LLM Zoomcamp FAQ

Editing guidelines:

* When adding a new FAQ entry, make sure the question is “Heading 2”
* Feel free to improve if you see something is off
* Don’t change the formatting in the document or add any visual “improvements.”
* Don’t change the pages format (it should be “pageless”)

# General course-related questions

## I just discovered the course. Can I still join?

Yes, but if you want to receive a certificate, you need to submit your project while we’re still accepting submissions.

## Course - I have registered for the LLM Zoomcamp. When can I expect to receive the confirmation email?

You don't need it. You're accepted. You can also just start learning and submitting homework (while the form is **Open**) without registering. It is not checked against any registered list. Registration is just to gauge interest *before* the start date.

## What is the video/zoom link to the stream for the “Office Hours” or live/workshop sessions?

The zoom link is only published to instructors/presenters/TAs.

Students participate via Youtube Live and submit questions to Slido (link would be pinned in the chat when Alexey goes Live). The video URL should be posted in the [announcements channel on Telegram & Slack](https://t.me/dezoomcamp) before it begins. Also, you will see it live on the DataTalksClub [YouTube Channel](https://www.youtube.com/c/DataTalksClub).

Don’t post your questions in chat as it would be off-screen before the instructors/moderators have a chance to answer it if the room is very active.

## Cloud alternatives with GPU

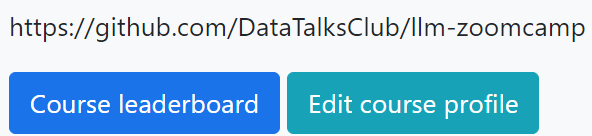
Check the quota and reset cycle carefully - is the free hours per month or per week? Usually if you change the configuration, the free hours quota might also be adjusted,or it might be billed separately.

1. Google Colab
2. Kaggle
3. Databricks (?), so many others.

Use GPTs to find out. Some might have restrictions on what you can and cannot install, so be sure to read what is included in a free vs paid tier.

## Leaderboard - I am not on the leaderboard / how do I know which one I am on the leaderboard?

When you set up your account you are automatically assigned a random name such as “Lucid Elbakyan” for example. Click on the *Jump to your record on the leaderboard* link to find your entry.

  
If you want to see what your Display name is, click on the Edit Course Profile button.

1. First field is your nickname/displayed-name, change it if you want to be known as your Slack username or Github username or whatever nickname of your choice, if you want to remain anonymous.
2. Unless you want “Lucid Elbakyan” on your certificate, it is **mandatory** that you change the second field to your **official name** as in your identification documents - passport, national ID card, driver’s license, etc. This is the name that is going to appear on your Certificate!

## Certificate - Can I follow the course in a self-paced mode and get a certificate?

No, you can only get a certificate if you finish the course with a “live” cohort.

We don't award certificates for the self-paced mode. The reason is you need to peer-review 3 capstone(s) after submitting your project.

You can only peer-review projects at the time the course is running; after the form is closed and the peer-review list is compiled.

## I missed the first homework - can I still get a certificate?

Yes, you need to pass the Capstone project to get the certificate. Homework is not mandatory, though it is recommended for reinforcing concepts, and the points awarded count towards your rank on the leaderboard.

## I was working on next week’s homework/content - why does it keep changing?

This course is being offered for the first time, and things will keep changing until a given module is ready, at which point it shall be announced. Working on the material/homework in advance will be at your own risk, as the final version could be different.

## When will the course be offered next?

Summer 2025 (via Alexey).

## Are there any lectures/videos? Where are they?

Please check the bookmarks and pinned links, especially DataTalks.Club’s YouTube account.

## WSL2 - ResponseError: model requires more system memory (X.X GiB) than is available (Y.Y GiB). My system has more than X.X GiB.

Your WSL2 is set to use Y.Y GiB, not all your computer memory. Create .wslconfig file under your Windows user profile directory (C:\Users\YourUsername\.wslconfig) with the desired RAM allocation:

[wsl2]

memory=8GB

Restart WSL: wsl --shutdown

Run the free command to verify the changes. For more details, read [this article](https://www.aleksandrhovhannisyan.com/blog/limiting-memory-usage-in-wsl-2/).

## **Server Error (500)** When logging in to course homework using GitHub

### Additional error text seen:

**Third-Party Login Failure**

An error occurred while attempting to login via your third-party account.

The current solution is to use Google or Slack to login and submit homework answers as the root cause analysis for the GitHub issue is sporadic and doesn’t impact all users.

## Why are we not using Langchain in the course?

Langchain is a framework for building LLM-powered apps. We're not using it to learn the basics; think of it like learning HTML, CSS, and JavaScript before learning React or Angular.

