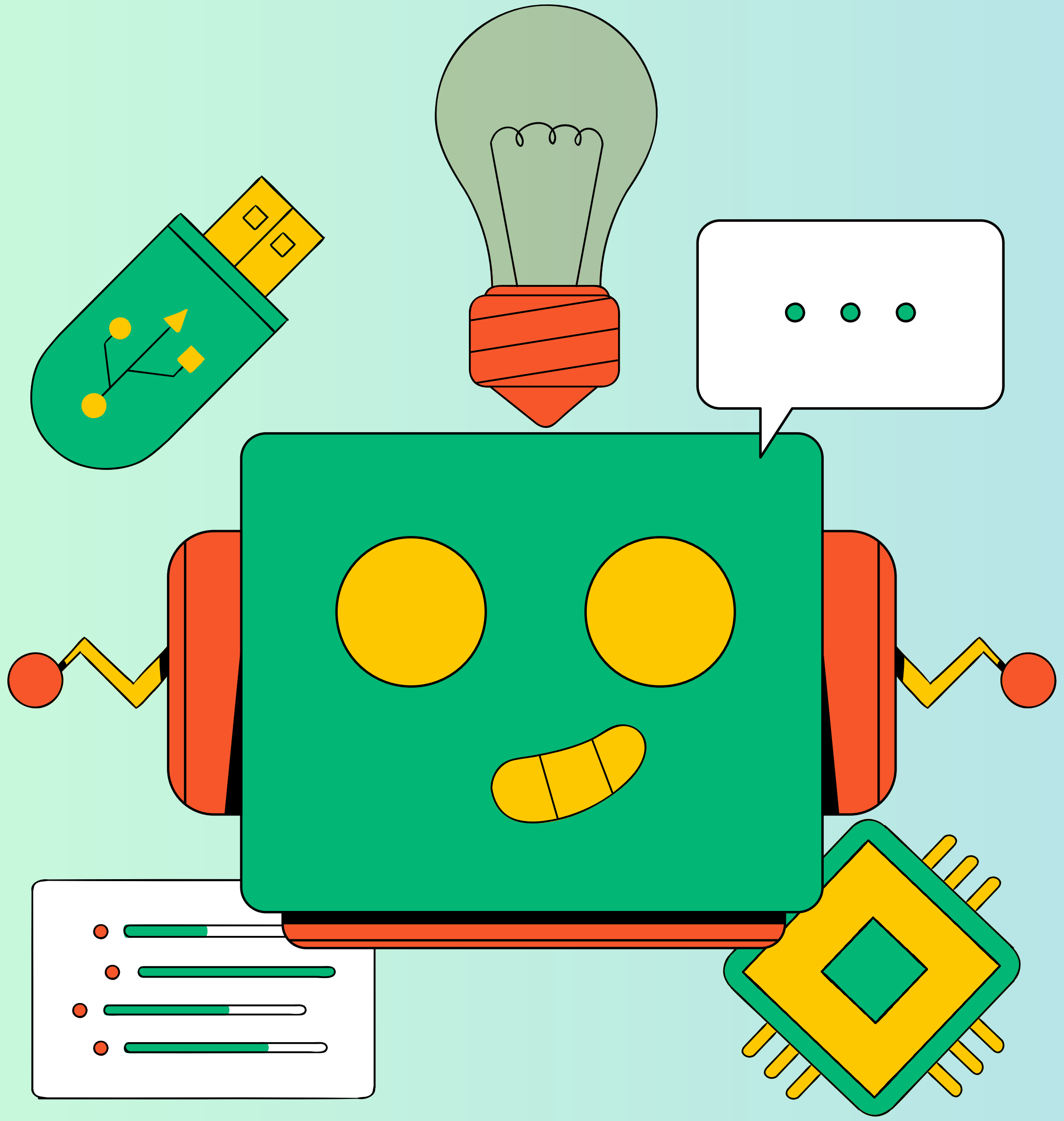
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHTN
ĐHQG-HCM
HCMUS
Viet Nam National University
Ho Chi Minh City
University of Science



SEMINAR REPORT

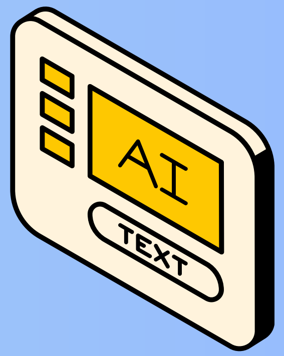
CHATBOT LLM FOR YOUTH UNION

Supervisor: **Associate Prof. Nguyen Thanh Binh**

Students:

Truong Quoc Trung - 21110427

Nguyen Minh Hung - 21110301



CONTENT

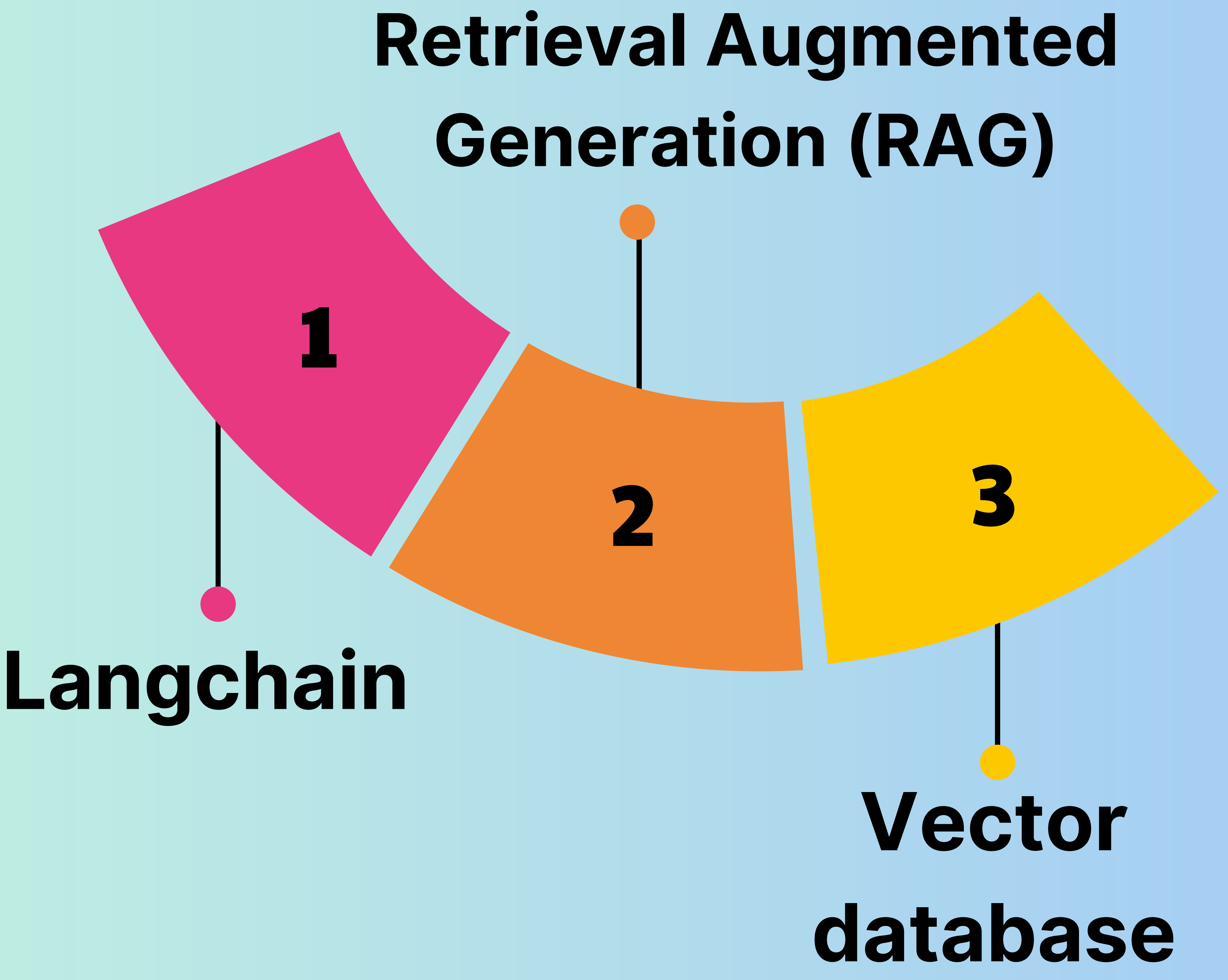
I. Introduction

Current chatbots **have not yet** effectively addressed questions about the Youth Union.

The Youth Union needs **a modern tool** to **support quick** and **effective communication** with members.

The Youth Union Chatbot was developed to help students **access information** and participate in Youth Union activities **more conveniently**.

II. Theoretical Basis



1. Langchain

LangChain is a **framework** for developing applications powered by language models.

