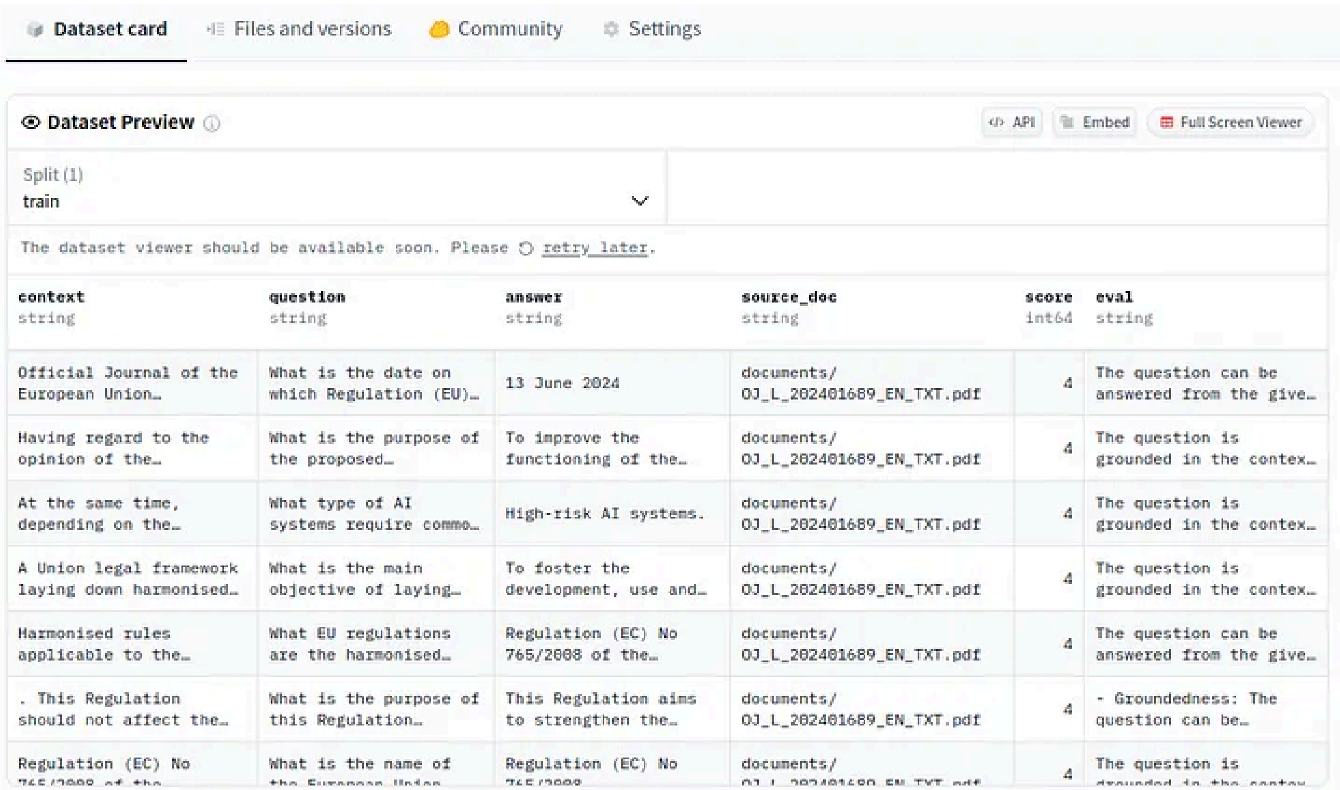
Member-only story

How to Create a RAG Evaluation Dataset From Documents

Automatically create domain-specific datasets in any language using LLMs

Dr. Leon Eversberg · Follow
Published in Towards Data Science
12 min read · 4 days ago

Listen Share More


Dataset card | Files and versions | Community | Settings

Dataset Preview | API | Embed | Full Screen Viewer

Split (1)
train

The dataset viewer should be available soon. Please retry later.

context string	question string	answer string	source_doc string	score int64	eval string
Official Journal of the European Union…	What is the date on which Regulation (EU)…	13 June 2024	documents/OJ_L_202401689_EN_TXT.pdf	4	The question can be answered from the give…
Having regard to the opinion of the…	What is the purpose of the proposed…	To improve the functioning of the…	documents/OJ_L_202401689_EN_TXT.pdf	4	The question is grounded in the contex…
At the same time, depending on the…	What type of AI systems require commo…	High-risk AI systems.	documents/OJ_L_202401689_EN_TXT.pdf	4	The question is grounded in the contex…
A Union legal framework laying down harmonised…	What is the main objective of laying…	To foster the development, use and…	documents/OJ_L_202401689_EN_TXT.pdf	4	The question is grounded in the contex…
Harmonised rules applicable to the…	What EU regulations are the harmonised…	Regulation (EC) No 765/2008 of the…	documents/OJ_L_202401689_EN_TXT.pdf	4	The question can be answered from the give…
. This Regulation should not affect the…	What is the purpose of this Regulation…	This Regulation aims to strengthen the…	documents/OJ_L_202401689_EN_TXT.pdf	4	- Groundedness: The question can be…
Regulation (EC) No	What is the name of	Regulation (EC) No	documents/	4	The question is