Added by Marcelo Nieva

# Module 1: Introduction

## `OpenAI: Error when running OpenAI chat.completions.create command

You may receive the following error when running the OpenAI chat.completions.create command due to insufficient credits in your OpenAI account:

| NotFoundError: Error code: 404 - {'error': {'message': 'The model `gpt-4o` does not exist or you do not have access to it.', 'type': 'invalid\_request\_error', 'param': None, 'code': 'model\_not\_found'}} |
| --- |

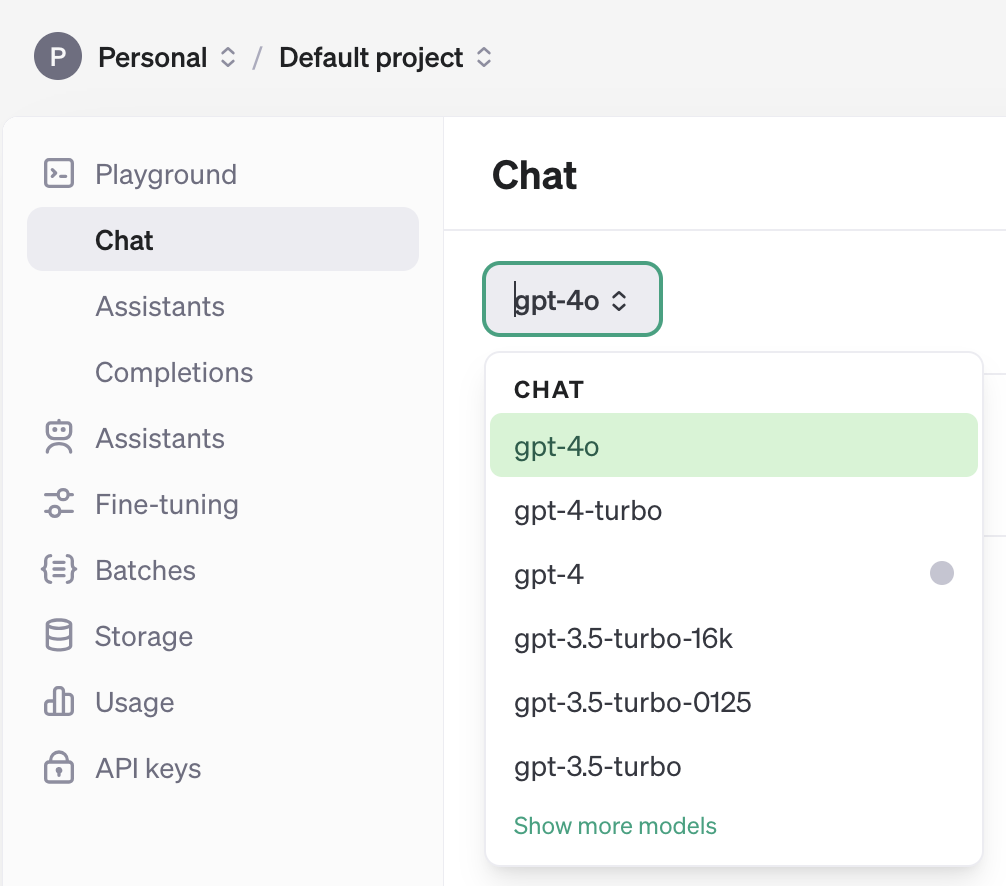
## ￼OpenAI: Error: RateLimitError: Error code: 429 -

RateLimitError: Error code: 429 - {'error': {'message': 'You exceeded your current quota, please check your plan and billing details. For more information on this error, read the docs: <https://platform.openai.com/docs/guides/error-codes/api-errors.>', 'type': 'insufficient\_quota', 'param': None, 'code': 'insufficient\_quota'}

The above errors are related to your OpenAI API account’s quota.  
There is no free usage of OpenAI’s API so you will be required to add funds using a credit card (see pay as you go in the OpenAI settings at [platform.openai.com](http://platform.openai.com)). Once added, re-run your python command and you should receive a successful return code.

**Steps to resolve:**

1. Add credits to your account [here](https://platform.openai.com/settings/organization/billing/overview) (min $5)
2. In chat.completions.create(model='**gpt-4o**', …) specify one of the available for you models:



1. You might need to recreate an API key after adding credits to your account and update it locally.

## OpenAI: Error: 'Cannot import name OpenAI from openai'; How to fix?

Update openai version from 0.27.0 -> any 1.x version

## OpenAI: How much will I have to spend to use the Open AI API?

Using the Openai API does not cost much, you can recharge from 5 dollars. At least for what I spent on the first unit it was barely 5 cents.

## OpenAI: Do I have to subscribe and pay for Open AI API for this course?

No, you don't have to pay for this service in order to complete the course homeworks, you could use some of the alternatives free from this list posted into the course Github.

[llm-zoomcamp/01-intro/open-ai-alternatives.md at main · DataTalksClub/llm-zoomcamp (github.com)](https://github.com/DataTalksClub/llm-zoomcamp/blob/main/01-intro/open-ai-alternatives.md)

## ElasticSearch: ERROR: BadRequestError: BadRequestError(400, 'media\_type\_header\_exception', 'Invalid media-type value on headers [Content-Type, Accept]', Accept version must be either version 8 or 7, but found 9.