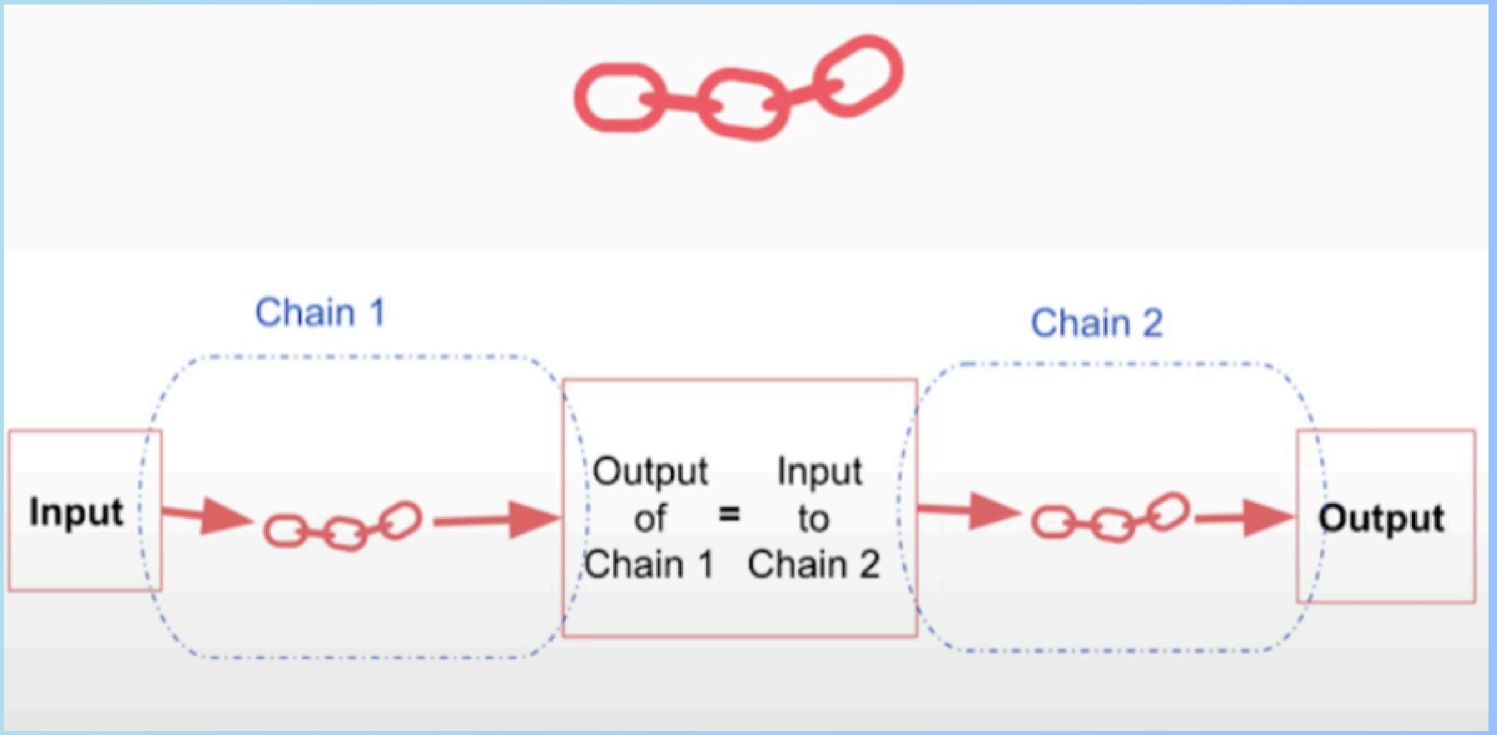


Figure 1: Simple Sequential Chain

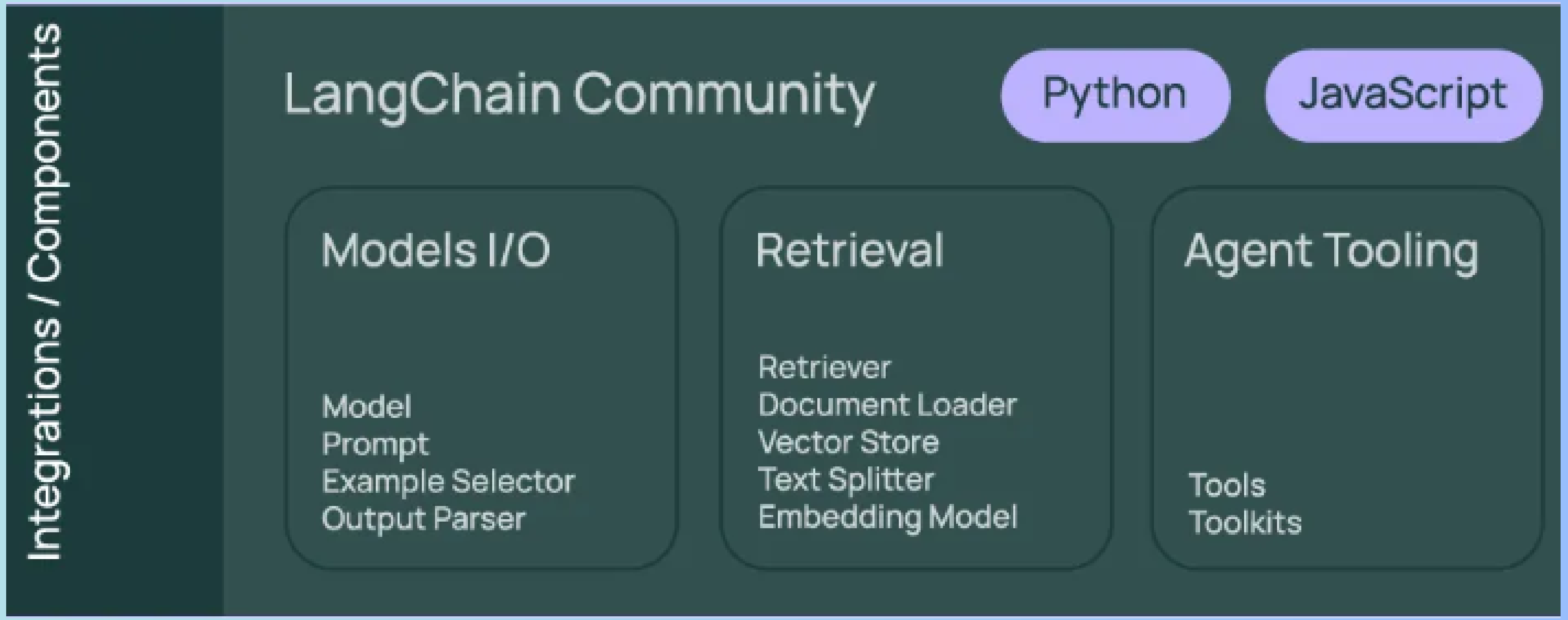


Figure 2: Langchain's components

2. Retrieval Augmented Generation (RAG)

RAG is a method that combines two important capabilities: **retrieval** and **generation**.