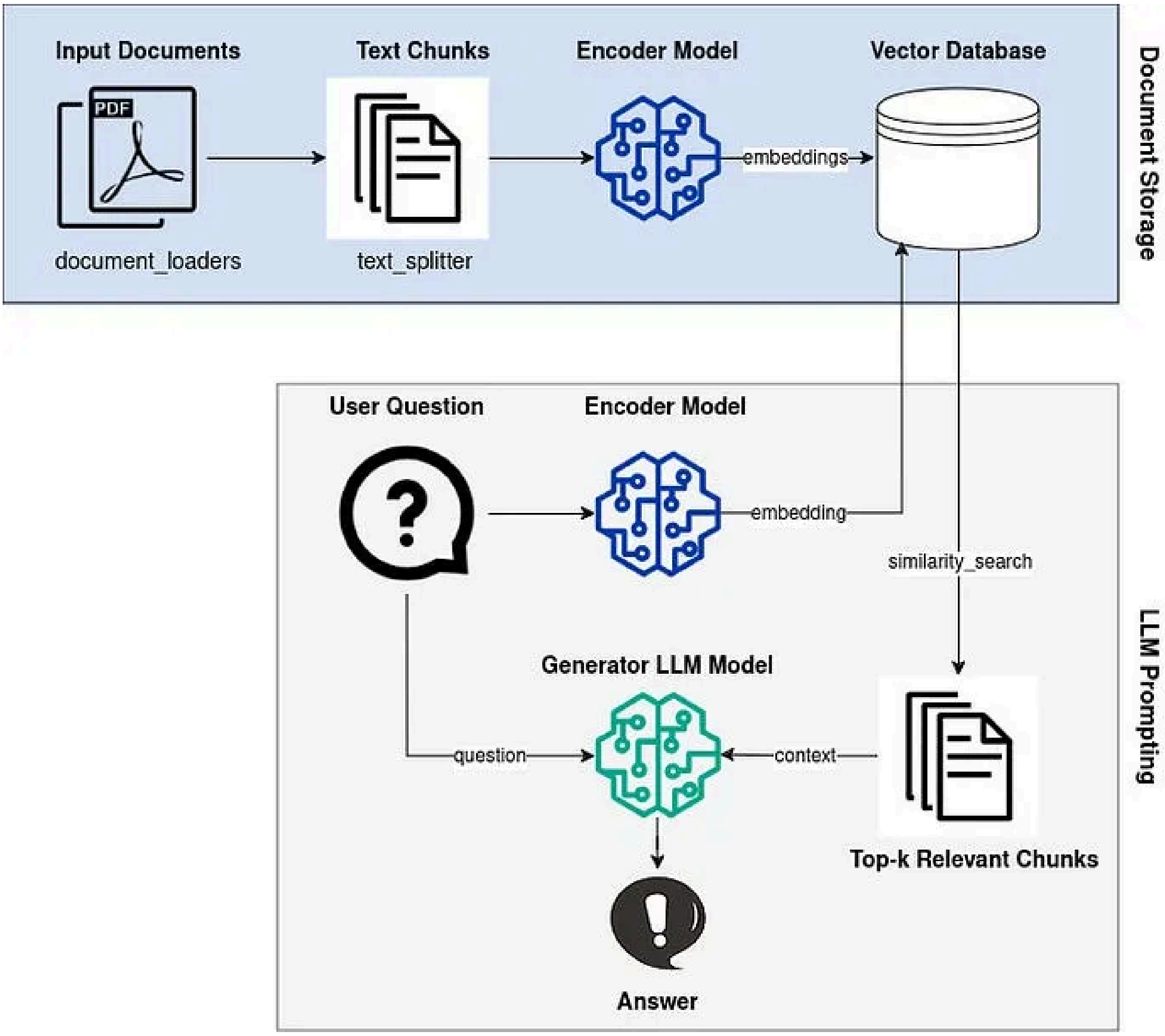
Our automatically generated RAG evaluation dataset on the Hugging Face Hub (PDF input file from the European Union licensed under CC BY 4.0). Image by the author

In this article I will show you how to create your own RAG dataset consisting of contexts, questions, and answers from documents in any language.

Retrieval-Augmented Generation (RAG) [1] is a technique that allows LLMs to access an external knowledge base.

By uploading PDF files and storing them in a vector database, we can retrieve this knowledge via a vector similarity search and then insert the retrieved text into the LLM prompt as additional context.

This provides the LLM with new knowledge and reduces the possibility of the LLM making up facts (hallucinations).



The basic RAG pipeline. Image by the author from the article "How to Build a Local Open-Source LLM Chatbot With RAG"

However, there are many parameters we need to set in a RAG pipeline, and researchers are always suggesting new improvements. How do we know which parameters to choose and which methods will really improve performance for our particular use case?

This is why we need a validation/dev/test dataset to evaluate our RAG pipeline. The dataset should be from the domain we are interested in and in the language we want to use.

Table Of Contents

· Deploying a Local LLM With VLLM
· Creating a RAG Evaluation Dataset
∘ Read Files
∘ Generating Question-Answer-Context Samples
∘ Filtering out Bad Question-Answer Pairs
∘ Saving The Dataset
∘ Creating a RAG Dataset in Another Language
· Conclusion
· References

· · ·

Deploying a Local LLM With VLLM

First, we get a local LLM up and running.

I used VLLM to set up an **OpenAI-compatible LLM server** with a quantized Llama-3.2–3B-Instruct. Make sure you use an LLM that has been trained on the language you want to use.

Deploying a local LLM with Docker and VLLM is quite simple:

With **Docker:**

```
docker run --runtime nvidia --gpus all \
    -v ~/.cache/huggingface:/root/.cache/huggingface \
    --env "HUGGING_FACE_HUB_TOKEN=<secret>" \
    -p 8000:8000 \
    --ipc=host \
    vllm/vllm-openai:latest \
    --model AMead10/Llama-3.2-3B-Instruct-AWQ \
    --quantization awq \
    --max-model-len 2048
```

With **Docker Compose:**

```
services:
  vllm:
    image: vllm/vllm-openai:latest
    command: ["--model", "AMead10/Llama-3.2-3B-Instruct-AWQ", "--max-model-len"
    ports:
      - 8000:8000
    volumes:
      - ~/.cache/huggingface:/root/.cache/huggingface
    environment:
      - "HUGGING_FACE_HUB_TOKEN=<secret>"
    deploy:
      resources:
        reservations:
          devices:
            - driver: nvidia
              count: 1
              capabilities: [gpu]
```

Now we can use our local LLM with the official OpenAI Python SDK.

If you want to use the official OpenAI models, just change the `base_url`, `api_key`, and `model` variables.

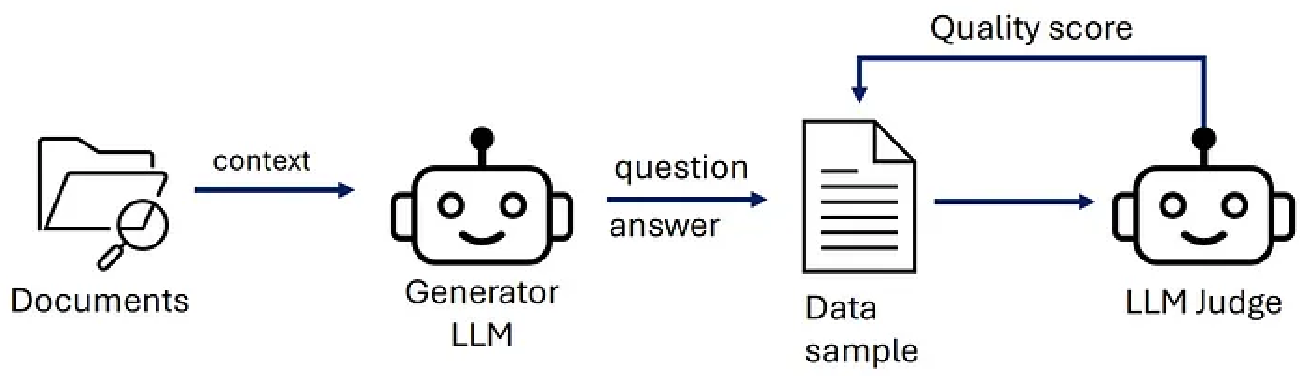
```
%pip install openai

from openai import OpenAI

# use local VLLM server
client = OpenAI(
    base_url="http://localhost:8000/v1",
    api_key="None",
)

chat_completion = client.chat.completions.create(
    messages=[
        {
            "role": "user",
            "content": "Say this is a test",
        }
    ],
```

```
    model="AMead10/Llama-3.2-3B-Instruct-AWQ",
)
```