Reason: Elastic search client and server are on different versions

Solution: Upgrade the Elastic search on Docker to version 9

docker run -it \

--rm \

--name elasticsearch \

-p 9200:9200 \

-p 9300:9300 \

-e "discovery.type=single-node" \

-e "xpack.security.enabled=false" \

elasticsearch:9.0.1

If upgrading to version 9 doesn’t work, check the client version (python module) using `pip show elasticsearch`. Then install that specific version of Elastic Search on Docker. Check if it worked using `curl<http://localhost:9200>`. Example output of `pip show elastic search`:

*Name: elasticsearch*

*Version: 9.0.2*

*Summary: Python client for Elasticsearch*

*Home-page: https://github.com/elastic/elasticsearch-py*

*Author:*

*Author-email: Elastic Client Library Maintainers <client-libs@elastic.co>*

*License-Expression: Apache-2.0*

*Location: /home/codespace/.python/current/lib/python3.12/site-packages*

*Requires: elastic-transport, python-dateutil, typing-extensions*

*Required-by:*

## Fix BadRequestError: BadRequestError(400, 'media\_type\_header\_exception', 'Invalid media-type value on headers [Content-Type, Accept]', Accept version must be either version 8 or 7, but found 9. Accept=application/vnd.elasticsearch+json; compatible-with=9)

When try to connect to the **Elasticsearch** server/node version 8.17.6 (as instructed for the Homework 1) running from with the Docker container with the python client **elasticsearch** version 9.x or more, we run into the **BadRequestError** mentioned above.

This happens because **pip install elasticsearch** install elasticsearch 9.x python client which runs into compatibility issues with the Elasticsearch 8.17.6. So we can use

**pip install "elasticsearch>=8,<9"** for mitigation of the problem.

(Added by Siddhartha Gogoi)

## ElasticSearch: ERROR: Elasticsearch exited unexpectedly

If you get this error, it’s likely that elasticsearch doesn’t get enough RAM

I specified the RAM size to the configuration (-m 4GB)

docker run -it \

--rm \

--name elasticsearch \

-m 4GB \

-p 9200:9200 \

-p 9300:9300 \

-e "discovery.type=single-node" \

-e "xpack.security.enabled=false" \

docker.elastic.co/elasticsearch/elasticsearch:8.4.3

(-m 2gb should also work)

Another possible solution may be to set the memory\_lock to false:

docker run -it \

--rm \

--name elasticsearch \

-p 9200:9200 \

-p 9300:9300 \

-e "discovery.type=single-node" \

-e see"xpack.security.enabled=false" \

-e "ES\_JAVA\_OPTS=-Xms512m -Xmx512m" \

-e "bootstrap.memory\_lock=false" \

docker.elastic.co/elasticsearch/elasticsearch:8.4.3

## ElasticSearch: ERROR: Elasticsearch.index() got an unexpected keyword argument 'document'

Instead of **document** as used in the course video, use **doc**

## Docker: How do I store data persistently in Elasticsearch?

When you stop the container, the data you previously added to elastic will be gone. To avoid it, we can add volume mapping:

docker volume create elasticsearch\_data

docker run -it \

--rm \

--name elasticsearch \

-p 9200:9200 \

-p 9300:9300 \

-v elasticsearch\_data:/usr/share/elasticsearch/data \

-e "discovery.type=single-node" \

-e "xpack.security.enabled=false" \

docker.elastic.co/elasticsearch/elasticsearch:8.4.3

## Authentication: Safe and easy way to store and load API keys

You can store your different API keys in a yaml file that you will add in your .gitignore file. Be careful to never push or share this file.

* For example, you can create a new file named “api\_keys.yml” in your repository.
* Then, do not forget to add it in your .gitignore file:

*#api\_keys*

api\_keys.yml

* You can now fill your api\_keys.yml file:

OPENAI\_API\_KEY: “sk[...]”

GROQ\_API\_KEY: “gqk\_[...]”

* Save your file.
* You will need the pyyaml library to load your yaml file, so run this command in your terminal:

pip install pyyaml

* Now, open your jupyter notebook.
* You can load your yaml file and the associated keys with this code:

import yaml

# Open the file

with open('api\_keys.yml', 'r') as file:

# Load the data from the file

data = yaml.safe\_load(file)

# Get the API key (Groq example here)

groq\_api\_key = data['GROQ\_API\_KEY']

* Now, you can easily replace the “api\_key” value directly with the loaded values without loading your environment variables.

Added by Mélanie Fouesnard

## How to store and load API keys using .env file

Store the API key in a .env file, then

import os

from dotenv import load\_dotenv

load\_dotenv(os.path.abspath("<path-to-.env>"))

os.getenv("API\_KEY\_abc")

Make sure to add the .env file in the .gitignore.

## Authentication: Why is my OPENAI\_API\_KEY not found in the jupyter notebook?

### Option1: using direnv

created the .envrc file & added my API key, ran direnv allow in the terminal

was getting an error: "OpenAIError: The api\_key client option must be set either by passing api\_key to the client or by setting the OPENAI\_API\_KEY environment variable"

resolution: install dotenv & add the following to a cell in the notebook. You can install dotenv by running: pip install python-dotenv.

from dotenv import load\_dotenv

load\_dotenv('.envrc')

### Option 2: using Codespaces Secrets

* Log in to your GitHub account and navigate to Settings > Codespaces
* There is a section called secrets where you can create Secrets like OPENAI\_API\_KEY and select for which repositories the secret is supposed to be available.
* Once you set this up, the key will be available in your codespaces session

## OpenSource: I am using Groq, and it doesn't provide a tokenizer library based on my research. How can we estimate the number of OpenAI tokens asked in homework question 6?

The question asks for the number of tokens in gpt-4o model. tiktoken is a python library that can be used to get the number of tokens. You don't need openai api key to to get the number of tokens. You can use the code provided in the question to get the number of tokens.

