

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)



Contents

- ① Motive
- ② Project Goal
- ③ Current Status
- ④ Future works
- ⑤ Schedule
- ⑥ References



Motive

- Project motivation

(1) Practical Challenges in Academic Life

► Challenges with Current systems in KWU

- Lack of integrated platform to retrieve lecture information

► The Rapid Growth of AI(Artificial Intelligence) industry

- Global AI market size is projected to reach \$1.8 trillion by 2030

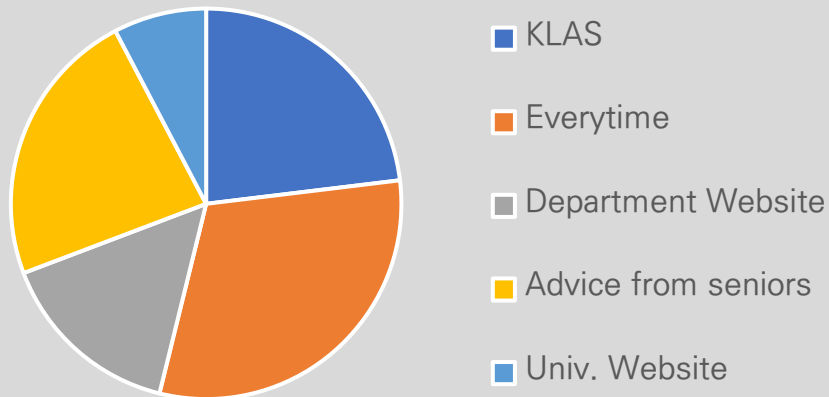


Fig1. General lecture information sources in KWU

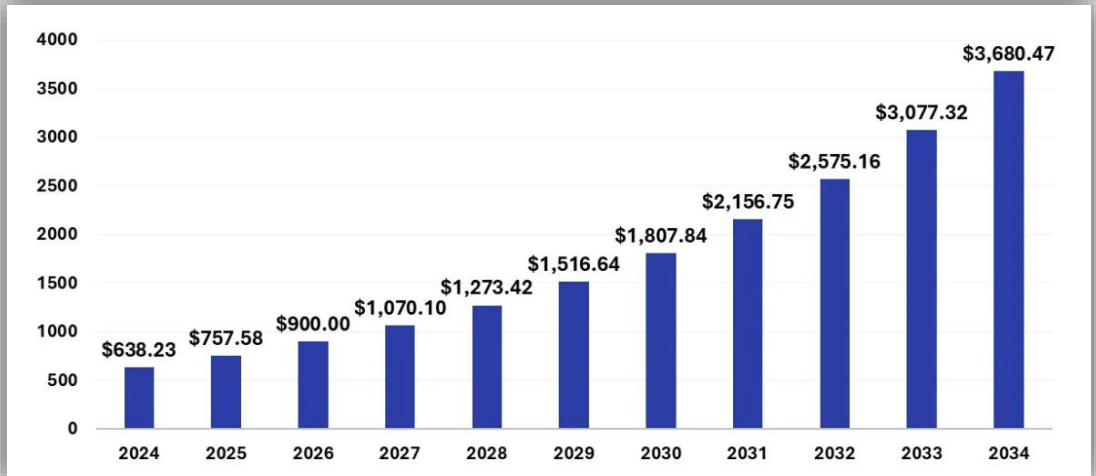
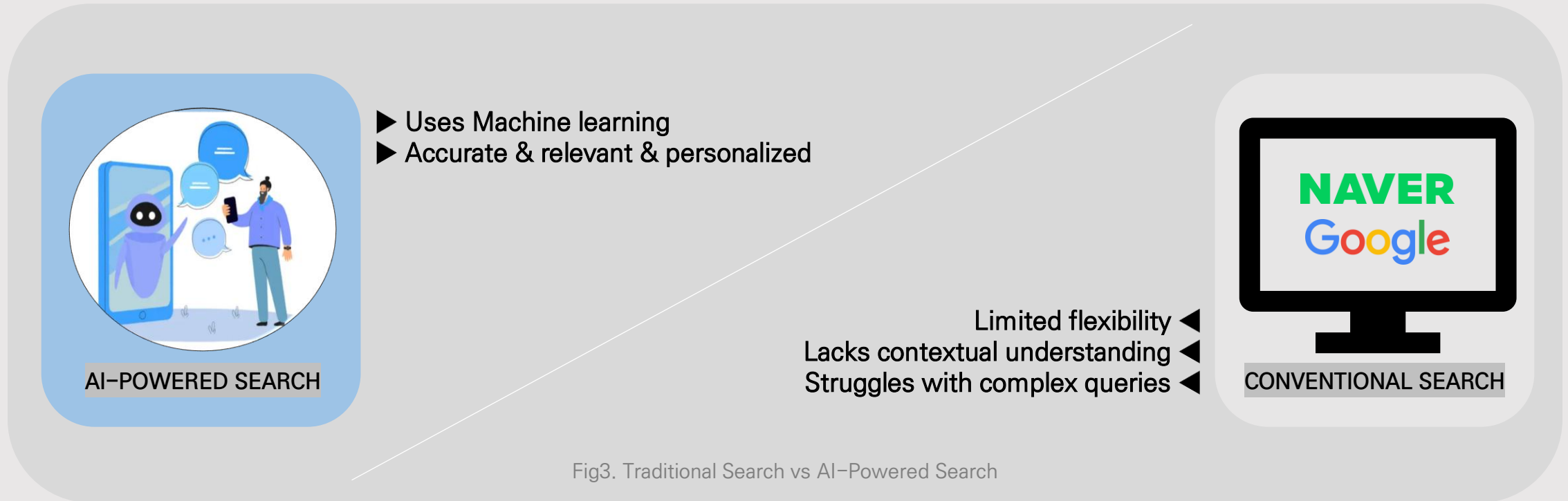


