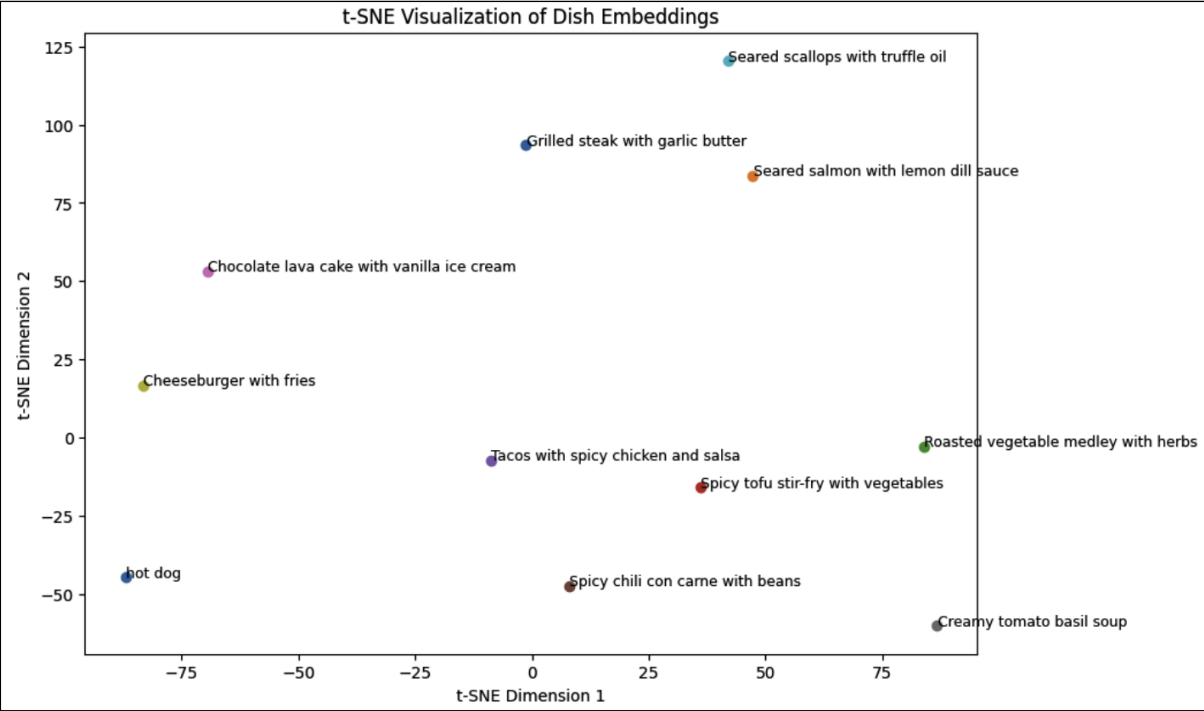
```

user_input = input("What do you prefer for dinner? ")
user_embed = client.embeddings(model="mistral-embed", input=user_input).data[0].embedding
user_embed[:10]

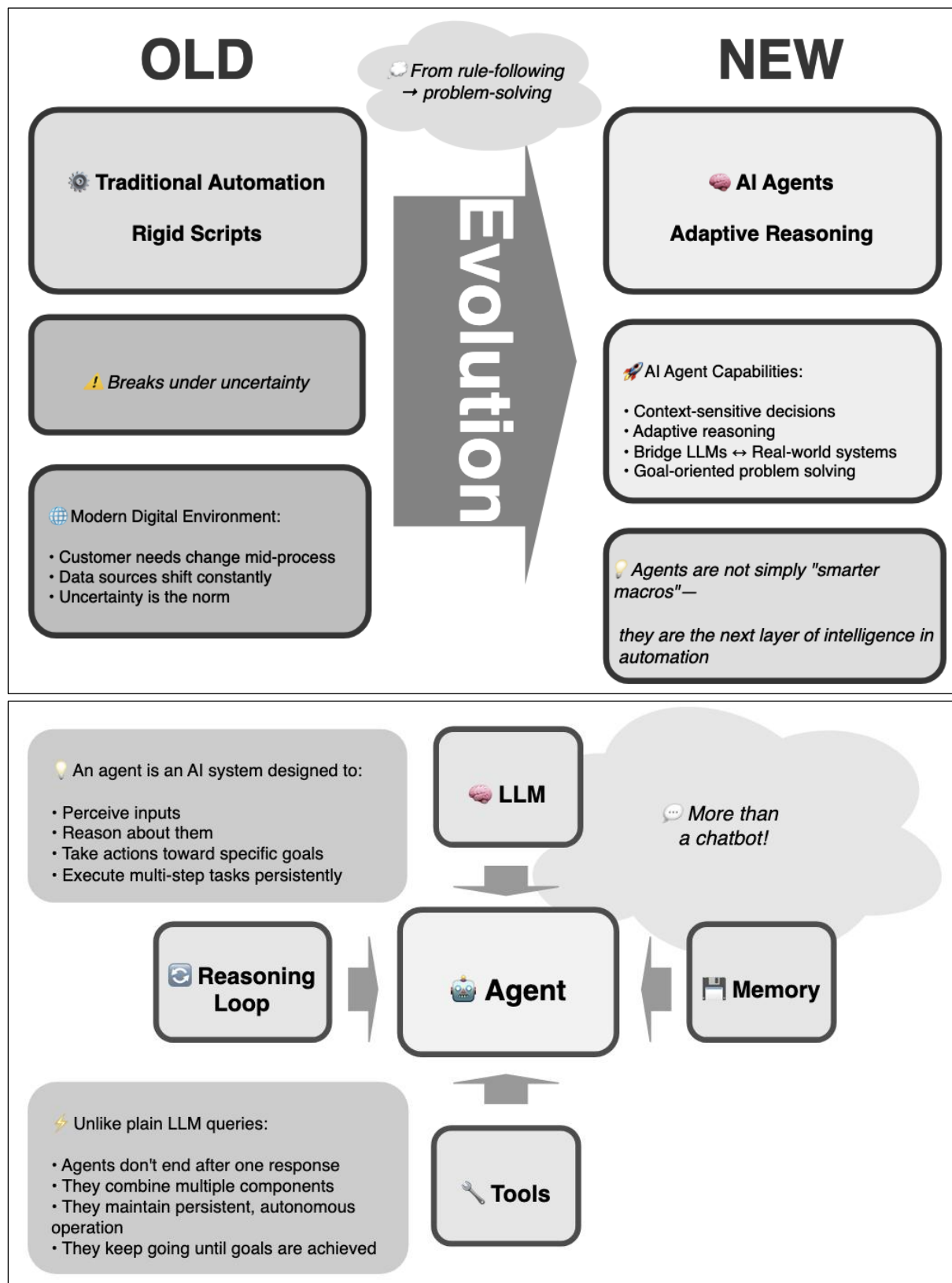
```

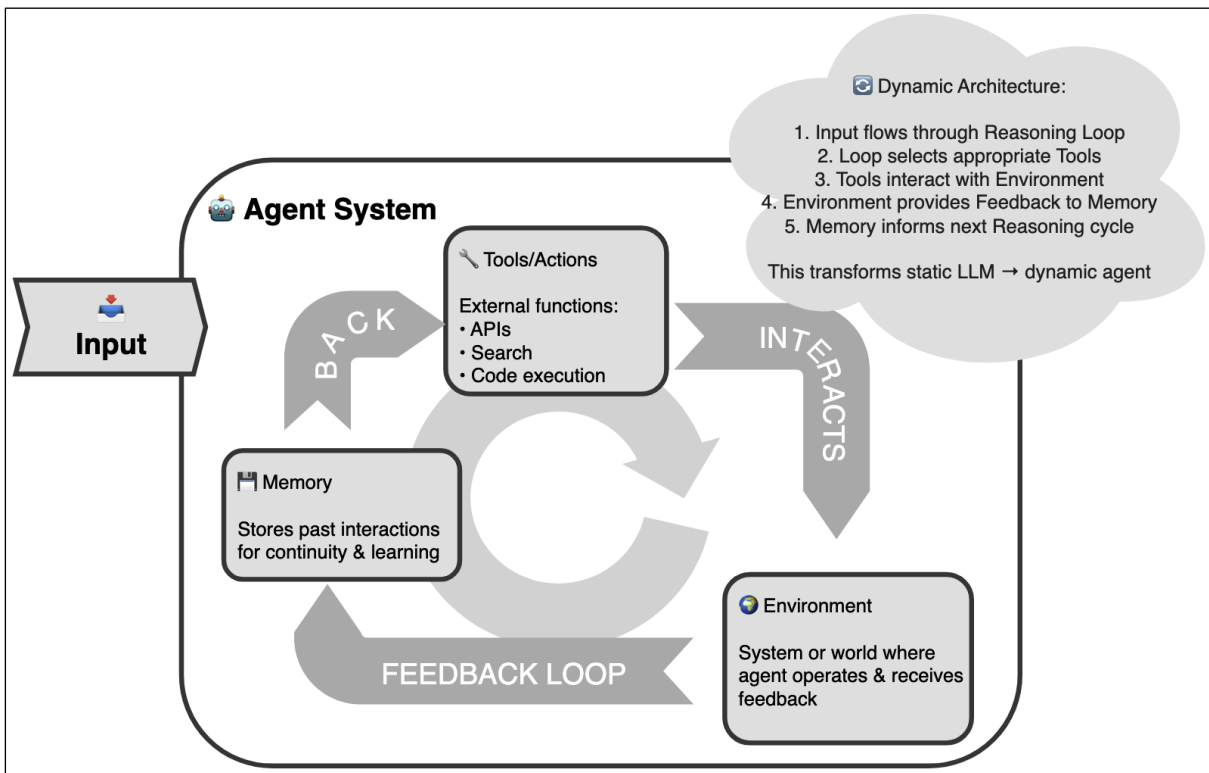
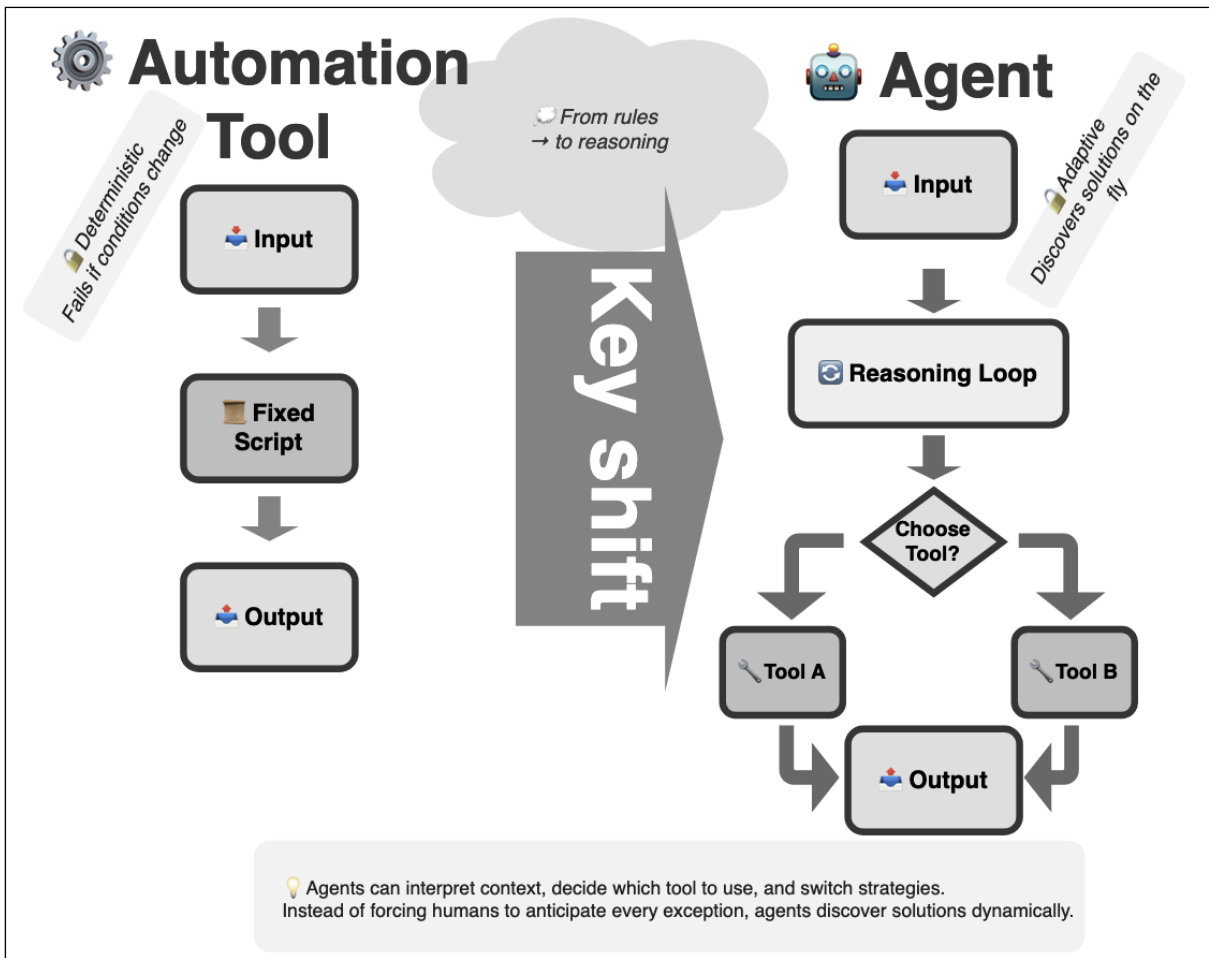
... What do you prefer for dinner?





Chapter 5: Agents: From Automation to Intelligence





✗ **Traditional Automation:**
Breaks with unexpected inputs

✓ **Agent-Based Systems:**
Adapt and find solutions

⚡ Agents are not hype

They are practical solutions for intelligent automation

🧠 Why Agents Matter

The backbone of next-generation intelligent automation

🏢 Business Relevance

- Customer service automation
- Adaptive data pipelines
- Knowledge management
- Direct ROI impact

🔄 Flexible

- Handle unexpected queries
- Adapt where rules fail
- Reason, retry, adapt
- Continue until task done

👤 Human-like Workflows

- Mimic human reasoning
- Multi-step processes
- Context awareness
- Intuitive interactions

📈 Scalable

- Work across multiple tools
- Enterprise-wide deployment
- Handle growing complexity
- Consistent performance

☁ From hype → practical impact

From Theory to Practice: A Hands-On Journey

1 🧠 Simple Agent

- Start with basics
- Use one tool
- Foundation concepts
- Build confidence

python

2 🏗️ Multi-agent System

- Specialized agents
- Collaboration patterns
- Complex workflows
- Scale up complexity

python

3 📊 Performance Evaluation

- Reality vs. hype
- Measure capabilities
- Identify limitations
- Best practices

python

4 🎮 Interactive Demo

- Real-world workflows
- Live demonstrations
- Q&A sessions
- Practical insights

python

🧠 Workshop Takeaways

- Deep understanding of agent systems
- Hands-on practical experience
- Real-world application knowledge
- Ready to implement solutions

Multi-Agent AI System Demo

Test all the agents we've built in this workshop!

Choose Agent:  Customer Service System 

Quick Examples:

Example 1

Example 2

Example 3

Your Query: My app keeps crashing when I upload files

 Submit Query

 Clear Output

Output:

 Processing with  Customer Service System agent...

 Query: My app keeps crashing when I upload files

=====

 Customer Service Response:

 Routed to: Technical Support



Set up owner account

Email *

First Name *

Last Name *

Password *

8+ characters, at least 1 number and 1 capital letter

☐ I want to receive security and product updates

Next

/newbot 3:43 PM ✓✓

Alright, a new bot. How are we going to call it? Please choose a name for your bot. 3:43 PM

MistralFinancialInsight 3:44 PM ✓✓

Good. Now let's choose a username for your bot. It must end in `bot`. Like this, for example: TetrisBot or tetris_bot. 3:44 PM

MistralFinancialInsightBot 3:44 PM ✓✓

Done! Congratulations on your new bot. You will find it at t.me/MistralFinancialInsightBot. You can now add a description, about section and profile picture for your bot, see [/help](#) for a list of commands. By the way, when you've finished creating your cool bot, ping our Bot Support if you want a better username for it. Just make sure the bot is fully operational before you do this.

Use this token to access the HTTP API:

`hP3Fk`

Keep your token **secure** and **store it safely**, it can be used by anyone to control your bot.