## OpenSource: Can I use Groq instead of OpenAI?

You can use any LLM platform for your experiments and your project. Also, the homework is designed in such a way that you don’t need to have access to any paid services and can do it locally. However, you would need to adjust the code for that platform. See their documentation pages.

## OpenSource: Can I use open-source alternatives to OpenAI API?

Yes. See module 2 and the [open-ai-alternatives.md](https://github.com/DataTalksClub/llm-zoomcamp/blob/main/01-intro/open-ai-alternatives.md) in module 1 folder.

## Returning Empty list after filtering my query (HW Q3)

This is likely to be an error when indexing the data. First you need to add the index settings before adding the data to the indices, then you will be good to go applying your filters and query.

## ModuleNotFoundError on import docx in parse-faq.ipynb

The correct package name for docx is python-docx, not docx.

## OpenAI: Why does my token count differ from what OpenAI reports?

When using tiktoken.encode() to count tokens in your prompt, you might get a number like 320, while OpenAI’s API response reports something like 327. This is expected and due to **internal tokens added by OpenAI’s chat formatting**.

Here’s what happens under the hood:

* Each message in a chat.completions.create() call (e.g., {role: "user", content: "..."}) adds **4 structural tokens** (role, content, separators).
* The API also adds **2 tokens globally** to mark the start of assistant response generation.
* Any **extra newlines, whitespace, or uncommon Unicode characters** in your content may slightly increase the token count.

So even if your visible text is 320 tokens, OpenAI may count 327 due to these internal additions.

Added by José Luis Martínez (Maxkaizo)

## Ollama: How to install Ollama?

First, install Ollama:

Go to <https://ollama.com/download>

Choose your operating system:

* macOS: Download the .pkg and install
* Windows: Download the .msi and install
* Linux: Run this in terminal:

*curl -fsSL https://ollama.com/install.sh | sh*

Open a terminal and type:

*ollama run llama3*

This will:

* Download theLLaMA 3 model (~4GB)
* Start the model locally
* Open a chat-like interface where you can type questions

To test the Ollama local server, execute the following command:  
  
*curl* [*http://localhost:11434*](http://localhost:11434)

You should receive something like:

*{"models": [...]}*

Then, install the Python client:

*pip install ollama*Here, you have a minimal python example:  
  
***import*** *ollama*

*response* ***=*** *ollama****.****chat(*

*model****=****'llama3',*

*messages****=****[{"role": "user", "content": your\_prompt}]*

*)*

*print(response['message']['content'])*

Added by Alexander Daniel Rios

## When I re-run the code in Jupyter notebook multiple times for homework#1, the index building code snippet fails.

The solution is to delete any potential existing index with the same name before attempting to create the index (see code snippet below).

# Check if the index exists and delete it if it does

if es\_client.indices.exists(index=index\_name):

print(f"Deleting existing index: {index\_name}")

es\_client.indices.delete(index=index\_name)

print(f"Index {index\_name} deleted.")

However, with this approach sometimes when you re-run the code multiple times, the index gets messed up and you for example will see a different score output each time you execute the code for question#3 in homework1. To fix this 1) Go to docker desktop and stop Elasticsearch container , delete the container image and re-initiate the Elasticsearch container by creating it from scratch per the instructions ‘1.6 Searching with ElasticSearch’ 2) Change the name of the index in your code to anything other than -> index\_name = "course-questions"

## Question

Answer

## Q3. Searching in the homework for Module 1,

## For Q3, you will observe that you need to remove filter from the search\_query dictionary definition within the elastic\_search\_filter function. Otherwise, you will not get answers that are in the options.

## SSL Error when connecting to locally running ElasticSearch instance via SDK:

The issue is likely that you’re trying to use HTTPS instead of HTTP when you call local.

To remove ES authentication constraints, set xpack.security.enabled=false in the ES docker settings.

## 

# Module 2: Vector Search

## 

## What are embeddings?

Embeddings = turning non-numerical data into numerical data, while preserving meaning and context. Similar non-numerical data, when entered into an embedding algorithm, should produce similar numerical data. The similar numerical data being close in mathematical values allows for mathematical semantic similarity algorithms to be leveraged. See, also, “vector space model” and “dimensionality reduction”.

## Find maximum of an numpy array (of any dimension):

max\_value = numpy\_array.max()

## What is the cosine similarity?

Cosine similarity is a measure used to calculate the similarity between two non-zero vectors, often used in text analysis to determine how similar two documents are based on their content. This metric computes the cosine of the angle between two vectors, which are typically word counts or TF-IDF values of the documents. The cosine similarity value ranges from -1 to 1, where 1 indicates that the vectors are identical, 0 indicates that the vectors are orthogonal (no similarity), and -1 represents completely opposite vectors.

## Can I use another vector db for running RAGs vector search?

Yes, there are other vector databases, for example Milvus, which is open sourced and you can see the documentation here: <https://milvus.io/docs/overview.md>

## Why does cosine similarity reduce to a matrix multiplication between the embeddings and the query vector?

Cosine similarity measures how aligned two vectors are, regardless of their magnitude. When all vectors (including the query) are normalized to unit length, their magnitudes no longer matter. In this case, cosine similarity is equivalent to simply taking the dot product between the query and each document embedding. This allows us to compute similarities efficiently using matrix multiplication.