Fig2. AI Market Size and Growth 2025 to 2034(USD billion)

(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





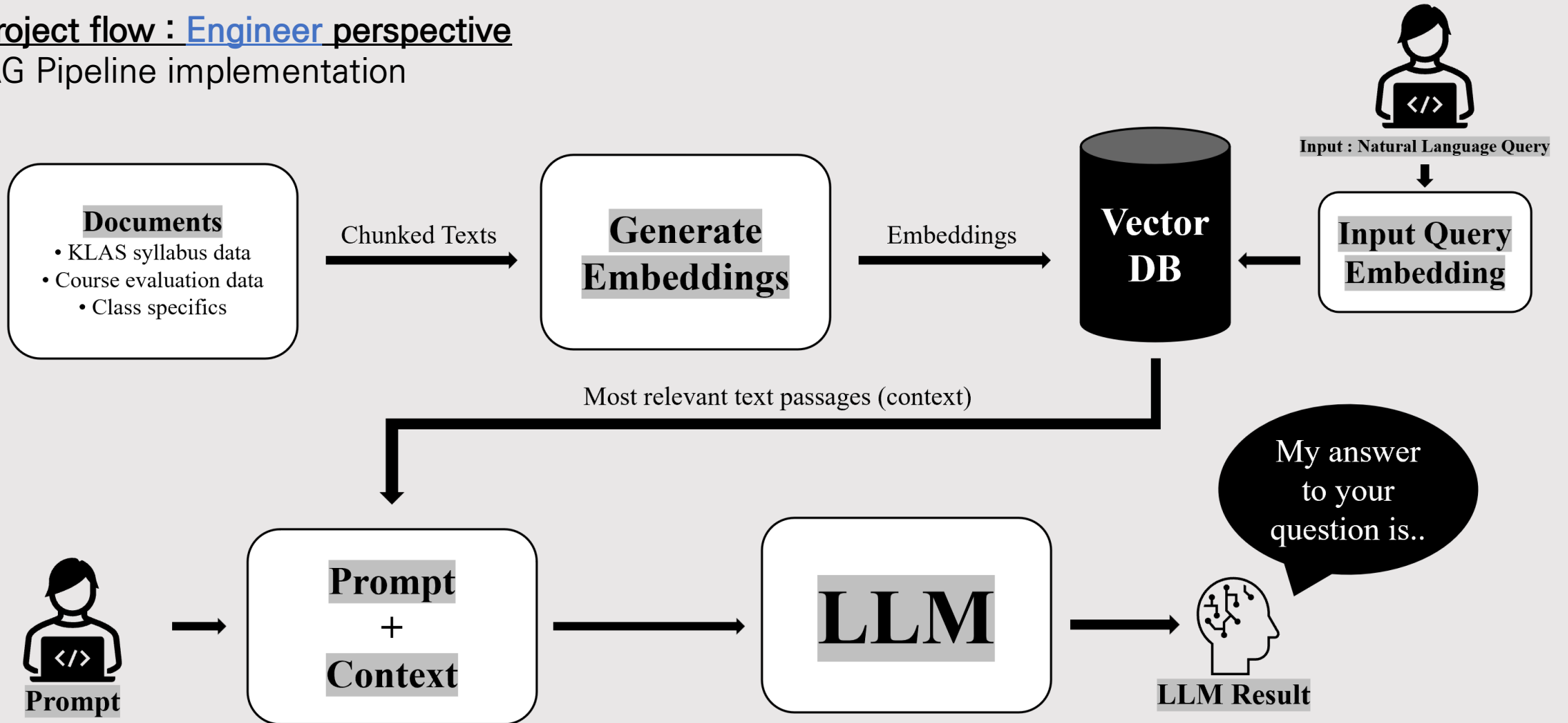
②

Project Goal

- Project flow

② Project Goal

- ▶ Construct RAG pipeline
- ▶ Project flow : [Engineer perspective](#)
 - RAG Pipeline implementation



② Project Goal

▶ Chatbot system based on RAG

▶ Project flow : **User** perspective

- Frontend Web UI sample



User(Human) : 통신이론2 과목에 대해 알려주십시오.



Chatbot(AI) : 통신이론2 과목은 전자공학과 3학년 전공선택 과목입니다. 위 과목에서는 통신이론1 과목에서 배운 아날로그 신호 처리 기술을 바탕으로 디지털 통신 기술에 대해 ...

