KwangWoon Univ.
Dept. of Electronic Engineering
2025 Capstone Design Proposal Presentation
Communication & Signal Processing Division

Development of a RAG-Based Conversational AI Using LLM

Advisor
Prof. Ji-sang Yoo

Members
Yoonsung Ji(2020706121)
Kyungtae Park(2020706061)
Jungin Lee(2021706127)
Geon Lee(2022117002)

KWANGWOON UNIVERSITY

Contents

①

Overview

- Project motivation
- Project goal

Overview : Project Motivation

(1) Practical Challenges in Academic Life

▶ <u>Challenges with Current systems in KWU</u>

- Students struggle to find **specified lecture information**
- Searching through multiple platforms is **inefficient**
- Lack of **integrated platform to retrieve lecture information**

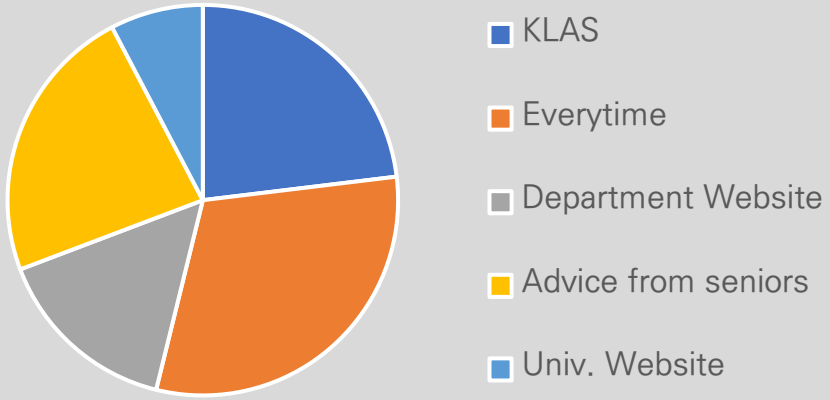


Fig1. General lecture information sources in KWU

(2) Growth of AI Industry

▶ <u>The Rapid Growth of AI(Artificial Intelligence) industry</u>

- Global AI market size is projected to reach $1.8 trillion by 2030
- Major companies(Google, OpenAI, Microsoft) heavily funding AI research

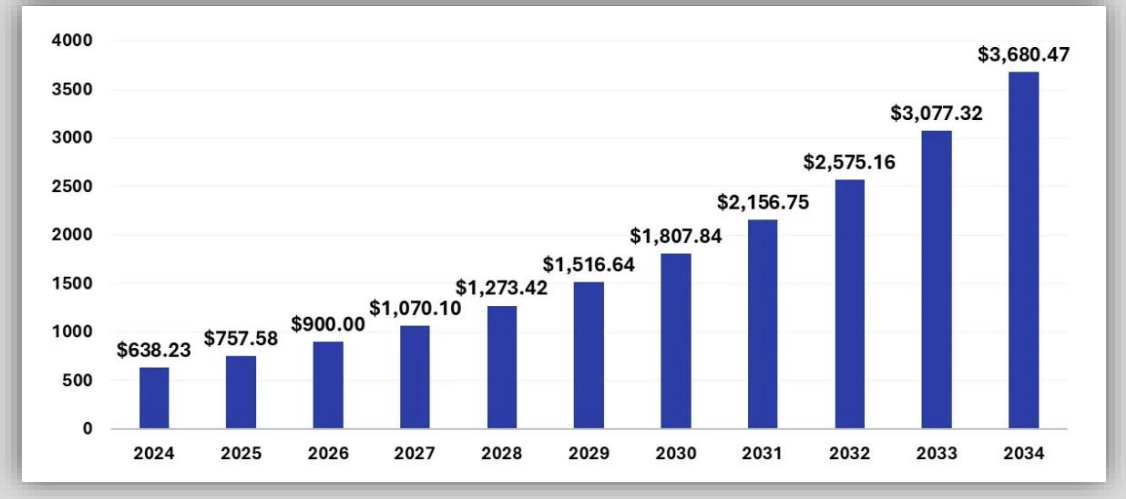


Fig2. AI Market Size and Growth 2025 to 2034(USD billion)

Fig3. Major AI Investments by Big Tech Companies(2023)

(2) Growth of AI Industry

▶ AI's Impact on Information Retrieval

• Traditional keyword-based search(Naver, Google) is being replaced by AI-powered semantic search

☞ AI-powered chatbots & virtual assistants are transforming education, customer support, and research