Added by José Luis Martínez

## Question: Why am I getting `docker: invalid reference format` when trying to run Qdrant with a volume in Windows?

If you're running the `docker run` command on \*\*Windows (especially Command Prompt or PowerShell)\*\* and you use `$(pwd)` to mount a volume, you'll likely get the following error:

**docker: invalid reference format.**

Answer: The expression **$(pwd)** is a Unix-style command used to get the current working directory. It \*\*won’t work in Windows\*\*, which causes Docker to misinterpret the image name or the `-v` argument, hence the “invalid reference format” error.

Solution:

1. Use the full absolute path instead of $(pwd), for example:

**docker run -p 6333:6333 -p 6334:6334 \**

**-v C:/Users/youruser/path/to/qdrant\_storage:/qdrant/storage:z \**

**qdrant/qdrant**

2. Alternatively, use a named volume, as shown in the video:

**docker volume create qdrant\_storage**

**docker run -p 6333:6333 -p 6334:6334 \**

**-v qdrant\_storage:/qdrant/storage \**

**qdrant/qdrant**

Added by José Luis Martínez

## ImportError: DLL load failed while importing onnxruntime\_pybind11\_state: A dynamic link library (DLL) initialization routine failed.

If you use Anaconda or Miniconda, you can try re-installing onnxruntime with conda

conda install -c conda-forge onnxruntime

One of the ways you can create an environment to use "qdrant-client[fastembed]>=1.14.2" which throws this error is to create the environment using Conda. Here are the steps -

1. Create a Conda environment using  
   conda create --name llm-zoomcamp-env python=3.10
2. Activate your new environment  
   conda activate llm-zoomcamp-env
3. Install the dependency  
   pip install "qdrant-client[fastembed]>=1.14.2"
4. Use this environment either using Jupyter notebook or in VSCode/Cursor  
   For VSCode/Cursor -> Ctrl+Shift+P/Cmd+Shift+P -> Select Python Enterpreter -> Select llm-zoomcamp-env

## Question: To set up a Qdrant client, when to use client = QdrantClient("http://localhost:6333") vs client = QdrantClient(":memory:")?

Use the former if you are running Qdrant in Docker locally and it connects your notebook to the Qdrant server running in Docker.

The latter option creates an in-memory Qdrant instance that runs inside your Python process (no Docker, no persistence, no networking). It’s:

* Only for testing or prototyping
* Not connected to your Docker-based Qdrant
* Wiped clean when the notebook or script stops

# Module 3: Evaluation

## I'm getting the error “cannot import name 'VectorSearch' from 'minsearch'” even though I installed the latest version of minsearch. How can I fix it?

If you're working with Jupyter notebooks, make sure the kernel you're using has the correct version of minsearch. You can check the version in your kernel with: minsearch.\_\_version\_\_

You can also try installing the latest version directly from a notebook cell using:

%pip install -U minsearch

%pip is a Jupyter magic command that makes sure the package gets installed in the same environment your notebook kernel is using (unlike !pip, which might install it somewhere else).

Added by Marcelo Nieva

## Question: Why was .dot(...) used directly to compute cosine similarity in the lesson, but normalization is emphasized in the homework?

Answer: In the lesson, .dot(...) was used under the assumption that the embeddings returned by the model (e.g. model.encode(...) from OpenAI) are already normalized to have unit length. In that case, the dot product is mathematically equivalent to cosine similarity.

In the homework, however, we use classic embeddings like TF-IDF + SVD, which are not normalized by default. This means that the dot product does not represent cosine similarity unless we manually normalize the vectors.

Added by José Luis Martínez

## Question

Answer

# Module 4: Monitoring

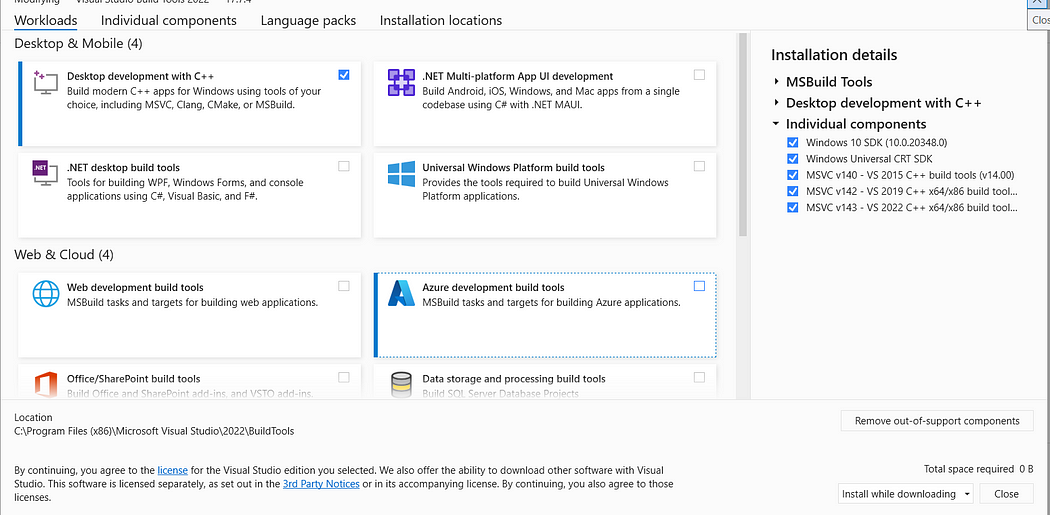
## Warning: 'model "multi-qa-mpnet-base-dot-v1" was made on sentence transformers v3.0.0 bet' how to suppress?