For a description of the Bot API, see this page:

<https://core.telegram.org/bots/api>

3:44 PM

4:04



MistralFinancialInsights Group

2 members

Notifications

Off



Add Members



online

Owner



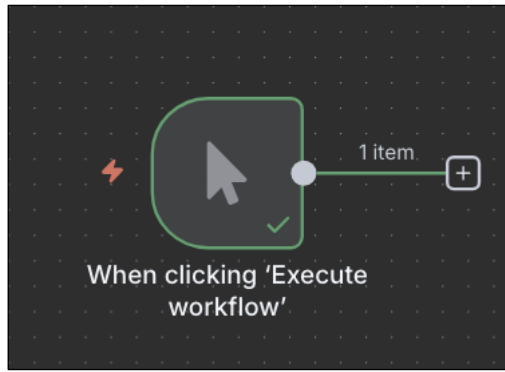
MistralFinancialInsight

has no access to messages



```
api.telegram.org/bot834...hP3Fk/getUpdates
Pretty-print ☒
{
  "chat": {
    "id": -4896759067,
    "title": "MistralFinancialInsights Group",
    "type": "group",
    "all_members_are_administrators": true,
    "accepted_gift_types": {
      "unlimited_gifts": false,
      "limited_gifts": false,
      "unique_gifts": false,
      "premium_subscription": false
    }
  },
  "date": 1756328651,
  "new_chat_participant": {
    "id": 8347743507,
    "is_bot": true,
    "first_name": "MistralFinancialInsight",
    "username": "MistralFinancialInsightBot"
  },
  "new_chat_member": {
    "id": 8347743507,
    "is_bot": true,
    "first_name": "MistralFinancialInsight",
    "username": "MistralFinancialInsightBot"
  },
  "new_chat_members": [
    {
      "id": 8347743507,
      "is_bot": true,
      "first_name": "MistralFinancialInsight",
      "username": "MistralFinancialInsightBot"
    }
  ]
}
```

```
"chat": {
  "accepted_gift_types": {
    "limited_gifts": false,
    "premium_subscription": false,
    "unique_gifts": false,
    "unlimited_gifts": false
  },
  "all_members_are_administrators": true,
  "id": -4896759067,
  "title": "MistralFinancialInsights Group",
  "type": "group"
},
```



Parameters

Settings

Execute step

Import cURL

Method

GET

URL

https://query1.finance.yahoo.com/v1/finance/screener/predefined/saved?count=100&scrls=most_actives

Authentication

None

Send Query Parameters

Send Headers

Send Body

Options

No properties

Add option

ParametersSettings

Execute step

Import cURL

Method

GET

URL

https://query1.finance.yahoo.com/v1/finance/screener/predefined/saved?count=100&scrlids=most_actives

Authentication

None

Send Query Parameters

Send Headers

Send Body

Options

No properties

Add option

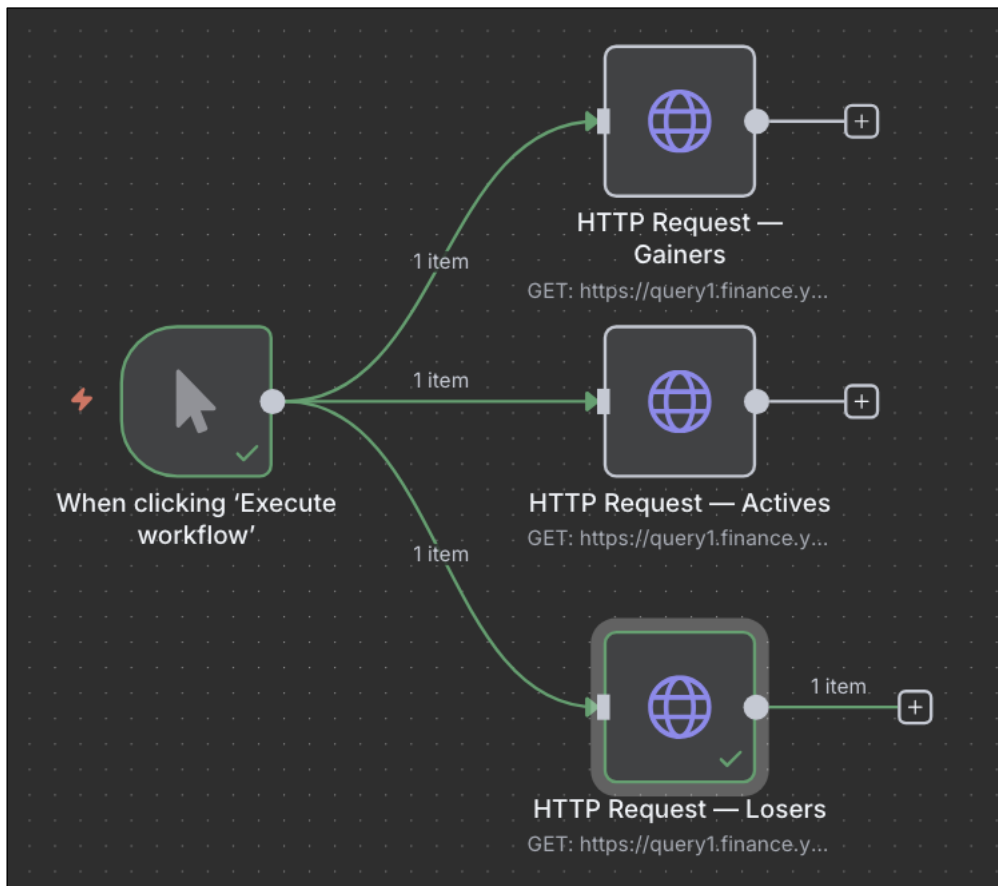
You can view the raw requests this node makes in your browser's developer console

OUTPUT

SchemaTableJSON

1 item

```
[{"finance": {"result": [{"id": "437465ef-980e-4d8c-a860-de7cbfbab373", "title": "Most Actives", "description": "Discover the most traded equities in the trading day.", "canonicalName": "MOST_ACTIVES", "criteriaMeta": {"size": 100, "offset": 0, "sortField": "dayvolume", "sortType": "DESC", "quoteType": "EQUITY", "includefields": ["change_in_number_of_institutional_holders", "trading_central_last_close_price_to_fair_value", "intradaypricechange", "estimated_revenue_growth", "intradaymarketcap", "morningstar_previous_rating", "fiftytwohigh", "pctheldinst", "morningstar_last_close_price_to_fair_value", "shares_bought_by_funds", "ror_percent"]}]}
```



ParametersSettings

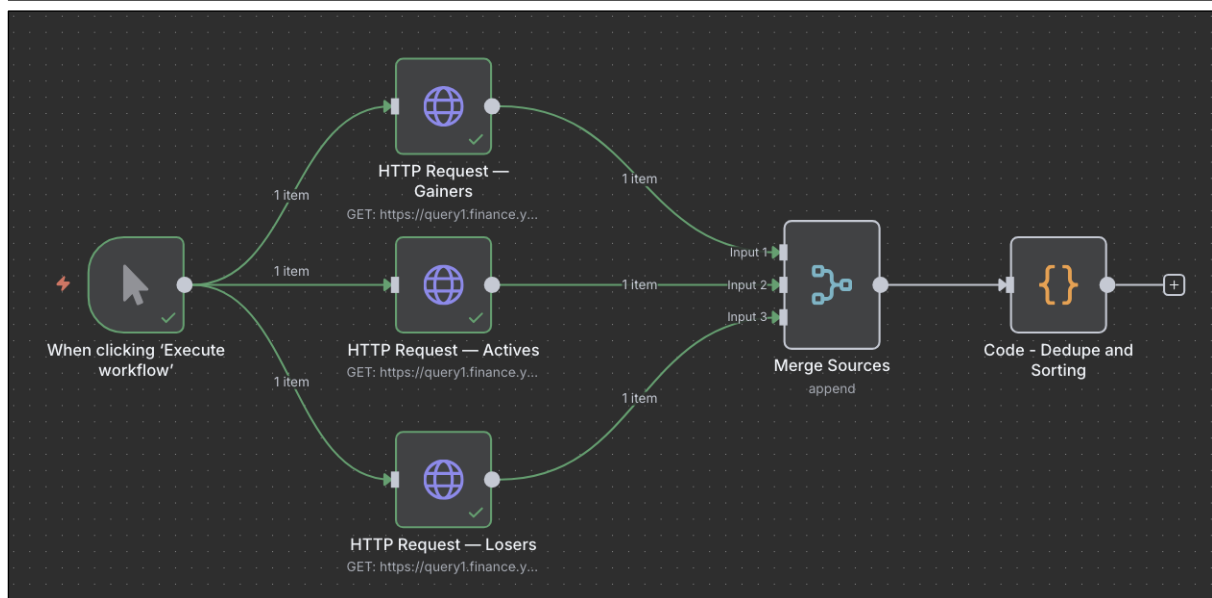
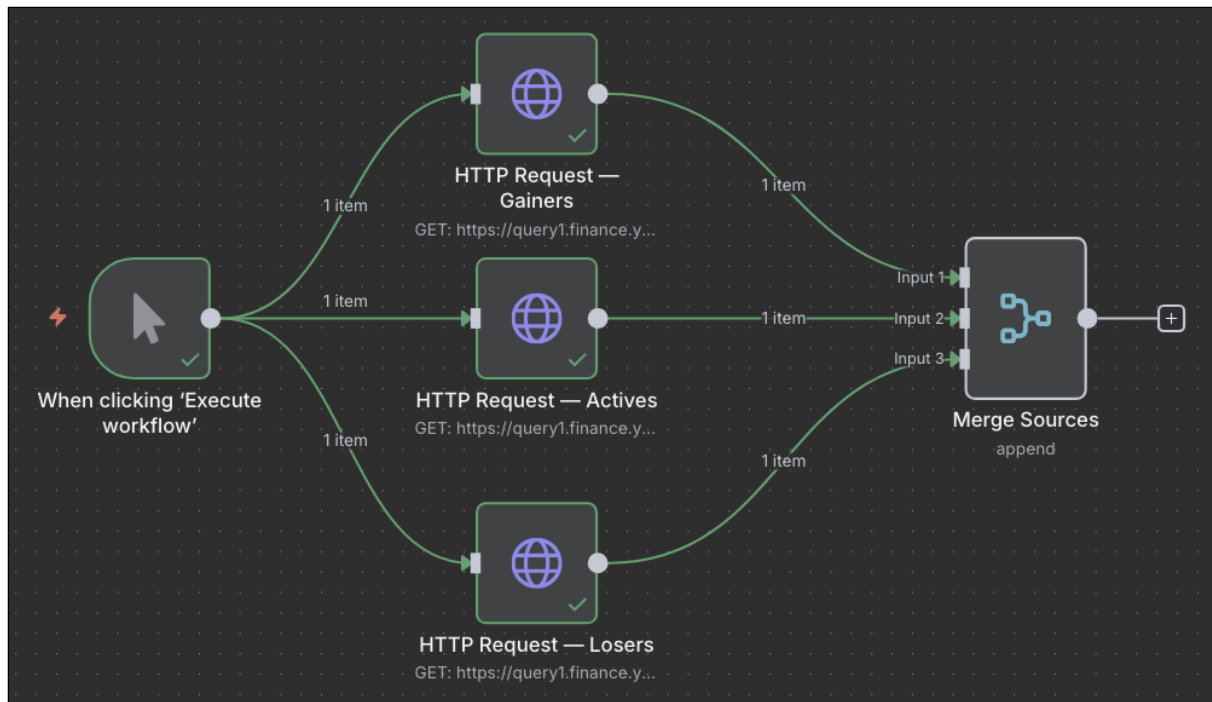
Execute step

Mode

Append

Number of Inputs

3



OUTPUT  



Schema

Table

JSON



To make sure expressions after this node work, return the input items that produced each output item. [More info](#)

5 items

```
[
  {
    "symbol": "PSTG",
    "changePct": 32.3365
  },
  {
    "symbol": "ARX",
    "changePct": -26.3571
  },
  {
    "symbol": "SNOW",
    "changePct": 20.2655
  },
  {
    "symbol": "BILL",
    "changePct": 18.184
  },
  {
    "symbol": "TLX",
    "changePct": -16.1157
  }
]
```



Parameters

Settings

Execute step

Import cURL

Method

GET

URL

fx

https://query1.finance.yahoo.com/v8/finance/chart/{{ \$json.symbol }}

[ERROR: No input connected]

Authentication

None

Send Query Parameters

Specify Query Parameters

Using Fields Below

Query Parameters

Name

interval

Value

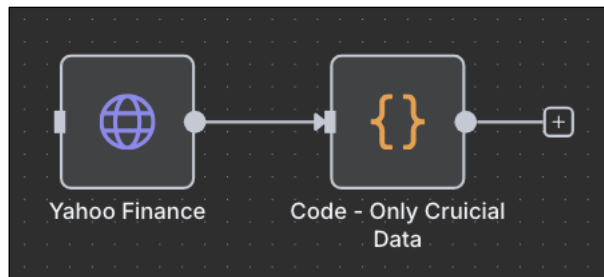
5m

Name

range

Value

5d



Parameters

Settings

Execute step

Import cURL

Method

GET

URL

fx

https://news.google.com/rss/search?q={{ \$json.symbol }}%20stock%20when:2d&hl=en-US&gl=US&ceid=US:en

[ERROR: No input connected]