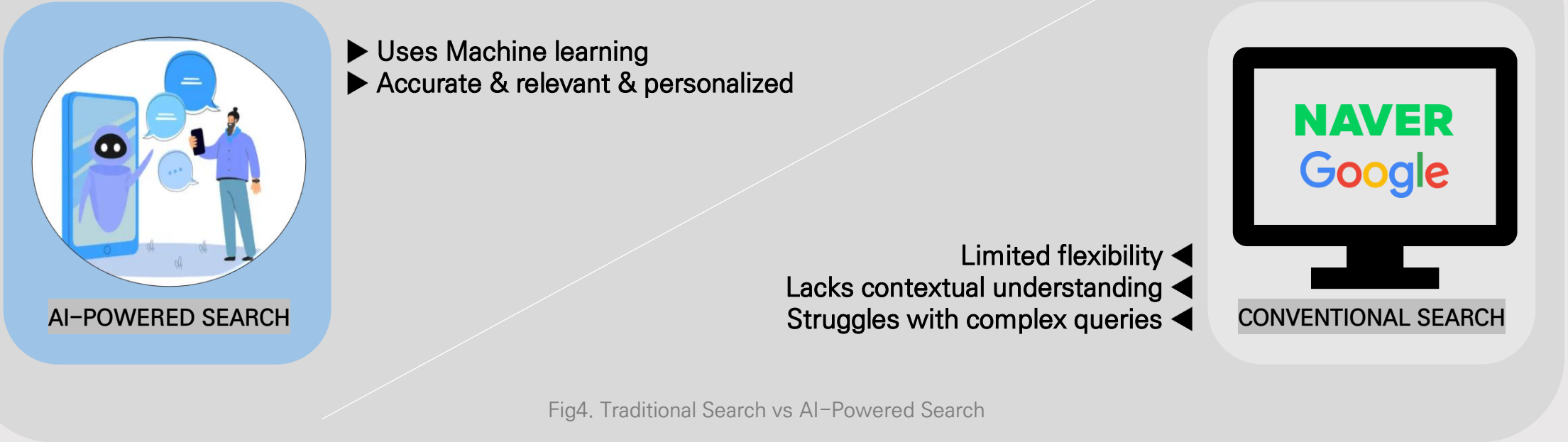


Fig4. Traditional Search vs AI-Powered Search

(3) How This Relates to Our Capstone Project?

▶ Project flow : User perspective

• Frontend Web UI sample

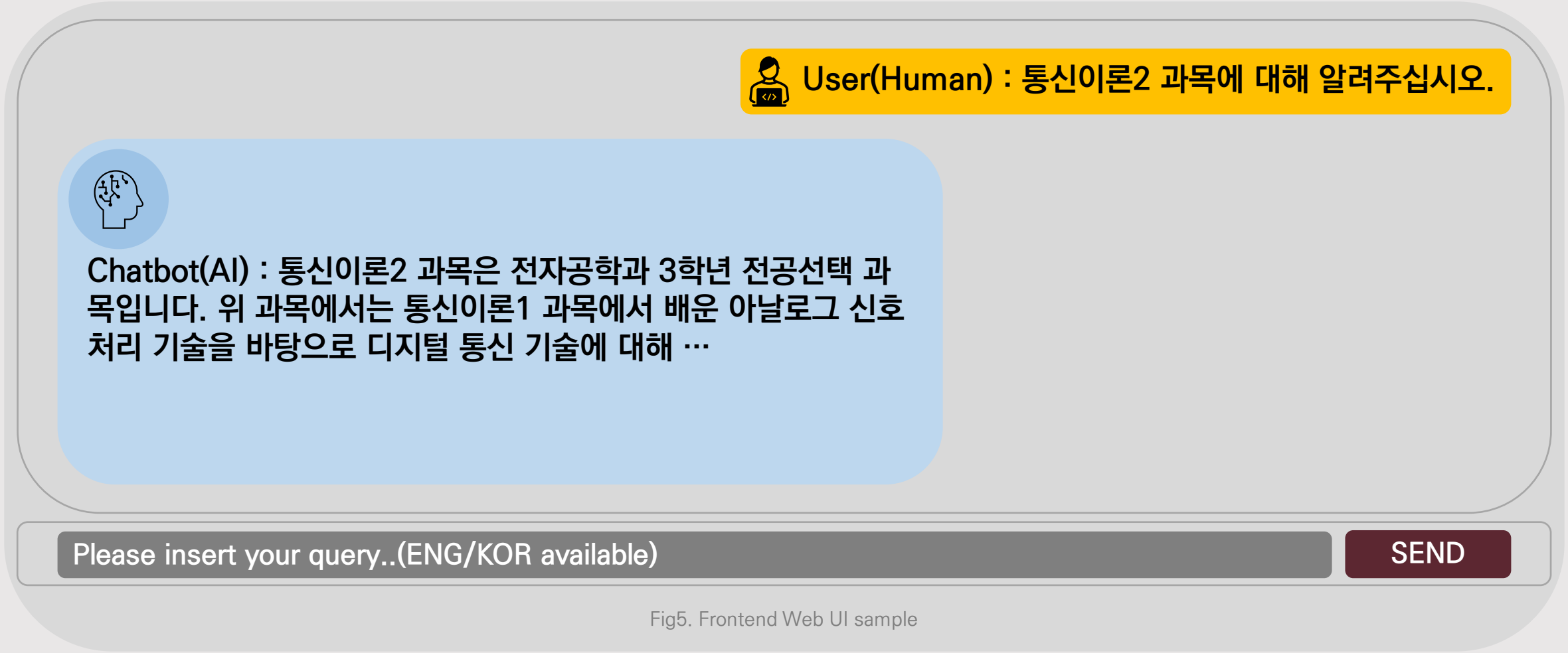


Fig5. Frontend Web UI sample

②

Related Theories

- RAG
- LLM
- Prompt Engineering

(1) What's RAG?

