Python Jogja
Lightning Talk

Jogja Cozy Space
26 July 2025

Introduction to **Retrieval Augmented Generation**

With **Langchain** & **PGVector**

Joko Eliyanto
Data Scientist at Teman Data

Python Jogja
Lightning Talk

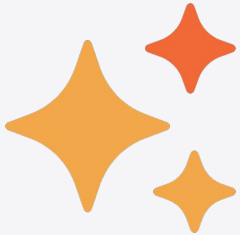
Jogja Cozy Space
26 July 2025

Outline

1. LLM Knowledge is limited
2. What is langchain?
3. What is RAG?
4. Vector Database & Vector Embedding
5. PG Vector for RAG
6. Demo session

LLM Knowledge Is Limited

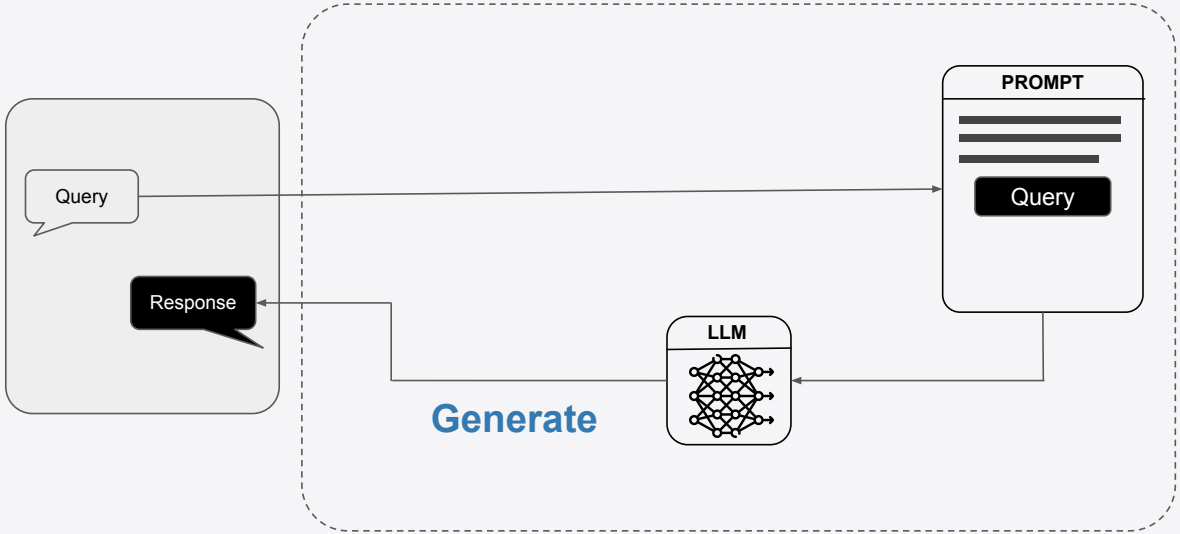
LLM models need to be updated and improved continuously.



We often require specific domain knowledge that general-purpose LLMs cannot access.

Fine-tuning LLMs introduces serious data privacy challenges.

How we get response using LLM?