Authentication

None

Send Query Parameters

Send Headers

Specify Headers

Using Fields Below

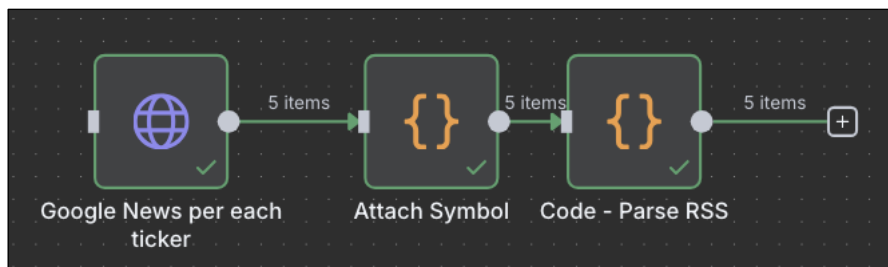
Header Parameters

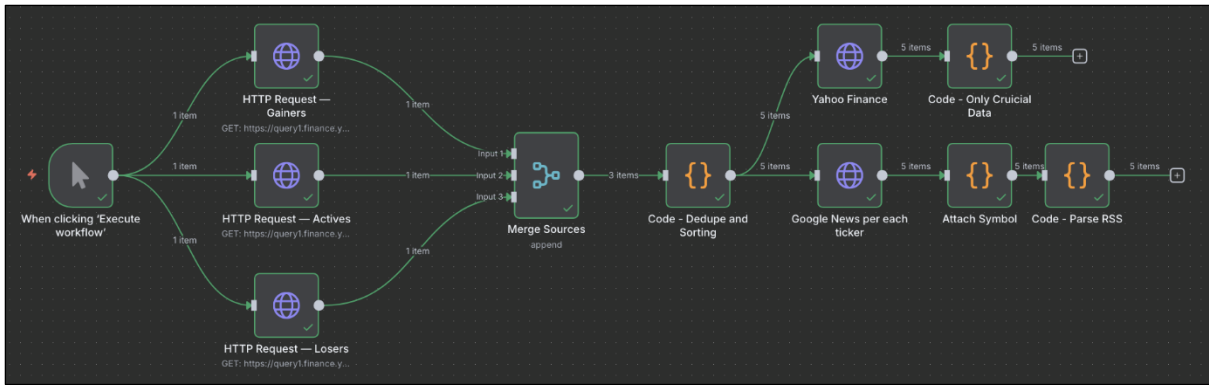
Name

Accept

Value

/





Parameters

Settings

Execute step

Mode

Combine

Combine By

Matching Fields

Fields To Match Have Different Names

☐

Fields to Match

symbol

Drag or type the input field name

Output Type

Keep Matches

Output Data From

Both Inputs Merged Together

Options

No properties

Add option

OUTPUT  



Schema Table **JSON**

5 items

```
[
  {
    "symbol": "PSTG",
    "price": 80.54,
    "change": -0.04,
    "changePercent": -0.05,
    "z1": -0.3,
    "zH": 1.78,
    "moveVsATR": 0.12,
    "volSpike": false,
    "volatilityFlag": false,
    "timestamp": "2025-08-28T20:00:00.000Z",
    "news": [
      "Why Pure Storage Stock Is Soaring After Its Results While Data Storage Rival NetApp Wavers - Investor's Business Daily - &lt;a href=\"https://news.google.com/rss/articles/CBMinwFBVV95cUxNS3VLTj1hTXVPdjhLZ3dHVGPly05TZzJJR2xvLbUpaV1UyWnk4a1JuVWVUc2Y5bVI5LULPNW1scExqSzl1aTnNmjd5Y0pPejk5XzhUYWc1blozZDJ0b0V1bWZZ0EVyY2p1VGh0YTllUNEJwem1YSVFRWEJsbC1BdHZ\", \"How to interpret RSI for PSTG stock - 2025 Biggest Moves &amp; Real-Time Sentiment Analysis - N &lt;a href=\"https://news.google.com/rss/articles/CBMieEFVX31xTE1rR1FmNG9EdEFTWlN1czRlWmhUWGtKbEU2TTloNQ3BUc1ppcDJDRDdZdE1FN2k3b2psNHhXZmZlcnJyVWJUazBtSzdkbXM1RlZmZmVWJFVUVNSDAXcVpsMDJGRnpPR1RpVHVSMNQ?oc=5\" target=\"_blank\"&gt\", \"$PSTG stock is up 28% today. Here's what we see in our data. - Quiver Quantitative - &lt;a href=\"https://news.google.com/rss/articles/CBMioAFBVV95cUxPaLZrSW52eDZwbzg3UjY5Y3pIU2l1eHhtczNXW eDdvSEVYT2VfS05la2FpaURJRktfUWdGd3A3c1F2N2swbTJD b1Q1ZXpXWkppQ0JlEVE0dTNPZVYwdkhVSzZlZXpDeDdfWXRHL T2NuZzhhMWNTMFg1Yldiem9VV0E\", \"Pure Storage Stock (PSTG) Rises 30% on News of Meta Platforms Deal - TipRanks - &lt;a href=\"https://news.google.com/rss/articles/CBMimAFBVV95cUxOdDFjYVvk5a1JuYTJYSHo4YkVJRWY0YlFzUS1DS WLB6T3JNdW5tZUhsVvdTejlpX052YkFoeVJ6R1NCSjl0RHFKbC1EcNv3eU1YT2thTk1HVk82aUJRbS1HTWhRZmctd0hQRlhLV V281V0FXZ29LX3RNM1lBaHNuTGJ\", \"Pure Storage (PSTG) Raises Revenue Guidance and Completes US$390 Million Share Buyback - Yahoo - &lt;a href=\"https://news.google.com/rss/articles/CBMihAFBVV95cUxPRHZvSzNzbDZDdkNmbkdqbz1XeEhwNDBRMXBvN NHpRU09lWXNOTkFs0FB0b2Fpc09vNmNwNF83aGZ0VTFKWnZ2NnIwc2VmWTFVMXctR3JaX1Z4UnhmLV9udZzZGNjF0cWdQcWMwS Z1lHVnhhdjk3ZUtUbky?oc=5\" t"
    ]
  },
]
```

Parameters

Settings

 Execute step

Import cURL

Method

POST

URL

https://api.mistral.ai/v1/chat/completions

Authentication

Predefined Credential Type

Credential Type

Mistral Cloud API

Mistral Cloud API

Mistral Cloud account

Send Query Parameters



Send Headers



Send Body



Body Content Type

JSON

Specify Body

Using JSON

JSON

 `{{ $json }}`



[ERROR: No input connected]

Mistral Cloud account

Mistral Cloud API

Connection

Sharing

Details

Connection tested successfully

Retry

Need help filling out these fields?

Open docs

API Key *

.....

Enterprise plan users can pull in credentials from external vaults.

More info

OUTPUT

5 items

SchemaTableJSON

```
[
  {
    "id": "92fef25cc4694710af3af99ae3257fed",
    "created": 1756429288,
    "model": "mistral-small-latest",
    "usage": {
      "prompt_tokens": 1046,
      "total_tokens": 1105,
      "completion_tokens": 59
    },
    "object": "chat.completion",
    "choices": [
      {
        "index": 0,
        "finish_reason": "stop",
        "message": {
          "role": "assistant",
          "tool_calls": null,
          "content": "*PSTG*\nScore: *7/10*\nRecommendation: *BUY*\nProjection (1w): *5.25%*\nRational, Positive news and strong performance indicate upward momentum.\nUncertainty: Volat and market conditions may impact short-term gains."
        }
      }
    ]
  }
],
```

Parameters

Settings

Execute step

Credential to connect with

Telegram account

Resource

Message

Operation

Send Message

Chat ID

-4896759067

Text

fx

{{ \$json.choices[0].message.content }}

PSTG Score: *7/10* Recommendation: *BUY* Projection (1w): *5.25%* Rationale: Positive news an...

Reply Markup

None

Additional Fields

No properties

Add Field

Telegram account

Telegram API

Connection

Sharing

Details

Connection tested successfully

Retry

Need help filling out these fields? [Open docs](#)

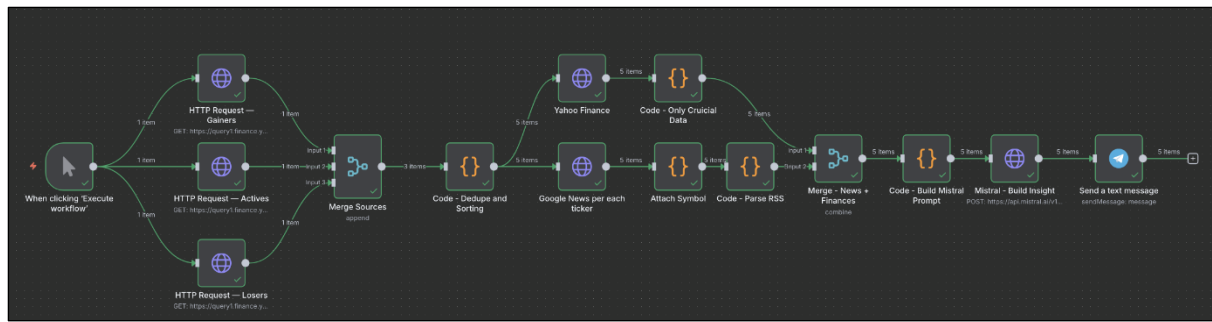
Access Token

.....

Base URL

https://api.telegram.org

Enterprise plan users can pull in credentials from external vaults. [More info](#)



4:31

84



MistralFinancialInsights

2 members



customer growth boost stock performance.
Uncertainty: Volatility may persist due to recent surge and high investor interest.

This message was sent automatically with n8n

4:29 PM

SATS

Score: 6/10

Recommendation: **HOLD**

Projection (1w): 1.25%

Rationale: Recent surge may have peaked, but fundamentals remain strong.

Uncertainty: Volatility and news-driven price swings add risk.

This message was sent automatically with n8n

4:29 PM

OPEN

Score: 6/10

Recommendation: **HOLD**

Projection (1w): 2.50%

Rationale: Mixed news and technical indicators suggest a neutral stance.

Uncertainty: Recent volatility may impact short-term projections.