Let’s perform a quick sanity check to see that everything works as expected:

```
print(chat_completion.choices[0].message.content)
>> "This appears to be a test. Is there anything specific you'd like to test or
```

Creating a RAG Evaluation Dataset



Workflow to automatically generate RAG evaluation data samples from documents. Image by the author

The basic workflow for automatically generating a RAG dataset starts with reading our knowledge base from documents, such as PDF files.

Then we ask a **generator LLM** to generate question-answer pairs from the given document context.

Finally, we use a **judge LLM** to perform quality control. The LLM will give each question-answer-context sample a score, which we can use to filter out bad samples.

Why not use a framework like Ragas to generate a synthetic test set for RAG? Because Ragas uses English LLM prompts under the hood. Using Ragas with non-English documents does not work at the moment.

I used the OpenAI cookbook “RAG Evaluation” [2] as the basis for my code in this article. However, I tried to simplify their sample code and changed the evaluation based on a few research findings [3, 4, 5].

Read Files

We will use LangChain to read a folder with all our files.

First, we need to install all the necessary packages. LangChain's DirectoryLoader uses the unstructured library to read all kinds of file types. In this article, I will only be reading PDFs so we can install a smaller version of `unstructured`.

```
pip install langchain==0.3.6 langchain-community==0.3.4 unstructured[pdf]==0.16
```

Now we can read our data folder to get the LangChain documents. The following code first loads all the PDF files from a folder and then chunks them into relatively large chunks of size 2000.

```
from langchain_text_splitters.character import RecursiveCharacterTextSplitter
from langchain_community.document_loaders.directory import DirectoryLoader

loader = DirectoryLoader("/path/to/data/folder", glob="**/*.pdf", show_progress
docs = loader.load()

text_splitter = RecursiveCharacterTextSplitter(
    chunk_size=2000,
    chunk_overlap=200,
    add_start_index=True,
    separators=["\n\n", "\n", ".", " ", ""],
)

docs_processed = []
for doc in docs:
    docs_processed.extend(text_splitter.split_documents([doc]))
```

The result is a list `docs_processed` with items of the type `Document`. Each document has some `metadata` and the actual `page_content`.

This list of documents is our knowledge base from which we will create question-answer pairs based on the context of the `page_content`.

Generating Question-Answer-Context Samples

Using the OpenAI client and the model we created earlier, we first write a generator function to create questions and answers from our documents.

```
def qa_generator_llm(context: str, client: OpenAI, model: str = "AMead10/Llama-
    generation_prompt = """
Your task is to write a factoid question and an answer given a context.
Your factoid question should be answerable with a specific, concise piece of fa
Your factoid question should be formulated in the same style as questions users
This means that your factoid question MUST NOT mention something like "accordir

Provide your answer as follows:

Output:::
Factoid question: (your factoid question)
Answer: (your answer to the factoid question)

Now here is the context.

Context: {context}\n
Output:::"""

    chat_completion = client.chat.completions.create(
        messages=[
            {
                "role": "system",
                "content": "You are a question-answer pair generator."
            },
            {
                "role": "user",
                "content": generation_prompt.format(context=context),
            }
        ],
        model=model,
        temperature=0.5,
        top_p=0.99,
        max_tokens=500
    )

    return chat_completion.choices[0].message.content
```

If you want to use a language other than English, you will need to translate the `generation_prompt` (and the system instruction).

Next, we simply loop through all of our document chunks in our knowledge base and generate a question and an answer for each chunk.

```
from tqdm.auto import tqdm

outputs = []
for doc in tqdm(docs_processed):
    # Generate QA couple
    output_QA = qa_generator_llm(doc.page_content, client)
    try:
        question = output_QA.split("Factoid question: ")[-1].split("Answer: ")[
        answer = output_QA.split("Answer: ")[-1]
        assert len(answer) < 500, "Answer is too long"
        outputs.append(
            {
                "context": doc.page_content,
                "question": question,
                "answer": answer,
                "source_doc": doc.metadata["source"],
            }
        )
    except Exception as e:
        print(e)
```

Depending on how many PDF files you have, this may take a while. Don't forget to translate the strings in `output_QA.split` if necessary.

To generate a RAG evaluation dataset, I used a PDF about the regulation of the EU AI Act from the European Union (licensed under CC BY 4.0). Here is my generated raw `outputs` dataset:

```
[{'context': 'Official Journal of the European Union\n\n2024/1689\n\nREGULATION
  'question': 'What is the date on which Regulation (EU) 2024/1689 of the Europ
  'answer': '13 June 2024',
  'source_doc': 'documents/OJ_L_202401689_EN_TXT.pdf'},
 {'context': 'Having regard to the opinion of the Committee of the Regions (3),
  'question': 'What is the purpose of the proposed Regulation on the developmen
  'answer': 'To improve the functioning of the internal market by laying down a
  'source_doc': 'documents/OJ_L_202401689_EN_TXT.pdf'},
 {'context': '(3)\n\nAI systems can be easily deployed in a large variety of se
  'question': 'What is the official journal number for the regulation related t
  'answer': '(4)',
  'source_doc': 'documents/OJ_L_202401689_EN_TXT.pdf'},
  ...
]
```

Filtering out Bad Question-Answer Pairs

Next, we use an **LLM as a judge** to automatically filter out bad samples.

When using an LLM as a judge to evaluate the quality of a sample, it is best practice to use a different model than the one that was used to generate it because of a **self-preference bias** [6] — you wouldn't grade your own paper, would you?

When it comes to judging our generated questions and answers, there are a lot of possible prompts we could use.

To build our prompt, there is a structure we can use from the **G-Eval** paper [3]:

- We start with the **task introduction**
- We present our **evaluation criteria**
- We want the model to perform **chain-of-thought (CoT)** reasoning [7] to improve its performance
- We ask for the **total score** at the end

For the evaluation criteria, we can use a list where each criterion adds one point if it is fulfilled [4].