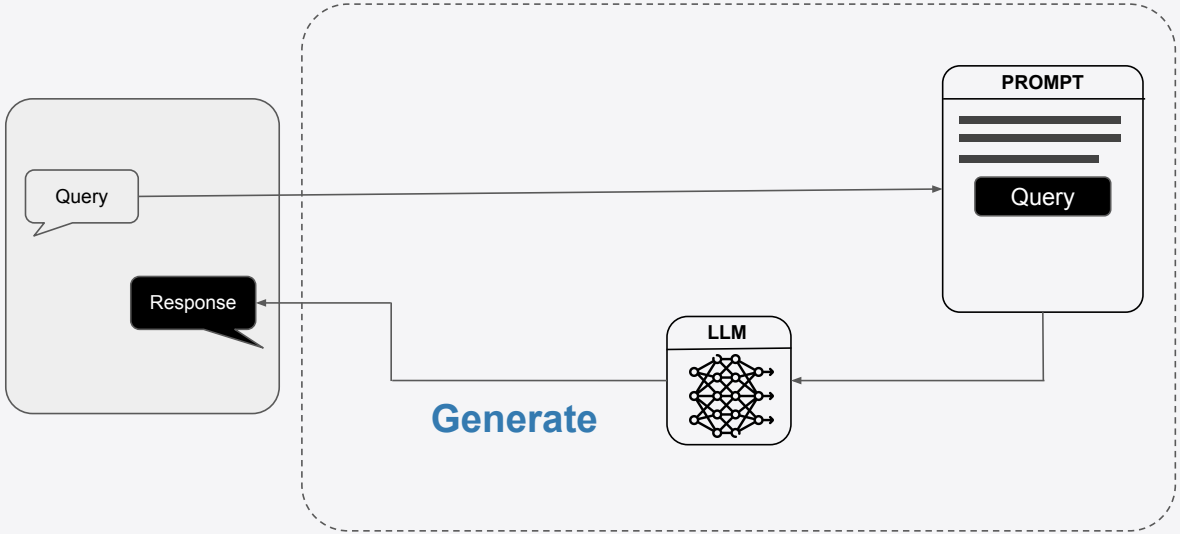


Joko Eliyanto
Data Scientist at Teman Data

Jogja Cozy Space
26 July 2025

What is LangChain?

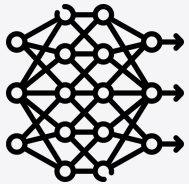
LangChain is a framework for developing applications powered by large language models (LLMs).