Enhances language model generation by incorporating external knowledge, **improving the accuracy** & **flexibility** of responses, **overcome** the problem of hallucinations

RAG significantly enhances LLM accuracy [1].

Model	mode	BCSC	Ophtho Questions	Mean
GPT-4-turbo	ZRS	80.38	77.69	79.03
	ZRS-CoT	81.54 (1.16↑)	79.62 (1.93↑)	80.58 (1.55↑)
	RAG	91.92 (11.54↑)	85.38 (7.69↑)	88.65 (9.62↑)
Llama-3-70B-Q4	ZRS	64.62	50.38	57.50
	ZRS-CoT	70.77 (6.15↑)	65.77 (15.39↑)	68.27 (10.77↑)
	RAG	84.62 (20.0↑)	78.08 (27.7↑)	81.35 (23.85↑)
Gemma-2-27B-Q4	ZRS	64.23	60.0	62.12
	ZRS-CoT	61.54 (-2.69↓)	56.92 (-3.08↓)	59.23 (-2.89↓)
	RAG	83.46 (19.23↑)	75.0 (15.0↑)	79.23 (17.11↑)
Mixtral-8x7B-Q4	ZRS	57.69	48.08	52.89
	ZRS-CoT	53.85 (-3.84↓)	52.69 (4.61↑)	53.27 (0.385↑)
	RAG	78.46 (20.77↑)	71.54 (23.46↑)	75.00 (22.11↑)
Antanki et al. 2023 [2] (GPT-4, temparature=0.3)	ZSR	75.8	70.8	71.7

Table 1: Compare accuracy without and with RAG [1]

[1] Nguyen, Q., Nguyen, D.-A., Dang, K., Liu, S., Nguyen, K., Wang, S. Y., Woof, W., Thomas, P., Patel, P. J., Balaskas, K., Thygesen, J. H., Wu, H., & Pontikos, N. (2024). Advancing Question-Answering in Ophthalmology with Retrieval Augmented Generations (RAG): Benchmarking Open-source and Proprietary Large Language Models. *medRxiv*, 2024-11. https://doi.org/10.1101/2024.11.18.24317510

3. Vector Database

- **Stores & manages data** in the form of multi-dimensional vectors, usually **embedding**.

- Each data point *(e.g. paragraph, document)* is converted into a vector, quickly searching for vectors with semantic similarity based on distance like **cosine similarity** or **Euclidean distance**