This message was sent automatically with n8n

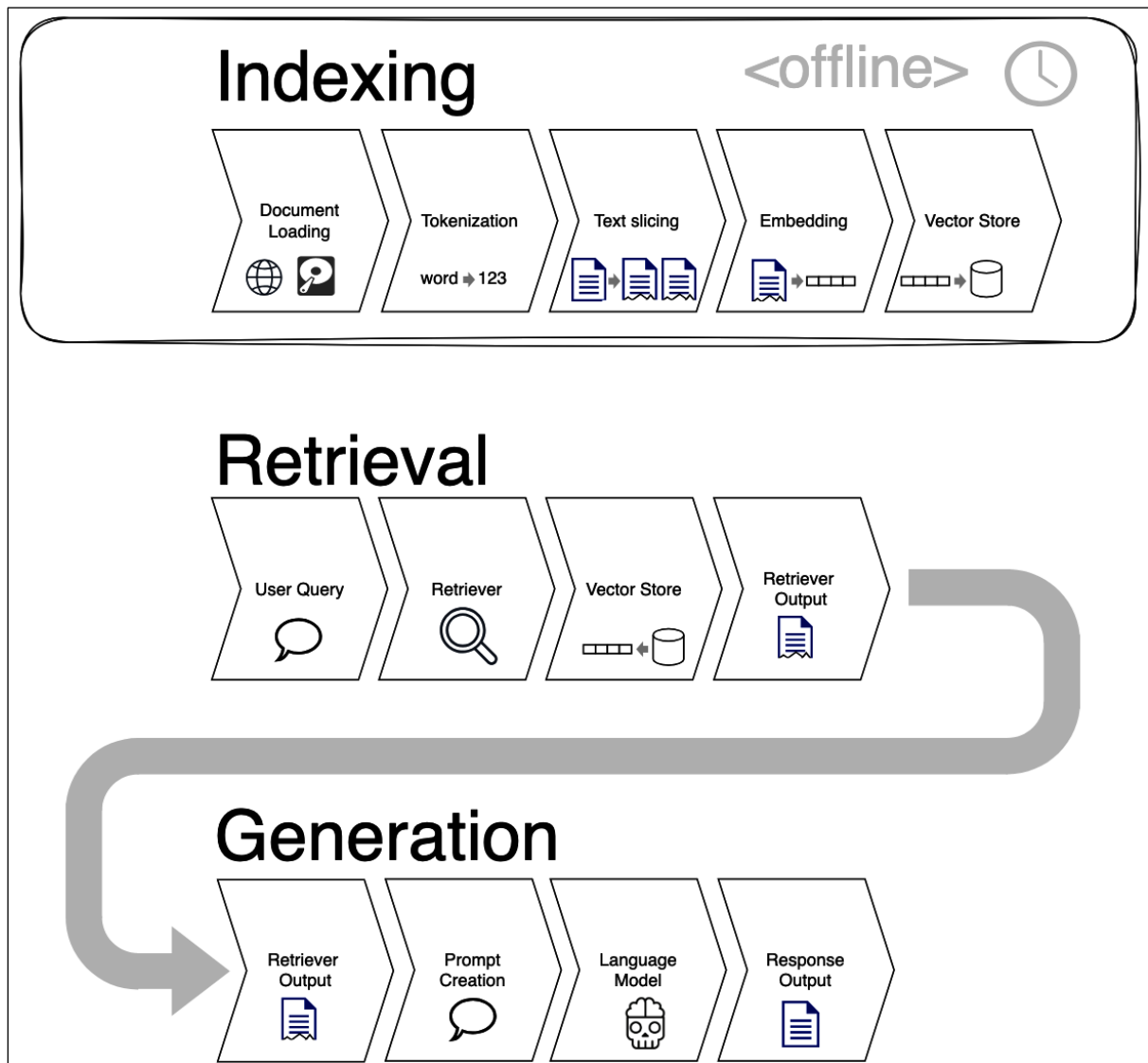
4:29 PM

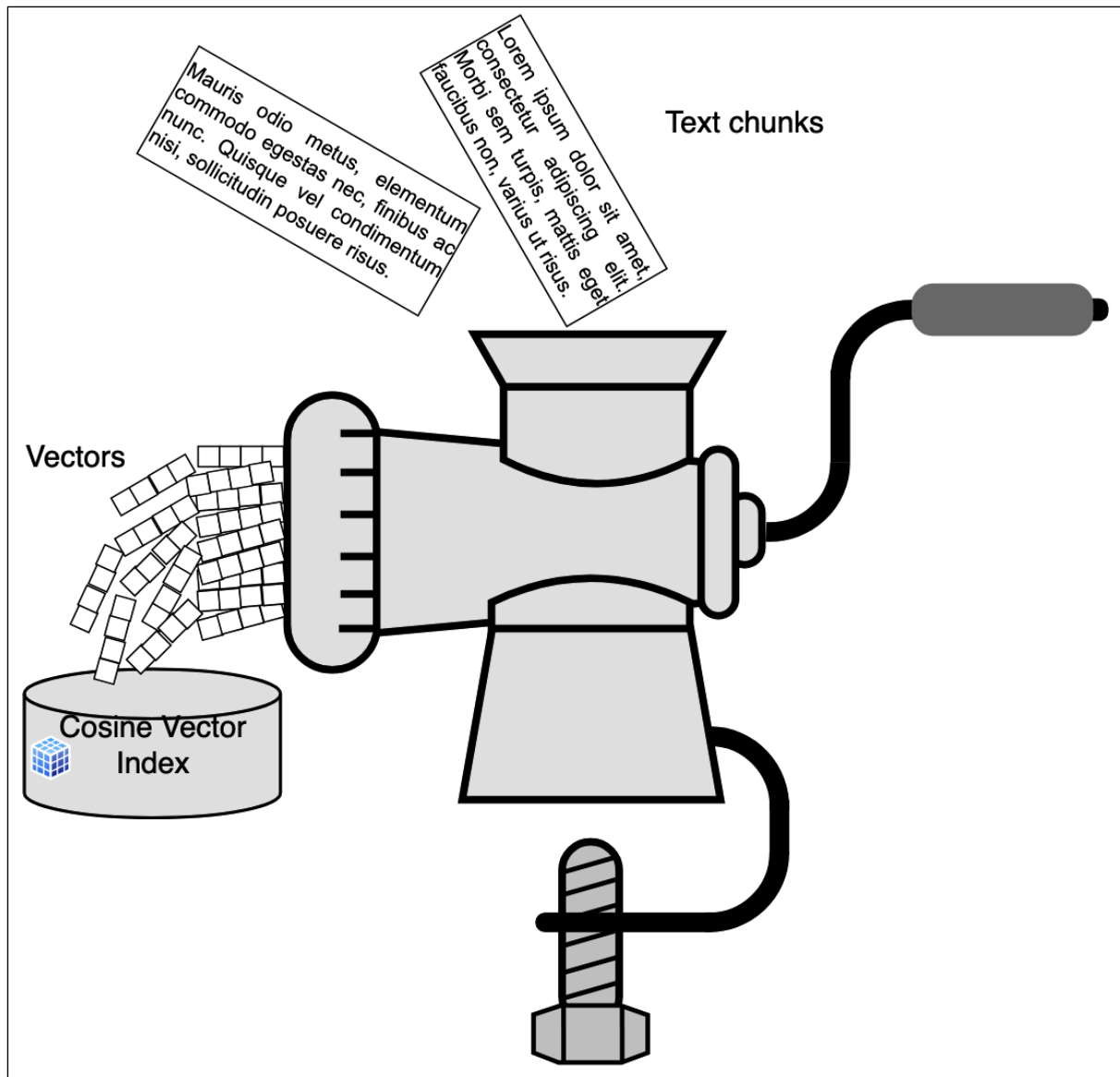


Message

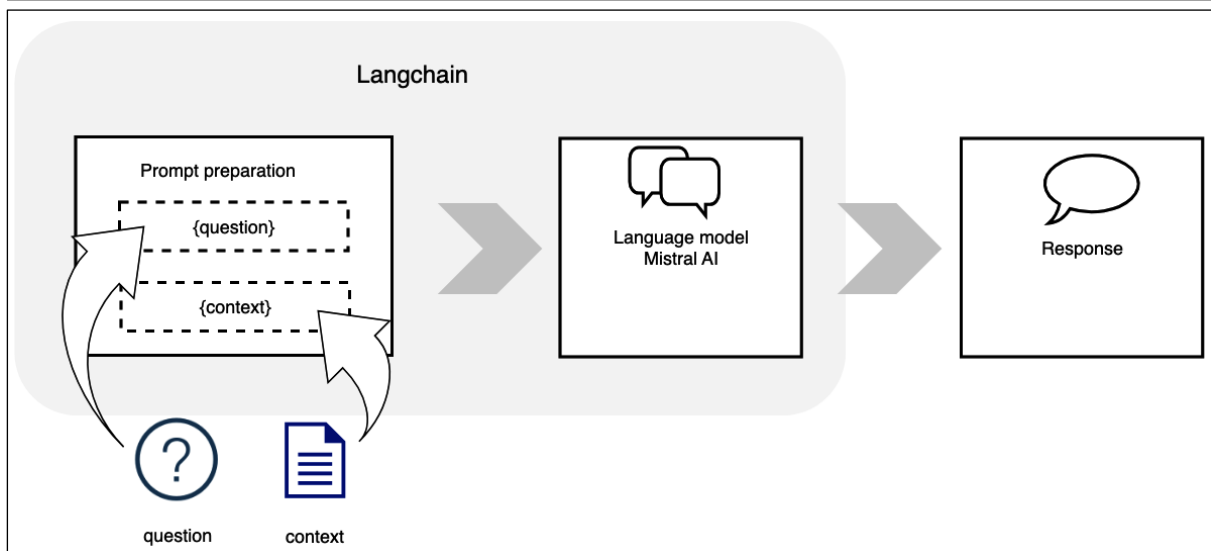
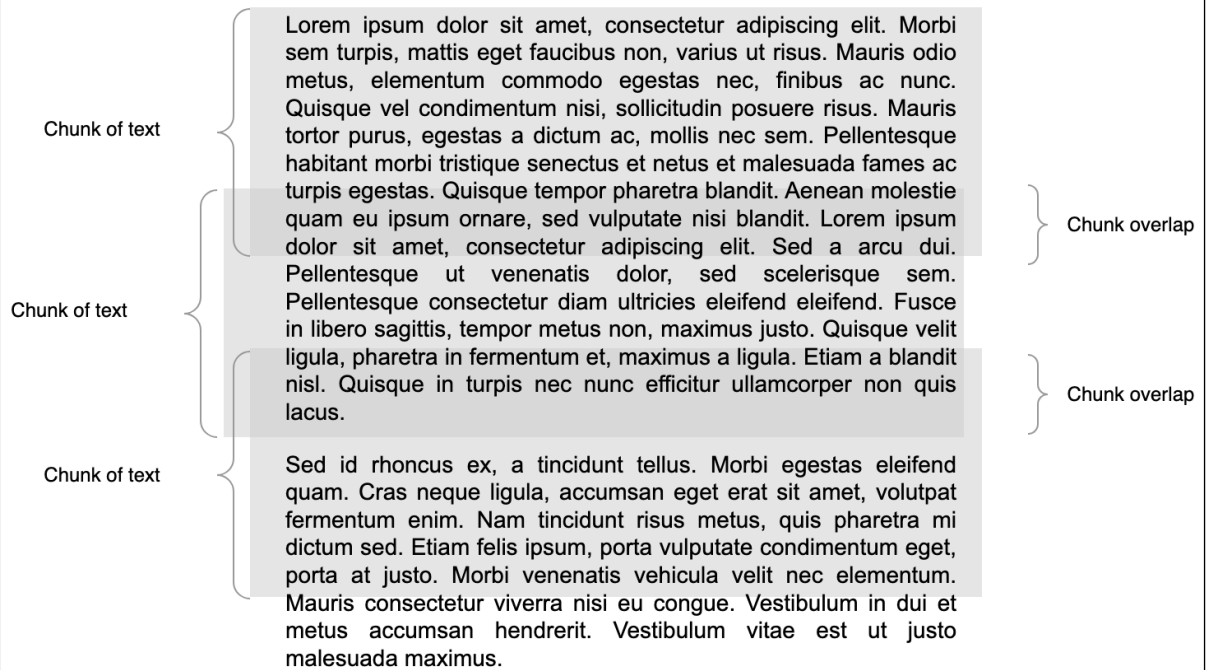


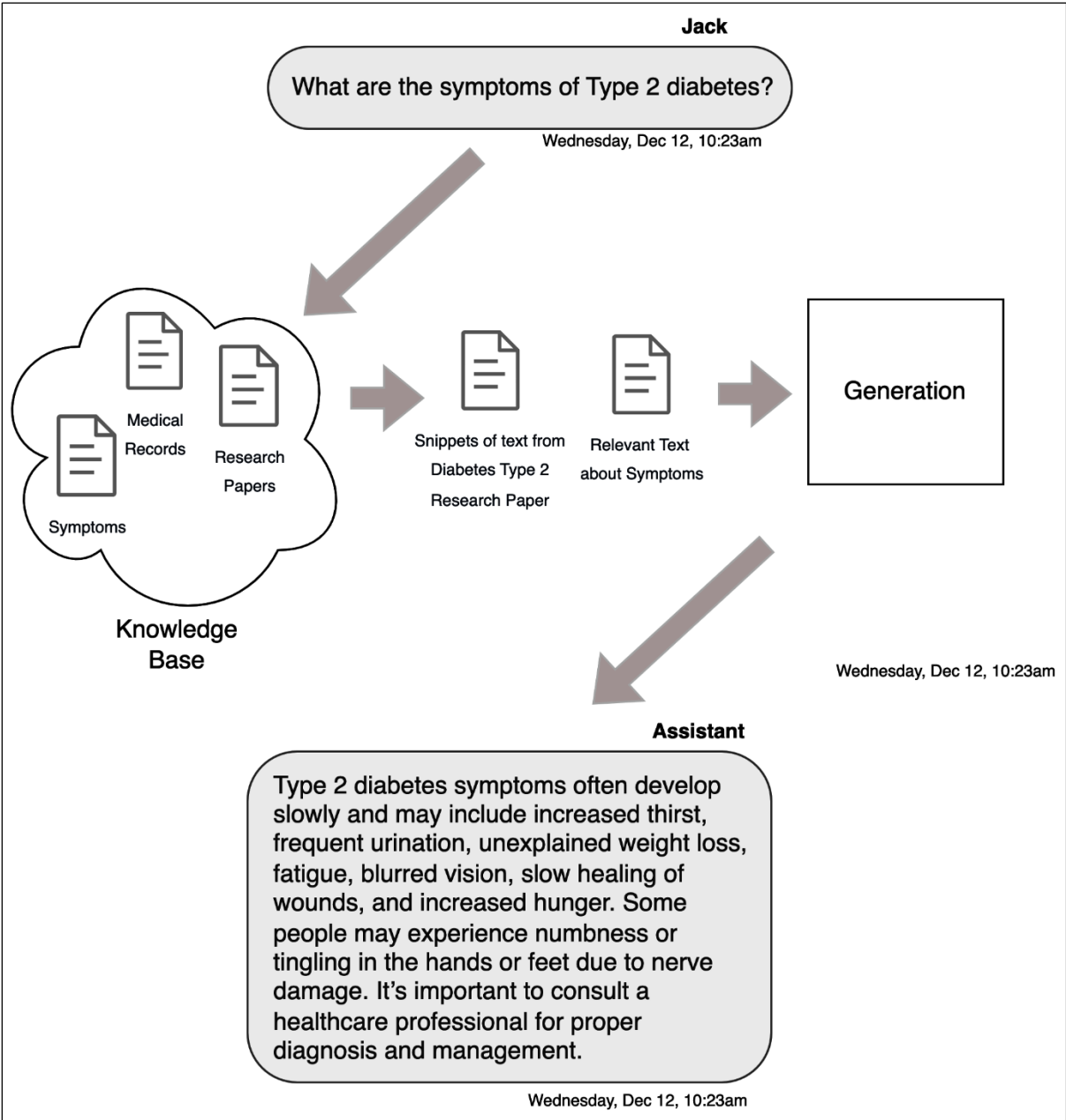
Chapter 6: Unpacking RAG Workflows

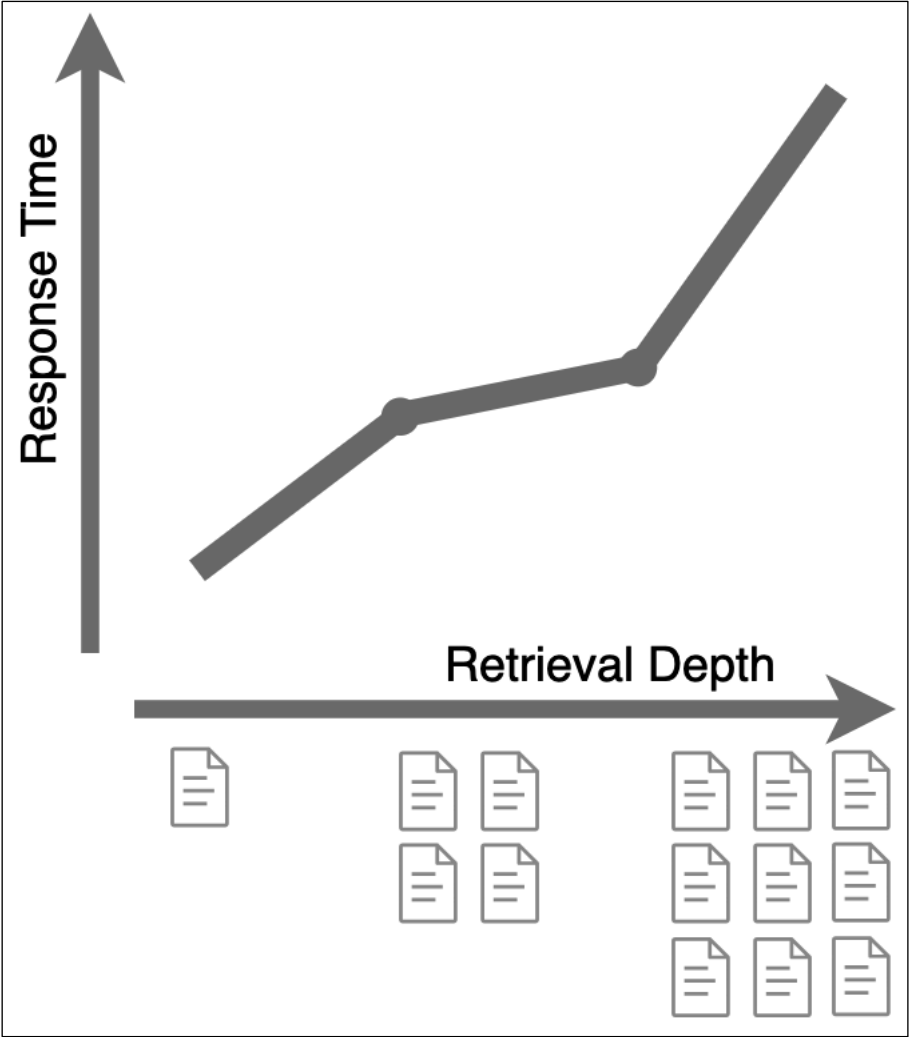




Long Text







book-passages ●

METRIC	DIMENSIONS	HOST
cosine	1024	https://book-passages-6k0twet.svc.aped-4627-b74a.pinecone.io

CLOUD	REGION	TYPE
AWS	us-east-1	Serverless

Personal > Tracing projects > default > RunnableSequence

TRACE

CollapseStatsFilterMost relevant

RunnableSequence

3.03s446

VectorStoreRetriever

0.65s

ChatMistralAI

mistral-large-latest

2.37s

Some runs have been hidden. [Show 4 hidden runs](#)

VectorStoreRetriever

PlaygroundAdd toShareEdit

RunFeedbackMetadata

Run IDTrace ID

Input

1query: Who is Ishmael, and why is he important?

YAML

Rendered Output

DOCUMENTS 1

I made no doubt, was all eagerness to vanish from before the awak... +1

START TIME11/01/2024, 10:48:56 PM

END TIME11/01/2024, 10:48:57 PM

TIME TO FIRST TOKENN/A

STATUS

Success

TOTAL TOKENS0 tokens

LATENCY0.65s

TYPERetriever

 ChatMistralAI

Playground

Add to

Run

Feedback

Metadata

Run ID

Trace ID

Input

HUMAN

Answer the question based only on the following context:
[Document(metadata={'source':
'https://www.gutenberg.org/cache/epub/2701/pg2701.txt'}, page_content='I made no doubt, was all eagerness to vanish from before the awakened\r\nwrath of Peleg. But to my astonishment, he sat down again on the\r\nransom very quietly, and seemed to have not the slightest intention of\r\nwithdrawing. He seemed quite used to impenitent Peleg and his ways. As\r\nfor Peleg, after letting off his rage as he had, there seemed no more\r\nleft in him, and he, too, sat down like a lamb, though he twitched a\r\nlittle as if still nervously agitated. "Whew!" he whistled at last—"the\r\nsquall's gone off to leeward, I think. Bildad, thou used to be good at\r\nsharpening a lance, mend that pen, will ye. My jack-knife here needs\r\nthe grindstone. That's he; thank ye, Bildad. Now then, my young man,\r\nIshmael's thy name, didn't ye say? Well then, down ye go here, Ishmael,\r\nfor the three hundredth lay." \r\n\r\n"Captain Peleg," said I, "I have a friend with me who wants to ship\r\ntoo—shall I bring him down to-morrow?")]

Question: Who is Ishmael, and why is he important?

START TIME

11/01/2024, 10:48:57 PM

END TIME

11/01/2024, 10:48:59 PM

TIME TO FIRST TOKEN

N/A

STATUS

Success

TOTAL TOKENS

446 tokens

LATENCY

2.37s

TYPE

LLM

TAGS

seq:step:3

Rendered Output

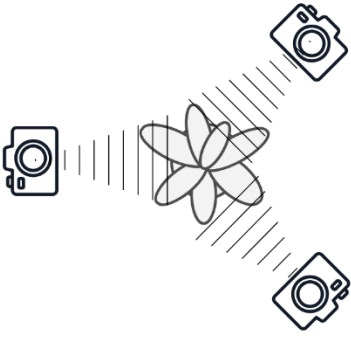
AI

Ishmael is the narrator of the story. He is important because the story is told from his perspective, and he is interacting with other characters, such as Captain Peleg, to potentially join a ship's crew.

Workshop 2

Multi-query Prompt

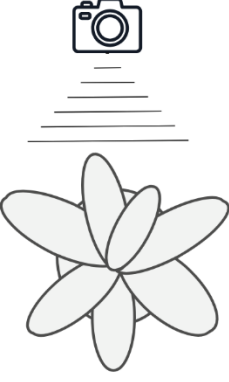
Multi-query prompts are like photographing a subject from different angles, capturing unique perspectives. Each rephrased question is a "new angle," revealing different aspects of the topic. Together, they provide a richer, more complete understanding by covering more ground than a single view.



Workshop 3

Detailed Sub-queries

Detailed subqueries are like zooming in on a subject to capture finer details. Each subquery focuses on specific aspects, revealing nuances and in-depth information that a broad question might overlook. This approach enhances the answer by adding intricate details.

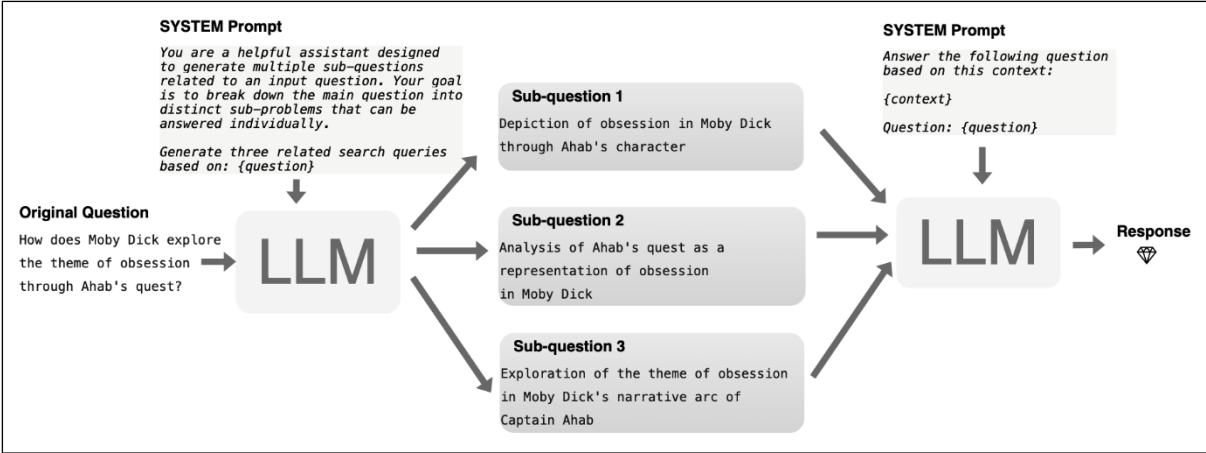


Workshop 4

Step-Back Generification

Step-back generification is like zooming out to capture the surrounding landscape. By rephrasing the question in broader terms, it helps situate the topic within a wider context, offering a comprehensive view that includes the big picture.