Please insert your query..(ENG/KOR available)

SEND

Fig4. Frontend Web UI sample

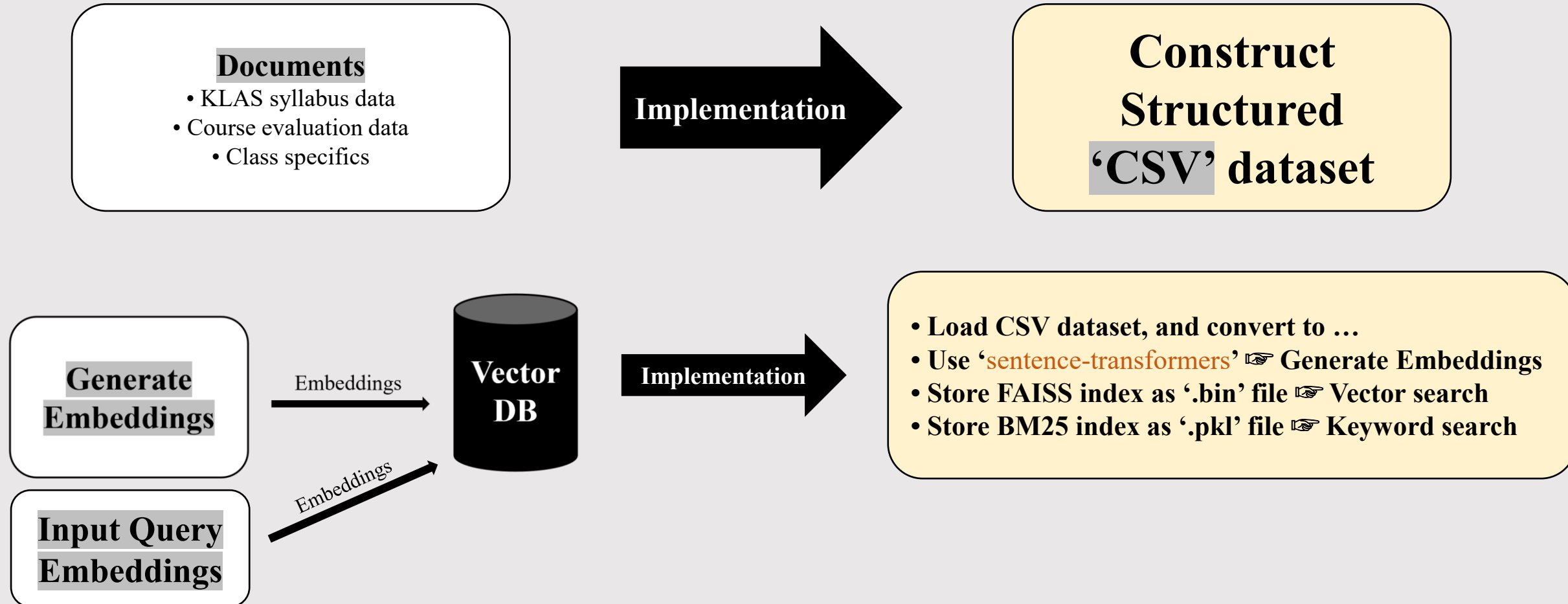


Current Status

- Backend system
- Frontend system
- Connection system

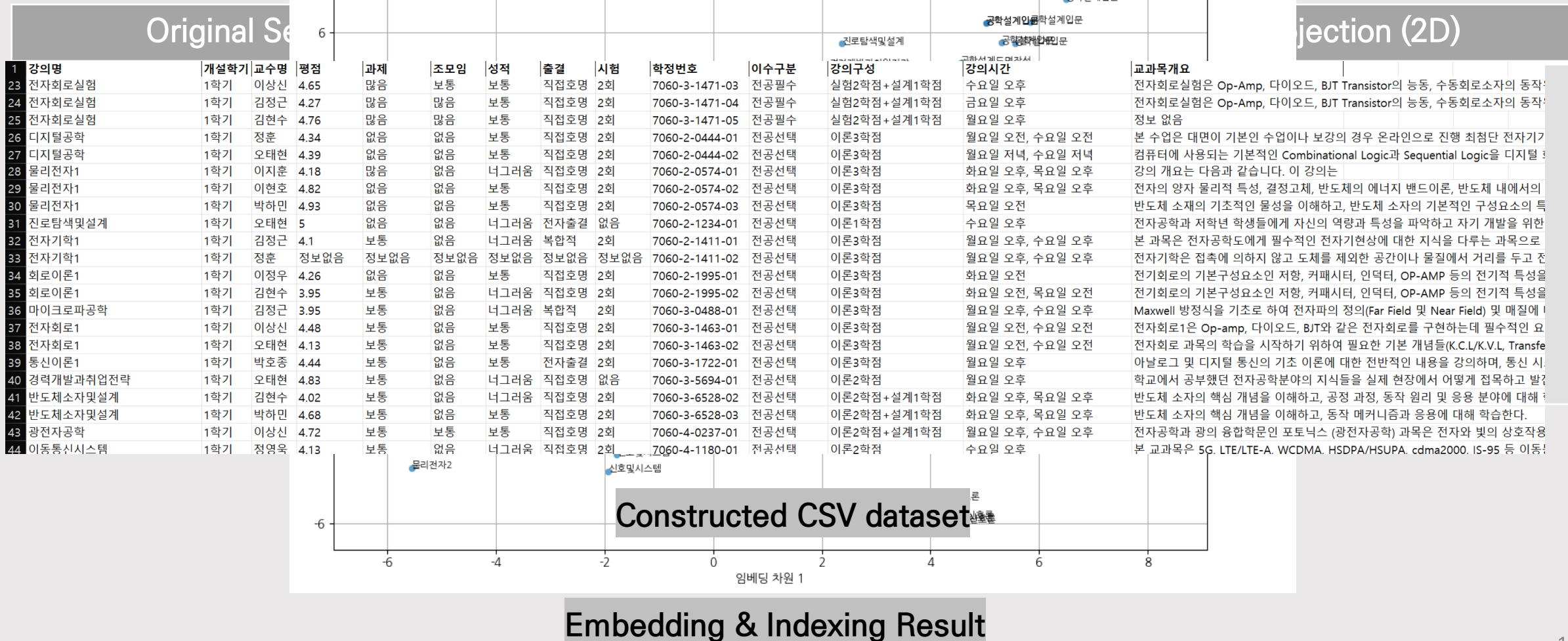
③ Current Status

(1) Backend system



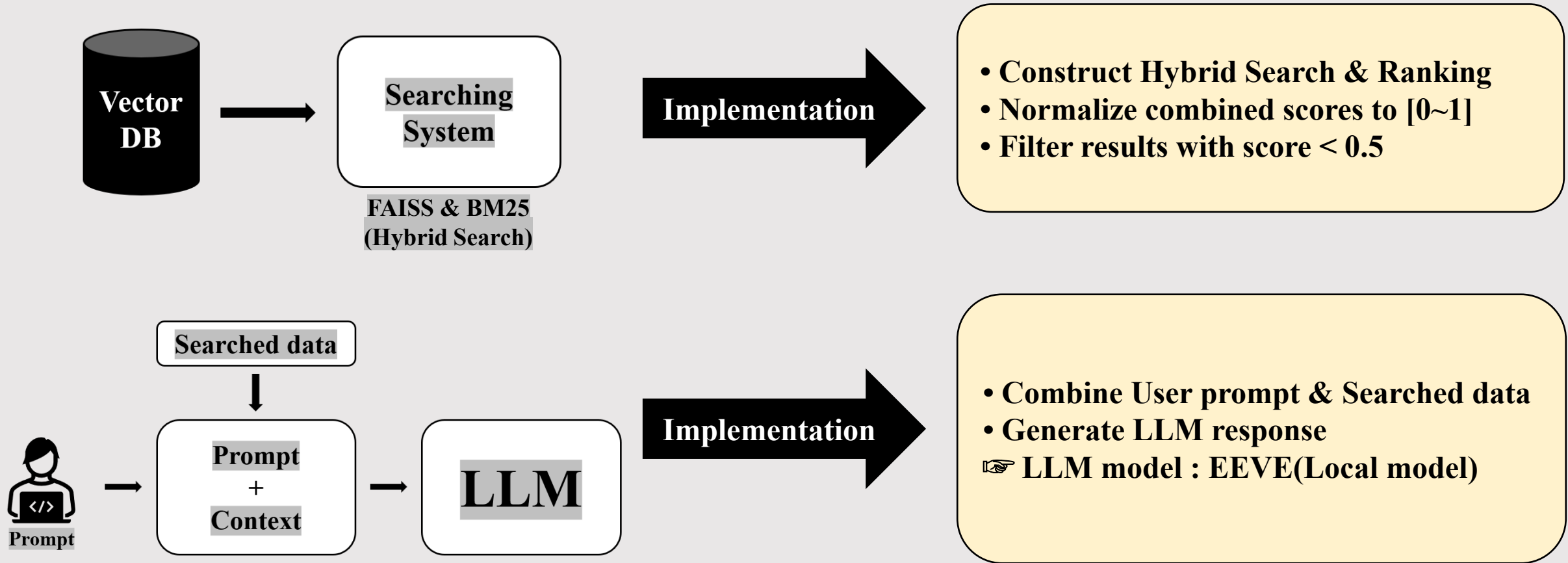
③ Current Status

(1) Backend system : Implementation



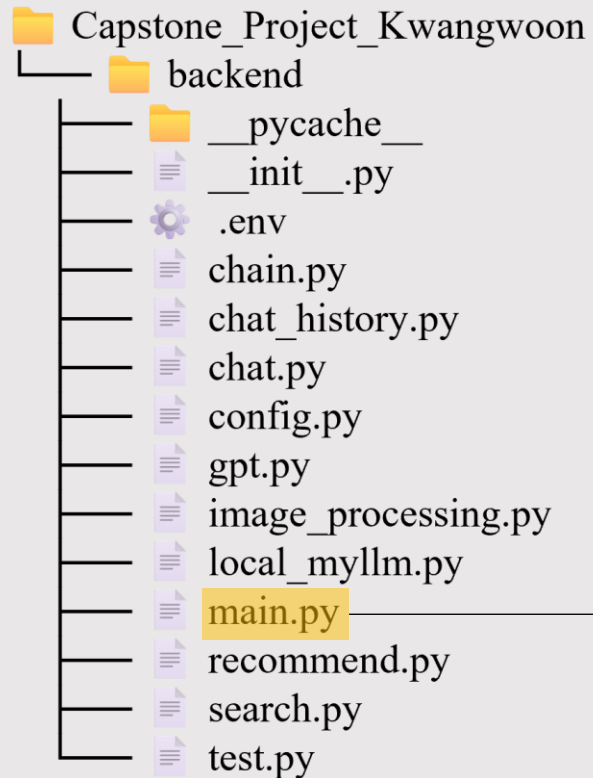
③ Current Status

(1) Backend system



③ Current Status

(1) Backend system : Implementation



Backend Architecture

```
from fastapi import FastAPI
from fastapi.staticfiles import StaticFiles
from fastapi.middleware.cors import CORSMiddleware
from backend.chain import chat_router
from backend.search import search_router
from backend.local_myllm import llm_router
from backend.recommend import recommend_router
from backend.image_processing import image_router
# from backend.gpt import gpt_router

app = FastAPI(title="광운대학교 챗봇 API", version="1.0")

# API 엔드포인트 등록
app.include_router(chat_router, prefix="/api/chat")
app.include_router(search_router, prefix="/api/search")
app.include_router(llm_router, prefix="/api/llm")
app.include_router(image_router, prefix="/api/image")
app.include_router(recommend_router, prefix="/api/recommend")
# app.include_router(gpt_router, prefix="/api/llm")
```

main.py

③ Current Status

local_llm.py

```
@chat_router.post("/api/chat")
async def chat(query: str):
    """ 사용자의 질문을 받아서 대화 기록을 기반으로 응답 """