The evaluation criteria should ensure that the question, the answer, and the context all fit together and make sense.

Here are two evaluation criteria from the OpenAI RAG evaluation cookbook [2]:

- **Groundedness:** can the question be answered from the given context?
- **Stand-alone:** is the question understandable without any context? (To avoid a question like `"What is the name of the function used in this guide?"`)

And two more evaluation criteria from the RAGAS paper [5]:

- **Faithfulness:** the answer should be grounded in the given context
- **Answer Relevance:** the answer should address the actual question posed

You can try to add more criteria or change the text for the ones that I used.

Here is the `judge_llm()` function, which critiques a question, answer, and context sample and produces a total rating score at the end:

```
def judge_llm(
    context: str,
    question: str,
    answer: str,
    client: OpenAI,
    model: str = "AMead10/Llama-3.2-3B-Instruct-AWQ",
):
    critique_prompt = """
You will be given a question, answer, and a context.
Your task is to provide a total rating using the additive point scoring system
Points start at 0 and are accumulated based on the satisfaction of each evaluat

Evaluation Criteria:
- Groundedness: Can the question be answered from the given context? Add 1 poir
- Stand-alone: Is the question understandable free of any context, for someone
- Faithfulness: The answer should be grounded in the given context. Add 1 point
- Answer Relevance: The generated answer should address the actual question tha

Provide your answer as follows:

Answer:::
Evaluation: (your rationale for the rating, as a text)
Total rating: (your rating, as a number between 0 and 4)

You MUST provide values for 'Evaluation:' and 'Total rating:' in your answer.

Now here are the question, answer, and context.

Question: {question}\n
Answer: {answer}\n
Context: {context}\n
Answer::: """

    chat_completion = client.chat.completions.create(
        messages=[
            {"role": "system", "content": "You are a neutral judge."},
            {
                "role": "user",
                "content": critique_prompt.format(
                    question=question, answer=answer, context=context
                ),
            },
        ],
        model=model,
        temperature=0.1,
        top_p=0.99,
        max_tokens=800
```

```
    )

    return chat_completion.choices[0].message.content
```

Now we loop through our generated dataset and critique each sample:

```
for output in tqdm(outputs):
    try:
        evaluation = judge_llm(
            context=output["context"],
            question=output["question"],
            answer=output["answer"],
            client=client,
        )
        score, eval = (
            int(evaluation.split("Total rating: ")[-1].strip()),
            evaluation.split("Total rating: ")[-2].split("Evaluation: ")[1],
        )
        output.update(
            {
                "score": score,
                "eval": eval
            }
        )
    except Exception as e:
        print(e)
```

Let’s filter out all the bad samples.

Since the generated dataset will be the ground truth for evaluation purposes, we should only allow very high-quality data samples. That’s why I decided to keep only samples with the highest possible score.

```
dataset = [doc for doc in outputs if doc["score"] >= 4]
```

And here is our final RAG evaluation dataset as a Pandas DataFrame:

```
import pandas as pd

pd.set_option("display.max_colwidth", 200)

df = pd.DataFrame(dataset)
display(df)
```

	context	question	answer	source_doc	score	eval
0	Official Journal of the European Union\n\n2024/1689\n\nREGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL\n\nof 13 June 2024\n\nlaying down harmonised rules on artificial inte...	What is the date on which Regulation (EU) 2024/1689 of the European Parliament and of the Council was laid down?\n	13 June 2024	documents/OJ_L_202401689_EN_TXT.pdf	4	The question can be answered from the given context as it directly mentions the date of the regulation. The question is also understandable without any context, as it is a factual question that ca...
1	Having regard to the opinion of the Committee of the Regions (3),\n\nActing in accordance with the ordinary legislative procedure (4),\n\nWhereas:\n\n(1)\n\nThe purpose of this Regulation is to im...	What is the purpose of the proposed Regulation on the development, placing on the market, putting into service, and use of artificial intelligence systems in the Union?\n	To improve the functioning of the internal market by laying down a uniform legal framework for the development, placing on the market, putting into service, and use of artificial intelligence syst...	documents/OJ_L_202401689_EN_TXT.pdf	4	\nThe question is grounded in the context as it can be answered from the given text. The question is also stand-alone and understandable without any context, as it is a clear and concise inquiry. ...
2	At the same time, depending on the circumstances regarding its specific application, use, and level of technological development, AI may generate risks and cause harm to public interests and funda...	What type of AI systems require common rules for protection of public interests as regards health, safety and fundamental rights?\n	High-risk AI systems.	documents/OJ_L_202401689_EN_TXT.pdf	4	\nThe question is grounded in the context as it can be answered from the provided information. However, the question's stand-alone quality is limited due to its dependence on the context. The answ...
3	A Union legal framework laying down harmonised rules on AI is therefore needed to foster the development, use and uptake of AI in the internal market that at the same time meets a high level of pr...	What is the main objective of laying down rules regulating the placing on the market, the putting into service and the use of AI systems in the European Union?\n	To foster the development, use and uptake of AI in the internal market while meeting a high level of protection of public interests.	documents/OJ_L_202401689_EN_TXT.pdf	4	\nThe question is grounded in the context as it directly refers to the main objective of laying down rules regulating the placing on the market, the putting into service and the use of AI systems ...

The filtered RAG evaluation dataset in English (PDF file licensed under CC BY 4.0). Image by the author

Saving The Dataset

We can convert our Pandas DataFrame into a Hugging Face dataset. Then, we can save it to disk and load it later when needed.

```
%pip install datasets==3.0.2

# save
from datasets import Dataset
dataset = Dataset.from_pandas(df, split="test")
dataset.save_to_disk("path/to/dataset/directory")

# load
from datasets import load_dataset
dataset = load_dataset("path/to/dataset/directory")
```

We can also upload the dataset to the Hugging Face Hub.

Creating a RAG Dataset in Another Language

I don't speak Spanish. However, I downloaded a Spanish legal document from the European Union law (licensed under CC BY 4.0) and converted my prompts using

DeepL Translate. I have no idea what the document says, but let's see if we can generate a new dataset.

