

# 중급 프로젝트 최종 발표

---

4팀 (RAGnRoll)



고인범  
유준영  
이종서  
공지연

# Contents

---

- Project Overview
- EDA
- Data Preprocessing
- Baseline & Experiments
- RAG Architecture
- Prompt Engineering
- Evaluation
- Wrap Up



# Project Overview

# Project Overview

---

- **목표:** B2G 입찰지원 스타트업 *입찰메이트*를 위한 사내 RAG 시스템 구축
  - 수백 건의 RFP를 자동 검색 → 요약 → 질문응답 → 비교 → 추천 까지 지원
  - 슬로건: **“입찰부터 낙찰까지!”**
- **핵심 가치**
  - 컨설턴트가 **핵심 정보(예산·기간·요구사항·평가기준)** 를 신속히 파악
  - **맞춤형 추천/비교**로 의사결정 속도 향상
  - 제안서 작성 보조로 **시간·비용 절감**

EDA

# Metadata & EDA

---

### 컬럼명: '입찰 참여 시작일' 결측치

- 데이터 타입: object
- 결측치 (NaN): 26개 (26.0%)
- 고유 값 개수: 73개
- 샘플 데이터: ['2024-10-14 10:00:00', '2024-08-29 09:00:00', '2024-05-02 10:00:00', '2024-04-26 09:00:00', '2025-01-08 14:30:00'] 등

-----

### 컬럼명: '입찰 참여 마감일'

- 데이터 타입: object
  - 결측치 (NaN): 8개 (8.0%)
  - 고유 값 개수: 87개
  - 샘플 데이터: ['2024-10-15 17:00:00', '2024-10-16 14:00:00', '2024-09-09 10:00:00', '2024-05-09 16:00:00', '2024-04-30 17:00:00'] 등
-

# Metadata & EDA

● 중복된 '사업명'을 가진 사업 목록:

	공고 번호	공고 차수	사업명	사업 금액	발주 기관	공개 일자	입찰 참여 시작 일	입찰 참여 마 감일	사업 요약	파일 형식	파일명	텍스트
15	NaN	NaN	통합정보시스템 고도화 응역	140000000.0	국가과학기술지식정보 서비스	2024-05-30 00:00:00	2024-05-30 00:00:00	2024-06-11 00:00:00	- 사업 개요: 통합정보시스템 고도화 응역, 사업기간 5개월 이내, 추정가 격 140...	hwp	국가과학기술지식정보서비스_통합정보시스템 고도화 응역.hwp	\r\n\r\n제안요청서\r\n\r\n\r\n통합정보시스템 고도 화 응역\r\n\r\n제...
53	20240535775	0.0	통합정보시스템 고도화 응역	140000000.0	한국한의학연구원	2024-05-30 09:04:12	2024-05-30 10:00:00	2024-06-11 11:00:00	- 사업 개요: 통합정보시스템 고도화 응역으로 기관생명윤리, 동물실험윤 리, 국가연구...	hwp	한국한의학연구원_통합정보 시스템 고도화 응역.hwp	\r\n\r\n제안요청서\r\n\r\n\r\n통합정보시스템 고도 화 응역\r\n\r\n제...

중복된 사업명

# Metadata & EDA

---

텍스트

```
0      \n \n2024년 특성화 맞춤형 교육환경 구축 - 트랙운영 학사정보시스템 ...
1      \r\n \r\n \r\n \r\n제 안 요 청 서\r\n[ 2024년 대학 ...
2      \r\n      \r\nEIP3.0 고압가스 안전관리\r\n시스템 구축 용역...\
3      \r\n \r\n\r\n도시계획위원회 통합관리시스템 구축\r\n제 안 요 청...
4      \r\n \r\n \r\n제안요청서\r\n \r\n사 업 명\r\n봉화...
```



# 메타데이터 EDA - 결론

---

- **결측치** : 존재 → 있는 그대로 유지 (추후 보강 로직 계획)
- **중복 사업명** : 일부 존재 → 컬렉션/파일명/광고번호 조합으로 구분.
  - 추가 확인 결과, **발주기관만 다른 동일한 사업**임을 확인
- **개행 문자(\n, \r)** : 다수 발견 → 전처리 시 정규화

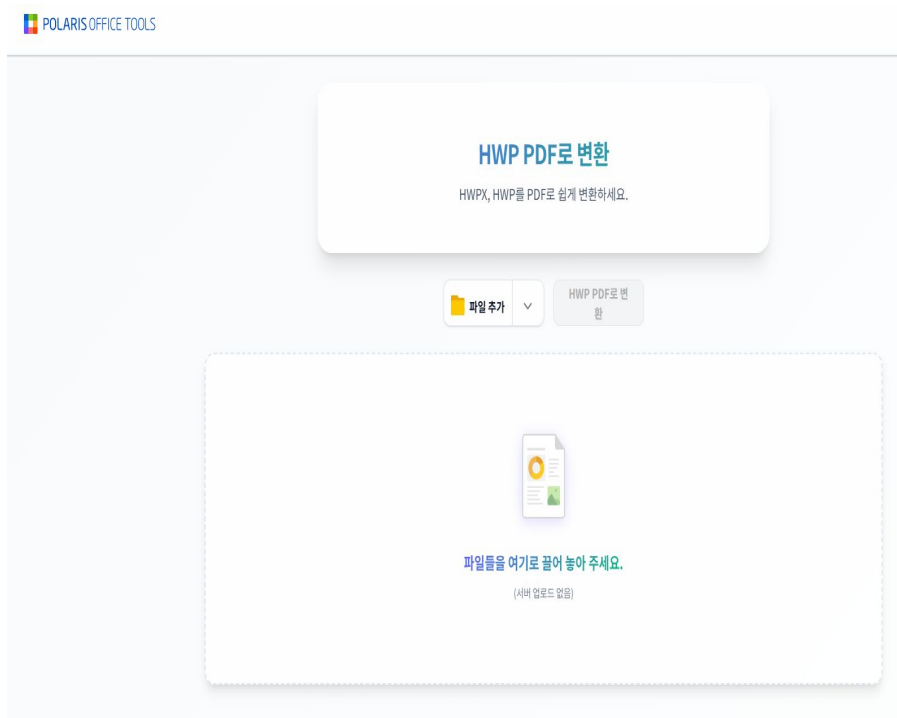
# Data Preprocessing

# 원본 데이터 특성(HWP/PDF)

---

- 이미지: 동일 / 유사 이미지 반복 → 텍스트 위주로 처리(이미지 제외)
- 추출 도구: **pdfplumber** 채택 (일관성 / 품질 / 속도 균형)
- 원본 확인 루프: Retrieval 성능 이상 시 원문 재검증
  - (특히 국민연금공단 / 케빈랩 / 보건산업진흥원 / 고려대 사례)
- 표 데이터: 테이블 구조 없이 텍스트로만 추출해도 성능에 큰 차이가 없었음
  - → 일반 텍스트로 처리