Upgrade `sentence-transformers` to v3.0.0>= e.x pip install sentence-transformers>=3.0.0 to avoid the warnings

## In Windows OS : OSError: [WinError 126] The specified module could not be found. Error loading "C:\Users\USER\AppData\Local\Programs\Python\Python310\lib\site-packages\torch\lib\fbgemm.dll" or one of its dependencies.

Solution 1 : Install Visual C++ Redistributable

Solution 2 : Install Visual Studio, not Visual Studio Code. Like in this depicted below and restart your system. For more details, please follow this link : https://discuss.pytorch.org/t/failed-to-import-pytorch-fbgemm-dll-or-one-of-its-dependencies-is-missing/201969

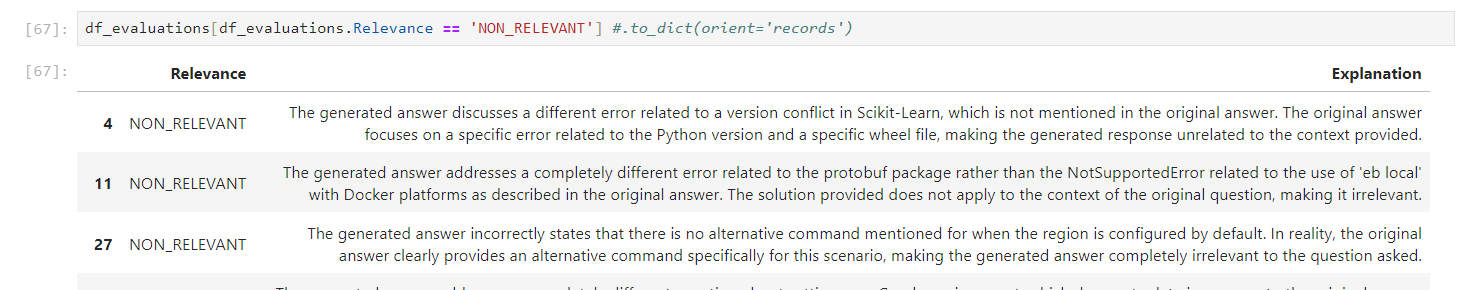


## OperationalError when running python prep.pypsycopg2. OperationalError: could not translate host name "postgres" to address: No such host is known. How do I fix this issue?

Inside .env file change POSTGRES\_HOST=localhost

## How set Pandas to show entire text content in a column. Useful to view the entire Explanation column content in the LLM-as-judge section of the offline-rag-evaluation notebook

By default, in the dataframe visualization, Pandas truncate the text content in a column to 50 characters. In order to view the entire explanation given by the judge llm for a NON RELEVANT answer, as in figure:



The instruction to show the results must be preceded by:

pd.set\_option('display.max\_colwidth', None)

Here are the specs for the display\_max\_colwidth option, as describide in the [official docs](https://pandas.pydata.org/docs/user_guide/options.html):

*display.max\_colwidth : int or None*

*The maximum width in characters of a column in the repr of*

*a pandas data structure. When the column overflows, a "..."*

*placeholder is embedded in the output. A 'None' value means unlimited.*

*[default: 50] [currently: 50]*

## How to normalize vectors in a Pandas DataFrame column (or Pandas Series)?

import numpy as np

normalize\_vec = lambda v: v / np.linalg.norm(v)

df["new\_col"] = df["org\_col"].apply(norm\_vec)

## How to compute the quantile or percentile of Pandas DataFrame column (or Pandas Series)?

To compute the 75% percentile or 0.75 quantile:

quantile: int = df["col"].quantile(q=0.75)

## How can I remove all Docker containers, images, and volumes, and builds from the terminal?

1. Delete all containers (including running ones):

```

docker rm -f

```

2. Remove all images:

```

docker rmi -f

```

3. Delete all volumes:

```

docker volume rm

```

## I want the user to only be able to give feedback once per submission (+1 or -1). For this I'm using a st.session attribute: submitted to enable or disable feedback buttons. When I submit text using the ask button. If I pressed +1, I'm allowed to repress +1 one more time (button not disabled). If I pressed -1, I'm allowed to repress +1 or -1 one more time (buttons not disabled). The buttons should be disabled if st.session.submitted is False. The issue is mainly in st.session.submitted, it gets somehow reassigned to True again, despite one feedback button being pushed.

Solved:

<https://discuss.streamlit.io/t/streamlit-session-attributes-reassigned-somewhere/76059/2?u=mohammed2>

## When trying to run a streamlit app using docker-compose, I get: Error response from daemon: failed to create task for container: failed to create shim task: OCI runtime create failed: runc create failed: unable to start container process: exec: "streamlit": executable file not found in $PATH: unknown. The app runs fine outside of docker-compose

Make sure to create a Dockerfile and run ‘docker-compose up –build’ when adding streamlit to docker-compose

# Module 6: X

## Question

Answer

## Question

Answer

# Capstone Project

## Is it a group project?

No, the capstone is a solo project.