This is what the filtered RAG dataset looks like after replacing the input document and translating the prompts from English to Spanish:

	context	question	answer	source_doc	score	eval
0	(3)\n\nLa Recomendación del Parlamento Europeo y del Consejo, de 23 de abril de 2008, relativa a la creación del Marco Europeo de Cualificaciones para el aprendizaje permanente (1) estableció un m...	¿Cuál es el título oficial de la Recomendación del Parlamento Europeo y del Consejo de 23 de abril de 2008?\n	Recomendación del Parlamento Europeo y del Consejo de 23 de abril de 2008.	documents/Legal text-ES.pdf	4	\n- Fundamentación: 1 punto, ya que la pregunta puede responderse a partir del contexto dado.\n- Independiente: 1 punto, ya que la pregunta es comprensible sin ningún contexto, para alguien con co...
1	(13)\n\nLos marcos y los sistemas nacionales de cualificaciones cambian a lo largo del tiempo, por lo que su correlación con el MEC debe revisarse y actualizarse cuando proceda.\n\n(14)\n\nLa conf...	¿Cuál fue el año en que la Recomendación del Parlamento Europeo y del Consejo relativa a la creación del Marco Europeo de Cualificaciones para el aprendizaje permanente se adoptó?\n	2008.	documents/Legal text-ES.pdf	4	\nLa pregunta puede responderse a partir del contexto dado, ya que la respuesta proporcionada (2008) se menciona explícitamente en el contexto. Además, la pregunta es independiente y puede respond...
2	(18) Aunque el acervo de la Unión en materia de migración legal y asilo establece la igualdad de trato con los nacio nales en términos de reconocimiento de las cualificaciones e incluso de medida...	¿Cuántos países han establecido vínculos más estrechos entre sus marcos de cualificaciones y el MEC?\n	Un número creciente de terceros países y regiones buscan vínculos más estrechos entre sus marcos de cualificaciones y el MEC.	documents/Legal text-ES.pdf	4	\n- Fundamentación: 1 punto, ya que la pregunta puede responderse a partir del contexto dado, ya que el contexto proporciona información sobre el MEC y su enfoque en los resultados del aprendizaje...
3	(25) Debe haber coherencia, complementariedad y sinergias a escala nacional y de la Unión entre la aplicación del MEC, los marcos o los sistemas nacionales de cualificaciones y las herramientas so...	¿Cuál es el título del documento de 31 de diciembre de 2004, por el cual se establece un marco comunitario único para la transparencia de las cualificaciones y competencias?\n\n	Decisión n.o 2241/2004/CE del Parlamento Europeo y del Consejo, de 15 de diciembre de 2004, relativa a un marco comunitario único para la transparencia de las cualificaciones y competencias (Europ...	documents/Legal text-ES.pdf	4	\n- Fundamentación: 1 punto, ya que la pregunta puede responderse a partir del contexto dado.\n- Independiente: 1 punto, ya que la pregunta es comprensible sin ningún contexto, para alguien con co...
4	Por el Consejo\n\nEl Presidente\n\nE. BARTOLO\n\nC 189/19\n\nC 189/20\n\nES\n\nDiario Oficial de la Unión Europea\n\nANEXO I\n\nDefiniciones\n\nA efectos de la presente Recomendación, se entenderá...	¿Qué es una "cualificación internacional" según la Recomendación del Consejo?\n	Una cualificación internacional es una cualificación otorgada por un organismo internacional establecido legalmente o por un organismo nacional que actúe en nombre de un organismo internacional.	documents/Legal text-ES.pdf	4	\nLa pregunta es independiente, ya que puede responderse por sí sola sin necesidad de contexto. La respuesta también se basa en el contexto dado, ya que se menciona explícitamente en el contexto q...
5	e)\n\n«resultados del aprendizaje»: declaraciones respecto de lo que una persona sabe, comprende y es capaz de hacer al culminar un proceso de aprendizaje; se define en términos de conocimientos, ...	¿Cuál es la definición de "resultados del aprendizaje" según el contexto del MEC?\n	Resultados del aprendizaje se definen como declaraciones respecto de lo que una persona sabe, comprende y es capaz de hacer al culminar un proceso de aprendizaje.	documents/Legal text-ES.pdf	4	\n- Fundamentación: 1 punto, ya que la pregunta puede responderse a partir del contexto dado.\n- Independiente: 1 punto, ya que la pregunta es comprensible sin ningún contexto, para alguien con co...

A domain-specific Spanish RAG dataset that was automatically created from a Spanish legal document from the European Union (licensed under CC BY 4.0).

By using our own dataset generation code, we can adapt it to any language and domain we want.

Conclusion

Automatically creating a RAG evaluation dataset from a collection of documents is easy. All we needed was a prompt for the LLM generator, a prompt for the LLM judge, and a little Python code in between.

To change the domain of our RAG evaluation dataset, we simply exchange the documents that we feed to the `DirectoryLoader`. The documents do not have to be PDF files, they can be CSV files, markdown files, etc.

To change the language of our RAG evaluation dataset, we simply translate the LLM prompts from English to another language.

If the generated data samples are not good enough for your use case, you can try to modify the prompts. Also, using bigger and better LLMs will increase the quality of the dataset.

• • •

References

[1] P. Lewis et al. (2021), Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, arXiv:2005.11401

[2] A. Roucher (2024), RAG Evaluation, Hugging Face AI Cookbook, accessed on 11–01–2024

[3] Y. Liu et al. (2023), G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment, arXiv:2303.16634

[4] W. Yuan et al (2024), Self-Rewarding Language Models, arXiv:2401.10020

[5] S. Es et al. (2023), RAGAS: Automated Evaluation of Retrieval Augmented Generation, arXiv:2309.15217

[6] K. Wataoka, T. Takahashi, and R. Ri (2024), Self-Preference Bias in LLM-as-a-Judge, arXiv:2410.21819

[7] J. Wei et al. (2022), Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, arXiv:2201.11903