# 데이터 처리 - 1차



HWP 파일 96개, PDF 파일 4개

POLARIS OFFICE를 통해  
HWP → PDF 변환

# 데이터 처리 - 최종 채택

---

## 단순 전처리 + Character Split 기반 청킹 (기본 전략으로 회귀)

- 전처리 전략: HWP 파일(96개) → PDF 파일로 변환 (Polaris Office) + PDF 파일 4개 = 총 100개 PDF 문서
- 청킹 전략: 후보 실험: 폰트 크기 기반 분할, 단순 Character Split, 테이블을 Markdown 형식으로 변환, 동적 폰트 크기 기반
- 추출 파이프라인 (베이스라인 기준)
  1. PDF 통일 → 2) **pdfplumber** 파싱 → 3) 텍스트 / 테이블 단순 분리 → 4) Recursive Character Text Split 청킹 → 5) 임베딩 & 메타데이터 저장

# 데이터 청킹 예시



2024년 이러닝시스템 운영 용역 제안요청서

## I 사업 안내

### 1. 사업개요

- ☐ 사업명: 『2024년 이러닝시스템 운영』 용역
- ☐ 사업기간: 계약체결일로부터 2025. 2월까지
- ☐ 사업예산

사업구분	사업예산
계	금 773,801천원
·공단 소유 콘텐츠·외부콘텐츠 등 교육과정 운영	금 620,801천원
·콘텐츠 개발·관리, 학습관리시스템 및 각종 서버 임대 등	금 153,000천원

※ 입찰 참여업체는 가격 제안 시 상기 2개 사업부문별 예산을 초과할 수 없음

### 2. 사업목적

- ☐ 비대면(온택트) 교육수요에 따른 니즈 충족
- ☐ 직급별 필요 역량진단 및 맞춤형 교육 제공으로 직원 역량 강화  
\* (상반기) 사전 역량진단 / (하반기) 사후 역량진단, 진단 문항은 공단에서 제공
- ☐ 현업 문제해결 능력 향상을 위한 직무 및 부서별 콘텐츠 개발
- ☐ 이러닝 시스템 운영을 통한 상시 학습 활성화·효율화

### 3. 사업범위

- ☐ 직무 콘텐츠 개발·운영 및 위탁교육(자기개발) 콘텐츠 운영
- ☐ 사이버(모바일)연수원 구축·운영 및 학습관리시스템 서버 임대
- ☐ 공단 콘텐츠 및 프로그램의 수정·보완 등 안정적 관리와 운영  
\* 운영에는 정보보안 및 개인정보보호 방안 및 법령준수 등이 포함됨



```
---- 8번째 chunk ----
메타데이터 (parent_header): I 사업 안내
###
1. 사업개요
□ 사업명 : 『 2024 년 이러닝시스템 운영 』 용역
□ 사업기간 : 계약체결일로부터 2025. 2 월까지
□ 사업예산
사업구분 사업예산
계 금 773,801천원
·공단 소유 콘텐츠·외부콘텐츠 등 교육과정 운영 금 620,801천원
·콘텐츠 개발·관리, 학습관리시스템 및 각종 서버 임대 등 금 153,000천원
※입찰 참여업체는 가격 제안 시 상기 2개 사업부문별 예산을 초과할 수 없음
2. 사업목적
□ 비대면 ( 온택트 ) 교육수요에 따른 니즈 충족
□ 직급별 필요 역량진단 * 및 맞춤형 교육 제공으로 직원 역량 강화
* (상반기) 사전 역량진단 / (하반기) 사후 역량진단, 진단 문항은 공단에서 제공
□ 현업 문제해결 능력 향상을 위한 직무 및 부서별 콘텐츠 개발
□ 이러닝 시스템 운영을 통한 상시 학습 활성화 . 효율화
3. 사업범위
□ 직무 콘텐츠 개발 . 운영 및 위탁교육 ( 자기개발 ) 콘텐츠 운영
---- 9번째 chunk ----
메타데이터 (parent_header): I 사업 안내
###
3. 사업범위
□ 직무 콘텐츠 개발 . 운영 및 위탁교육 ( 자기개발 ) 콘텐츠 운영
□ 사이버 ( 모바일 ) 연수원 구축 . 운영 및 학습관리시스템 서버 임대
□ 공단 콘텐츠 및 프로그램의 수정 . 보완 등 안정적 관리와 운영 *
* 운영에는 정보보안 및 개인정보보호 방안 및 법령준수 등이 포함됨
```

# 데이터 처리

---

- 의사결정 배경

- 한 차례는 최종 추출 파이프라인(테이블 Markdown 변환 + 동적 헤더 기반 청킹)을 적용했으나, 성능 저하가 발생.
- 성능 저하 원인 추정: 테이블을 Markdown으로 바꾸고, 동적 헤더를 메타데이터에 포함시키는 과정에서 **semantic하지 않은 헤더 정보(특수문자, 의미 없는 텍스트)**가 많이 유입됨. 이로 인해 검색 및 임베딩 품질이 떨어진 것으로 판단됨.
- 따라서 최종적으로는 단순 전처리 및 **Character Split** 청킹 전략(베이스라인)이 가장 안정적이라고 결론 내림.
- 결론적으로, 데이터 전처리나 청킹을 통한 **RAG** 성능 향상보다는, **retriever**와 **generator**의 성능 개선을 통해 **RAG** 시스템 고도화를 목표로 삼음

참고: 실험 기록 및 의사결정 근거는 노션 문서(데이터 전처리 / 청킹 전략)에 상세화.