## Do we submit 2 projects, what does attempt 1 and 2 mean?

* You only need to submit 1 project.   
  If the submission at the first attempt fails, you can improve it and re-submit during **attempt#2** submission window.
* If you want to submit 2 projects for the experience and exposure, you must use different datasets and problem statements.
* If you can’t make it to the **attempt#1** submission window, you still have time to catch up to meet the **attempt#2** submission window
* Remember that the submission does not count towards the certification if you do not participate in the peer-review of 3 peers in your cohort

## Does the competition count as the capstone?

No, it does not (answered in office hours Jul 1st, 2024). You can participate in the [math-kaggle-llm-competition](https://datatalks-club.slack.com/archives/C0791HB4A58) as a group if you want to form teams; but capstone is an individual attempt.

## How is my capstone project going to be evaluated?

* Each submitted project will be evaluated by 3 (three) randomly assigned students who have also submitted the project.
* You will also be responsible for grading the projects from 3 fellow students yourself. Please be aware that: not complying to this rule also implies you failing to achieve the
* Certificate at the end of the course.
* The final grade you get will be the median score of the grades you get from the peer reviewers. And of course, the peer review criteria for evaluating or being evaluated must follow the guidelines defined here (TBA for link).

## When and how will we be assigned projects for review/grading?

After the submission deadline has passed, an Excelsheet will be shared with 3 projects being assigned to each participant.

## I’ve already submitted my project. Why can’t I review any projects?

Once the project submission deadline has passed, we will be assigned projects to evaluate. You can't choose which projects to evaluate, and you can’t review before the list has been released.

## How can I find some good ideas or datasets for the project?

Answer: Please check <https://github.com/DataTalksClub/llm-zoomcamp/blob/main/project.md> for several ideas and datasets that could be used for the project, along with tips and guidelines.

## Do I have to use ElasticSearch or X library?

Answer: No, you don’t have to use ElasticSearch. You can use any library you want. Just make sure it is documented so your peer-reviewers can reproduce your project.

## What other alternatives to ElasticSearch are there?

You could use some of this free alternatives for elastic search

* [Milvus](https://milvus.io/): an open source library that has the same functionalities that has elastic
* [OpenSearch](https://opensearch.org/): also a free open source library that has the same functionalities as elastic

Let's imagine, today I start using multi-qa-distilbert-cos-v1 (https://huggingface.co/sentence-transformers/multi-qa-distilbert-cos-v1).

I create embeddings and index them.

Tomorrow, the author of the model decides to update it because of some reason.

What happens with all indexed embeddings? Do they become incompatible and I will need to re-index everything?

There is an option to save the model locally also. This way even if the cloud model changes your code should work.

# Certificates

## !] See [names on certificates](#_naukgwju1z8z)

## Question

Answer

# Workshops: dlthub

## Can I use the workshop materials for my own projects or share them with others?

Since dlt is open-source, we can use the content of this workshop for a capstone project. Since the main goal of dlt is to load and store data easily, we can even use it for other zoomcamps (mlops zoomcamp project for example). Do not hesitate to ask questions or use it directly in your projects.

Added by Mélanie Fouesnard

## How to set up a new dlt project when loading from cloud?

Start with “dlt init filesystem duckdb” on the command line.

More directions: https://dlthub.com/docs/tutorial/filesystem

## How much free time does Google Colab gives for T4 GPU resource type?

## Google colab offers only 1 hour every 24h for using the T4 GPU resource type. But you can still use the CPU which is a bit slower than T4, especially while running the RAG.

## There is an error when opening the table using dbtable = db.open\_table("notion\_pages\_\_\_homework"): FileNotFoundError: Table notion\_pages\_\_\_homework does not exist.Please first call db.create\_table(notion\_pages\_\_\_homework, data)

The error indicates that you have not changed all instances of “employee\_handbook” to “homework” in your pipeline settings

## There is an error when running main(): FileNotFoundError: Table notion\_pages\_\_\_homework does not exist.Please first call db.create\_table(notion\_pages\_\_\_homework, data)

Make sure you open the correct table in line 3: dbtable = db.open\_table("notion\_pages\_\_\_homework")T

## How do I know which tables are in the db

You can use the db.table\_names() to list all the tables in the db

## Does DLT have connectors to ClickHouse or StarRocks?

Currently, DLT does not have connectors for ClickHouse or StarRocks but are open to contributions from the community to add these connectors.

## Notebook does not have secret access or 401 Client Error: Unauthorized for url: https://api.notion.com/v1/search

If you get this error

## 

Or 401 Client Error , then you either need to grant access to the key or the key is wrong.

## Error: How to fix requests library only installs v2.28 instead of v2.32 required for lancedb?

Install directly from source E.g `pip install "requests @ <https://github.com/psf/requests/archive/refs/tags/v2.32.3.zip>"`

## Question: What is the approximate cost of running the workshop notebook (DLT + Cognee)?

Answer:The total cost is approximately $0.09 USD, based on pricing as of July 7, 2025.

This estimate includes all API calls to OpenAI for generating embeddings and relationship extraction, as well as local operations for loading data into Qdrant and Kuzu.

Added by José Luis Martínez

# Workshops: Agents

## Connection refused error on prompting the ollam RAG?

If you get this error while doing the homework , simply restart the ollama server using nohup y running this line of the notebook !nohup ollama serve > nohup.out 2>&1 &

If you do stop and restart the cell, you will need to rerun the cell containing ollama serve first.