Llm Rag Programming Data Science Hands On Tutorials

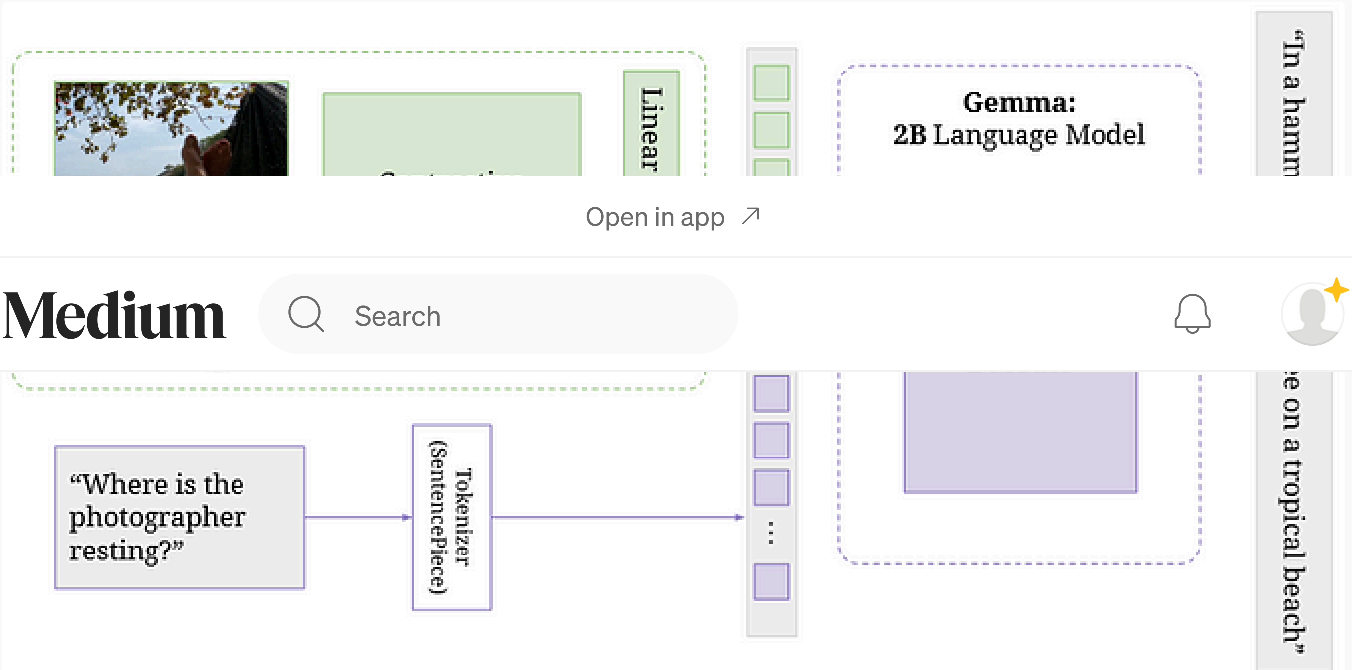
Follow

Written by Dr. Leon Eversberg

3.5K Followers · Writer for Towards Data Science

Machine Learning PhD | AI Software Engineer | Research & Development Specialist | Data Scientist | LLM Enthusiast

More from Dr. Leon Eversberg and Towards Data Science



Dr. Leon Eversberg in Towards Data Science

Revisiting Karpathy's "State of Computer Vision and AI"

Looking back at AI progress since the 2012 blog post "The state of Computer Vision and AI: we are really, really far away"

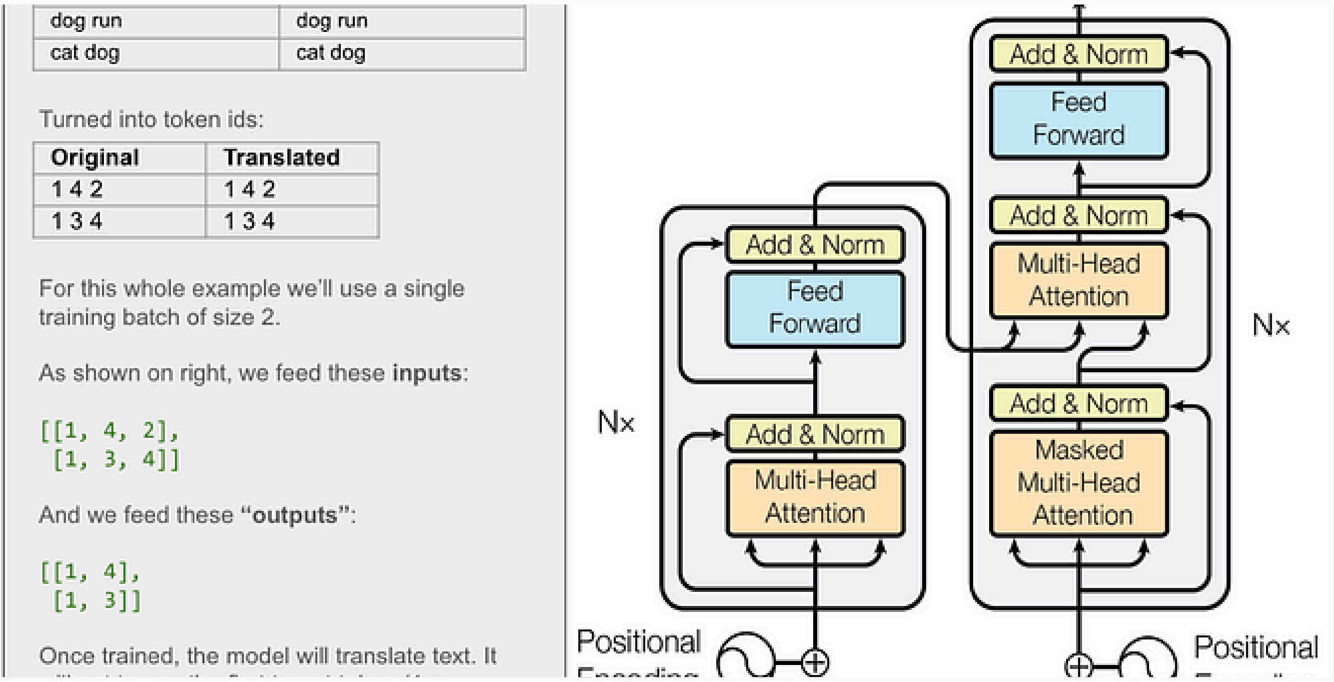
Oct 18 · 582 · 10

Meghan Heintz in Towards Data Science

Watermarking for AI Text and Synthetic Proteins: Fighting Misinformation and Bioterrorism