# Baseline & Experiments



# 베이스라인과 실험 기록

---

- 초기 계획: 단일 RAG 파이프라인(벡터검색 + 단순 QA)으로 시작
- 개선 흐름
  - 지연 : k값 튜닝, **Re-ranking** 도입으로 빠른 성능 향상
  - 종서 : 텍스트 추출 다양화, **Hybrid Search**, **Multi-turn** 실험
  - 인범 : 기본 기능 확보한 베이스라인 RAG 구현, 평가 루프 정비
  - 준영 : PymuPDF, hwp5html, pyhwp와 pymupd4llm 등을 사용한 전처리, EasyOCR 이용한 이미지 처리에 집중

# 베이스라인과 실험 기록

---

- Baseline 구조를 기준으로 기능 추가하며 실험 진행

	Faithfulness	Answer Similarity	Context Recall	Context Precision
Baseline	0.6321	0.5462	0.6200	0.4894
Baseline + Re-Ranking	0.7940	0.5485	0.5800	0.5033
Baseline + Routing	0.7489	0.5570	0.6000	0.6316
Baseline + Multi Retrieval	0.5753	0.4900	0.6400	0.6800

# 베이스라인과 실험 기록

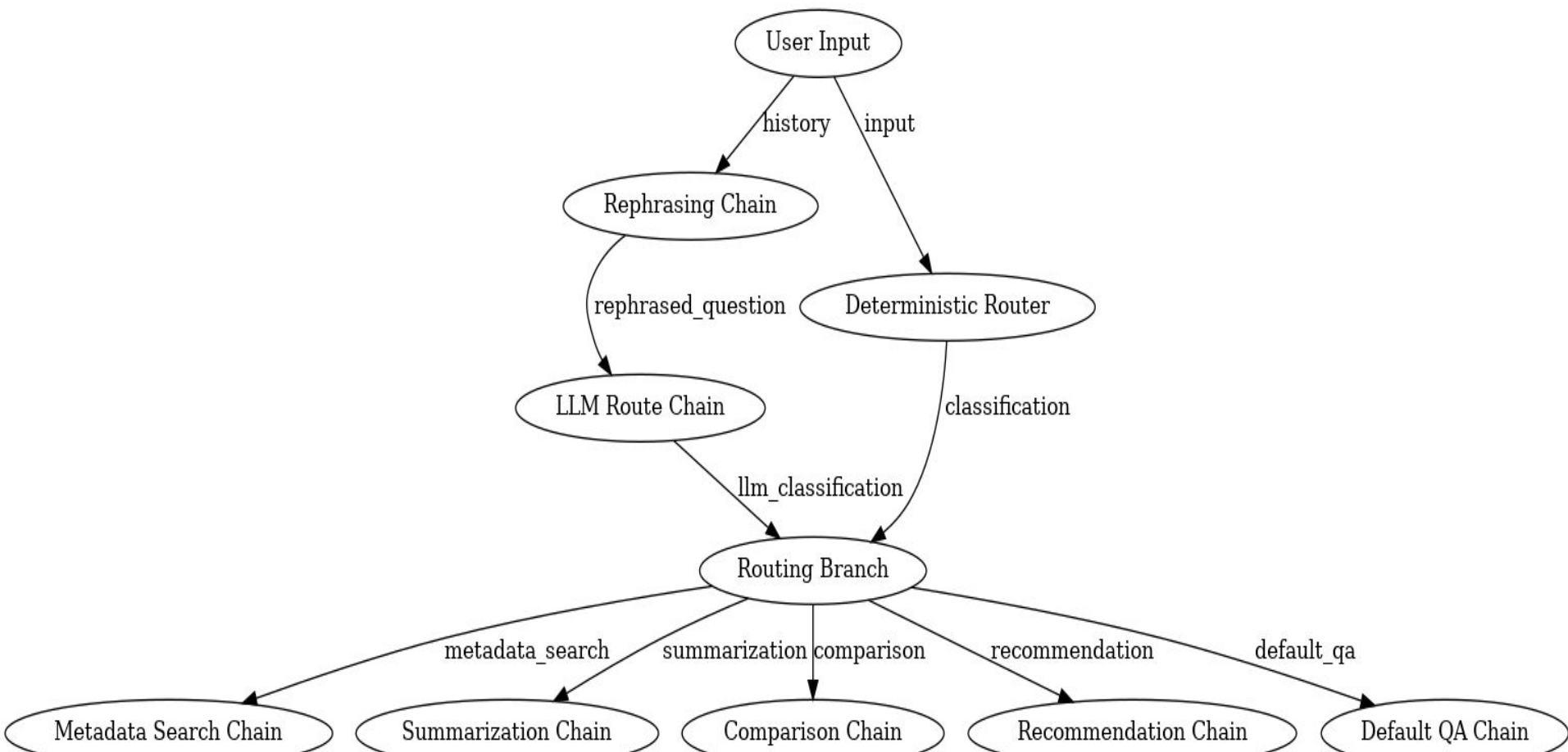
---

- 주요 결정

- 일부 라이브러리 호환 문제 및 Mac/Window 간의 문제로 인해 Linux 환경으로 통일
- 파이썬 `venv`로 통일(Conda 미사용)
- Git main 중심 잦은 머지 → 충돌 감소 / 속도 향상
- HWP는 일괄 PDF 변환 후 일관 파이프라인 적용

# RAG Architecture

# 전체 다이어그램



# Vector Store

- **FAISS** → **Chroma** 전환: 메타데이터 필터/업데이트 편의성
- 임베딩 : OpenAI Embeddings (Config에서 모델 주입: )
  - EMBEDDING\_MODEL = "text-embedding-3-small"

Model	Usage
text-embedding-3-small	\$0.00002 / 1K tokens
text-embedding-3-large	\$0.00013 / 1K tokens
ada v2	\$0.00010 / 1K tokens

Eval benchmark	ada v2	text-embedding-3-small
MIRACL average	31.4	44.0
MTEB average	61.0	62.3

# Rephrased Query를 통한 멀티턴

---

- 프롬프트를 통해 주어진 대화 이력을 기반으로 입력받은 쿼리를 검색에  
용이한 독립적인 질문으로 재구성
- 세션의 이전 대화 맥락을 반영하여 쿼리 강화, 멀티턴 구현