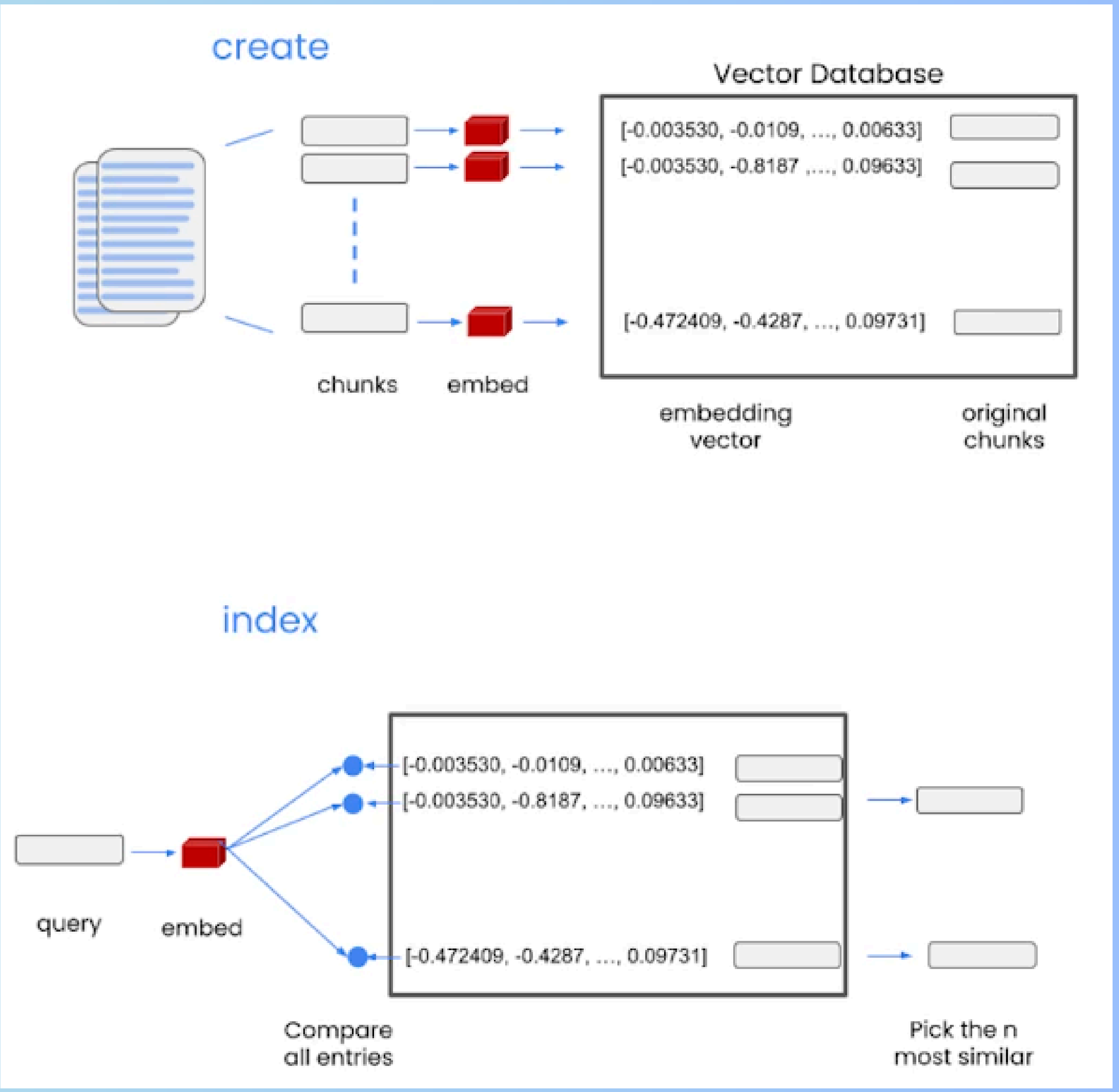
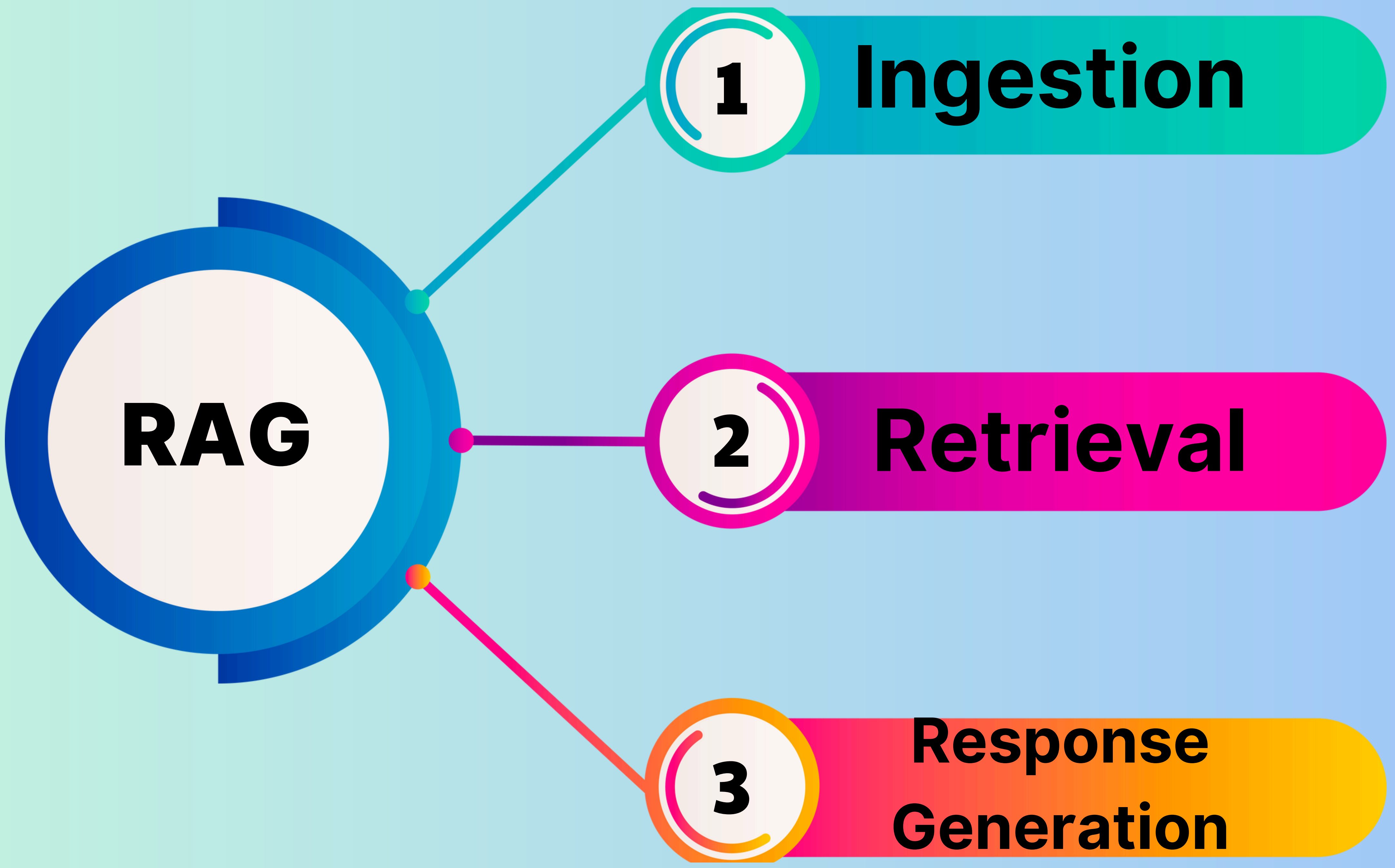