    try:
        chat_history = load_chat_history() # 기존 대화 기록 불러오기
        history_text = "\n".join([f"User: {item['user']}\nBot: {item['bot']}" for item in chat_history])

        # ✅ LLM(챗봇 모델)에 기존 대화 기록 포함하여 요청
        bot_response = generate_response(history_text + f"\nUser: {query}")

        add_to_chat_history(query, bot_response) # 대화 기록 저장

        return {"response": bot_response, "chat_history": load_chat_history()}

    except Exception as e:
        raise HTTPException(status_code=500, detail=str(e))
```

chat.py

main.py

search.py

```
D, I = faiss_index.search(query_vector, top_k)
tokenized_query = query.split()
bm25_scores = bm25_index.get_scores(tokenized_query)
bm25_top_scores = [bm25_scores[idx] for idx in I[0]]

faiss_norm = normalize_scores(D[0])
bm25_norm = normalize_scores(np.array(bm25_top_scores))
combined_scores = (1 - BM25_WEIGHT) * faiss_norm + BM25_WEIGHT * bm25_norm
sorted_indices = np.argsort(combined_scores)[::-1]
```

```
def generate_answer(query, retrieved_docs):
    try:
        # 1. 강의명이 정확히 입력된 경우 → csv에서 직접 검색 후 반환
        course_info = search_course_directly(query)
        if course_info:
            return f"✅ 요청한 강의 정보입니다:\n{course_info}"

        # 2. 교수명 검색 처리
        professor_names_in_query = [prof for prof in PROFESSOR_DICT.keys() if prof in query]
        if professor_names_in_query:
            professor_name = professor_names_in_query[0]
            related_courses = PROFESSOR_DICT.get(professor_name, [])