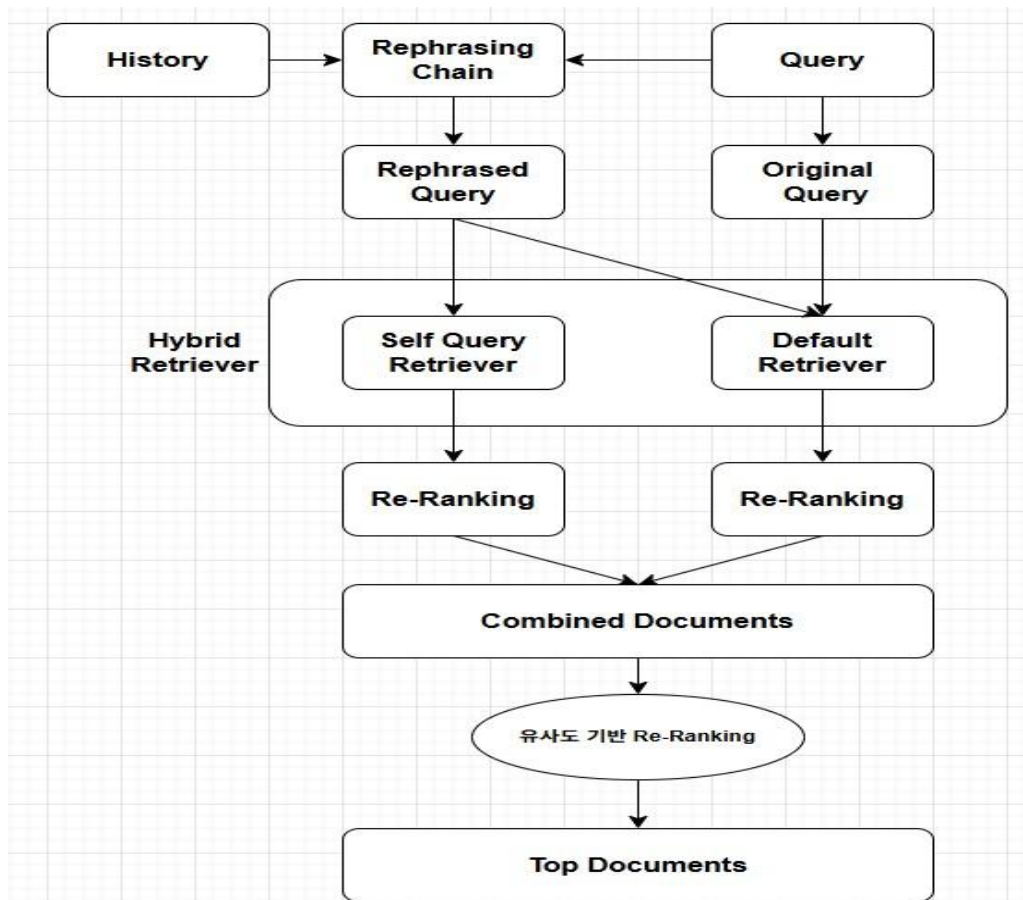
# Hybrid Retrieval / Re-Ranking(문맥 압축)

---

- 2개의 Retriever를 통해 검색 성능 극대화, 멀티턴 구현
- **SelfQueryRetriever, Default Retriever**에 각각 **HuggingFaceCrossEncoder + ContextualCompressionRetriever**를 적용해 리랭킹 과정으로 1차 검색 과정에서 **정확도** , **노이즈**  → faithfulness/answer\_correctness 개선에 기여
  - RERANK\_MODEL = 'cross-encoder/ms-marco-MiniLM-L-6-v2'
  - **SelfQueryRetriever + Chroma(metadata filter)**: 메타데이터 필터링을 통해 쿼리 강화, 정밀 검색
  - **Default Retriever(MMR)**: 의미기반 유사도 검색(다양성/중복 억제)
- **Hybrid Retrieval**: (원문 질문 + 재구성 질문 + 이전 대화 맥락) 결과 통합·중복 제거 후 **score** 기반으로 상위 5개 필터링



# Hybrid Retrieval / Re-Ranking(문맥 압축)

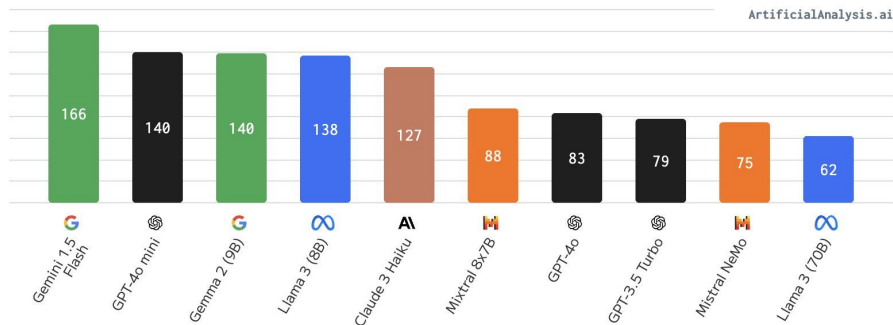


# Router 기반 RAG

- LLM\_MODEL = "gpt-4o-mini"