Understanding AI applications in bio for machine learning engineers

16h ago 54 1

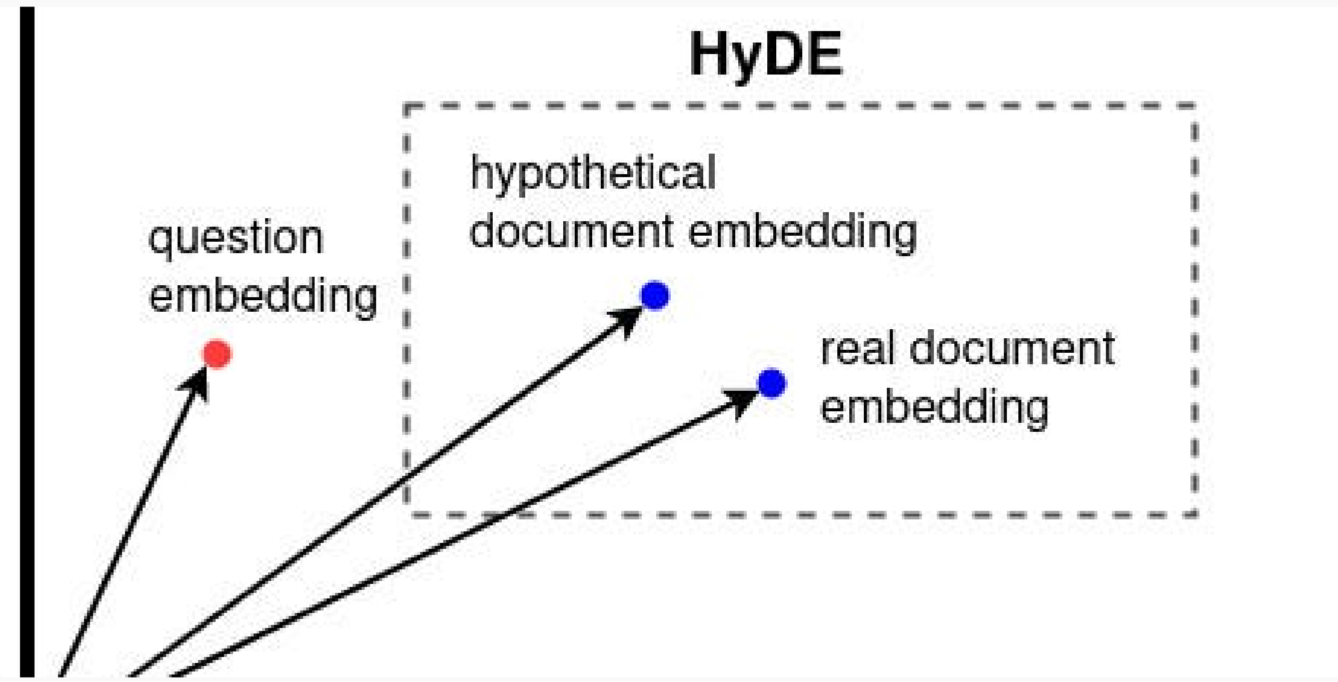


Eric Silberstein in Towards Data Science

Tracing the Transformer in Diagrams

What exactly do you put in, what exactly do you get out, and how do you generate text with it?

1d ago 176 1



Dr. Leon Eversberg in Towards Data Science

How to Use HyDE for Better LLM RAG Retrieval

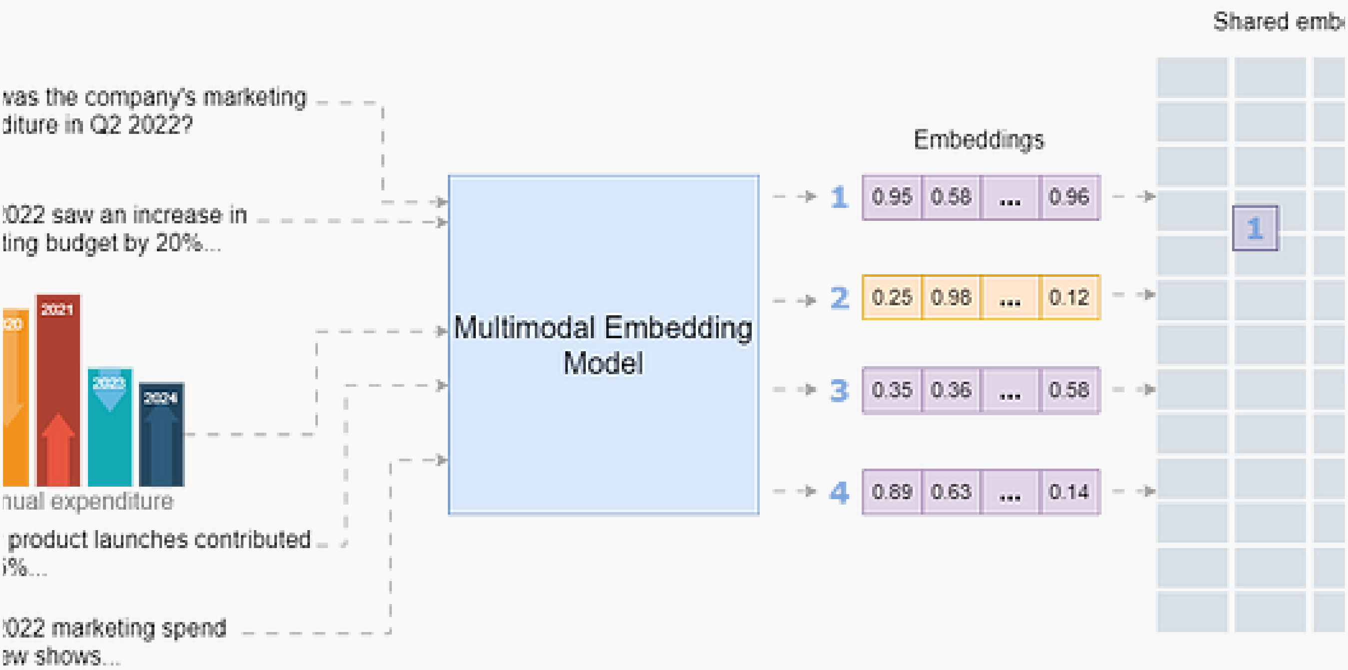
Building an advanced local LLM RAG pipeline with hypothetical document embeddings

Oct 4 535 1

See all from Dr. Leon Eversberg

See all from Towards Data Science

Recommended from Medium



Umair Ali Khan in Towards Data Science

Multimodal AI Search for Business Applications

Enabling businesses to extract real value from their data

6d ago 449 5

Shaw Talebi in The Data Entrepreneurs

I Built an AI App in 4 Days—here's how I did it.