            course_descriptions = [
                f"♦ 강의명: {doc['강의명']} - {doc['교과목개요'][:150]}..."
                for doc in retrieved_docs if doc['강의명'] in related_courses
            ]

            if not related_courses:
                return f"❌ '{professor_name}' 교수님의 강의를 찾을 수 없습니다."

        # 3. 일반적인 강의 검색 (벡터 검색 활용 복원)
        context = []
        for doc in retrieved_docs:
            lecture_name = doc.get('강의명', '정보 없음')
            professor_name = doc.get('교수명', '정보 없음')
            rating = doc.get('평점', '정보 없음')
            assignment = doc.get('과제', '정보 없음')
            attendance = doc.get('출결', '정보 없음')
            exam = doc.get('시험', '정보 없음')
            summary = doc.get('교과목개요', '정보 없음')

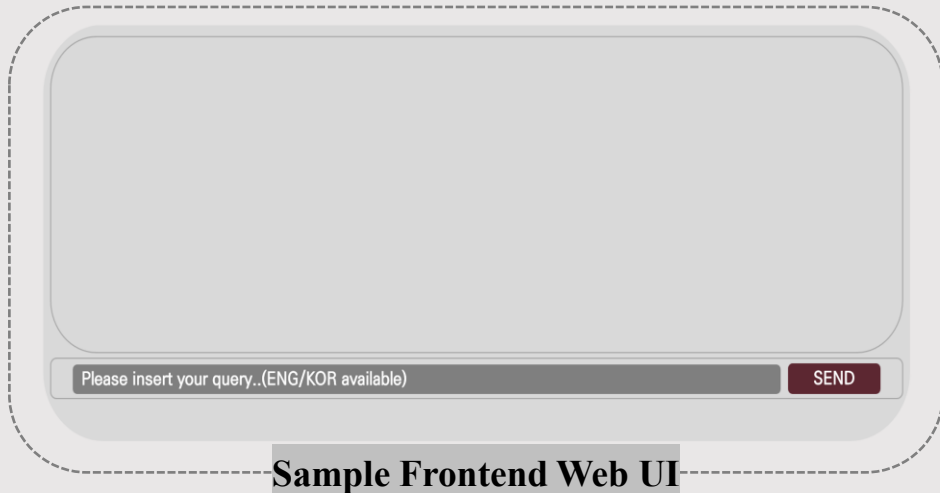
            details = []
            if rating != "정보 없음":
                details.append(f"평점은 {rating}점입니다.")
            if assignment != "정보 없음":
                details.append(f"과제는 {assignment}입니다.")
            if attendance != "정보 없음":
                details.append(f"출결은 {attendance} 방식입니다.")
            if exam != "정보 없음":
                details.append(f"시험은 {exam}로 진행됩니다.")

            details_text = " ".join(details) if details else "추가적인 정보가 제공되지 않았습니다."

            lecture_info = (
                f"강의명: {lecture_name}, 교수명: {professor_name}, {details_text} "
                f"강의 개요: {summary[:150]}..."
            )
            context.append(lecture_info)
```