## Output Speed

Output Tokens per Second; Higher is better



## MMLU vs. Price, Smaller models

MMLU: General reasoning quality benchmark, Price: USD per 1M Tokens



# Router 기반 RAG

---

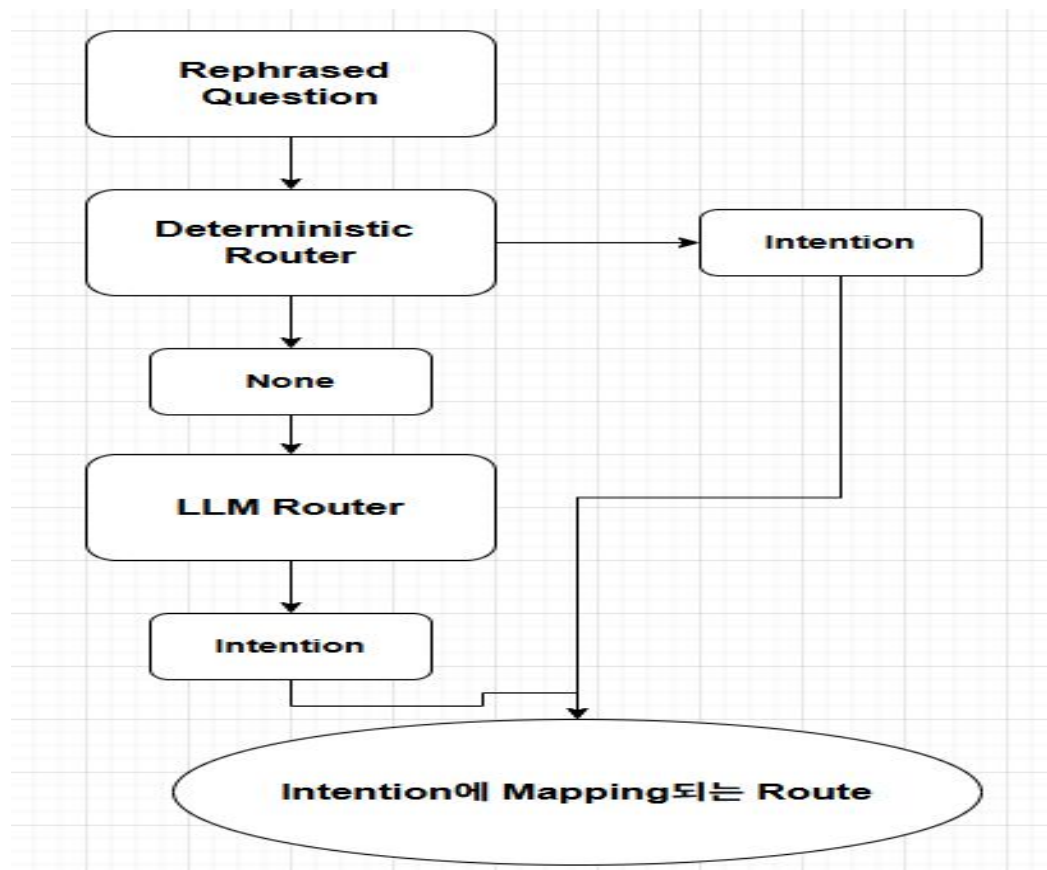
- 컨설턴트의 입장에서 해당 **RAG** 시스템을 이용할 때, 어떤 의도(intention)가 있을 것이라고 가정
  - 입력한 쿼리의 의도를 추출하고, 이에 맞춰 적합한 검색/생성 방법을 구현
- **채택한 방법론 : Router**
  - 변환된 쿼리를 적절한 처리 경로나 데이터 소스로 안내하는 과정

# Router 기반 RAG

---

- **분류 카테고리**: RFP 문서를 다루는 입장에서 자주 입력하게 되는 질문 패턴을 분석
  - "국민연금공단 이러닝 사업의 예산이 얼마인가요?" → **metadata\_search**
  - "부산관광공사 사업의 핵심 과업을 요약해줘." → **summarization**
  - "고려대학교 포털 사업과 국민연금 이러닝 사업의 사업 기간을 비교해줘." → **comparison**
  - "AI 기반 콜센터 구축과 비슷한 사업을 찾아 추천해줄래?" → **recommendation**
  - "국민연금공단 RFP에서, 제안서 평가는 어떤 방식으로 진행되나요?" → **default\_qa**

# Router 기반 RAG



# metadata\_search

---

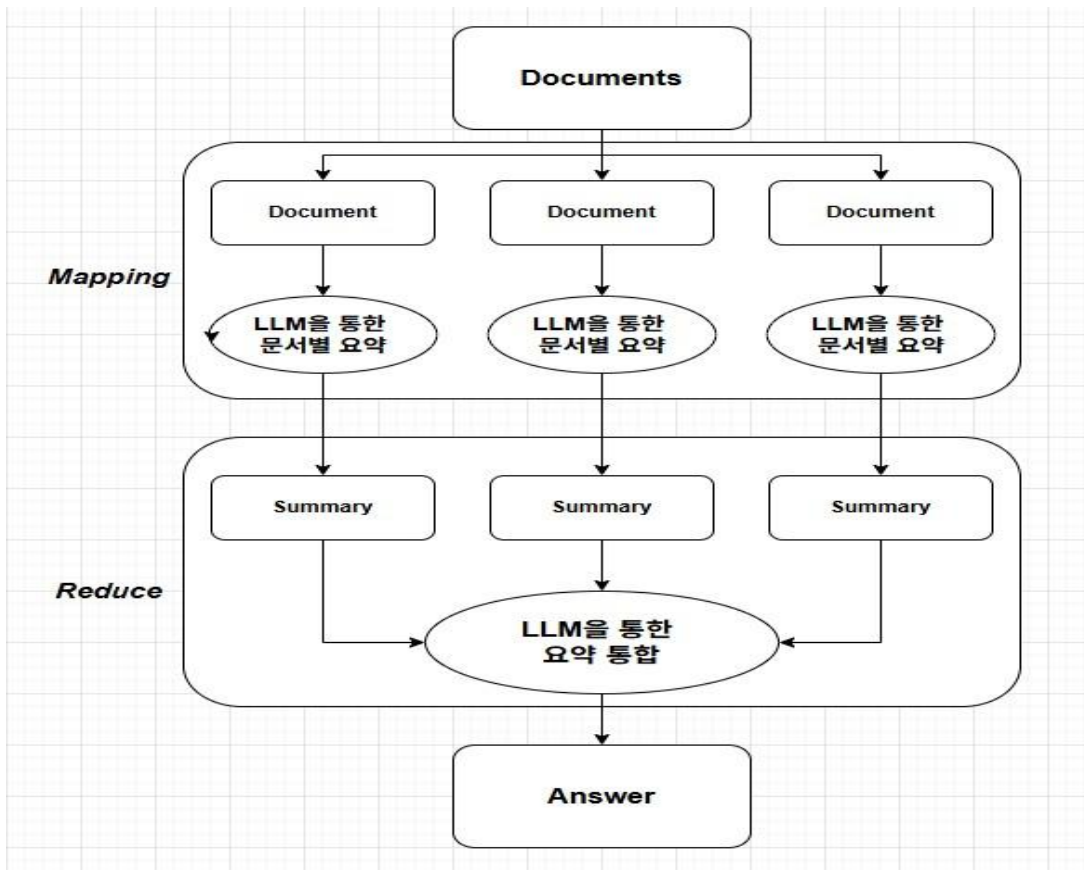
- 기관/사업명/예산/광고번호/마감 등 **정형 메타데이터 조회**
- SelfQueryRetriever로 **정밀 메타데이터 필터링** + Default Retriever로  
메타데이터의 **결측치 보완**