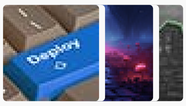
(as a data scientist with no web dev experience)

Oct 30 762 25

Lists

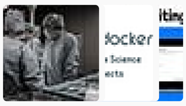
General Coding Knowledge

20 stories · 1704 saves



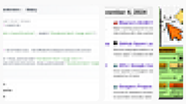
Predictive Modeling w/ Python

20 stories · 1641 saves



Coding & Development

11 stories · 890 saves



Natural Language Processing

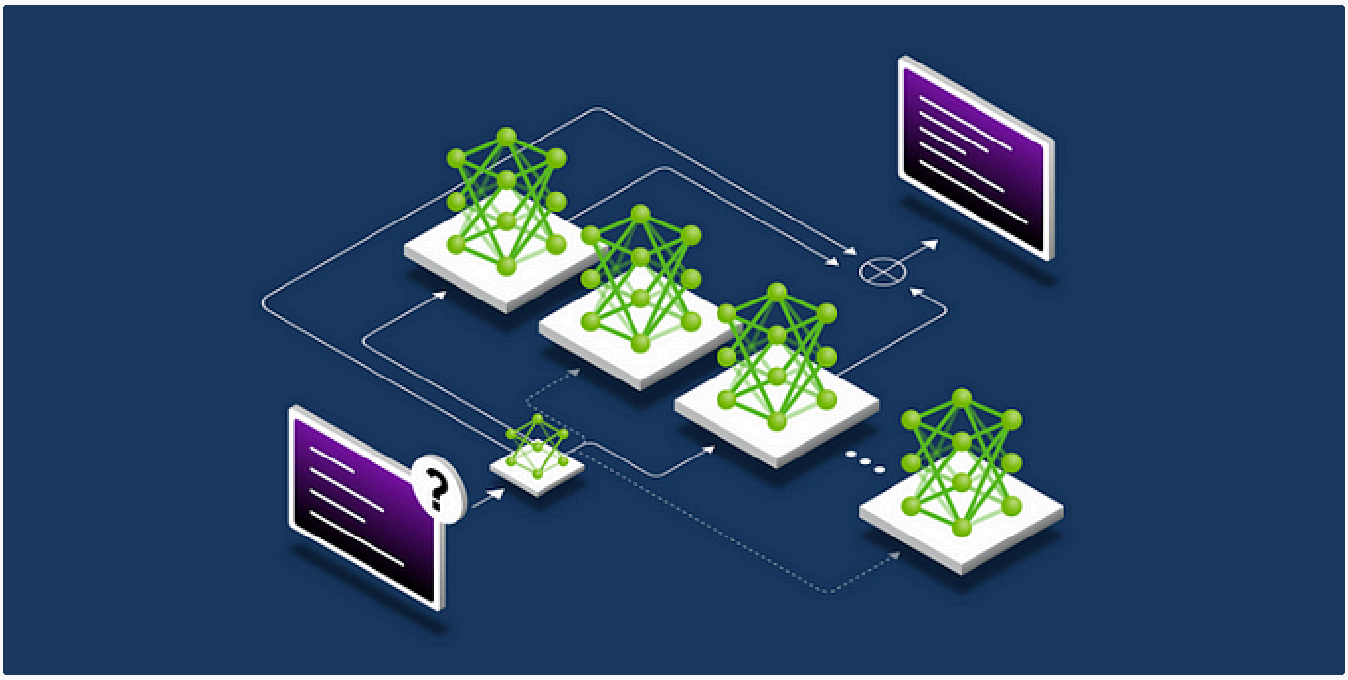
1798 stories · 1408 saves

Harendra

How I Am Using a Lifetime 100% Free Server

Get a server with 24 GB RAM + 4 CPU + 200 GB Storage + Always Free

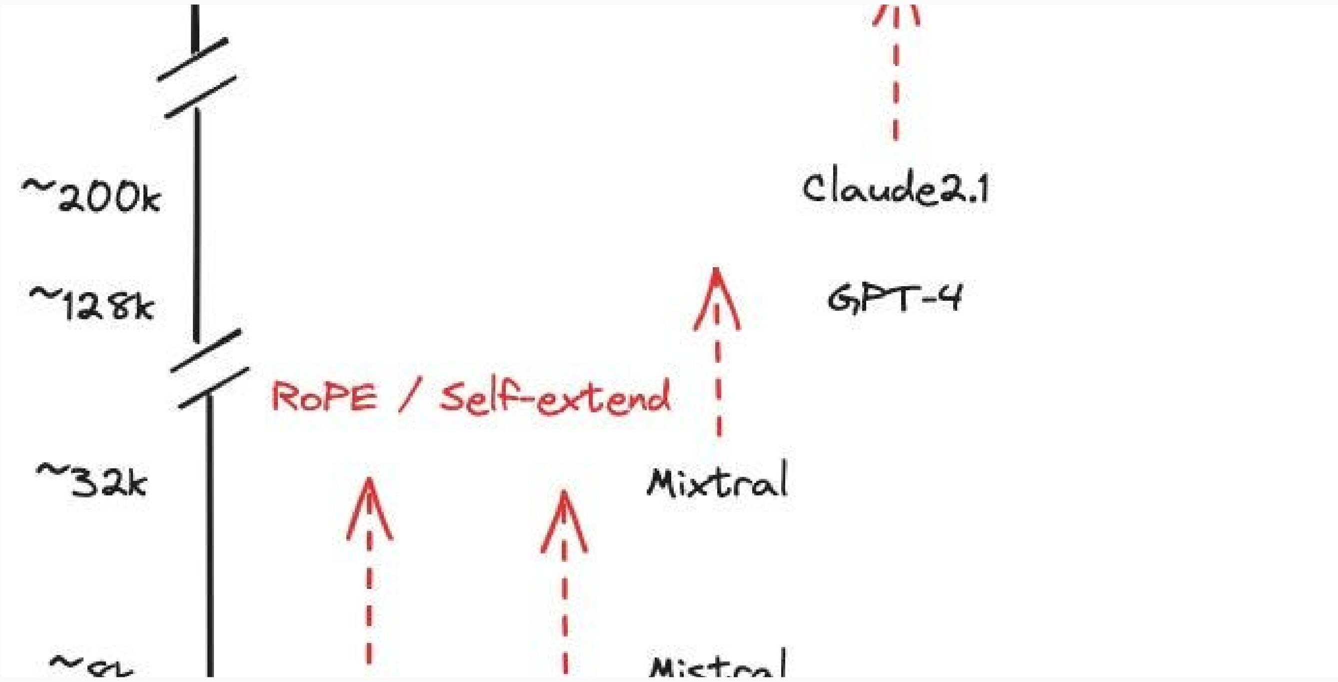
Oct 26 2.8K 34



Yuki Shizuya in Generative AI

Unlocking Mixture-of-Experts (MoE) LLM : Your MoE model can be embedding model for free

Mixture-of-experts (MoE) LLM can be used as an embedding model for free.



Barhoumi Mosbeh in Towards AI

RAG From Scratch

I'm working as a machine learning engineer, and I frequently use Claude or ChatGPT to help me write code. However, in some cases, the model...

Eva Jurado Cortés in Data Science at Microsoft

Creating specialized AI models: Fine-tuning GPT-4o mini for a medical assistance model

When embarking on the journey of creating a chat solution powered by generative AI, efficiency should be our guiding star. From leveraging…

Oct 29 49 2

See more recommendations