Runs

Threads

Monitor

Setup

≡ Filters

🕒 Last 7 days

Root Runs

LLM Calls

All Runs

☰ Columns

☐	☒	Name	Input	Output	Error
☐	☒	 ChatMistralAI	human: Answer the fo...	ai: In Moby Dick, Captain Ahab is portrayed as both a hero and a villain in se...	
☐	☒	 ChatPromptTemplate	How is Captain Ahab ...	{"messages":[{"content":"Answer the following question based on this cont...	
☐	☒	 get_unique_union	[{"metadata":{"sourc...	[{"metadata":{"source":"https://www.gutenberg.org/cache/epub/2701/pg27...	
☐	☒	 VectorStoreRetriever	5. How can the portra...	[{"metadata":{"source":"https://www.gutenberg.org/cache/epub/2701/pg27...	
☐	☒	 VectorStoreRetriever	4. How does Melville ...	[{"metadata":{"source":"https://www.gutenberg.org/cache/epub/2701/pg27...	
☐	☒	 VectorStoreRetriever	3. Could we analyze C...	[{"metadata":{"source":"https://www.gutenberg.org/cache/epub/2701/pg27...	
☐	☒	 VectorStoreRetriever	2. How does the char...	[{"metadata":{"source":"https://www.gutenberg.org/cache/epub/2701/pg27...	
☐	☒	 VectorStoreRetriever	1. In what ways does ...	[{"metadata":{"source":"https://www.gutenberg.org/cache/epub/2701/pg27...	
☐	☒	 RunnableEach<VectorStoreRetriever:	[{"1. In what ways does...	[{"metadata":{"source":"https://www.gutenberg.org/cache/epub/2701/pg2...	

TRACE

Collapse Stats Filter

Most relevant ▾

- RunnableSequence 8.23s 1,594
 - map:key:context 2.38s
 - ChatMistr... mistral-s... 1.85s
 - VectorStoreRetriever 0.49s
 - VectorStoreRetriever 0.42s
 - VectorStoreRetriever 0.47s
 - VectorStoreRetriever 0.47s
 - VectorStoreRetriever 0.51s
 - get_unique_union 0.00s
 - ChatMistral... mistral-sm... 5.84s

Some runs have been hidden. [Show 8 hidden runs](#)

VectorStoreRetriever

Run Feedback Metadata

Run ID Trace ID

Input ▾

1 query: 1. In what ways does Herman Melville's Moby Dick present Captain Ahab as a heroic figure and as an antagonist?

YAML ↕

Rendered Output ▾

DOCUMENTS 1

of his manner. "Captain Ahab." "What! the cap... +1

START TIME
11/13/2024, 05:37:47 PM

END TIME
11/13/2024, 05:37:47 PM

TIME TO FIRST TOKEN
N/A

STATUS
 Success

TOTAL TOKENS
0 tokens

LATENCY
0.49s

TRACE

Collapse

Stats

Filter

Most relevant ▾

RunnableSequence

8.23s 1,594

map:key:context 2.38s ▾

ChatMistr... mistral-s... 1.85s

VectorStoreRetriever 0.49s

VectorStoreRetriever 0.42s

VectorStoreRetriever 0.47s

VectorStoreRetriever 0.47s

VectorStoreRetriever 0.51s

get_unique_union 0.00s

ChatMistr... mistral-sm... 5.84s

Some runs have been hidden. [Show 8 hidden runs](#)

ChatMistralAI

Playground

Add to ▾

⬆

✎

Run

Feedback

Metadata

Run ID

Trace ID

11/13/2024, 05:37:45 PM

HUMAN

Generate five unique variations of the given user question to improve document retrieval from a vector database.
The goal is to create diverse perspectives on the question, helping to mitigate limitations of distance-based similarity search.
Separate each alternative question with a newline. Original question: How is Captain Ahab portrayed as both a hero and a villain in Moby Dick?

Rendered Output ▾

AI
1. In what ways does Herman Melville's Moby Dick present Captain Ahab as a heroic figure and as an antagonist?
2. How does the character of Captain Ahab in Moby Dick embody elements of both heroism and villainy?
3. Could we analyze Captain Ahab from Moby Dick as a hero and a villain simultaneously? If so, how would we go about it?
4. How does Melville use the character of Captain Ahab to explore the complexities of heroism and villainy in his novel Moby Dick?
5. How can the portrayal of Captain Ahab in Moby Dick be interpreted from two viewpoints: one, seeing him as a hero, and the other, viewing him as a villain?

END TIME

11/13/2024, 05:37:47 PM

TIME TO FIRST TOKEN

N/A

STATUS

Success

TOTAL TOKENS

257 tokens

LATENCY

1.85s

TYPE

LLM

TAGS

seq:step:2

SYSTEM Prompt

You are a helpful assistant designed to generate multiple sub-questions related to an input question. Your goal is to break down the main question into distinct sub-problems that can be answered individually.

Generate three related search queries based on: {question}

Original Question

How does Moby Dick explore the theme of obsession through Ahab's quest?

LLM

Sub-question 1

Depiction of obsession in Moby Dick through Ahab's character

Sub-question 2

Analysis of Ahab's quest as a representation of obsession in Moby Dick

Sub-question 3

Exploration of the theme of obsession in Moby Dick's narrative arc of Captain Ahab

SYSTEM Prompt

Answer the following question based on this context:

{context}

Question: {question}

LLM

Response

SYSTEM Prompt

You are a knowledgeable assistant.

Your task is to transform specific questions into broader, more general questions that are easier to answer and provide a wider perspective. This approach is known as creating 'step-back' questions.

Original question: {question}

Original Question

What does Moby Dick say about humanity's struggle against nature?

LLM

Step-back question

How does literature portray humanity's relationship with and struggle against nature? Specifically, what does Moby Dick reveal about this theme?

SYSTEM Prompt

Answer the following question thoroughly. Use relevant information from the provided context but disregard any irrelevant details.

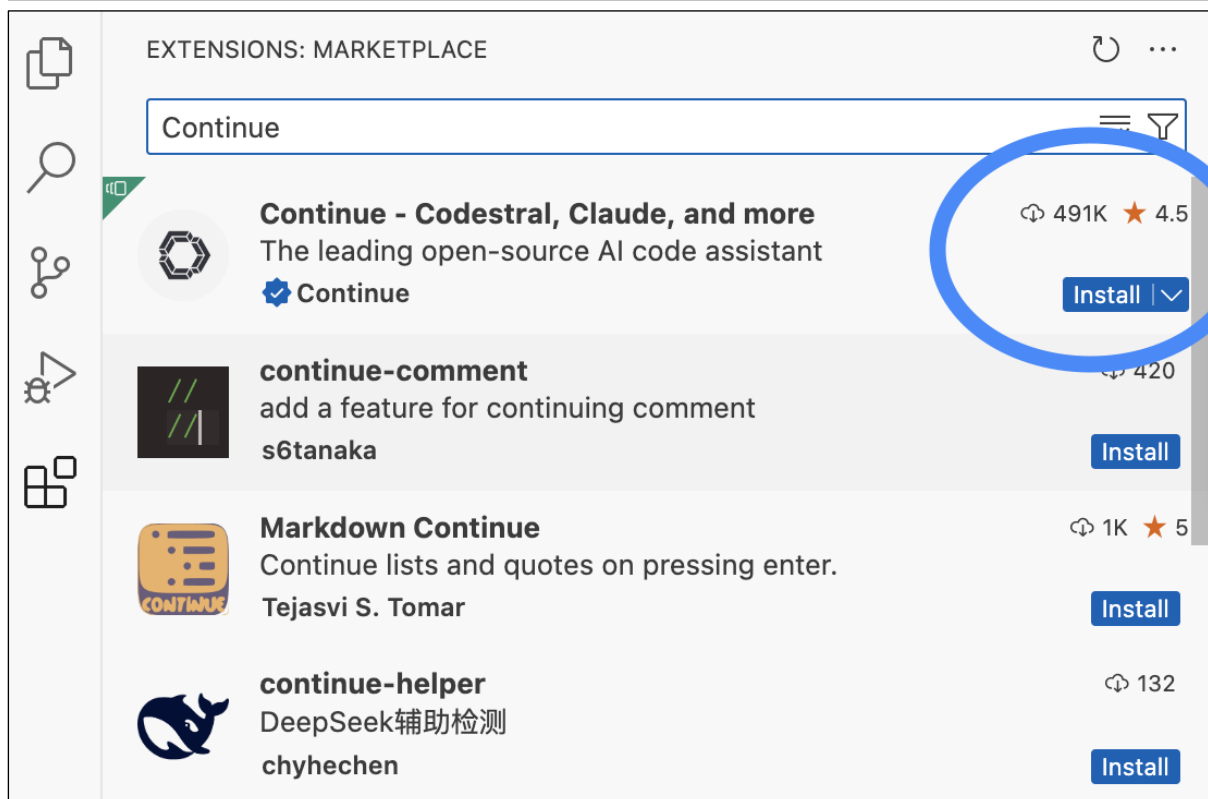
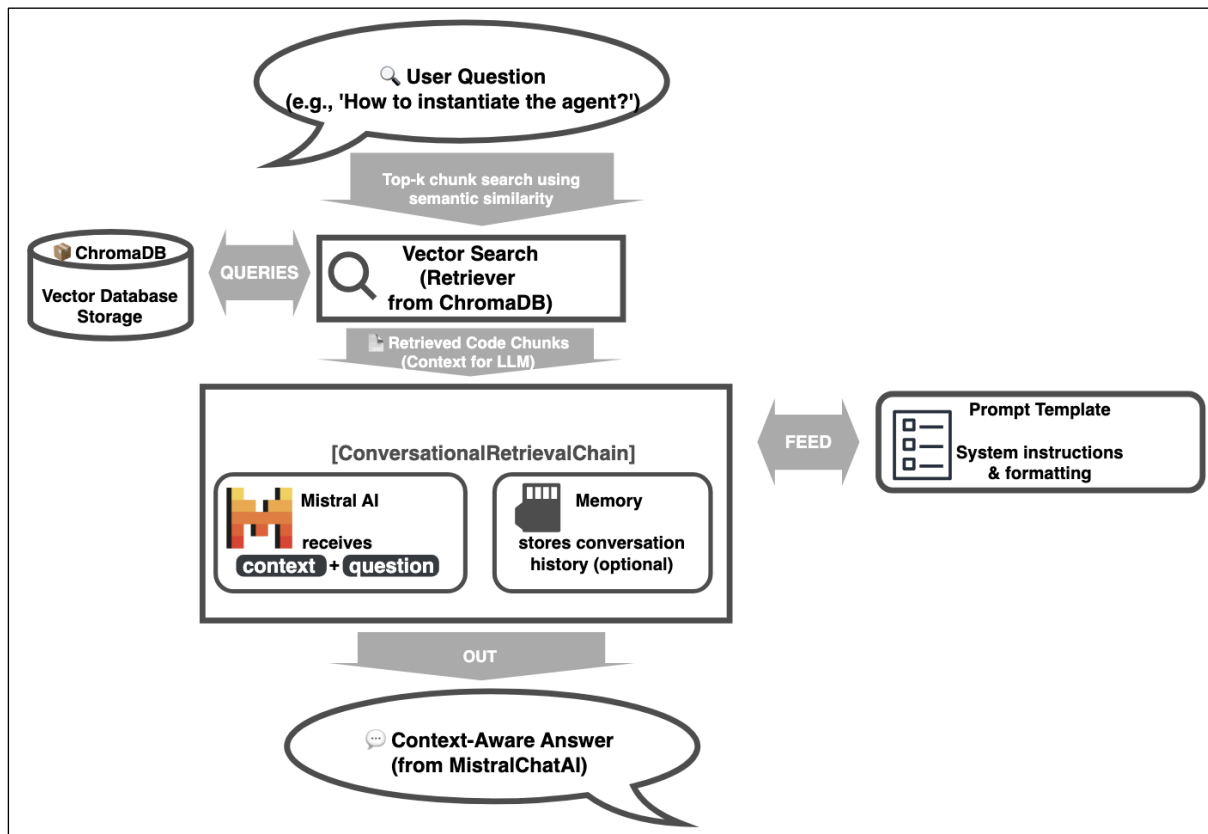
Context: {normal_context} {step_back_context}

Question: {question}

LLM

Response

Chapter 7: Coding with Mistral



Extension: Continue - Codestral, Claude, and more



Continue - Codestral, Claude, and more

v0.8.63

Continue [continue.dev](#) | 491,549 | ★★★★★ (64)

The leading open-source AI code assistant

Disable

Uninstall

Switch to Pre-Release Version

☒ Auto Update

DETAILS

FEATURES

CHANGELOG



Categories

Programming Languages

Snippets

Machine Learning

Education AI Chat

A vertical sidebar containing six icons from top to bottom: a document icon, a magnifying glass (search) icon, a graph/network icon, a play button icon, a grid icon, and the Continue logo (a hexagon with internal geometric patterns).

```
{ } config.json •
Users > pavlocherkashyn > .continue > { } config.json > [ ] models > { } 5 > apiKey
2   "models": [
27   {
29     "provider": "anthropic",
30     "apiKey": "",
31     "title": "Claude 3.5 Sonnet"
32   },
33   {
34     "apiKey": "xxxxxxxxxxxxxxxxxxxxxx",
35     "title": "Codestral",
36     "model": "codestral-latest",
37     "provider": "mistral",
38     "systemMessage": "You are an expert software developer. You give helpful and concise responses."
39   }
]
```

write a script to retrieve the joke from api.chucknorris.io using http client

Sure, I'll make the necessary changes to retrieve a joke from api.chucknorris.io using an HTTP client.

1. I'll modify the config.json file to include the API URL:

> { } config.json

Copy

Applied ✓

2. Next, I'll modify the main.py file to use this API URL and fetch a joke:

> main.py

Copy

Applied ✓

Summary:

- Modified config.json to include the API URL
- Modified main.py to fetch a joke from the API URL and print it out
- This solution uses the requests library to make HTTP requests and json to parse the response.



1. I'll modify the config.json file to include the API URL:

⌵ {} config.json

📄 Copy

Applied ✓

```
{
  "api_url": "https://api.chucknorris.io/jokes/random"
}
```

2. Next, I'll modify the main.py file to use this API URL and fetch a joke:

⌵ 🐍 main.py

📄 Copy

Applied ✓

```
import json
import requests

# Load the API URL from the config file
with open('/Users/pavlocherkashyn/.continue/config.json') as config_file:
    config = json.load(config_file)

# Fetch a joke from the API
response = requests.get(config['api_url'])
joke = response.json()['value']

# Print the joke
print(joke)
```

{} config.json ●

Users > pavlocherkashyn > .continue > {} config.json >  joke_api

```
76     "slashCommands": [
84         },
85         {
86             "name": "commit",
87             "description": "Generate a git commit message"
88         }
89     ],
90     "embeddingsProvider": {
91         "provider": "free-trial"
92     },
93     "reranker": {
94         "name": "free-trial"
95     },
96     "joke_api": "https://api.chucknorris.io/jokes/random?category=science"
97 }
```

Add to chat (Cmd+L) | Edit highlighted code (Cmd+I).

main.py > ...

Add to chat (Cmd+L) | Edit highlighted code (Cmd+I).

```
1 import requests
2
3 import json
4
5
6 with open('/Users/pavlocherkashyn/.continue/config.json') as config_file:
7     config = json.load(config_file)
8
9 response = requests.get(config['joke_api'])
10
11 joke = response.json()['value']
12 print(joke)
```

main.py (1-11)

```
import requests
import json

with open('/Users/pavlocherkashyn/.continue/config.json') as config_file:
    config = json.load(config_file)

response = requests.get(config['joke_api'])

joke = response.json()['value']

print(joke)
```

wrap this into web service endpoint

Codestral (Free Trial) @

@codebase Enter

Sure, I can help you with that. Here's a simple example of how you might wrap this into a Flask web service endpoint:

```
from flask import Flask, jsonify
import requests
import json

app = Flask(__name__)

@app.route('/joke', methods=['GET'])
def get_joke():
    with open('/Users/pavlocherkashyn/.continue/config.json') as config_file:
        config = json.load(config_file)

        response = requests.get(config['joke_api'])
        joke = response.json()['value']

        return jsonify({'joke': joke})

if __name__ == '__main__':
    app.run(debug=True)
```

In this code, we're creating a new Flask web application. We then define a route `/joke` that, when accessed with a GET request, will call the `get_joke` function. This function fetches a joke from the API and returns it as a JSON response.