Joko Eliyanto
Data Scientist at Teman Data

Jogja Cozy Space
26 July 2025

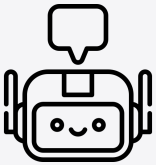
LangChain Main* Component



LLM

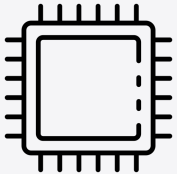
Prompt Template

Chain



Agent

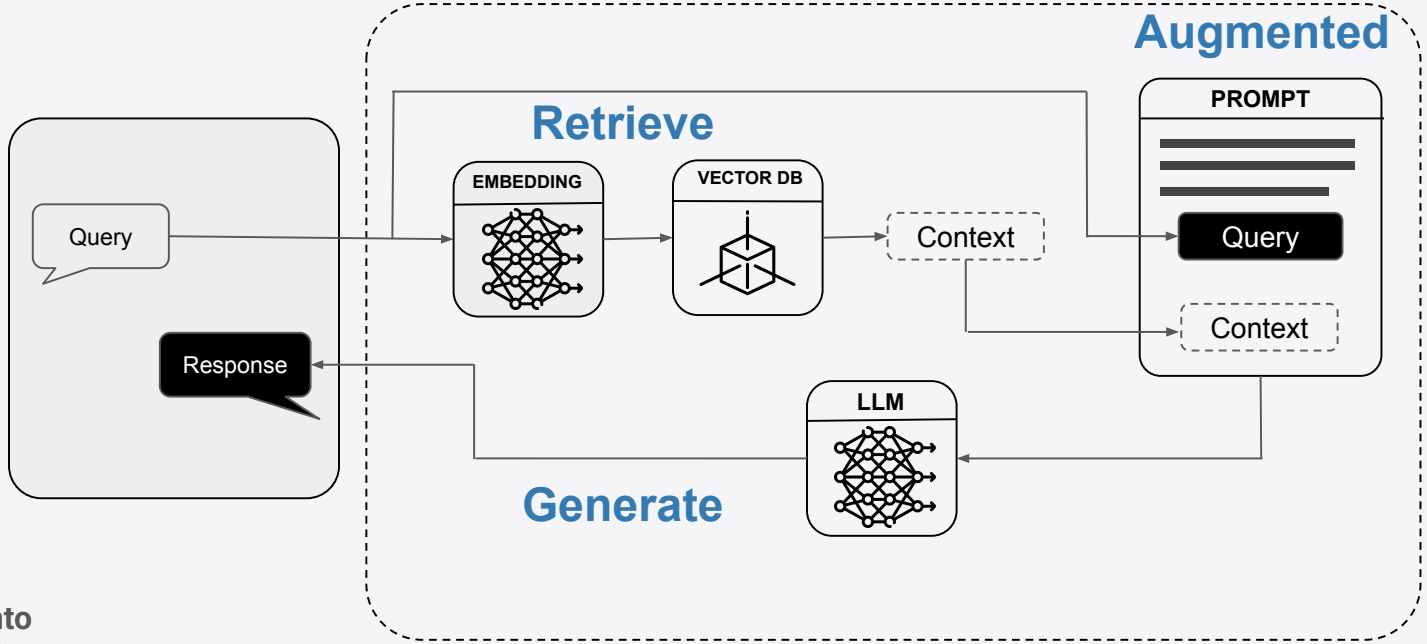
Tools



Memory

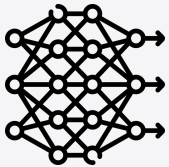
Joko Eliyanto
Data Scientist at Teman Data

Retrieval Augmented Generation



Joko Eliyanto
Data Scientist at Teman Data

RetrievalQA Chain



LLM

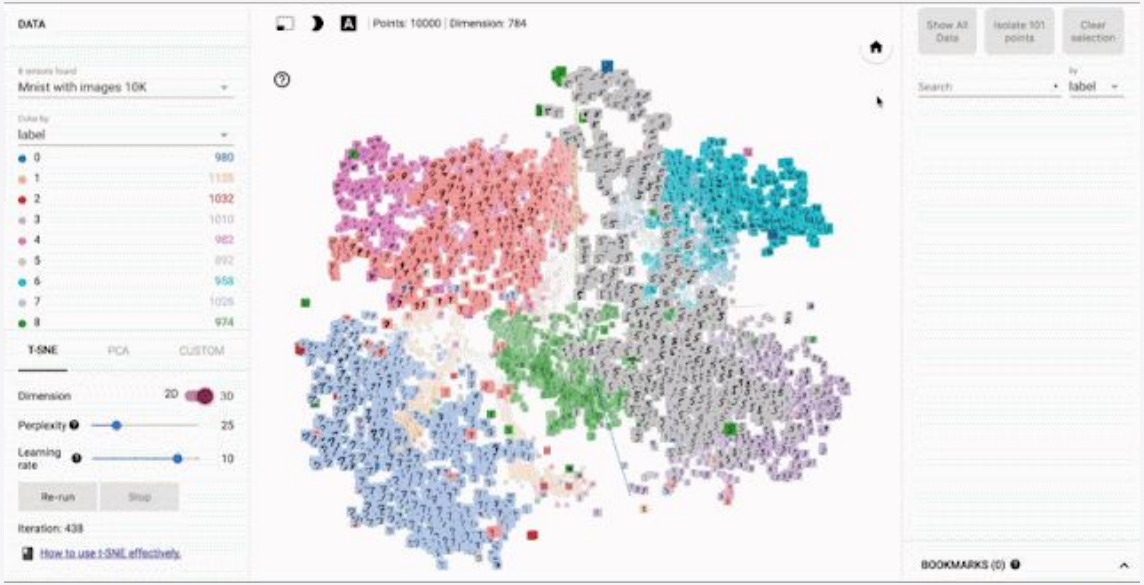
Vector Store

Retriever

```python
vectorstore = PGVector.from_documents(
    documents=docs,
    embedding=embedding,
    collection_name="pdf_test",
    connection_string="postgresql+psycopg2://postgres:1234567890@172.24.194.113:5432/postgres"
)

llm = ChatOpenAI(
    openai_api_base="http://localhost:1234/v1",
    openai_api_key="lm-studio",  # bisa dummy
    model="nama-model-dari-lm-studio"
)

retriever = vectorstore.as_retriever(
    search_type="similarity",
    search_kwargs={"k": 3}
)

qa_chain = RetrievalQA.from_chain_type(
    llm=llm,
    retriever=retriever,
    return_source_documents=True
)
```