▶ Concept

- Retrieval Augmented Generation ☞ Combines search & generation
- AI approach that **retrieves relevant external information** before generating responses

▶ RAG Pipeline

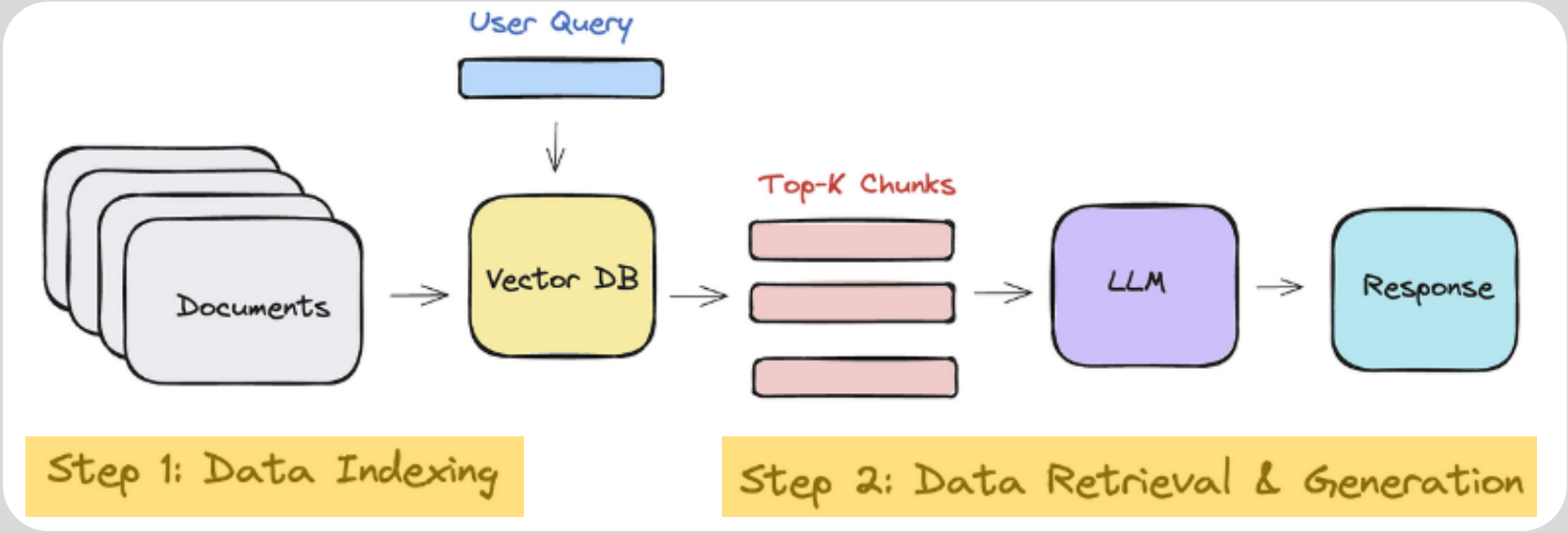
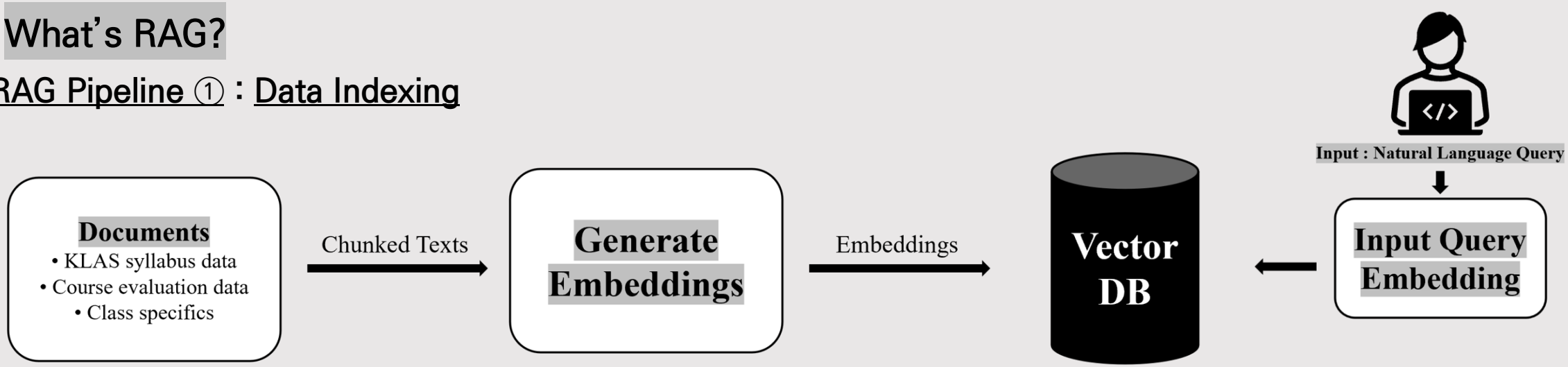


Fig6. RAG Pipeline

(1) What's RAG?

▶ RAG Pipeline ① : Data Indexing



- Construct a vector dataset ☞ store it in the backend database
- Retrieve relevant documents using FAISS(vector search) & BM25(keyword search)
- Apply cosine similarity method for ranking retrieved results

(1) What's RAG?