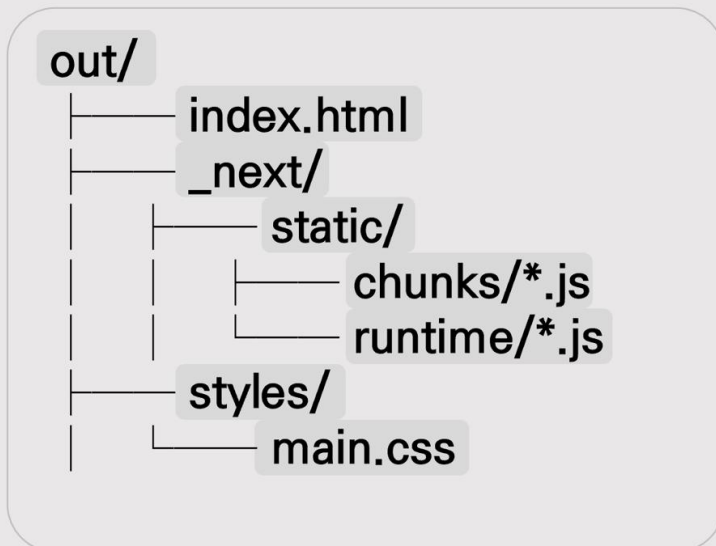
③ Current Status

(2) Frontend system, Connection between Frontend ↔ Backend



Implementation

- Chat interface 🖱️ React & TypeScript
- Text queries available
- Linked with backend API