Joko Eliyanto
Data Scientist at Teman Data

What is vector embedding & vector DB?



pgvector

PGVector X LangChain for RAG

langchain_pg_collection | Enter a SQL expression to filter results (use Ctrl+Space)

	name	cmetadata	uuid
1	local_test	{}	42e23638-02bc-4a17-a658-432aac71b614
2	staging_test_locally	{}	ae6e12b9-66f2-433d-aefd-e165b08c6188
3	staging_test_locally_test	{}	c7594301-8828-4245-b100-49b3a9a54983
4	load_test_kb	{}	bf1a1c3c-6993-4dfe-b66a-127c843e6674
5	cookbook_images	{}	0285c223-4ec0-4704-83a1-deeceb3bda4b
6	laporan-perekonomian-indonesia-2024	{}	043b8a9c-af4d-4266-a0ef-6e70c22993cb
7	pdf_cohere_v3	null	c2272811-c52e-4c99-a8b6-6135288b1ece
8	csv	{}	abb2d6dc-7c03-43dc-a22c-bade52543ec1
9	csv_pdf	{}	a8de42fe-89f1-42c9-999d-4a8e681fee19
10	multiple_csv	{}	589ccbc6-a68f-43b6-a49b-640e7da9669b
11	csv-pdf	{}	cfa3b39d-c6db-4b08-93c2-d45aee00a27d

```
retriever = vector store.as_retriever(search_type="mmr",
                                      search_kwargs={"k": 1})

retriever.invoke("kitty")
```

```
[Document(metadata={'id': 1, 'topic': 'animals', 'location':
'pond'}, page_content='there are cats in the pond')]
```

Properties | Data | Diagram

langchain_pg_embedding | Enter a SQL expression to filter results (use Ctrl+Space)