## Error: Connecting to Elasticsearch at http://elasticsearch:9200

Try removing driver bridge from there?

Added by Abiodun Gbadamosi

## Question

Multiple retrieval approaches are evaluated, and the best one is used (2 points) I am trying to evaluate a project. The person use only minsearch for evaluated but did boosting and got boostin parameter posted. Do they get one mark?

Answer

Here you go:

The evaluation criteria state that to receive 2 points, multiple RAG approaches must be evaluated, and the best one must be used. Since the individual in question is using only minsearch for evaluation, despite applying boosting, this would not qualify as evaluating multiple RAG approaches.

Therefore, they would receive only 1 point for utilizing a single RAG approach (minsearch) in their evaluation, even though they incorporated a boosting parameter. The boosting itself does not constitute a separate approach; it is simply an enhancement applied to the single method being used.

Added by Wali Mohamed

## Elasticsearch version error

Error : elasticsearch.BadRequestError: BadRequestError(400, 'media\_type\_header\_exception', 'Invalid media-type value on headers [Content-Type, Accept]', Accept version must be either version 8 or 7, but found 9. Accept=application/vnd.elasticsearch+json; compatible-with=9)

Fix :

pip uninstall elasticsearch

pip install elasticsearch==8.10.0

## AppendableIndex error in minsearch

Error: 'ImportError: cannot import name 'AppendableIndex' from 'minsearch''

Fix: pip install --upgrade minsearch

(minsearch 0.0.4 works well; jupyter kernel restarting is needed after this upgrade)

## AppendableIndex error in minsearch (not resolved by upgrading minsearch)

Error: 'ImportError: cannot import name 'AppendableIndex' from 'minsearch''

Fix: from minsearch.append import AppendableIndex

## AppendableIndex error in minsearch (not resolved by upgrading minsearch or importing from minsearch.append)

Error: 'ImportError: cannot import name 'AppendableIndex' from 'minsearch''

Fix: Rename the previously downloaded minsearch.py file to avoid conflicts, then reinstall minsearch using pip so the import works correctly.

## Any free models with tool use support?

Several Groq models offer tool use, such as Deepseek R1 or Llama 4, all of which can be used for free for development

<https://console.groq.com/docs/tool-use>

Added by Marcelo Nieva

## Question: I passed a float to my tool, but got a validation error saying it expected a number. Isn’t float a number?

Yes — in Python, float is a numeric type. But when working with FastMCP, tool inputs are validated against JSON Schema, which uses the term "number" to represent any numeric value (integers or floats).

The important thing is not the type you use in Python, but whether the JSON you send matches the tool's declared input schema.

Example:

"inputSchema": {

"type": "object",

"properties": {

"temp": {

"type": "number"

}

},

"required": ["temp"]

}

Make sure the values in "arguments" match the types declared in the tool’s schema — not Python types, but JSON types (string, number, boolean, etc.).

Added by Maxkaizo - José Luis Martínez

# 

## Testing MCP Servers with MCP Inspector

## Install MCP Inspector

1. Node should be installed

2. To install the MCP Inspector, simply run the following command in your terminal:

npm i @modelcontextprotocol/inspector

## Run MCP Inspector

· Run from the terminal the following

npx @modelcontextprotocol/inspector

## Inspect MCP Server

· Connect to the MCP Server

A screenshot of a phone

AI-generated content may be incorrect.

· The inspector can list tools, templates, resources and prompts from the MCP Server

A screenshot of a computer

AI-generated content may be incorrect.

**Reference**

[**https://medium.com/@anil.goyal0057/how-to-test-your-mcp-server-using-mcp-inspector-c873c417eec1**](https://medium.com/@anil.goyal0057/how-to-test-your-mcp-server-using-mcp-inspector-c873c417eec1)

Added - Sundara Kumar Padmanabhan

## How to Solve "RuntimeError: Already running asyncio in this thread"

Jupyter notebooks already run an event loop in the main thread to handle asynchronous code. For this reason, when you try to call asyncio.run() inside a cell, you get the following error:

RuntimeError: asyncio.run() cannot be called from a running event loop

Instead of using asyncio.run(), simply use await directly in the notebook cell.

Incorrect:

import asyncio

async def main():

async with Client(weather\_server.mcp) as mcp\_client:

# your code here

# This will cause the RuntimeError

result = asyncio.run(main())

Correct:

async def main():

async with Client(weather\_server.mcp) as mcp\_client:

# your code here

# Use await directly

result = await main()

Jupyter notebooks automatically create an asyncio event loop when they start. Since asyncio.run() attempts to create a new event loop, it conflicts with the existing loop. By using await directly, you leverage the already running event loop.

Added by Marcelo Nieva

## Question: I am using Azure OpenAI and I am still getting an error of Error code: 400 - {'error': {'message': "Missing required parameter: 'tools[0].function'.", 'type': 'invalid\_request\_error', 'param': 'tools[0].function', 'code': 'missing\_required\_parameter'}}?

Modify the get\_weather\_tool json to be the following:   
get\_weather\_tool = {

"type": "function",

"function": {

"name": "get\_weather",

"description": "Get the current weather for a specific city or generate fake data",

"parameters": {

"type": "object",

"properties": {

"city": {

"type": "string",

"description": "The name of the city to get the weather for."

},

},

"required": ["city"],

"additionalProperties": False

},

}

}