▶ RAG Pipeline ① : Data Indexing

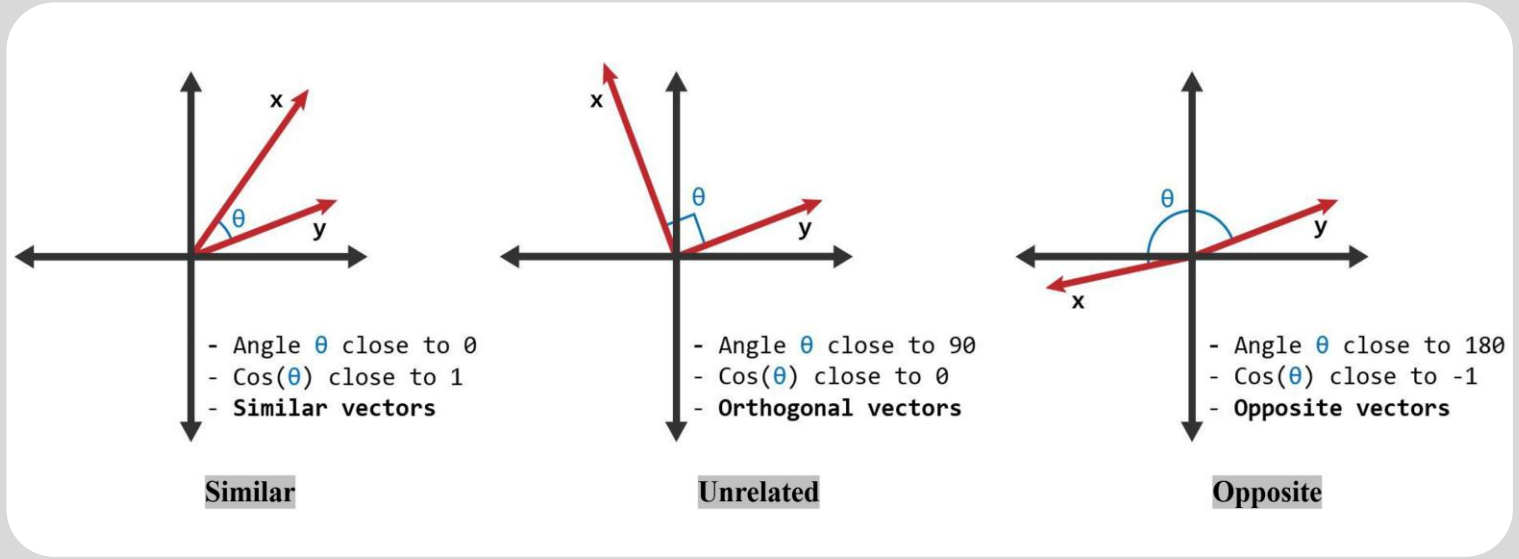
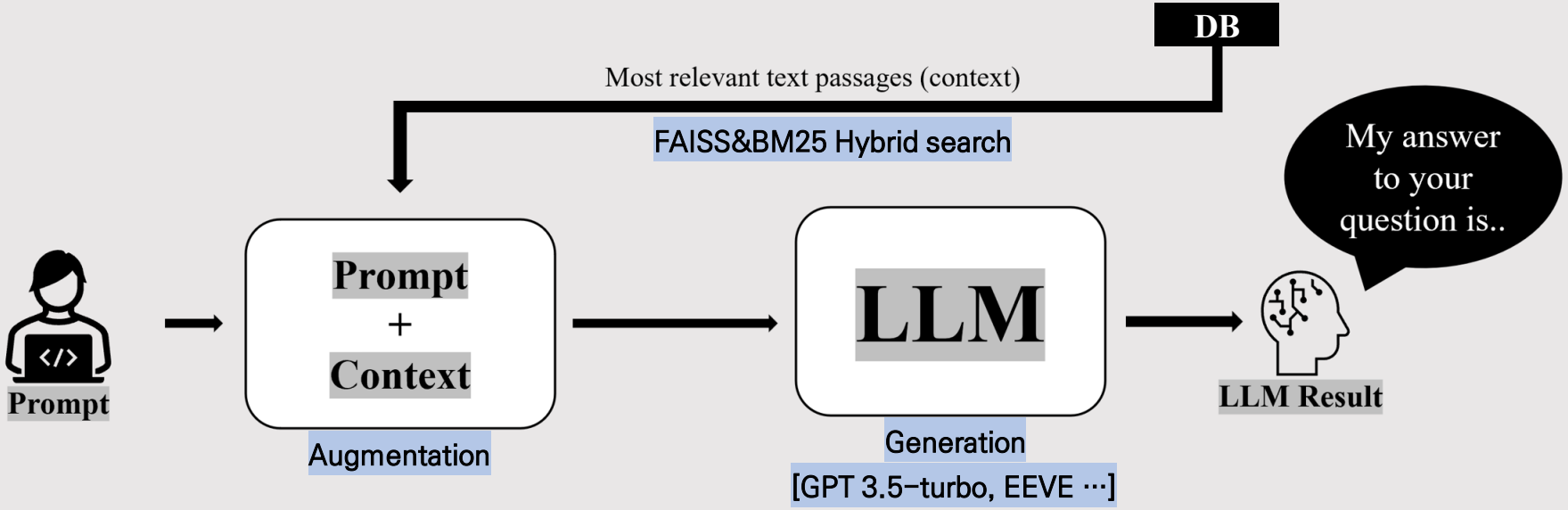


Fig8. Similarity Metrics in Vector DB

(1) What's RAG?