# summarization

---

- 문서(또는 섹션) **핵심 요약/브리핑**
- Hybrid Retriever를 통해 메타데이터 필터링을 통한 검색 성능 향상
- **Map-Reduce** 방식 적용 → 요약 성능 강화
  - Mapping: 검색한 문서에 대한 개별 요약
  - Reducing: Mapping 과정에서 도출된 각 문서에 대한 개별 요약을  
통합/요약

# summarization



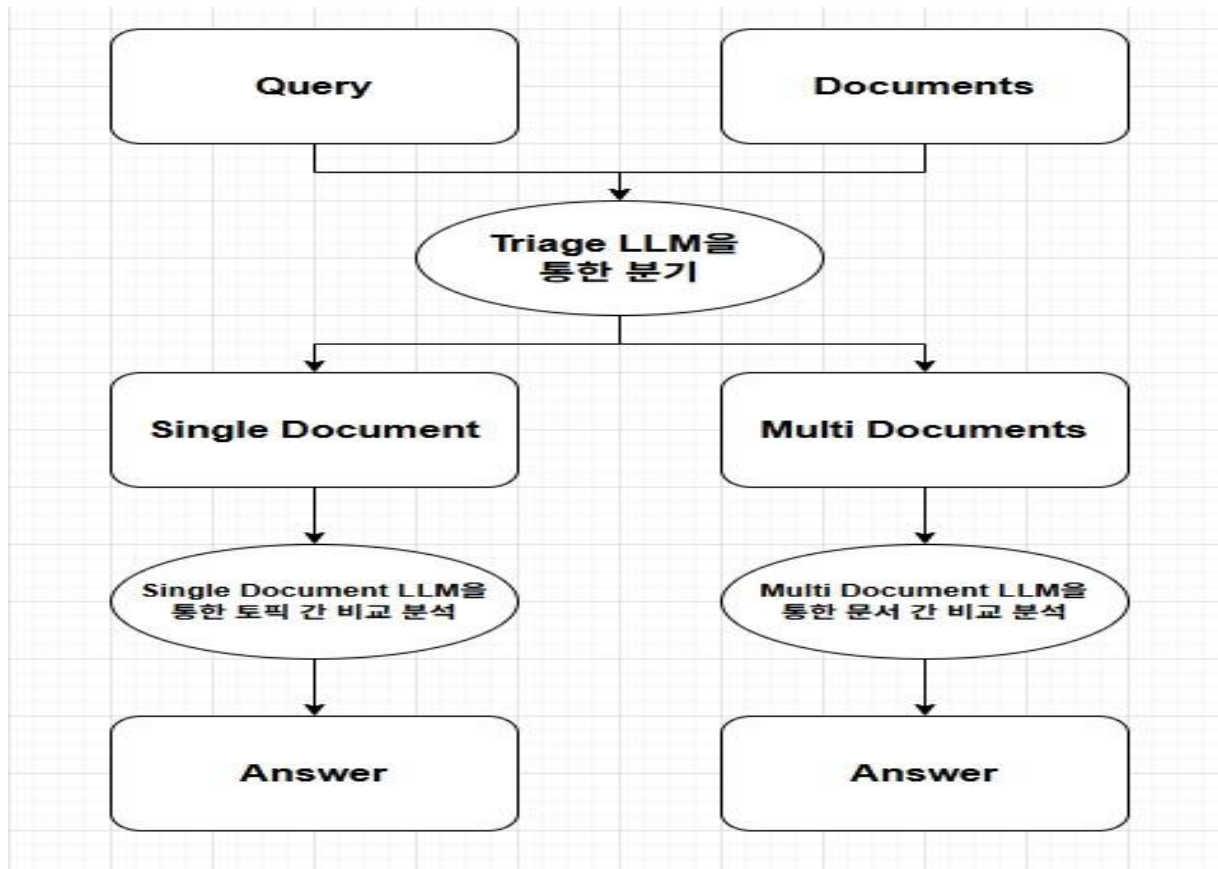


# comparison

---

- 단일 문서 내의 내용 비교/다중 문서 간 내용 비교
- Self-Query Retriever를 통해 메타데이터 필터링을 통한 검색 성능 강화
- Single Document Comparison
  - 문서 내 개념 비교: 기준 문서 컨텍스트에서 **토픽 A/B 스니펫 추출** →  
**대조**
- Multi Document Comparison
  - 문서 간 비교: **Self Query**로 각 대상 회수 후 비교 테이블 생성

# comparison



# recommendation

---

- 유사 사업 탐색/추천(확장 쿼리·의미 유사)
- Query Expansion 통해서 검색에 효과적인 키워드 추출
- 도출된 키워드를 바탕으로 Hybrid Retriever를 통해 사업 추천
- **recommendation: Query Expansion** → Hybrid Retriever 검색 → 근거 기반 추천

# default\_qa

---

- 위 4종 외 구체 질의응답, **fallback** 라우트
- Hybrid Retriever를 통해 메타데이터 필터링을 통한 검색 성능 향상

# Prompt Engineering

# 프롬프팅 전략

---

- LLM에게 기능에 따른 명확한 역할 부여
- 규칙과 작업 지침을 통해 환각(hallucination) 통제
- Few-Shot을 통해 답변 생성 품질 향상

# 프롬프트 예시 (Routing)

""당신은 사용자의 질문 의도를 정확하게 분석하여 5개의 카테고리 중 하나로 분류하는 전문가입니다.  
\*\*'원본 질문'을 통해 사용자의 최종 목표(요약, 비교, 추천 등)를 파악하세요.\*\*  
\*\*오직 아래 5개의 카테고리 이름 중 하나만! 답변해야 합니다.\*\* 다른 설명은 절대 추가하지 마세요.

