Tanimawo AI Academy

ADVANCED

Give AI Your Own Documents (RAG)

Retrieval-augmented generation: answers grounded in your data

Free guide

About 18 minutes to read

tanimawoai.com

Free AI training every Sunday
with Dr. Daniel Tanimawo

TANIMAWO ACADEMY
tanimawoai.com · Dr. Daniel Tanimawo

© 2026 Dr. Daniel Tanimawo · Tanimawo AI Academy · tanimawoai.com · Free to download and share with credit. Selling, rebranding or removing this notice is not allowed.

The problem RAG solves

A model only knows its training data. It does not know your company handbook, your class notes or last week's price list. **Retrieval-augmented generation (RAG)** fixes this: before answering, your app **finds the most relevant passages** from your documents and gives them to the model as context.

How RAG works

1. **Chunk:** split documents into small passages (for example 300 to 800 words).
2. **Embed:** turn each chunk into an **embedding**, a list of numbers that captures its meaning.
3. **Store:** save chunks and embeddings (a file, a database, or a vector database).
4. **Retrieve:** embed the user's question and find the chunks with the closest meaning.
5. **Generate:** send the question plus those chunks to the model, with instructions to answer only from them.

> **No-code option:** NotebookLM does this for you. Upload documents and ask questions; answers cite your files.

Minimal example in Python

```
pip install openai numpy
```

TANIMAWO AI ACADEMY
tanimawoai.com · Dr. Daniel Tanimawo

© 2026 Dr. Daniel Tanimawo · Tanimawo AI Academy · tanimawoai.com · Free to download and share with credit. Selling, rebranding or removing this notice is not allowed.

```
import numpy as np
from openai import OpenAI

client = OpenAI()

chunks = [
    "Classes hold every Sunday at 8 PM Nigeria time on Microsoft Teams.",
    "Replays are posted on the Replays page after each class.",
    "Registration is free and only needs a name, email and country.",
]

def embed(texts):
    r = client.embeddings.create(model="text-embedding-3-small", input=texts)
    return np.array([d.embedding for d in r.data])

chunk_vectors = embed(chunks)

def answer(question, k=2):
    q = embed([question])[0]
    scores = chunk_vectors @ q / (
        np.linalg.norm(chunk_vectors, axis=1) * np.linalg.norm(q))
    best = [chunks[i] for i in np.argsort(scores)[::-1][:k]]
    context = "\n".join(best)
    r = client.chat.completions.create(
        model="gpt-4.1-mini",
        max_tokens=200,
        messages=[
            {"role": "system", "content":
             "Answer only from the context. If the answer is not there, "
             "say you don't know.\n\nContext:\n" + context},
            {"role": "user", "content": question},
        ],
    )
    return r.choices[0].message.content

print(answer("When is the class?"))
```

Design choices that matter

Choice	Guidance
Chunk size	Small enough to be specific, big enough to keep meaning; add a little overlap
Number of chunks (k)	Start with 3 to 5; more context costs more
Storage	A list in memory for tests; a vector database for thousands of documents
Updates	Re-embed documents when they change
Citations	Return which chunk each answer used so users can check

© 2026 Dr. Daniel Tanimawo · Tanimawo AI Academy · tanimawoai.com · Free to download and share with credit. Selling, rebranding or removing this notice is not allowed.

Common problems

- **Wrong passages retrieved:** improve chunking, add titles to chunks, or combine keyword and meaning search.
- **Model ignores the context:** strengthen the instruction and put the context clearly labelled.
- **Private data:** only index documents users are allowed to see, and filter by user permissions at retrieval time.

Checklist

- [ ] Documents are cleaned (no headers or footers repeated on every page)
- [ ] Embeddings are stored, not recomputed on every question
- [ ] The prompt says “answer only from the context”
- [ ] Answers show their sources

Keep learning, free

Join the free live AI class every Sunday at 8 PM Nigeria time with Dr. Daniel Tanimawo. Register at **tanimawoai.com**, watch past classes on the Replays page, and download more free guides.

Share this guide with a friend who wants to learn AI.

TANIMAWO AI ACADEMY
tanimawoai.com · Dr. Daniel Tanimawo

© 2026 Dr. Daniel Tanimawo · Tanimawo AI Academy · tanimawoai.com · Free to download and share with credit. Selling, rebranding or removing this notice is not allowed.