	collection_id	embedding	document	cmetadata	custom_id
1	42e23638-02bc-4a17-a658-432aac71b6	[-0.008794164,-0.011105486,-0.022489445,-0.0037259632,0.0111500425,-(	¶The image is the cover of a book titled "Understanding Machine Learnin	{"source": "arn:aws:s3:::sample-bucket/2 mb - 1.pdf", "chunk": 0, "base64":	2 mb - 1.pdfchunk0
2	42e23638-02bc-4a17-a658-432aac71b614	[0.01670586,-0.033852313,0.03138009,0.00834681,-0.008432481,0.009099	UNDERSTANDING¶MACHINE LEARNING¶From Theory to¶Algorithms¶Sha	{"source": "arn:aws:s3:::sample-bucket/2 mb - 1.pdf", "chunk": 1, "base64":	2 mb - 1.pdfchunk1
3	42e23638-02bc-4a17-a658-432aac71b614	[0.008182924,-0.011308913,0.018240184,0.017586052,-0.012906502,-0.025	32 Avenue of the Americas, New York, NY 10013-2473, USA¶Cambridge U	{"source": "arn:aws:s3:::sample-bucket/2 mb - 1.pdf", "chunk": 2, "base64":	2 mb - 1.pdfchunk2
4	42e23638-02bc-4a17-a658-432aac71b614	[0.0077830125,0.0052473233,0.0455941,0.027530337,0.003926005,0.01170	10.5 Summary¶141¶Figure 10.2 The first and second features selected by A	{"source": "arn:aws:s3:::sample-bucket/2 mb - 1.pdf", "chunk": 3, "base64":	2 mb - 1.pdfchunk3
5	42e23638-02bc-4a17-a658-432aac71b614	[0.014359891,-0.009379176,0.0216626,0.03490709,0.0027955137,0.003181	10.5 Summary¶141¶Figure 10.2 The first and second features selected by A	{"source": "arn:aws:s3:::sample-bucket/2 mb - 1.pdf", "chunk": 4, "base64":	2 mb - 1.pdfchunk4
6	42e23638-02bc-4a17-a658-432aac71b614	[0.031306356,0.003353333,0.045337576,0.043664128,-0.022527188,0.0131	328¶Dimensionality Reduction¶x¶x¶x¶x x¶x¶x¶o¶o¶o¶o o¶o¶o¶*¶*¶*¶***¶*¶	{"source": "arn:aws:s3:::sample-bucket/2 mb - 1.pdf", "chunk": 10, "base64":	2 mb - 1.pdfchunk10
7	42e23638-02bc-4a17-a658-432aac71b614	[0.011526412,-0.014393889,0.055287227,0.036500305,0.012119683,0.0236	10.5 Summary¶141¶Figure 10.2 The first and second features selected by A	{"source": "arn:aws:s3:::sample-bucket/2 mb - 1.pdf", "chunk": 5, "base64":	2 mb - 1.pdfchunk5
8	42e23638-02bc-4a17-a658-432aac71b614	[0.036251042,0.0025084978,0.044787977,0.044664253,-0.013819936,0.008	328¶Dimensionality Reduction¶x¶x¶x¶x x¶x¶x¶o¶o¶o¶o o¶o¶o¶*¶*¶*¶***¶*¶	{"source": "arn:aws:s3:::sample-bucket/2 mb - 1.pdf", "chunk": 6, "base64":	2 mb - 1.pdfchunk6
9	42e23638-02bc-4a17-a658-432aac71b614	[0.0302293,0.0026771901,0.048489265,0.046726912,-0.015702102,0.00880	328¶Dimensionality Reduction¶x¶x¶x¶x x¶x¶x¶o¶o¶o¶o o¶o¶o¶*¶*¶*¶***¶*¶	{"source": "arn:aws:s3:::sample-bucket/2 mb - 1.pdf", "chunk": 7, "base64":	2 mb - 1.pdfchunk7
10	42e23638-02bc-4a17-a658-432aac71b614	[0.0354065,0.0020113033,0.041605785,0.043672215,-0.022629919,0.01350	328¶Dimensionality Reduction¶x¶x¶x¶x x¶x¶x¶o¶o¶o¶o o¶o¶o¶*¶*¶*¶***¶*¶	{"source": "arn:aws:s3:::sample-bucket/2 mb - 1.pdf", "chunk": 8, "base64":	2 mb - 1.pdfchunk8
11	42e23638-02bc-4a17-a658-432aac71b614	[0.028565522,0.012746571,0.042784806,0.04509544,-0.02173519,0.016644	328¶Dimensionality Reduction¶x¶x¶x¶x x¶x¶x¶o¶o¶o¶o o¶o¶o¶*¶*¶*¶***¶*¶	{"source": "arn:aws:s3:::sample-bucket/2 mb - 1.pdf", "chunk": 9, "base64":	2 mb - 1.pdfchunk9