▶ RAG Pipeline ② : Augmentation & Generation



- Use **LLM** to enhance the quality and contextuality of generated responses
- **LLM** Utilize retrieved documents as context
- In each part, apply **prompt engineering** to LLM process to generate structured and informative answers

(1) What's RAG?

▶ **Advantages**

- More up-to-date information compared to standard LLMs
- Reduces hallucinations (false information) in AI-generated responses
- Improved relevance and factual grounding for specific domains (e.g., academic Q&A)

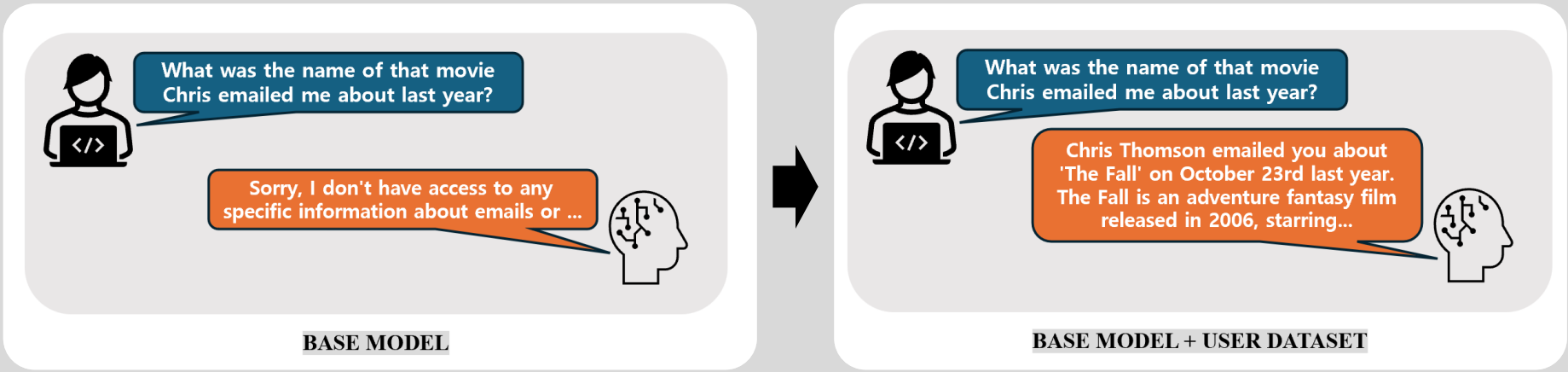


Fig9. Comparison of General LLM & RAG system

(2) What's LLM?

▶ **Concept**

- Large Language Model
- Deep neural networks trained on massive datasets to understand & generate human-like language
- Using LLM in the generation part has become a dominant approach in RAG systems since 2022-2023
- We'll use & analyze pre-trained LLMs in project