You can run this script and then access the joke by navigating to `http://localhost:5000/joke` in your web browser, or by sending a GET request to this URL with a tool like curl or Postman.



main.py > get_joke

```
1  from flask import Flask, jsonify
2  import requests
   Accept | Reject | %! to add instructions
3
4  import json
5
   Accept | Reject
6  app = Flask(__name__)
7
   Accept | Reject
8  @app.route('/joke', methods=['GET'])
9  def get_joke():
10     with open('/Users/pavlocherkashyn/.continue/config.json') as config_file:
11         config = json.load(config_file)
12
13     response = requests.get(config['joke_api'])
   Accept | Reject
14     print(response)
15     joke = response.json()['value']
16
   Accept | Reject
17     print(joke)
18     return jsonify({'joke': joke})
19
20 if __name__ == '__main__':
21     app.run(debug=True)
22
```

Sure, here's an example of how you might modify the Flask application to serve an HTML page that displays the joke:



```
from flask import Flask, render_template
import requests
import json

app = Flask(__name__)

@app.route('/joke', methods=['GET'])
def get_joke():
    with open('/Users/pavlocherkashyn/.continue/config.json') as
        config = json.load(config_file)

    response = requests.get(config['joke_api'])
    joke = response.json()['value']

    return render_template('joke.html', joke=joke)

if __name__ == '__main__':
    app.run(debug=True)
```

In this code, we're using Flask's `render_template` function to render a template called `joke.html`. This template should be located in a folder called `templates` in the same directory as your Python script.

Here's an example of what `joke.html` might look like:

```
<!doctype html>
<html>
  <body>
    <h1>Joke of the Day</h1>
    <p>{{ joke }}</p>
  </body>
</html>
```

In this HTML, `{{ joke }}` is a Jinja2 template variable that will be replaced with the actual joke text when the template is rendered.

With these changes, when you navigate to `http://localhost:5000/joke` in your web browser, you should see an HTML page that displays the joke of the day.

← → ↻ ⓘ 127.0.0.1:5000/joke

TemplateNotFound

jinja2.exceptions.TemplateNotFound: joke.html

Traceback (most recent call last)

File "/opt/anaconda3/lib/python3.12/site-packages/flask/app.py", line 1498, in __call__

return self.wsgi_app(environ, start_response)

File "/opt/anaconda3/lib/python3.12/site-packages/flask/app.py", line 1476, in wsgi_app

response = self.handle_exception(e)

File "/opt/anaconda3/lib/python3.12/site-packages/flask/app.py", line 1473, in wsgi_app

response = self.full_dispatch_request()

File "/opt/anaconda3/lib/python3.12/site-packages/flask/app.py", line 882, in full_dispatch_request

rv = self.handle_user_exception(e)

TemplateNotFound

jinja2.exceptions.TemplateNotFound: joke.html

Codestral (Free Trial) ▾ @

⌘↵ @codebase

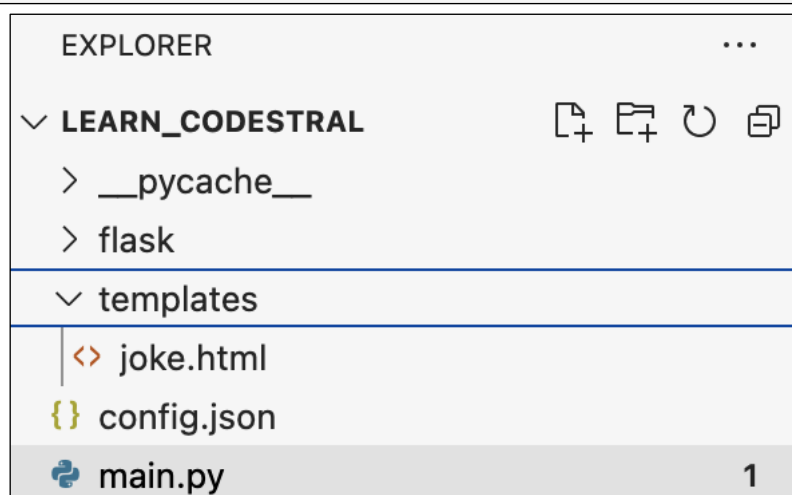
↵ Enter

It seems like you're having trouble because Flask can't find the `joke.html` template. This could be because the template file is missing, or because it's not located in the `templates` folder in the same directory as your Python script.

To fix this issue, make sure that the `joke.html` file is located in a folder called `templates` that's in the same directory as your Python script. So, if your Python script is located at `/path/to/myscript.py`, the `joke.html` file should be located at `/path/to/templates/joke.html`.

If the file is already in the correct location, make sure that the filename and extension are exactly as specified in your Python script. File names are case-sensitive, so make sure that the filename and extension match exactly.

If you're still having trouble, please let me know and I'll do my best to help you further.



Joke of the Day

Time waits for no man. Unless that man is Chuck Norris.

This is a Python script that uses the Flask web framework to create a simple web application. The script defines a route for the URL '/joke' that accepts GET requests. When a GET request is made to this URL, the function `get_joke()` is executed.

Here's a breakdown of the code:

1. `@app.route('/joke', methods=['GET'])`: This is a decorator that Flask provides to assign URLs in our app to functions easily. In this case, the function `get_joke()` is assigned to the '/joke' URL and can only be accessed via GET requests.
2. `with open('/Users/pavlocherkashyn/.continue/config.json') as config_file:`: This line opens the 'config.json' file located at the specified path and assigns the file object to 'config_file'. The 'with' statement is used for resource management and ensures that the file is properly closed after it is no longer needed.
3. `config = json.load(config_file)`: This line uses the 'json' module to parse the JSON data from the 'config.json' file and assigns the resulting Python object to 'config'.
4. `response = requests.get(config['joke_api'])`: This line uses the 'requests' library to send a GET request to the URL specified in 'config['joke_api']'. The response from the server is assigned to 'response'.
5. `joke = response.json()['value']`: This line parses the JSON data from the response and assigns the value of the 'value' key to 'joke'.
6. `return render_template('joke.html', joke=joke)`: Finally, the function returns a rendered HTML template called 'joke.html'. The 'joke' variable is passed to the template as a parameter, which means that it can be accessed and used in the template to display the joke to the user.



Mistral Code Enterprise

Mistral AI | 📄 1,589 installs | ★★★★★ (3) | Free

Your AI coding copilot powered by state-of-the-art Mistral coding models

[Install](#)

[Trouble Installing?](#) [🔗](#)



MISTRAL CODE ENTERPRISE



Mistral Code Enterprise

Tab Autocomplete

Get code completions with Codestral

Chat

Chat with Mistral Medium 3, reference code

Highlighting

Highlight code and give instructions

[Login to Mistral Code Enterprise](#)

Chapter 8: Building Smarter Defenses with Mistral

1. Identified Misconfigurations:

- Overly Permissive Ingress Rules: The security group allows SSH (port 22), HTTP (port 80), and HTTPS (port 443) access from any IP address (0.0.0.0/0). This can pose a security risk as it exposes these services to the internet, potentially allowing unauthorized access or attacks.
 - Line numbers: 16-18, 22-24
 - Impact: Unauthorized access or denial-of-service attacks.
- Overly Permissive Egress Rules: The security group has an open egress rule that allows all outbound traffic (protocol -1) to any IP address (0.0.0.0/0). This is a security risk as it allows the instance to communicate with any external resource, which may not be necessary or desirable.
 - Line number: 30-33
 - Impact: Unintended data leakage or communication with malicious resources.
- Overly Permissive IAM Role: The IAM role has an inline policy that grants full permissions ("") to all resources (""). This violates the principle of least privilege and exposes the instance to potential unauthorized actions.
 - Line numbers: 40-50
 - Impact: Unauthorized actions on AWS resources.

2. Mitigation Recommendations:

- Limit Ingress Rules: Restrict the ingress rules to only allow necessary traffic from trusted sources. Use specific IP addresses or security group IDs instead of using the broad 0.0.0.0/0 CIDR block.
 - Example: Replace "0.0.0.0/0" with the specific IP addresses or CIDR blocks of trusted sources.
- Limit Egress Rules: Restrict the egress rules to only allow necessary outbound traffic. Consider using security group rules or network ACLs to control outbound traffic.
 - Example: Remove the open egress rule and add specific rules for required outbound traffic.
- Limit IAM Role Permissions: Follow the principle of least privilege and grant the IAM role only the necessary permissions. Avoid using the "*" action or resource.
 - Example: Specify the specific actions and resources required by the instance.

3. Secure Configuration Recommendations:

- Use Security Groups for Instance-Level Access Control: Use security groups to control inbound and outbound traffic at the instance level.
- Implement Network ACLs for Subnet-Level Access Control: Use network ACLs to add an additional layer of security at the subnet level.
- Use IAM Roles and Policies for Access Control: Use IAM roles and policies to manage access to AWS resources. Follow the principle of least privilege and grant only the necessary permissions.
- Enable VPC Flow Logs: Enable VPC flow logs to capture information about traffic flowing in and out of your instances. This can help identify unauthorized or suspicious traffic.
- Regularly Review and Update Security Configurations: Regularly review and update security configurations to ensure they are still aligned with the organization's security policies and best practices.

1. Identified Vulnerabilities:

- SQL Injection: Line 37 - The SQL query is crafted by concatenating the user input directly into the query string. This allows an attacker to manipulate the query and potentially execute unintended SQL commands, leading to data leakage or modification.
- Hardcoded Credentials: Lines 24-26 - The database credentials are hardcoded into the servlet, which poses a significant security risk. If the source code is compromised, the attacker can easily access the database.
- Improper Error Handling: Line 41 - The exception message is directly sent to the client in the case of an error. This could potentially leak sensitive information about the application or the database to the attacker.

2. Mitigation Recommendations:

- SQL Injection:
 - Use Prepared Statements: Instead of concatenating the user input into the query string, use a PreparedStatement. This separates the SQL code from the data, preventing SQL injection.
 - Example:

```
String sql = "SELECT * FROM users WHERE username = ?";
PreparedStatement pstmt = conn.prepareStatement(sql);
pstmt.setString(1, param);
ResultSet rs = pstmt.executeQuery();
```
- Hardcoded Credentials:
 - Store credentials securely: Store the credentials in a secure location, such as environment variables or a secure vault. Do not hardcode them in the source code.
- Improper Error Handling:
 - Implement a custom error page: Instead of sending the exception message directly to the client, implement a custom error page. This page should display a generic error message to the user without leaking sensitive information.

3. Secure Coding Recommendations:

- Use parameterized queries or prepared statements to protect against SQL injection.
- Never store sensitive information in the source code. Use secure methods to store and retrieve credentials.
- Implement proper error handling to prevent the leakage of sensitive information to the client.
- Validate and sanitize user input to ensure it is in the expected format and does not contain malicious content.
- Use secure communication protocols, such as HTTPS, to encrypt data in transit.
- Implement access control mechanisms to restrict access to sensitive resources and data.
- Regularly update and patch the application to protect against known vulnerabilities.

```
3      $input = $_GET['input'];
4      echo "<div>Error: $input</div>";
5  } else {
```

1. Identified Vulnerabilities:

- Cross-Site Scripting (XSS) Vulnerability: The script directly outputs user-provided data without proper sanitization or encoding.
 - Line numbers: 4
 - Impact: An attacker could inject malicious JavaScript code into the 'input' parameter, which would then be executed in the browser of any user who views the page. This could lead to various attacks, such as session hijacking, defacement of the website, or the theft of sensitive data.

2. Mitigation Recommendations:

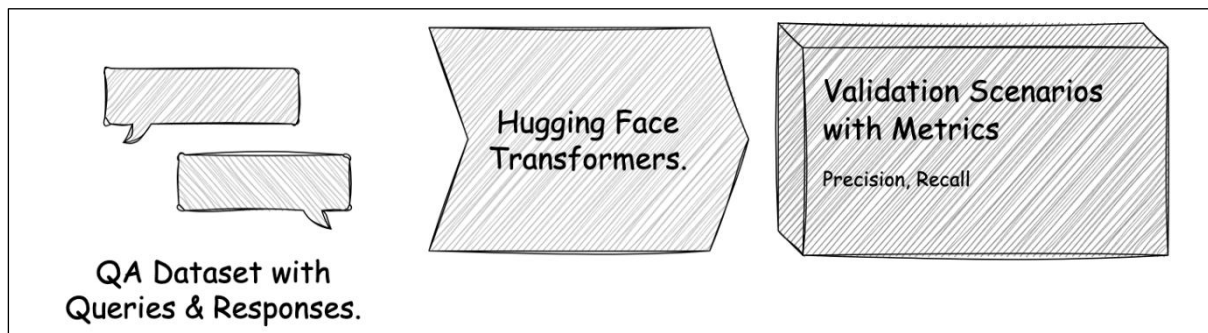
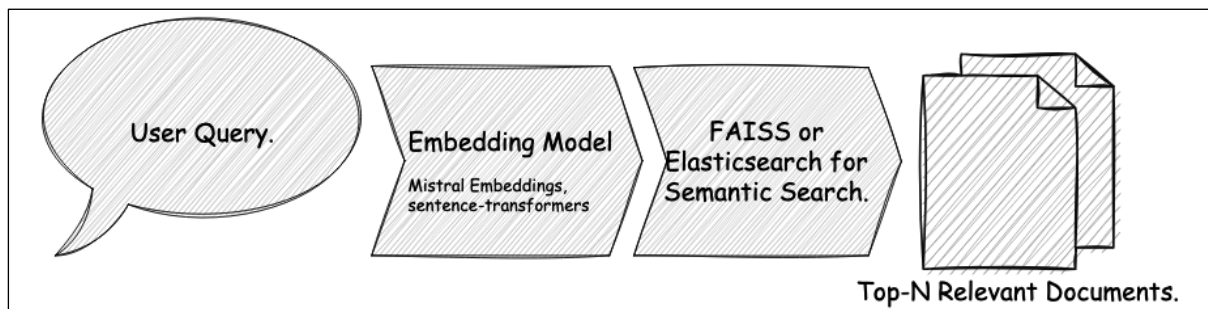
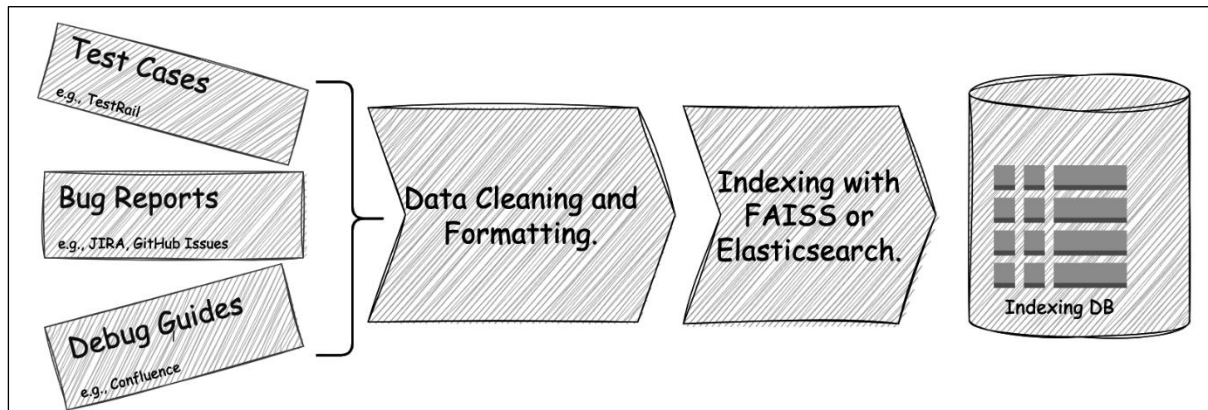
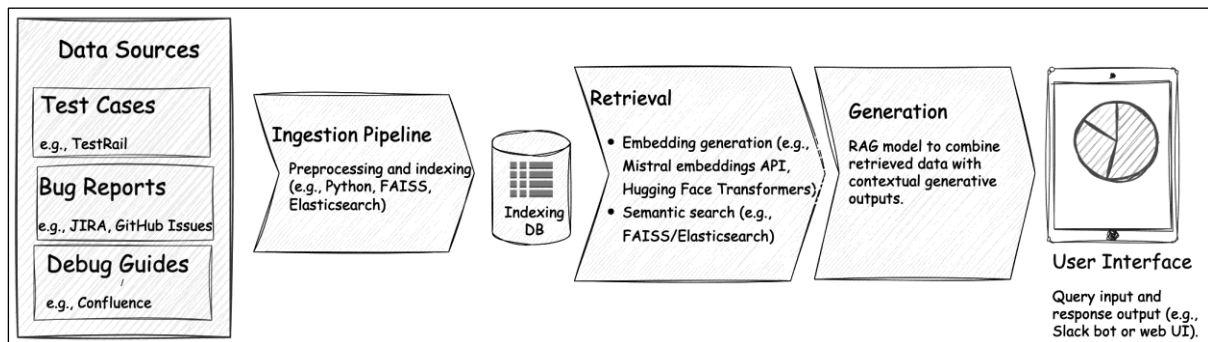
- Use the `htmlspecialchars()` function to encode special characters in the user-provided data, preventing them from being interpreted as HTML or JavaScript.
 - Suggested code improvements:

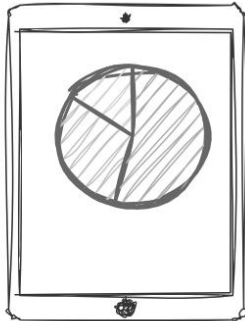
```
$input = htmlspecialchars($_GET['input'], ENT_QUOTES, 'UTF-8');
```
- Consider using a more secure method of input handling, such as prepared statements or parameterized queries, when dealing with databases. In this case, since the script does not interact with a database, this is not applicable.