12	42e23638-02bc-4a17-a658-432aac71b614	[0.050366234,-0.04871488,0.009991969,-0.016887654,0.012133566,0.0225	Technical Lemmas¶211¶Using Stirling's approximation we further have tha	{"source": "arn:aws:s3:::sample-bucket/2 mb - 1.pdf", "chunk": 479, "base64	2 mb - 1.pdfchunk479
13	42e23638-02bc-4a17-a658-432aac71b614	[0.03775887,0.00018039312,0.048406214,0.042943444,-0.009844367,0.005	328¶Dimensionality Reduction¶x¶x¶x¶x x¶x¶x¶o¶o¶o¶o o¶o¶o¶*¶*¶*¶***¶*¶	{"source": "arn:aws:s3:::sample-bucket/2 mb - 1.pdf", "chunk": 11, "base64"	2 mb - 1.pdfchunk11
14	42e23638-02bc-4a17-a658-432aac71b614	[0.0324695,0.0055720424,0.04507974,0.047722608,-0.009879285,0.010961	328¶Dimensionality Reduction¶x¶x¶x¶x x¶x¶x¶o¶o¶o¶o o¶o¶o¶*¶*¶*¶***¶*¶	{"source": "arn:aws:s3:::sample-bucket/2 mb - 1.pdf", "chunk": 12, "base64"	2 mb - 1.pdfchunk12
15	42e23638-02bc-4a17-a658-432aac71b614	[0.034125797,0.004243785,0.045810312,0.045183457,-0.017902879,0.0055	328¶Dimensionality Reduction¶x¶x¶x¶x x¶x¶x¶o¶o¶o¶o o¶o¶o¶*¶*¶*¶***¶*¶	{"source": "arn:aws:s3:::sample-bucket/2 mb - 1.pdf", "chunk": 13, "base64"	2 mb - 1.pdfchunk13
16	42e23638-02bc-4a17-a658-432aac71b614	[0.033418305,0.006528947,0.0535188,0.041884277,-0.020484185,0.006751	328¶Dimensionality Reduction¶x¶x¶x¶x x¶x¶x¶o¶o¶o¶o o¶o¶o¶*¶*¶*¶***¶*¶	{"source": "arn:aws:s3:::sample-bucket/2 mb - 1.pdf", "chunk": 14, "base64"	2 mb - 1.pdfchunk14
17	42e23638-02bc-4a17-a658-432aac71b614	[0.036519434,0.0005258238,0.04691439,0.04803615,-0.026224198,0.01348	328¶Dimensionality Reduction¶x¶x¶x¶x x¶x¶x¶o¶o¶o¶o o¶o¶o¶*¶*¶*¶***¶*¶	{"source": "arn:aws:s3:::sample-bucket/2 mb - 1.pdf", "chunk": 15, "base64"	2 mb - 1.pdfchunk15
18	42e23638-02bc-4a17-a658-432aac71b614	[0.028604832,0.005878907,0.04254373,0.049187273,-0.014904095,0.00804	328¶Dimensionality Reduction¶x¶x¶x¶x x¶x¶x¶o¶o¶o¶o o¶o¶o¶*¶*¶*¶***¶*¶	{"source": "arn:aws:s3:::sample-bucket/2 mb - 1.pdf", "chunk": 16, "base64"	2 mb - 1.pdfchunk16
19	42e23638-02bc-4a17-a658-432aac71b614	[0.031674806,0.005197806,0.04785811,0.04610389,-0.02434908,0.0032644	328¶Dimensionality Reduction¶x¶x¶x¶x x¶x¶x¶o¶o¶o¶o o¶o¶o¶*¶*¶*¶***¶*¶	{"source": "arn:aws:s3:::sample-bucket/2 mb - 1.pdf", "chunk": 17, "base64"	2 mb - 1.pdfchunk17
20	42e23638-02bc-4a17-a658-432aac71b614	[0.033891622,0.0040294197,0.04689453,0.04115941,-0.01496818,0.007830	328¶Dimensionality Reduction¶x¶x¶x¶x x¶x¶x¶o¶o¶o¶o o¶o¶o¶*¶*¶*¶***¶*¶	{"source": "arn:aws:s3:::sample-bucket/2 mb - 1.pdf", "chunk": 18, "base64"	2 mb - 1.pdfchunk18

Python Jogja
Lightning Talk

Jogja Cozy Space
26 July 2025

References

https://medium.com/data-science/retrieval-augmented-generation-rag-from-theory-to-langchain-implementation-4e9bd5f6a4f2

https://python.langchain.com/docs/concepts/rag/

https://python.langchain.com/docs/concepts/retrievers/

DEMO TIMES

Python Jogja
Lightning Talk

Jogja Cozy Space
26 July 2025

Python Jogja
Lightning Talk

Jogja Cozy Space
26 July 2025

Introduction to **Retrieval Augmented Generation**

With **Langchain** & **PGVector**

Joko Eliyanto
Data Scientist at Teman Data

jokoeliyanto@gmail.com Joko Eliyanto