# 카테고리 목록:

`metadata\_search`: 특정 조건(광고 번호, 광고 차수, 사업명, 사업 금액, 발주 기관, 공개 일자, 입찰 참여 시작일, 입찰 참여 마감일)  
`summarization`: 문서의 전체 내용, 특정 부분, 또는 핵심 요구사항 등을 요약해달라는 요청.  
`comparison`: 두 개 이상의 RFP 문서를 비교하거나, 한 문서 내의 두 가지 항목을 비교/대조해달라는 요청.  
`recommendation`: 특정 사업과 유사한 다른 사업을 추천해달라는 요청.  
`default\_qa`: 특정 RFP 문서 내용에 대한 구체적인 세부 정보를 묻는 질문. (위 4가지에 해당하지 않는 모든 질문)

# 분류 기준:

- "사업 찾아줘", "목록 알려줘" -> `metadata\_search`
- "요약해줘", "브리핑해줘", "정리해줘" -> `summarization`
- "비교해줘", "~랑 ~의 차이점 알려줘" -> `comparison`
- "추천해줘", "비슷한 사업 찾아줘" -> `recommendation`
- 그 외 특정 정보 질문 (e.g., "평가 방식은 뭐야?", "유지보수 기간 알려줘") -> `default\_qa`

# --- 분석할 질문 ---

# 원본 질문: {input}

# 검색용 질문: {rephrased\_question}


# -----

# 분류 예시 (Few-shot Examples):

질문: "국민연금공단 이러닝 사업의 예산이 얼마인가요?"

분류: metadata\_search

# 프롬프트 예시 (metadata\_search)

```
("system",  
  "당신은 대한민국 B2G(정부 대상 사업) RFP(제안요청서) 데이터베이스 검색 전문가 비디(Bidy)입니다.\n"  
  "당신은 주어진 [검색 결과] 목록을 바탕으로 사용자의 [질문]에 대한 답변을 생성해야 합니다.\n\n"  
  "**작업 지침:**\n"  
  "1. **검색 결과 확인:** [검색 결과]에 내용이 있는지 확인합니다.\n"  
  "2. **결과 기반 답변:**\n"  
  "   - **결과가 있을 경우:** \"요청하신 조건에 맞는 사업 목록입니다.\"라고 서두를 시작한 뒤, [검색 결과]에 있는 사업 목록을 빠짐없이, 순서대로 제시하세요  
  "   - **결과가 없을 경우:** \"요청하신 조건에 맞는 사업을 찾을 수 없었습니다.\"라고만 답변하세요. 다른 말을 덧붙이지 마세요.\n"  
  "3. **정보 추가 금지:** [검색 결과]에 없는 내용은 절대로 언급해서는 안 됩니다."),  
  MessagesPlaceholder("history"), #  답변 단계에서만 history 반영  
  ("human", "[질문]: {input}\n\n[컨텍스트]:\n{context}"))
```



# 프롬프트 예시 (summarization)

```
map_prompt = ChatPromptTemplate.from_template(
    "당신은 B2G 사업 전문 컨설턴트 비디(Bidy)입니다. 주어진 [문서 정보]와 [문서 본문]을 모두 참고하여, 제안 결정에 영향을 미칠 수 있는 다음 핵심 정보들을 항목별로 요약해 주십시오.
    _ **핵심 과업/요구사항:** (기술, 기능, 보안 등)\n"
    "_ **예산/기간:** (금액, 계약 기간 등)\n"
    "_ **일정:** (제안 마감일, 평가일 등)\n"
    "_ **평가방식/참여조건:** (기술/가격 배점, 필수 자격 등)\n\n"
    "--- 문서 내용 ---\n"
    "{context}\n"
    "--- 끝 ---\n\n"
    "항목별 핵심 정보 요약:"
)
```

```
reduce_prompt = ChatPromptTemplate.from_messages([
    ("system",
     "당신은 B2G 사업 수주 전략을 수립하는 수석 컨설턴트 비디(Bidy)입니다. "
     "아래에 흩어져 있는 정보들을 종합하여, 의사결정을 위한 최종 '사업 요약 브리핑'을 작성해 주십시오.\n\n"
     "***브리핑 작성 가이드라인:**\n"
     "1. **핵심 요약 (Executive Summary):** 가장 먼저 사업명, 발주기관, 예산, 기간, 핵심 기술/과업을 한두 문장으로 요약하여 제시하세요.\n"
     "2. **본문:** '사업 목표', '주요 과업 범위', '예산 및 기간', '제안 시 주요 고려사항(평가방식, 참여자격, 특이사항 등)' 순서로 구조화하여 상세히 설명하세요.\n"
     "3. **확인 필요한 정보:** 만약 예산, 기간 등 **의사결정에 필수적인 정보가 누락되었다면, 반드시 '※ 확인 필요한 핵심 정보' 항목을 만들어 명시**해야 합니다."
    ),
    # ✅ 여기서 과거 대화 이력을 반영
    MessagesPlaceholder("history"),
    ("human",
     "--- 부분 정보 목록 ---\n"
     "{context}\n"
     "--- 끝 ---\n\n"
     "최종 사업 요약 브리핑:"
    )
])
```