3. Secure Coding Recommendations:

- Always sanitize and validate user input before using it in your application.
- Use output encoding functions to prevent XSS attacks.
- Use least privilege principle, i.e., only grant necessary permissions to the application.
- Keep your software up-to-date with the latest security patches.
- Follow the principle of defense in depth, i.e., use multiple layers of security.
- Limit the amount of sensitive data that is stored or processed by the application.
- Implement a secure error handling mechanism that does not leak sensitive information to the user.
- Use a secure communication protocol (HTTPS) to protect data in transit.

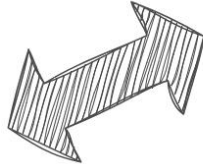
Chapter 9: Take-Home RAG Challenges





User Interface

Slack Bot,
Web App, or JIRA
Plugin.



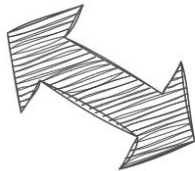
Slack API, JIRA API

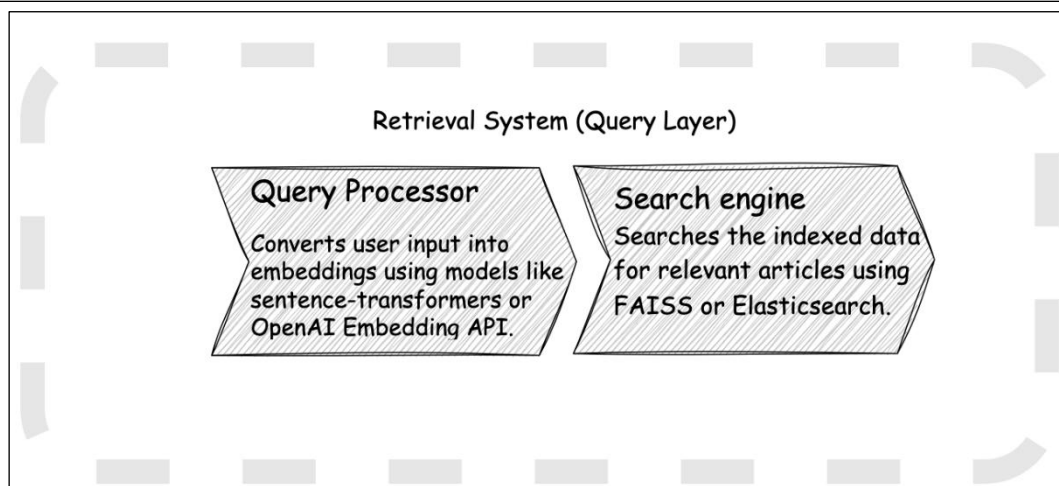
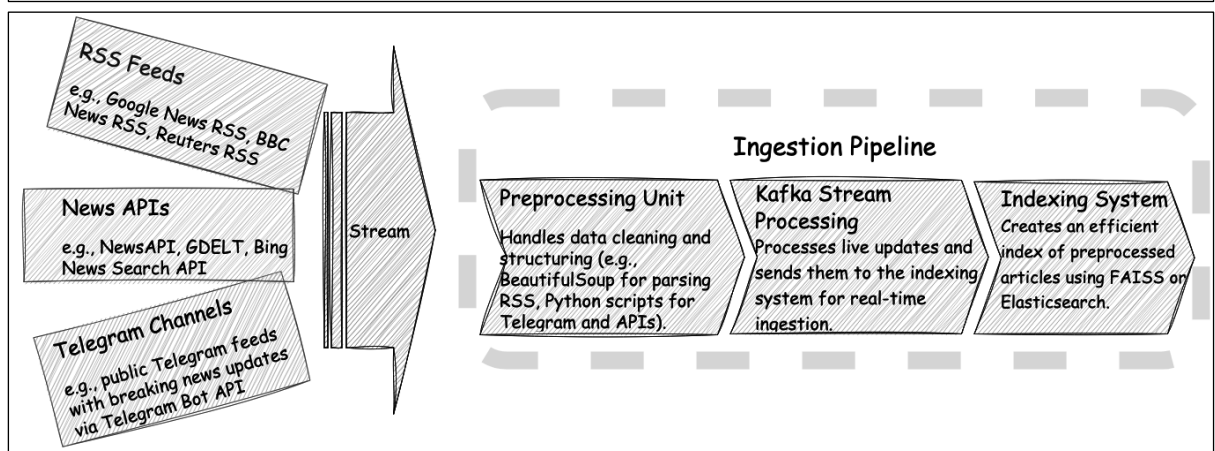
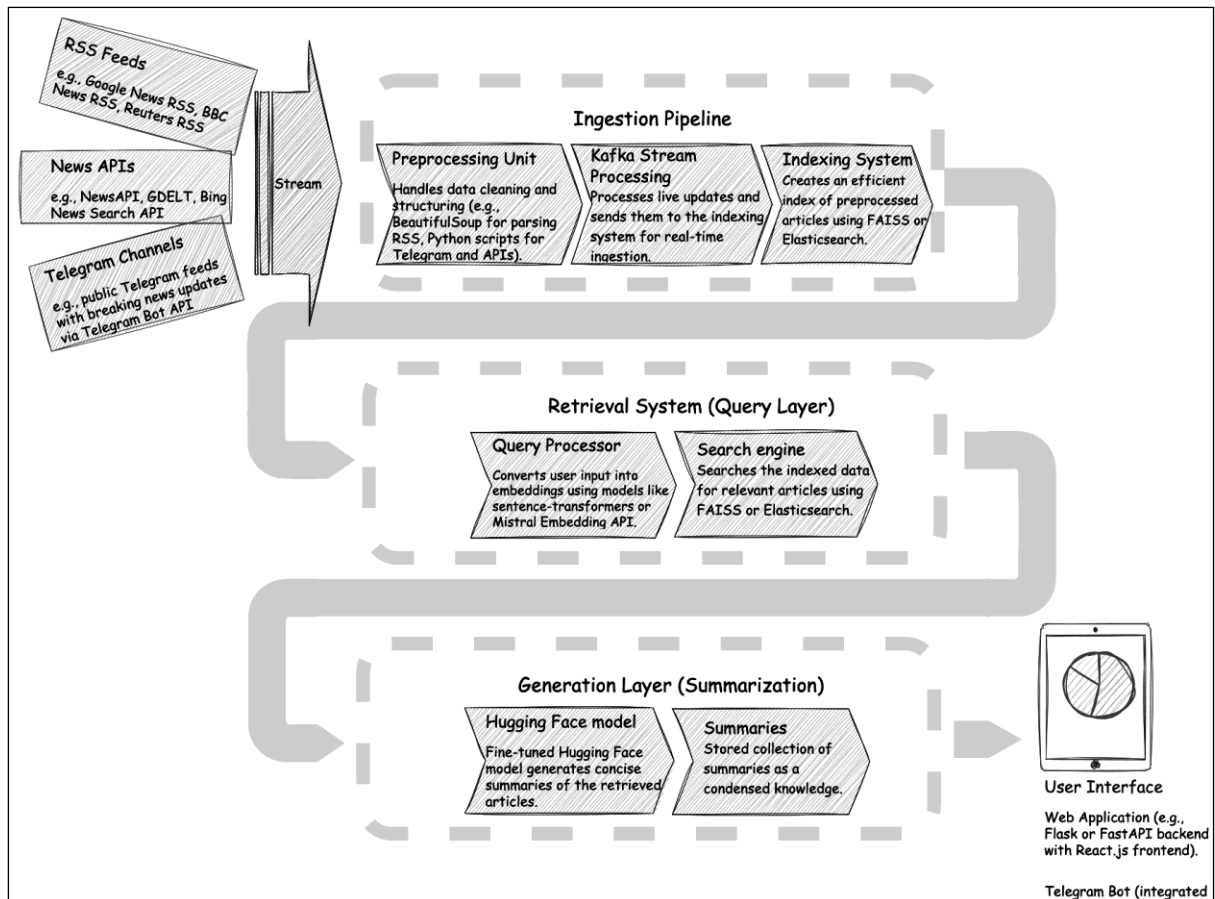


RAG

Retrieval

Generation





Generation Layer (Summarization)

Hugging Face model

Fine-tuned Hugging Face model generates concise summaries of the retrieved articles.

Summaries

Stored collection of summaries as a condensed knowledge.

Specify user details

User details

User name

bedrock-cli-user

The user name can have up to 64 characters. Valid characters: A-Z, a-z, 0-9, and + = , . @ _ - (hyphen)

☐ **Provide user access to the AWS Management Console - *optional***

If you're providing console access to a person, it's a [best practice](#) to manage their access in IAM Identity Center.

i If you are creating programmatic access through access keys or service-specific credentials for AWS CodeCommit or Amazon Keyspaces, you can generate them after you create this IAM user. [Learn more](#)

Cancel

Next

Permissions options

☐ **Add user to group**
Add user to an existing group, or create a new group.
We recommend using groups to manage user permissions by job function.

☐ **Copy permissions**
Copy all group memberships, attached managed policies, and inline policies from an existing user.

☒ **Attach policies directly**
Attach a managed policy directly to a user. As a best practice, we recommend attaching policies to a group instead. Then, add the user to the appropriate group.

Permissions policies (1/1319)

Choose one or more policies to attach to your new user.

✕ Filter by Type All types 6 matches < 1 > ⚙️

☐	Policy name ↗	▲ Type ▼	Attached entities ▼
☒	 AmazonBedrockFullAccess	AWS managed	0
☐	 AmazonBedrockReadOnly	AWS managed	0

Permissions options

☐ **Add user to group**
Add user to an existing group, or create a new group. We recommend using groups to manage user permissions by job function.

☐ **Copy permissions**
Copy all group memberships, attached managed policies, and inline policies from an existing user.

☒ **Attach policies directly**
Attach a managed policy directly to a user. As a best practice, we recommend attaching policies to a group instead. Then, add the user to the appropriate group.

Permissions policies (2/1319)

Choose one or more policies to attach to your new user.

Filter by Type All types 45 matches < 1 2 3 > ⚙️

☐	Policy name 🔗	Type	Attached entities
☒	 AdministratorAccess	AWS managed - job function	0
☐	 AdministratorAccess-Amplify	AWS managed	1
☐	 AdministratorAccess-AWSElasticBea...	AWS managed	0

Review and create

Review your choices. After you create the user, you can view and download the autogenerated password, if enabled.

User details

User name bedrock-cli-user	Console password type None	Require password reset No
-------------------------------	-------------------------------	------------------------------

Permissions summary

< 1 >

Name 	Type	Used as
AdministratorAccess	AWS managed - job function	Permissions policy
AmazonBedrockFullAccess	AWS managed	Permissions policy

bedrock-cli-user [Info](#)

Delete

Summary

ARN  arn:aws:iam::203918873727:user/bedrock-cli-user	Console access Disabled	Access key 1 Create access key
Created January 28, 2025, 13:11 (UTC-06:00)	Last console sign-in -	

Permissions Groups Tags **Security credentials** Last Accessed

Console sign-in

[Enable console access](#)

Console sign-in link  https://203918873727.signin.aws.amazon.com/console	Console password Not enabled
---	---------------------------------

Multi-factor authentication (MFA) (0)

Remove

Resync

[Assign MFA device](#)

Use MFA to increase the security of your AWS environment. Signing in with MFA requires an authentication code from an MFA device. Each user can have a maximum of 8 MFA devices assigned. [Learn more](#) 

Type	Identifier	Certifications	Created on
No MFA devices. Assign an MFA device to improve the security of your AWS environment			

[Assign MFA device](#)

Access keys (0)

[Create access key](#)

Use access keys to send programmatic calls to AWS from the AWS CLI, AWS Tools for PowerShell, AWS SDKs, or direct AWS API calls. You can have a maximum of two access keys (active or inactive) at a time. [Learn more](#) 

No access keys. As a best practice, avoid using long-term credentials like access keys. Instead, use tools which provide short term credentials. [Learn more](#) 

[Create access key](#)

Access key best practices & alternatives [Info](#)

Avoid using long-term credentials like access keys to improve your security. Consider the following use cases and alternatives.

Use case

☒ **Command Line Interface (CLI)**

You plan to use this access key to enable the AWS CLI to access your AWS account.

☐ **Local code**

You plan to use this access key to enable application code in a local development environment to access your AWS account.

☐ **Application running on an AWS compute service**

You plan to use this access key to enable application code running on an AWS compute service like Amazon EC2, Amazon ECS, or AWS Lambda to access your AWS account.

☐ **Third-party service**

You plan to use this access key to enable access for a third-party application or service that monitors or manages your AWS resources.

Retrieve access keys [Info](#)

Access key

If you lose or forget your secret access key, you cannot retrieve it. Instead, create a new access key and make the old key inactive.

Access key

Secret access key

 AK[REDACTED]E

 ***** [Show](#)

Access key best practices

- Never store your access key in plain text, in a code repository, or in code.
- Disable or delete access key when no longer needed.
- Enable least-privilege permissions.
- Rotate access keys regularly.

For more details about managing access keys, see the [best practices for managing AWS access keys](#).

[Download .csv file](#)

Done

```
AWS Access Key ID [None]: AK[REDACTED]44E
AWS Secret Access Key [None]: xdG[REDACTED]3k4
Default region name [None]: us-east-1
Default output format [None]: json
```

bedrock

Services

Services

Resources New

Documentation

Knowledge articles

Marketplace

Blog posts

Tutorials

 **Amazon Bedrock**
The easiest way to build and scale generative AI applications with foundation models (...)

 **Amazon SageMaker**
The center for data, analytics, and AI

Amazon Bedrock

▼ Getting started

Overview

Providers

▼ Foundation models

Model catalog New

Marketplace deployments New

Imported models

Prompt Routers Preview

▼ Playgrounds

Chat / Text

Image / Video

Amazon Bedrock > Providers

Providers (7)

Choose from a range of serverless providers to find the model that is best suited for your use case.