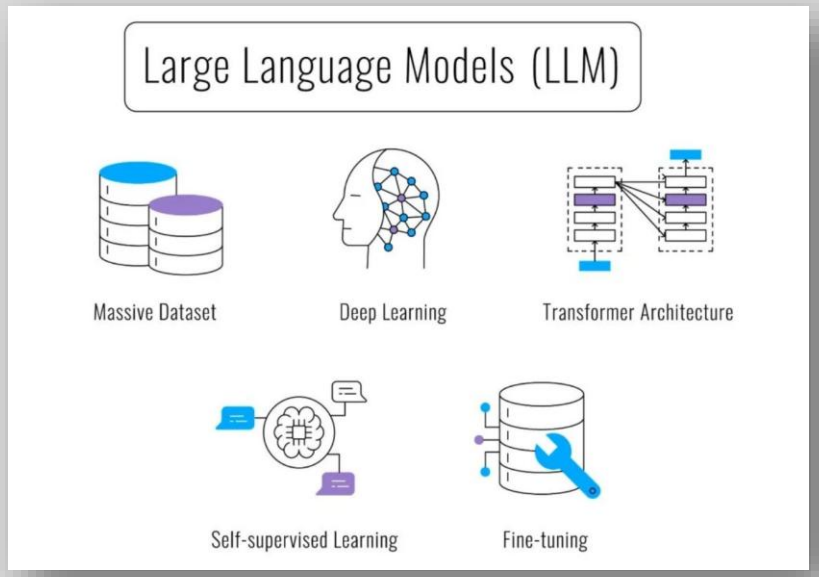


Fig10. LLM features

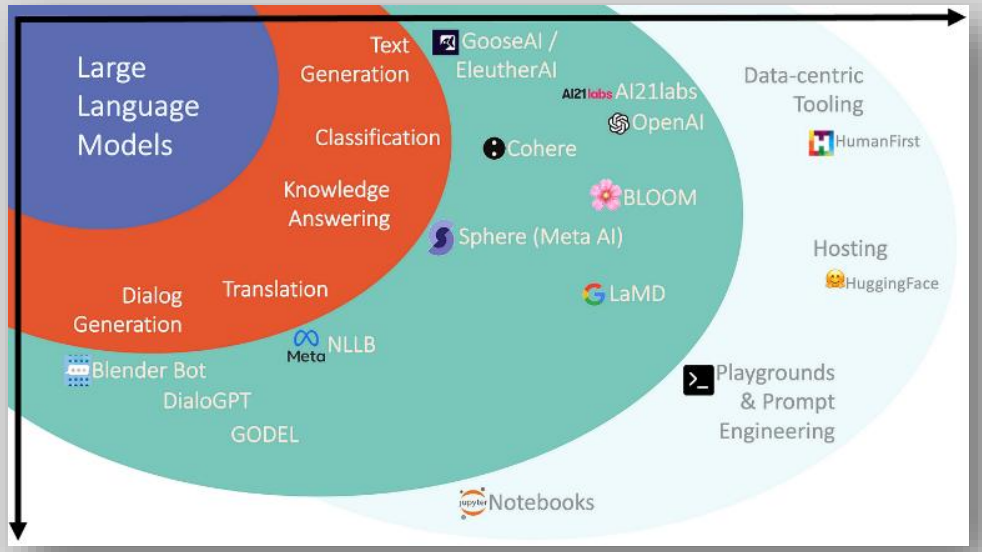


Fig11. LLM landscape

(3) What's Prompt Engineering?

▶ Workflow

- Use prompt engineering to optimize AI models
- Process of crafting inputs(prompts) to guide an AI model's response in a desired direction
- This will help LLMs generate accurate, relevant, and context-aware responses in our project