Figure 3: Vector Databases's workflow

III. Development Workflow



III. Development Workflow

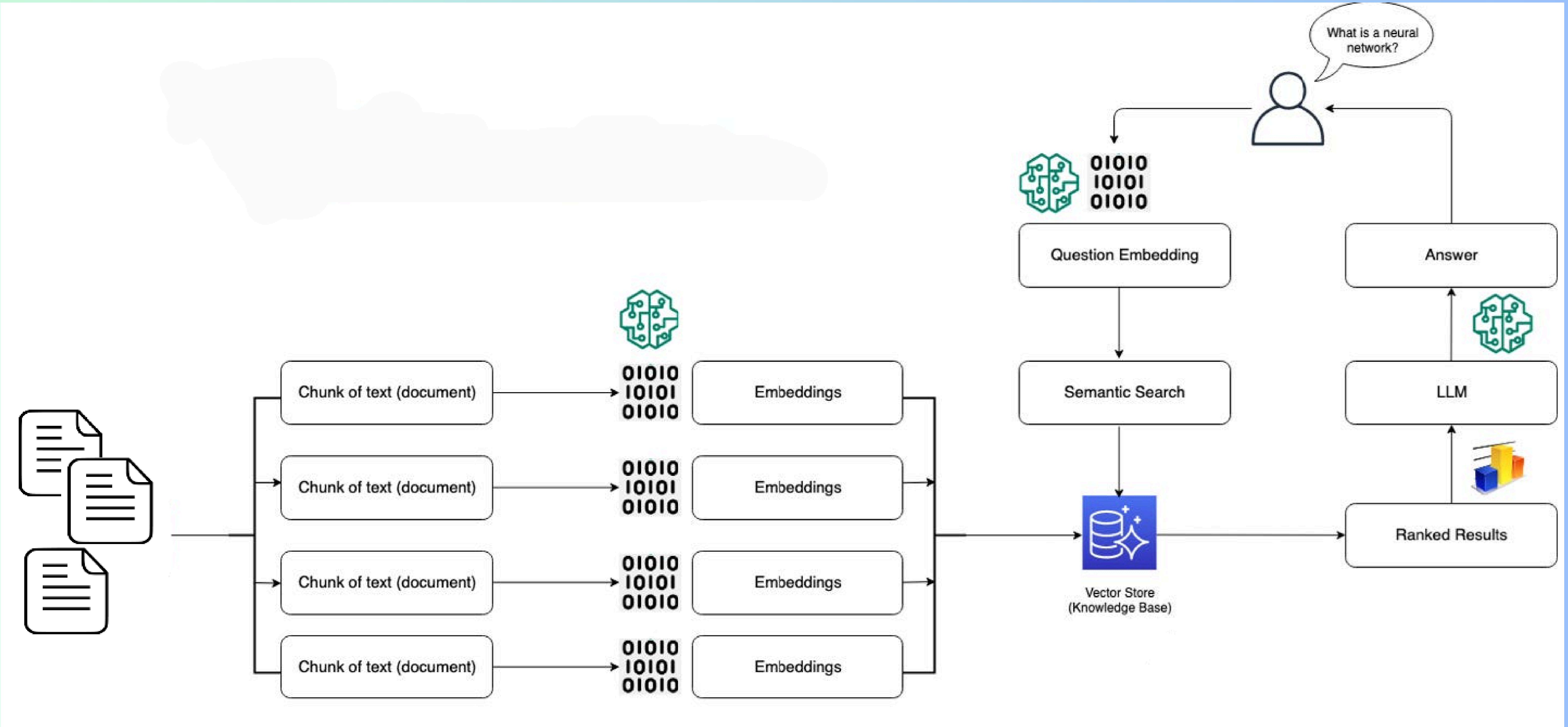
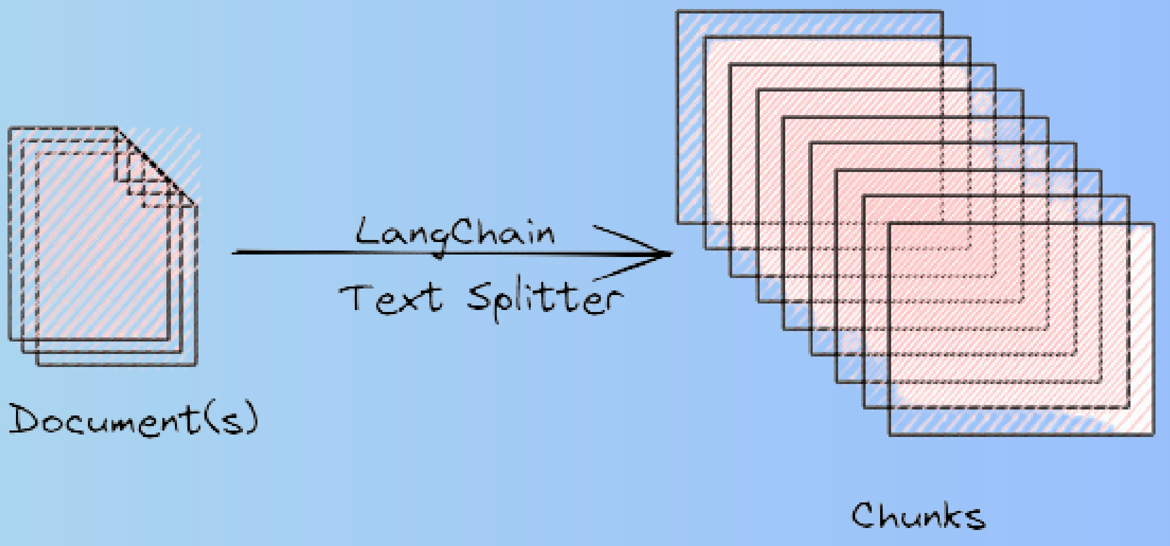


Figure 4: Our workflow

1. Ingestion

Collect data & Chunking with Langchain

Use **.docx** data containing documents about the Youth Union, student handbooks, information about President Ho Chi Minh from official websites.
(ex: https://doanthanhnien.vn/tai-lieu)



Criteria:
- Text only;
- Paragraphs on a page & use headings;
- Use special characters to separate pages *(ex: "###")*.