Find providers

< 1 > ⚙️

Provider name	Number of available models	Deployment type	Latest release
AI21 Labs	5	Serverless	Thu, 05 Sep 2024 08:00:00 GMT
Amazon	13	Serverless	Tue, 03 Dec 2024 08:00:00 GMT
Anthropic	9	Serverless	Tue, 22 Oct 2024 08:00:00 GMT
Cohere	6	Serverless	Mon, 29 Apr 2024 08:00:00 GMT
Meta	9	Serverless	Wed, 18 Dec 2024 08:00:00 GMT
Mistral AI	4	Serverless	Tue, 02 Apr 2024 08:00:00 GMT
Stability AI	1	Serverless	Wed, 26 Jul 2023 08:00:00 GMT

Models (5)

Filter for a model

Most popular

< 1 >



Pixtral Large (25.02)
Available to request

Complex reasoning analysis, Text generation, Image, Code generation, RAG, Agents

Serverless



Mistral Large (24.02)
Available to request

Complex reasoning analysis, Text generation, Code generation, RAG, Agents

Serverless



Mistral 7B Instruct
Access granted

Classification, Text generation, Code generation

Serverless



Mixtral 8x7B Instruct
Available to request

Complex reasoning analysis, Text generation, Code generation

Serverless



Mistral Small (24.02)
Available to request

Text generation, Code generation, Classification, RAG, Conversation

Serverless

Open in playground

Purchase provisioned throughput

Modify access

Customize

☐	▼ Mistral AI (5)	1/5 access granted		
☐	Mistral 7B Instruct	✓ Access granted	Text	EULA
☐	Mixtral 8x7B Instruct	⋮ Available to request	Text	EULA
☐	Mistral Large (24.02)	⋮ Available to request	Text	EULA
☒	Mistral Small (24.02)	⋮ Available to request	Text	EULA
☐	Pixtral Large (25.02) Cross-region inference	⋮ Available to request	Text & Vision	EULA

▼ Mistral AI (5)	2/5 access granted		
Mistral 7B Instruct	⋮ Available to request	Text	EULA
Mixtral 8x7B Instruct	⋮ Available to request	Text	EULA
Mistral Large (24.02)	⋮ Available to request	Text	EULA
Mistral Small (24.02)	✓ Access granted	Text	EULA
Pixtral Large (25.02) Cross-region inference	⋮ Available to request	Text & Vision	EULA

CloudTrail

Event history (49+) Info

Download events

Create Athena table

Event history shows you the last 90 days of management events.

Lookup attributes

Event source bedrock.amazonaws.com

Filter by date and time

Clear filter

Event name	Event time	User name	Event source	Resource type	Re
InvokeModel	June 18, 2025, 12:...	bedrock-cli-user	bedrock.amazonaws.com	-	-
ListFoundationModels	June 18, 2025, 12:...	bedrock-cli-user	bedrock.amazonaws.com	-	-
InvokeModel	June 18, 2025, 12:...	bedrock-cli-user	bedrock.amazonaws.com	-	-
ListFoundationModels	June 18, 2025, 12:...	bedrock-cli-user	bedrock.amazonaws.com	-	-

Chapter 11: Harnessing Mistral’s Power via Google Cloud Vertex AI

Select a project

NEW PROJECT

Search projects and folders

 |

RECENT

STARRED

ALL

Name	ID
------	----

Project name *

LearnMistral



Project ID: learnmistral. It cannot be changed later.

Edit

Location *

 No organization

Browse

Parent organization or folder

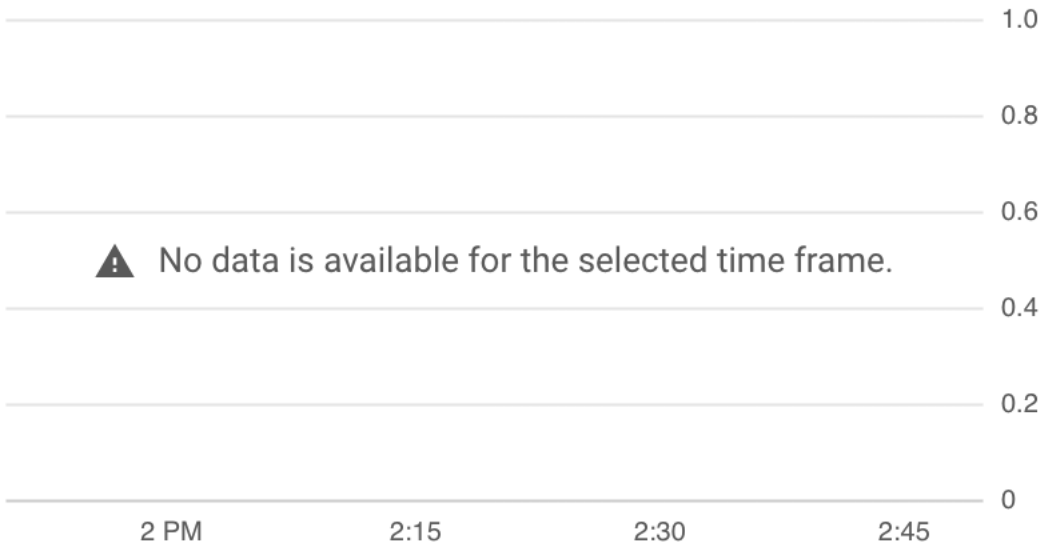
Create

Cancel

API APIs



Requests (requests/sec)



→ Go to APIs overview



Welcome

You're working in [LearnMistral](#)

Project number: 524070227409



Project ID: learnmistral



[Dashboard](#)

[Recommendations](#)

+ Create a VM

+ Run a query in BigQuery

+ Deploy an application

+ Create a storage bucket

Join us for the latest AI innovations at Google Cloud Next, April 9-11 in Vegas.

Register now



Quick access

API APIs & Services

IAM & Admin

Billing

Compute Engine

Cloud Storage

BigQuery

VPC network

Kubernetes Engine

Google Cloud

Search (/) for resources, docs, products, and more

Billing

My billing accounts

My projects

Create account

Filter

Status : Active

Enter property name or value

X

?

	Billing account name	Billing account ID	Status	Last 30 days' spend	Account type		Organi
	My Billing Account 1		Active	\$0	Direct		—

Create a new billing account

Name *

new account for Mistral workshop

The name of this billing account is only used to help you remember what it is.

Country *

United States

Currency

USD

Continue

Cancel

Set up your billing profile

new account for Mistral workshop ▾

Payments profile

[Redacted]

Individual • Also used with Google Cloud, Google Play Apps, Google Offers, Google Play Devices, ... • United States • ID:

[Redacted]

[Change](#)



Your payment information is saved in a payments profile, which is associated with your Google Account and shared across Google services. [Learn more about payments profile](#)

Payment method



Mastercard •••• [Redacted]

[Change](#)

[Submit and enable billing](#)



Welcome

You're working in [LearnMistral](#)

Project number: 524070227409 Project ID: learnmistral

[Dashboard](#) [Recommendations](#)

[+ Create a VM](#)

[+ Run a query in BigQuery](#)

[+ Deploy an application](#)

[+ Create a storage bucket](#)

Join us for the latest AI innovations at Google Cloud Next, April 9-11 in Vegas.

[Register now](#)



Quick access

API & Services

IAM & Admin

Billing

Compute Engine

Cloud Storage

BigQuery

VPC network

Kubernetes Engine

Billing

This project has no billing account

This project is not linked to a billing account

[Link a billing account](#)

[Manage billing accounts](#)

Set the billing account for project “LearnMistral”

Billing account *

My Billing Account 1



Any charges for this project will be billed to the account you select here.

Cancel

Set account



Google Cloud



LearnMistral

Search (/) for resources, docs, products, and more



Search

API

APIs & Services



APIs & Services

[+ Enable APIs and services](#)



Google Cloud



LearnMistral



Product details



Vertex AI API

[Google Enterprise API](#)

Train high-quality custom machine learning models with minimal machine learning expertise and...

ENABLE

[TRY THIS API](#)

Billing required

Vertex AI API requires a project with a billing account.

CANCEL

ENABLE BILLING



Google Cloud



LearnMistral



Product details



Cloud Resource Manager API

[Google Enterprise API](#)

Creates, reads, and updates metadata for Google Cloud Platform resource containers.

ENABLE

TRY THIS API [↗](#)

Garden|



Search

Search Results



Model Garden

Vertex AI



Model Garden on Vertex AI

Search models

Mistral



Mistral Small 3.1 (25.03)

Mistral Small 3.1 (25.03) is the enhanced version of Mistral Small 3, featuring multimodal capabilities and an extended context length of...



Mistral Large (24.11)

Mistral Large (24.11) is the latest version of the Mistral Large model now with improved reasoning and function calling capabilities.



Mistral Nemo

Mistral Nemo is a cutting-edge Language Model (LLM) boasting state-of-the-art reasoning, world knowledge, and coding capabilities wit...



Mistral Small 3.1 (25.03)

Mistral Small 3.1 (25.03) is the enhanced version of Mistral Small 3, featuring multimodal capabilities and an extended context length of up to 128k.

Enable

 Open Notebook

View Code

Purchase summary

Agree to the following details and terms to enable this product

Project

LearnMistral

Update a project from the project selector on top navigation bar

Pricing

Usage fee

Usage fee is billed every month

Input Tokens 	USD 0.10 /1M tokens
Output Tokens 	

If you pay in a currency other than USD, the prices listed in your currency on [Google Cloud SKUs](#)  apply. Please refer to [Google Cloud Price List](#)  for the latest pricing.

Terms

Terms and agreements

- ☒ By purchasing, deploying, accessing, or using this product, you acknowledge that Google is the merchant of record and Vendor's reseller with respect to this transaction and you agree to comply with the [Google Cloud Marketplace Terms of Service](#) , [Mistral AI Terms of Service](#)  and the terms of applicable open source software licenses bundled with the product.

AGREE

Successfully purchased Mistral Small 3.1 (25.03)



Now that you've agreed to the terms, you can enable Mistral Small 3.1 (25.03) in Vertex AI

[MANAGE ON VERTEX AI](#)



Mistral Small 3.1 (25.03)

Mistral Small 3.1 (25.03) is the enhanced version of Mistral Small 3, featuring multimodal capabilities and an extended context length of up to 128k.

[CO Open Notebook](#)

[View Code](#)

Enable required APIs



To use Colab Enterprise, you need to enable the following APIs.

[Vertex AI API](#)  

 Enabled

[Dataform API](#)  

 Enabled

[Compute Engine API](#)  

 Not enabled

[Send feedback](#)

[Cancel](#)

[Enable](#)

Colab Enterprise ▾ mistral... ynb × +

File Edit View Insert Runtime Tools ☁

Share

Commands + Code + Text

✓ RAM Disk ▾ Switch to L4 ^

☰

🔍

{x}

📁

▾

Getting Started with Mistral AI Models

CO

Open in Colab

CO

Open in Colab Enterprise

🏠

Open in Workbench

🔗

View on GitHub

Overview

Mistral AI on Vertex AI

Mistral AI models on Vertex AI offer fully managed and serverless models are offered as managed APIs. To use a Mistral AI model on Vertex AI, send a request directly to the Vertex AI API endpoint.

You can stream your Mistral AI model responses to reduce the end-user latency perception. A streamed response uses server-sent events (SSE) to incrementally stream the response.

Learn more about [Vertex AI](#).

▾

Authenticate your notebook environment (Colab only)

✓ 0s

[1]

import sys

if "google.colab" in sys.modules:
 from google.colab import auth

 auth.authenticate_user()

▾

Select one of Mistral AI models

▶

MODEL = "mistral-small-2503" # @param ["mistral-small-2503", "codestral-large-2407", "codestral-2501"]
if MODEL == "mistral-small-2503":
 available_regions = ["europe-west4", "us-central1"]
 available_versions = ["latest"]
elif MODEL == "mistral-large-2411":
 available_regions = ["europe-west4", "us-central1"]
 available_versions = ["latest"]
elif MODEL == "mistral-large":
 available_regions = ["europe-west4", "us-central1"]
 available_versions = ["2407"]
elif MODEL == "mistral-nemo":
 available_regions = ["europe-west4", "us-central1"]
 available_versions = ["2407"]
elif MODEL == "codestral":
 available_regions = ["europe-west4", "us-central1"]
 available_versions = ["2405"]
elif MODEL == "codestral-2501":
 available_regions = ["europe-west4", "us-central1"]
 available_versions = ["latest"]

MODEL:

mistral-small-2503 ▾

🔗

▶

import ipywidgets as widgets
from IPython.display import display

dropdown_loc = widgets.Dropdown(
 options=available_regions,
 description="Select a location:",
 font_weight="bold",
 style={"description_width": "initial"},
)



Select a location: us-central1



Select the model version (optional): latest



Selected: us-central1

```
PROJECT_ID = "learnmistral" # @param {type:"string"}
ENDPOINT = f"https://{LOCATION}-aiplatform.googleapis.com"
SELECTED_MODEL_VERSION = "" if MODEL_VERSION == "latest" else f"@{MODEL_VERSION}"

if not PROJECT_ID or PROJECT_ID == "[your-project-id]":
    raise ValueError("Please set your PROJECT_ID")
```

PROJECT_ID: learnmistral

Sign in with Google



Sign in to Google Cloud SDK

pavlocherkashyn@gmail.com

By continuing, Google will share your name, email address, and profile picture with Google Cloud SDK. See Google Cloud SDK's Privacy Policy and Terms of Service.

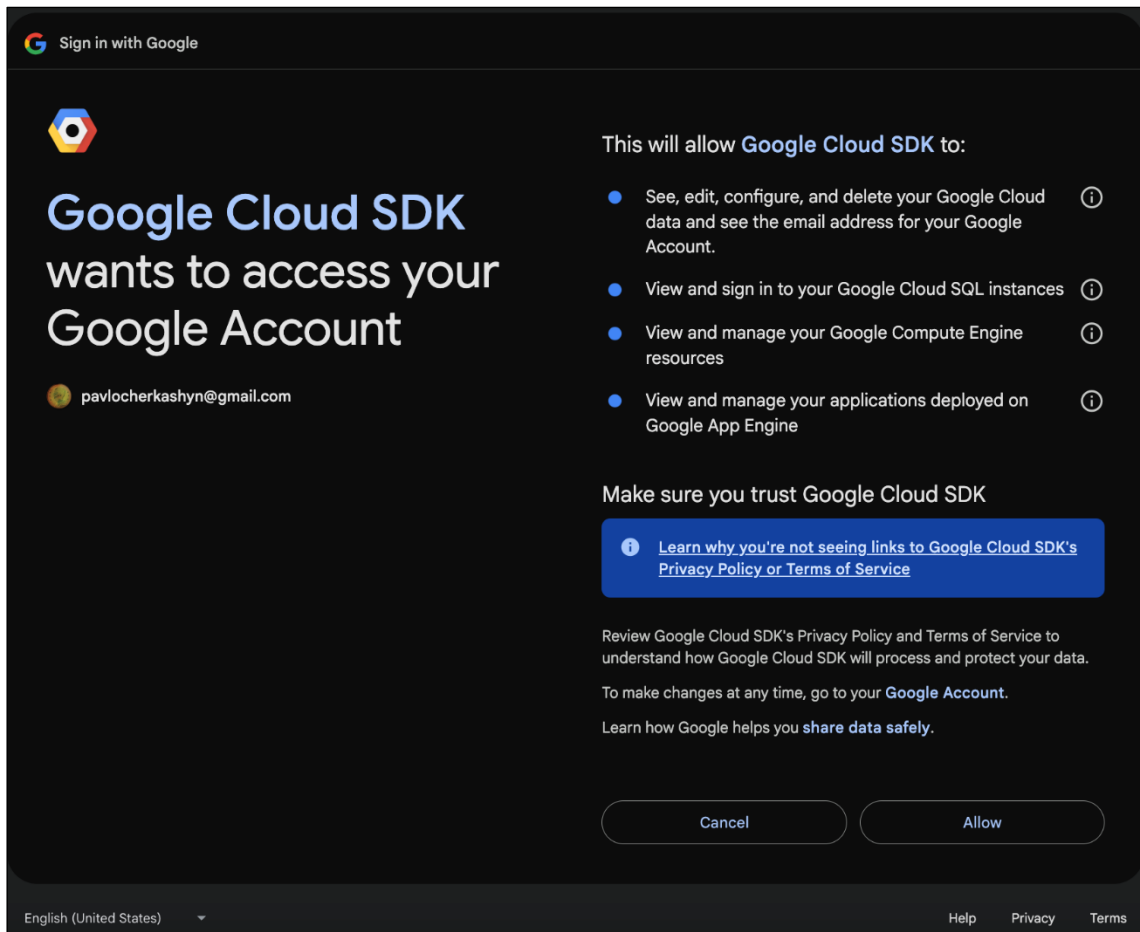
You can manage Sign in with Google in your [Google Account](#).

Cancel

Continue

English (United States)

[Help](#) [Privacy](#) [Terms](#)



You are now authenticated with the gcloud CLI!

[Send feedback](#)



The authentication flow has completed successfully. You may close this window, or check out the resources below.

Information about command-line tools and client libraries

To learn more about Google Cloud CLI commands, see the [gcloud CLI guide](#).

To learn more about the command-line tools for App Engine, Compute Engine, Cloud Storage, BigQuery, Cloud SQL, and Cloud DNS (which are all bundled with the gcloud CLI), see [Accessing services with the gcloud CLI](#).

If you're a client application developer and want to find out more about accessing Google Cloud services with a programming language or framework, see [Client Libraries Explained](#).