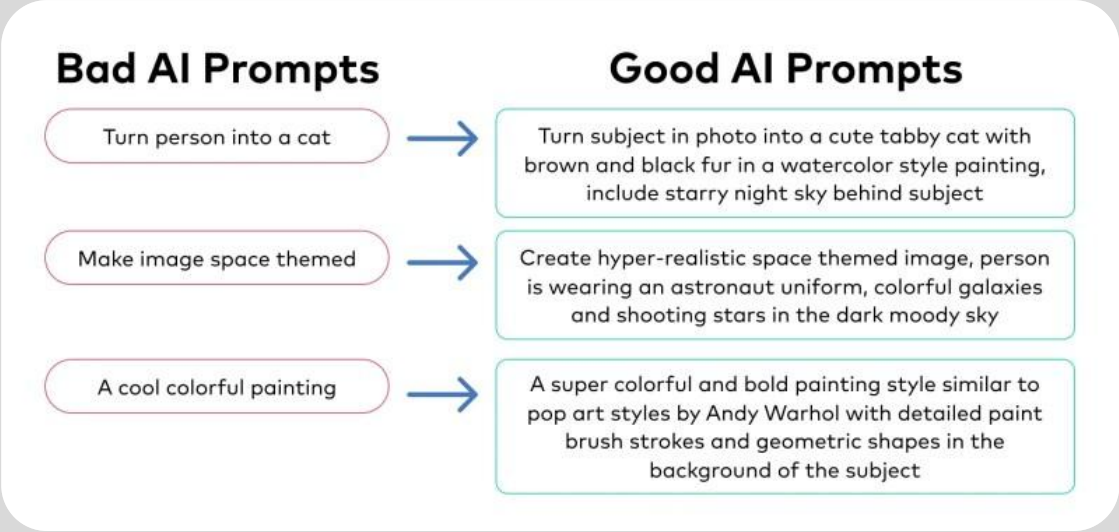


Fig12. Good & Bad AI prompts

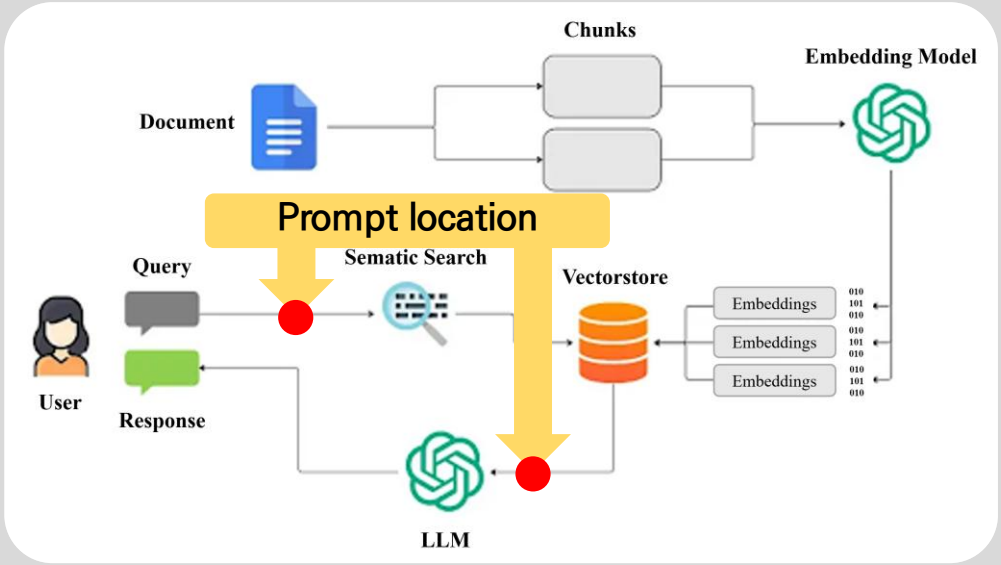
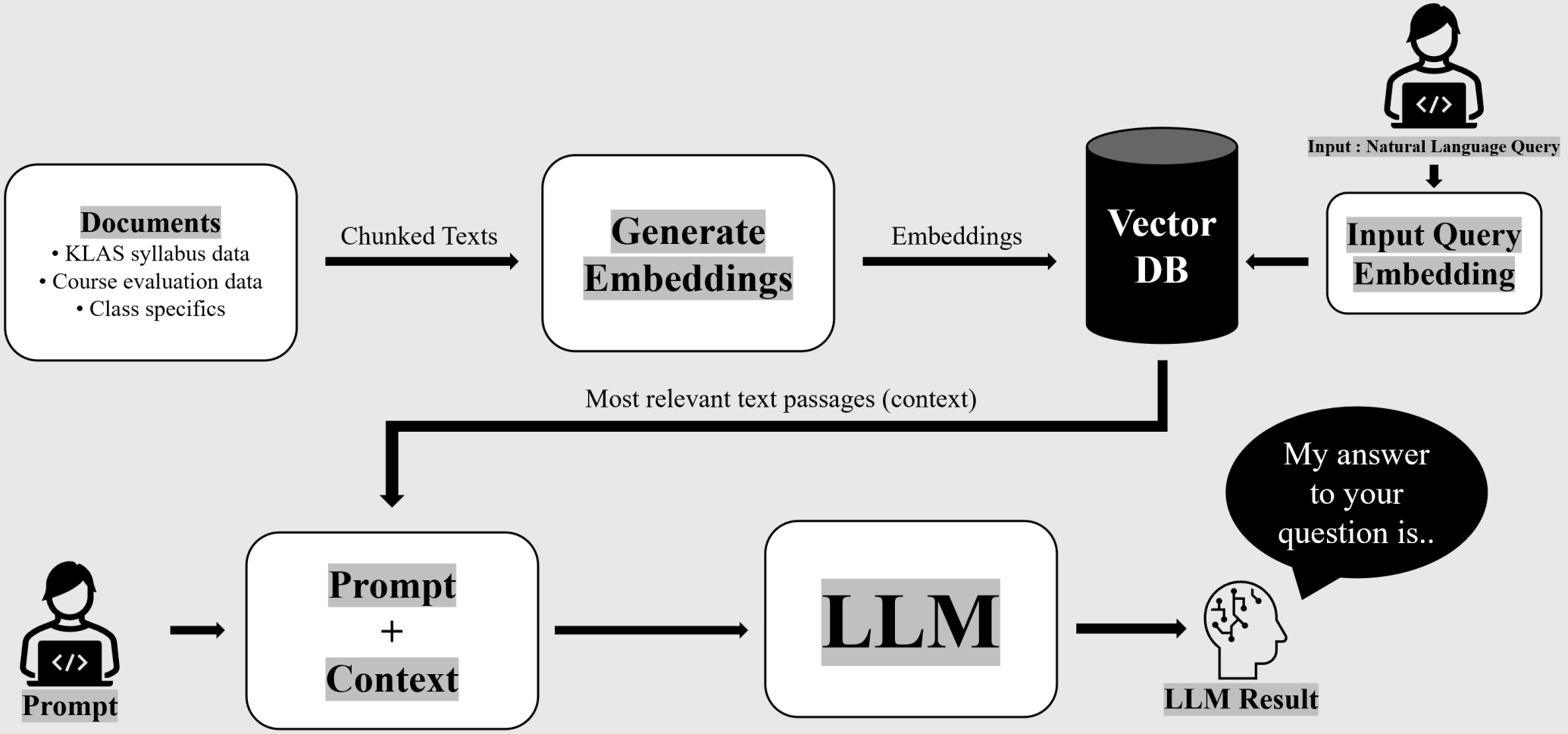


Fig13. Prompts in our project

Related Theories : Conclusion

▶ How This Relates to Our Capstone Project?

▶ Project flow



Related Theories : Conclusion

▶ How This Relates to Our Capstone Project?

▶ Main Focus in our project

Retrieval part	Backend part	Frontend part
• Implement FAISS vs BM25 • Construct ideal algorithm for our data • Retrieval code implementation • Programming language : Python	• API development(☞FastAPI) • RAG pipeline integration • LLM prompt engineering • Programming language : Python	• Chatbot UI development(React&Typescript) • Integration with backend API • Improve UI/UX design • Programming language : TypeScript

▶ Expected effects

- Development of an Integrated Academic Search Platform for Web-based Access
- Implementation of an AI Chatbot for Academic Assistance at Kwangwoon University
- Optimization of Search Performance using Machine Learning Techniques
- Practical Application and Evaluation of AI and NLP in a Capstone Project