Recursive Character Text Splitter

Split text by character, making sure each paragraph is less than a certain length.

Useful for documents with natural paragraph / sentence breaks.

1. Ingestion

Embeddings & Vector Database

Chunks are **encoded** into **embedding vectors** using modern models then stored in a **vector database**.

Milvus stands out as the most **comprehensive** solution among the databases evaluated, **meeting all the essential criteria** and **outperforming** other open-source options [2].

Database	Multiple Index Type	Billion-Scale	Hybrid Search	Cloud-Native
Weaviate	✗	✗	✓	✓
Faiss	✓	✗	✗	✗
Chroma	✗	✗	✓	✓
Qdrant	✗	✓	✓	✓
Milvus	✓	✓	✓	✓

Table 2: Comparison of Various Vector Databases [2]

[2] Wang, X., Wang, Z., Gao, X., Zhang, F., Wu, Y., Xu, Z., Shi, T., Wang, Z., Li, S., Qian, Q., Yin, R., Lv, C., Zheng, X., & Huang, X. J. (2024). Searching for best practices in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, (pp. 17716–17736), Miami, Florida, USA: Association for Computational Linguistics.

2. Retrieval

Retrieval is a **core component** of the RAG system, responsible for **retrieving relevant information** from large databases, acting as the **chatbot's "external memory"**.

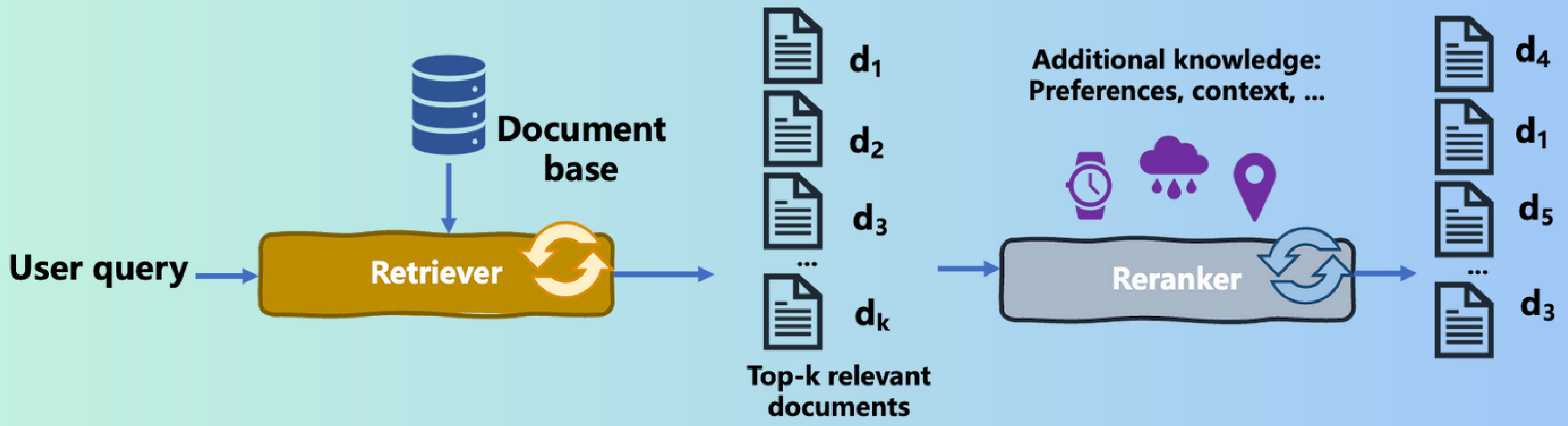


Figure 5: The retrieve & rerank pipeline

3. Respone Generation

Generates user responses by combining **retrieved information** with the model's **pre-trained knowledge**. This ensures **coherent**, **contextual**, **conversational** responses, and **advoids negativity**.

Strategic prompt design, such as placing important information at the beginning or end of an **input sequence**, enhances the **system efficiency** [3].