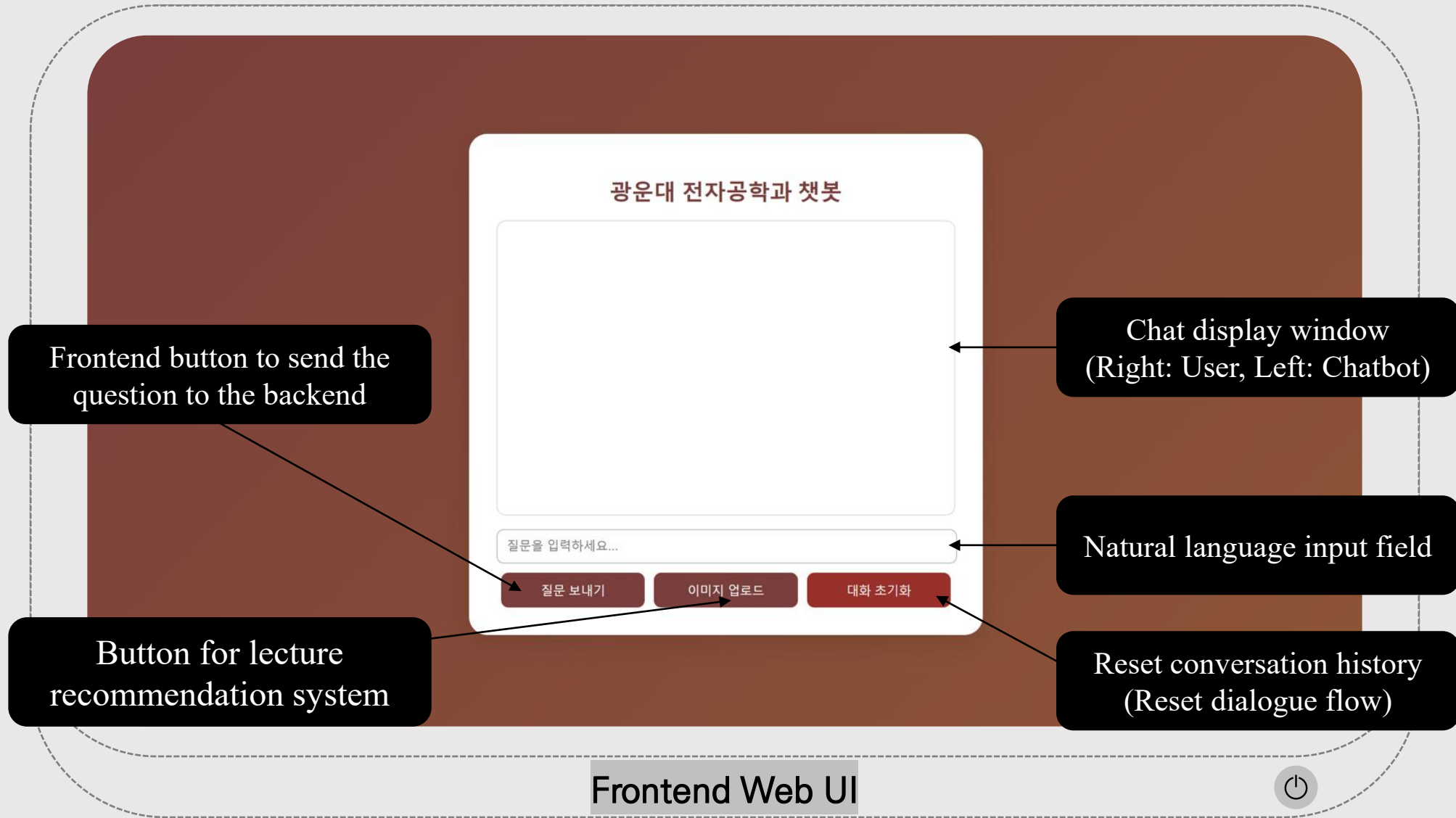


Static Asset Structure Generated by next export in Next.js

Implementation


- Frontend(React) sends HTTP requests
 - ▶▶▶ to backend(FastAPI)
- Ngrok exposes local backend to the Web
 - ▶▶▶ Using public tunnel URL


(3) Implementation Result



(3) Implementation Result

광운대 전자공학과 챗봇

 **사용자:** 통신이론2 과목에 대해 알려주세요,

 **챗봇:** 통신이론2 과목은 광운대학교 전자공학과에서 배우는 필수 과목으로 디지털 통신 기술을 다룹니다. 이 과목을 가르치는 교수님들은 박호종님과 채성호님이며, 두 강의 모두 평점이 매우 좋으며 평균 4.68점과 4.57점을 유지하고 있습니다. 이 수업에서는 통신이론1에서 배운 신호 처리 기술과 아날로그 통신 개념을 활용하여 디지털 통신을 심층적으로 다룹니다. 변복조 방법을 배우고 여러 방법들을 비교 분석하며, 잡음에 강한 수신기의 구조와 그 해석에 사용되는 수학적 기술에 대해 배웁니다. 또한, 디지털 신호의 Baseband 및 Passband 전송 기법과 같은 주제도 다루며, 이러한 개념을 실습하기 위한 프로그래밍 과정도 포함되어 있습니다. 강의 방식으로는 '직접 호명' 출


질문을 입력하세요...

질문 보내기

이미지 업로드

대화 초기화

광운대 전자공학과 챗봇

 **챗봇:** 통신이론2 과목은 광운대학교 전자공학과에서 배우는 필수 과목으로 디지털 통신 기술을 다룹니다. 이 과목을 가르치는 교수님들은 박호종님과 채성호님이며, 두 강의 모두 평점이 매우 좋으며 평균 4.68점과 4.57점을 유지하고 있습니다. 이 수업에서는 통신이론1에서 배운 신호 처리 기술과 아날로그 통신 개념을 활용하여 디지털 통신을 심층적으로 다룹니다. 변복조 방법을 배우고 여러 방법들을 비교 분석하며, 잡음에 강한 수신기의 구조와 그 해석에 사용되는 수학적 기술에 대해 배웁니다. 또한, 디지털 신호의 Baseband 및 Passband 전송 기법과 같은 주제도 다루며, 이러한 개념을 실습하기 위한 프로그래밍 과정도 포함되어 있습니다. 강의 방식으로는 '직접 호명' 출

질문을 입력하세요...

질문 보내기


이미지 업로드


대화 초기화

- A Q&A service for major courses
- Example query: “Tell me about Communication Theory 2”
- Answer generated by the LLM via RAG system

(3) Implementation Result

광운대 전자공학과 챗봇

 사용자: 정영욱 교수님 수업에 대해 알려주세요.

 챗봇: 안녕하세요! 정영욱 교수님이 담당하는 강의는 이동통신시스템과 데이터통신 두 과목입니다. 1. 이동통신시스템: 이 과목을 통해 학생들은 셀룰러 및 무선 네트워크의 원리, 설계, 성능 측정에 대해 배우게 됩니다. 주된 내용은 LTE(Long Term Evolution) 기술, 선행 코드 분할 다중 접속(CDMA), FDD(Frequency Division Duplexing), TDD(Time Division Duplexing) 등입니다. 학생들은 이동통신 네트워크의 설계 및 분석 기법을 배우고, 모바일 트래픽을 최적화하는 방법을 이해하게 됩니다. 2. 데이터통신: 이 과목은 디지털 통신 시스템과 프로토콜에 중점을 둡니다. 주로 디지털 신호 처리, 데이터 압축, 오류 제어 등에 대해 다룹니다. 학생들은 TCP/IP 프로토콜 스위트, 이더넷 네트워크, 광대역 통신을 비롯한 다양한 통신

질문을 입력하세요...