③

Project Plan

- Project timeline
- Role sharing

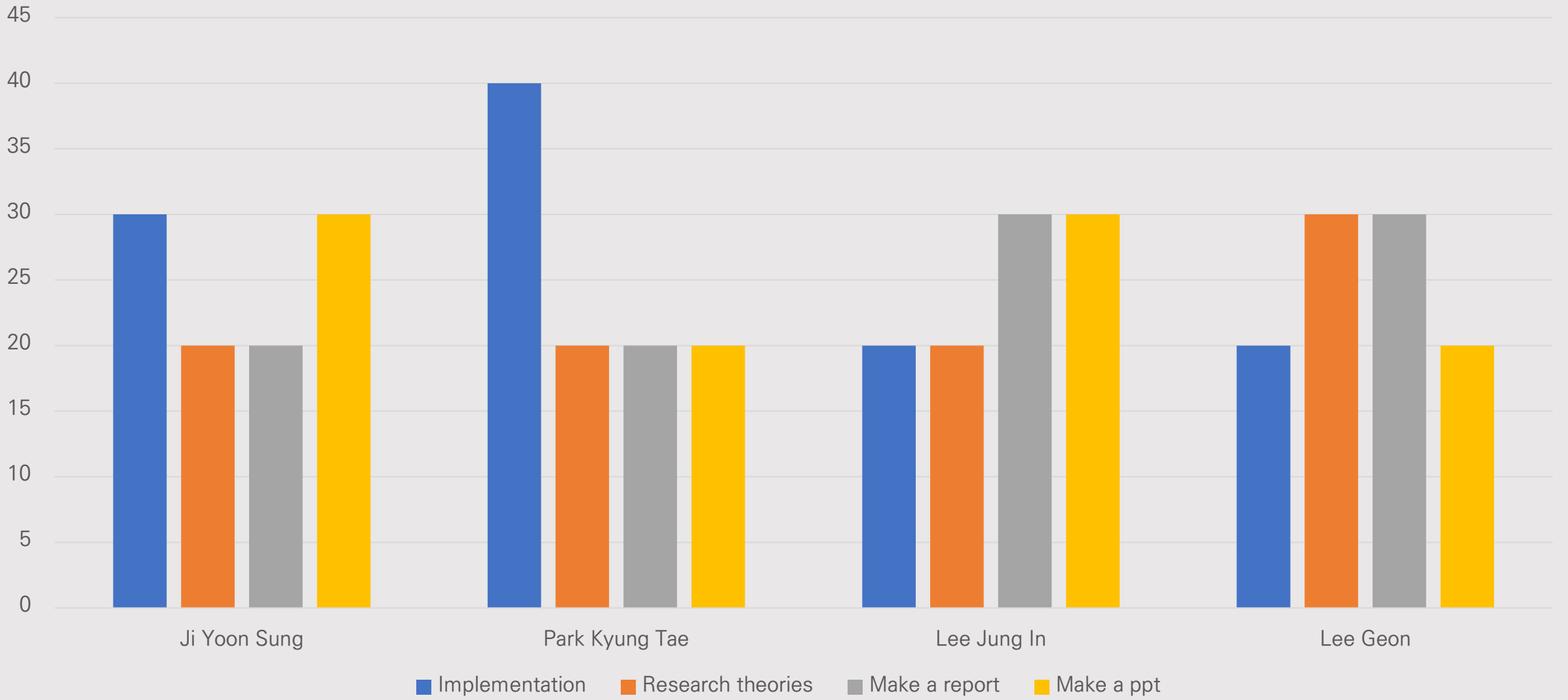
Project Plan : Project timeline

▶ Timetable

	Week												
	1	2	3	4	5	6	7	8	9	10	11	12	13
Selecting topic													
Data processing													
Build the backend													
Build the frontend													
Deploy to the web													
Data expansion													
Test & Feedback													
Final Presentation													

Project Plan : Role sharing

▶ Role Sharing



④

References

- Academic references

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, and S. Riedel, "Retrieval-augmented generation for knowledge-intensive NLP tasks," University College London, New York University, 2020.

[2] Sanghyuk Choi, Jinseok Seol and Sang-goo Lee, On Word Embedding Models and Parameters Optimized for Korean, Korean Language Information Science Society, 2016.

[3] G.-W. Yi and S. K. Kim, "Design of a question-answering system based on RAG model for domestic companies," Computer Engineering, Jeju National University, 2024.

[4] C.-G. Hwang, C.-P. Yoon, and Y. D. Yeol, "Sentence similarity analysis using ontology based on cosine similarity," Kwangwoon University, Gyeonggi University of Science and Technology, 2021.

[5] J.-I. Lee, J.-H. Ahn, K.-T. Koh, and Y.-S. Kim, "A study on the optimal search keyword extraction and retrieval technique generation using word embedding," Korea Institute of Civil Engineering and Building Technology, 2023.

[6] Ha-Young Joo, Hyeontaek Oh and JinHong Yang, A Survey on Open Source based Large Language Models, Korea Information Electronic Communication Technology, 2023.

[7] Gyeong-Won Jang and Seong-Soo Han, Prompt Engineering Technique for efficient use of ChatGPT, Kang-Won National University, 2023.

THANK YOU

Thank you for listening to my presentation

KwangWoon Univ.
Dept. of Electronic Engineering
2025 Capstone Design Proposal Presentation

KWANGWOON UNIVERSITY