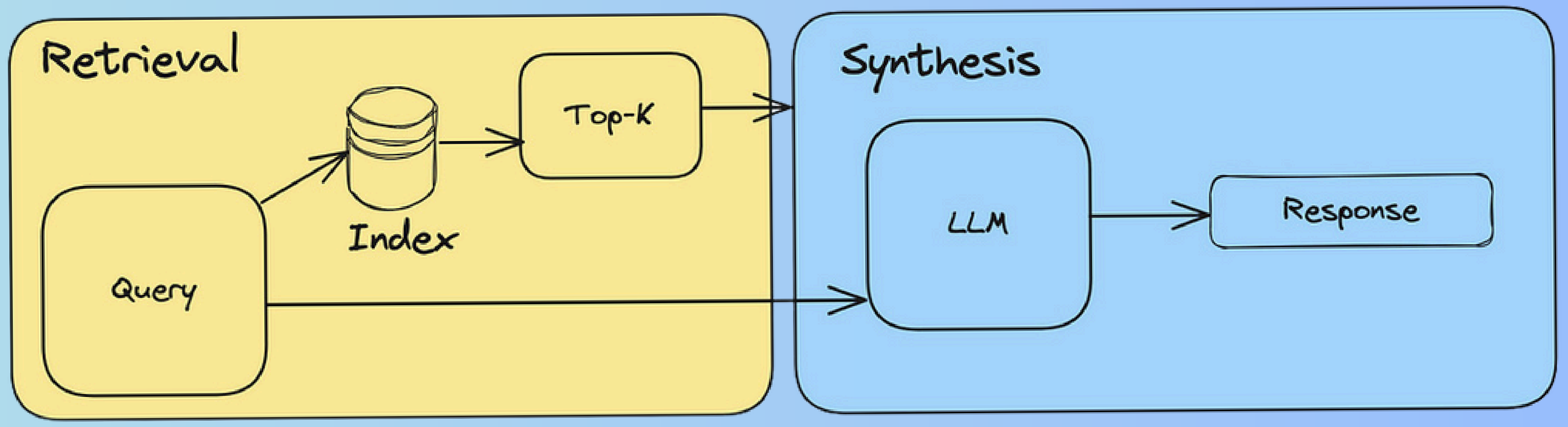


Figure 6: Respone Generation

[3] Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157-173.

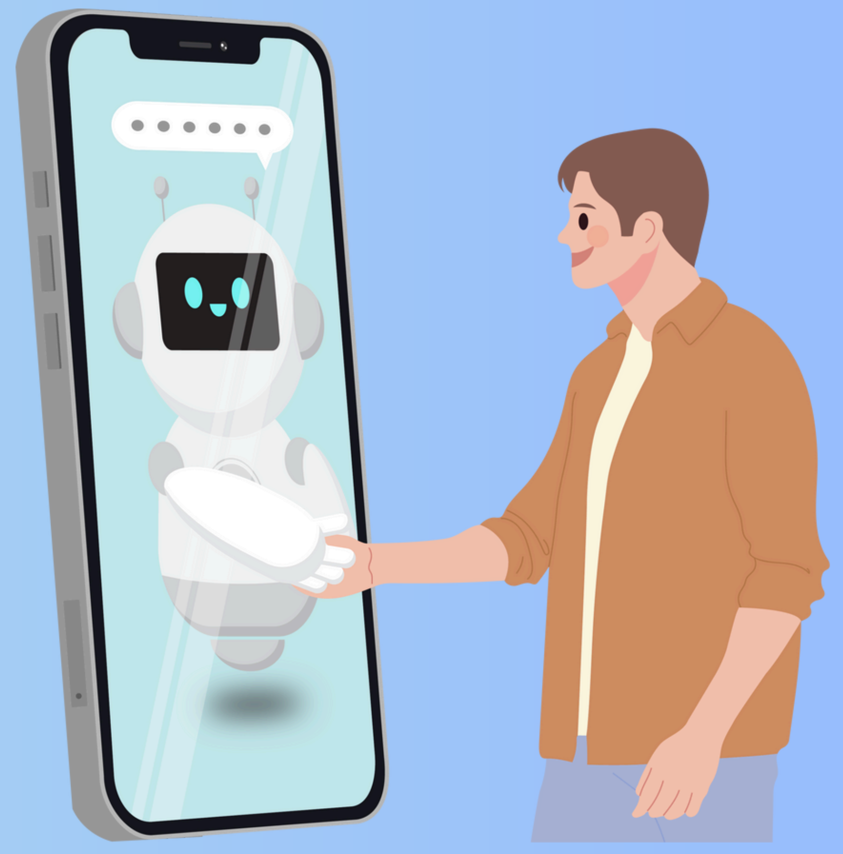
IV. Challenges & Future Works

Challenges

- Data processing still needs to be done manually.
- Chunk size.
- Ability to filter negative questions.
- Need to optimize RAG.

Future Works

- Advanced technology.
- Integrating UI.
- Optimize RAG.



V. References

[1] Nguyen, Q., Nguyen, D.-A., Dang, K., Liu, S., Nguyen, K., Wang, S. Y., Woof, W., Thomas, P., Patel, P. J., Balaskas, K., Thygesen, J. H., Wu, H., & Pontikos, N. (2024). Advancing Question-Answering in Ophthalmology with Retrieval Augmented Generations (RAG): Benchmarking Open-source and Proprietary Large Language Models. *medRxiv*, 2024-11. https://doi.org/10.1101/2024.11.18.24317510

[2] Wang, X., Wang, Z., Gao, X., Zhang, F., Wu, Y., Xu, Z., Shi, T., Wang, Z., Li, S., Qian, Q., Yin, R., Lv, C., Zheng, X., & Huang, X. J. (2024). Searching for best practices in retrieval-augmented generation. In P*roceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, (pp. 17716–17736), Miami, Florida, USA: Association for Computational Linguistics.

[3] Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157-173.

THANK YOU

for your attention