질문 보내기 이미지 업로드 대화 초기화

광운대 전자공학과 챗봇

이 과목을 통해 학생들은 셀룰러 및 무선 네트워크의 원리, 설계, 성능 측정에 대해 배우게 됩니다. 주된 내용은 LTE(Long Term Evolution) 기술, 선행 코드 분할 다중 접속(CDMA), FDD(Frequency Division Duplexing), TDD(Time Division Duplexing) 등입니다. 학생들은 이동통신 네트워크의 설계 및 분석 기법을 배우고, 모바일 트래픽을 최적화하는 방법을 이해하게 됩니다. 2. 데이터통신: 이 과목은 디지털 통신 시스템과 프로토콜에 중점을 둡니다. 주로 디지털 신호 처리, 데이터 압축, 오류 제어 등에 대해 다룹니다. 학생들은 TCP/IP 프로토콜 스위트, 이더넷 네트워크, 광대역 통신을 비롯한 다양한 통신 기술을 배우고, 이러한 시스템의 설계 및 성능 평가 방법을 익힙니다. 요약하자면, 정영욱 교수님의 강의는 이동통신 및 데이터통신 영역에 중점을 두고 있으며, 관련 기술과 프로토콜을 이해하고 분석하는 데 필요한 지식을 학생들에게 제공합니다.

질문을 입력하세요...

질문 보내기 이미지 업로드 대화 초기화

- A Q&A test for professor-specific courses
- Example query: “Tell me about Professor Jung’s classes”
- Answer generated by the LLM via RAG system

(3) Implementation Result

광운대 전자공학과 챗봇

사용자: 전자회로2

챗봇: ✅ 요청한 강의 정보입니다: 강의명: 전자회로2
교수명: 이상신 이수구분: 전공선택 평점: 4.45 과제: 보통
출결: 직접호명 시험: 2회 강의 개요: 전자회로2는 BJT (Bipolar Junction Transistor)와 FET (Field Effect Transistor)를 이용하여 다양한 형태의 증폭기, 논리회로, 스위치 등을 설계하고 해석하는 학문이다. 따라서 트랜지스터의 기본적인 특성, 바이어스, 증폭기 ...

질문을 입력하세요...

질문 보내기 이미지 업로드 대화 초기화

광운대 전자공학과 챗봇

사용자: 디지털공학

챗봇: ✅ 요청한 강의 정보입니다: 강의명: 디지털공학
교수명: 정훈 이수구분: 전공선택 평점: 4.34 과제: 없음
출결: 직접호명 시험: 2회 강의 개요: 본 수업은 대면이 기본인 수업이나 보강의 경우 온라인으로 진행 최첨단 전자기기 내 디지털 칩의 Processor, memory 등 - 을 구성하는 논리회로의 이론과 설계 방법을 배우고, 터득한 지식을 통해 digital system을 구성해본다. 첫째로, 다양한 지능...

질문을 입력하세요...

질문 보내기 이미지 업로드 대화 초기화

- Input/output test for exact course name search (e.g., *Electronic Circuits 2*, *Digital Engineering*)
- This is a simple lookup, not based on the RAG system



Future Works

- Problems & Solutions
- Improvement plan

► Problems & Solutions

<u>Problems</u>	<u>Solutions</u>
<ul style="list-style-type: none">• Low modularity• Complex structure	<ul style="list-style-type: none">• Improved modularity with package structure• Implement ‘__init__.py’ in backend
<ul style="list-style-type: none">• Unstable Cloudflared tunnel	<ul style="list-style-type: none">• Switched to ngrok tunnel• Built static frontend
<ul style="list-style-type: none">• API fetch failed due to CORS policy	<ul style="list-style-type: none">• Added CORS middleware• So, cross-origin access allowed

► Improvement Plan

<u>Retrieval part</u>	<u>Backend part</u>	<u>Frontend part</u>
<ul style="list-style-type: none">• Retrieval algorithm improvement<ul style="list-style-type: none">• Query expansion• BM25 & FAISS optimization• Evaluation using IR metrics	<ul style="list-style-type: none">• Integration of GPT-3.5 Turbo API(LLM)• Store search, chat history 📁 for LLM• Prompt engineering optimization• Implement Timetable recommend system	<ul style="list-style-type: none">• Unified UI Design• Time Selection Functionality• Enhanced Chat History UX• UX/UI Improvements



Schedule

- Timetable

⑤ Schedule

► Timetable

 : Finished  : Planned

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



References

- Academic references

⑥ 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 our presentation

KwangWoon Univ.

Dept. of Electronic Engineering

2025 Capstone Design Midterm Presentation