# 프롬프트 예시 (comparison)

```
triage_prompt = ChatPromptTemplate.from_template(
```

```
    """당신은 두 개 또는 한 개의 사업(RFP) 정보를 받아서, 주어진 기준에 따라 명확하게 비교 분석하는 전문 컨설턴트입니다.
    사용자의 비교 질문을 분석하여 'multi_document' 또는 'single_document' 유형으로 분류하고 관련 정보를 추출하세요.
    - 'multi_document': 서로 다른 두 문서를 비교. 'item_A', 'item_B'를 추출.
    - 'single_document': 하나의 문서 내에서 두 개념을 비교. 기준이 되는 'base_document'와 두 개념 'topic_A', 'topic_B'를 추출.
    - 'criteria': 비교 기준을 추출. 명확하지 않으면 "전반적인 특징"으로 설정.
    오직 JSON 객체로만 답변하세요.
```

```
    예시 1 (multi_document):
```

```
    질문: "A사업과 B사업을 비교해줘"
```

```
    JSON: [{"type": "multi_document", "item_A": "A사업", "item_B": "B사업", "criteria": "전반적인 특징"}]
```

```
    예시 2 (single_document):
```

```
    질문: "부산관광공사 사업에서 '대결 기능'과 '협조 기능'을 비교해줘"
```

```
    JSON: [{"type": "single_document", "base_document": "부산관광공사 사업", "topic_A": "대결 기능", "topic_B": "협조 기능", "criteria": "전반적인 특징"}]
```

```
    사용자 질문: {input}
```

```
    JSON 출력: ""
```

```
)
```

```
single_doc_prompt = ChatPromptTemplate.from_messages([
```

```
    ("system",
```

```
    | "당신은 한 문서 내의 두 가지 주제를 명확하게 비교 분석하는 전문가 비디(Bidy)입니다. 주어진 정보를 바탕으로 [비교 기준]에 따라 두 주제의 공통점과 차이점을 설명해야 합니다.
```

```
    MessagesPlaceholder("history"), # ☒ 답변 생성에서만 history 반영
```

```
    ("human", "***기준:** {criteria}\n\n{extracted_snippets}")
```

```
])
```

```
multi_doc_prompt = ChatPromptTemplate.from_messages([
```

```
    ("system",
```

```
    | "당신은 두 B2G 사업의 전문 비교 분석가 비디(Bidy)입니다. 각 사업의 정보를 바탕으로, 사용자가 요청한 [비교 기준]에 따라 명확하게 차이점과 공통점을 설명해야 합니다.
```

```
    MessagesPlaceholder("history"), # ☒ 답변 생성에서만 history 반영
```

```
    ("human",
```

```
    | "***비교 기준:** {criteria}\n\n"
```

```
    | "***[사업 A: {item_A}]*\n\n{context_A}\n\n"
```

```
    | "***[사업 B: {item_B}]*\n\n{context_B}")
```

# 프롬프트 예시 (recommendation)

```
("system",
'D당신은 B2G 사업 분석가 비디(Bidy)이며, '검색된 유사 사업 목록'을 바탕으로 사용자에게 맞춤형 사업을 추천하는 전문가입니다.\n\n'
'***작업 지침:***\n'
'1. **사용자 요청 분석:** '사용자 원본 요청'을 파악하여 어떤 종류의 사업을 원하는지 이해합니다.\n'
'2. **유사도 판단:** '검색된 유사 사업 목록'의 각 사업이 사용자 요청과 얼마나 유사한지 비교 분석합니다.\n'
'3. **추천 목록 생성:**\n'
'   - **결과가 있을 경우:** 유사도가 가장 높다고 판단되는 사업을 **최대 3개까지** 추천합니다. 추천 목록은 **번호(1., 2., 3.)**를 붙여주세요.\n'
'   - 각 추천 항목에는 반드시 **'사업명'과 '발주기관'을 포함해야 합니다.\n'
'   - **가장 중요한 것은, '어떤 점에서 유사한지' 구체적인 이유와 근거를 명확하게 설명**해야 합니다.\n'
'   - **결과가 없을 경우:** '검색된 유사 사업 목록'에 '추천할 만한 유사 사업을 찾지 못했습니다.'라는 내용이 있다면, \n'요청하신 내용과 유사한 사업을 찾을 수 없었습니다.\n'
'***매우 중요한 규칙:***\n'
'- **근거 기반 추천:** 당신의 모든 추천은 반드시 '검색된 유사 사업 목록'에 있는 정보에만 근거해야 합니다.'
'--- (규칙 끝) ---'
),
# ✅ 여기서 멀티턴 맥락(history) 반영
messagesPlaceholder("history"),
("human",
'***사용자 원본 요청:**\n{input}\n\n'
'***검색된 유사 사업 목록:**\n{context}\n\n'
'***추천 목록 (위 지침에 따라 작성):**')
)
```

# 프롬프트 예시 (default\_qa)

```
"system",
"당신은 대한민국 B2G(정부 대상 사업) RFP(제안요청서) 분석을 전문으로 하는 AI 컨설턴트 비디(Bidy)입니다.\n"
"당신은 주어진 [컨텍스트]에서 사용자의 [질문]에 대한 정답을 찾는 정보 추출 전문가입니다.\n\n"
"**작업 절차:**\n"
"1. **질문 분석:** 사용자의 [질문] 의도를 명확히 파악합니다.\n"
"2. **정보 탐색:** [컨텍스트] 전체를 꼼꼼히 읽고, 질문에 답할 수 있는 **정확한 근거 문장이나 구절**을 찾습니다.\n"
"3. **답변 생성:** 찾은 근거를 바탕으로, 질문에 대해 명확하고 간결하게 답변합니다. 답변은 항상 근거가 된 문서의 'project_title'을 먼저 언급하며 시작해야 합니다.\n\n"
"**답변 스타일 가이드:**\n"
"- 핵심 내용을 먼저 말하고, 필요시 부가 설명을 덧붙이는 **두괄식**으로 답변해주세요.\n"
"- 가능한 경우, 정보를 **볼렛 포인트(•)**나 **번호 매기기**를 사용하여 구조화해주세요.\n\n"
"**매우 중요한 규칙:**\n"
"- **근거 기반 답변:** 모든 답변은 반드시 [컨텍스트]에 기반해야 합니다. 당신의 사전 지식을 사용해서는 안 됩니다.\n"
"- **출처 명시:** 답변의 시작 부분에 반드시 어떤 문서에서 정보를 찾았는지 명시하세요. (예: '「2024년 이러닝시스템 운영 용역」 문서에 따르면...')\n"
"- **정보 부재 시:** [컨텍스트]를 여러 번 확인했음에도 답변의 근거를 정말 찾을 수 없을 때만, '제공된 문서에서는 질문에 대한 명확한 정보를 찾을 수 없었습니다.'라고 답변
```

```
messagesPlaceholder("history"), # ☒ 답변 단계에서만 history 반영
("human", "[질문]: {input}\n\n[컨텍스트]:\n{context}")
```

# Evaluation

# 평가 체계

---

## LLM as a Judge

- 질문·답안 생성을 LLM에 의뢰 → 시스템 출력과 함께 **LLM 평가**
- 프롬프트/체인 변경 시 **신속 비교** 용이

## RAGAS

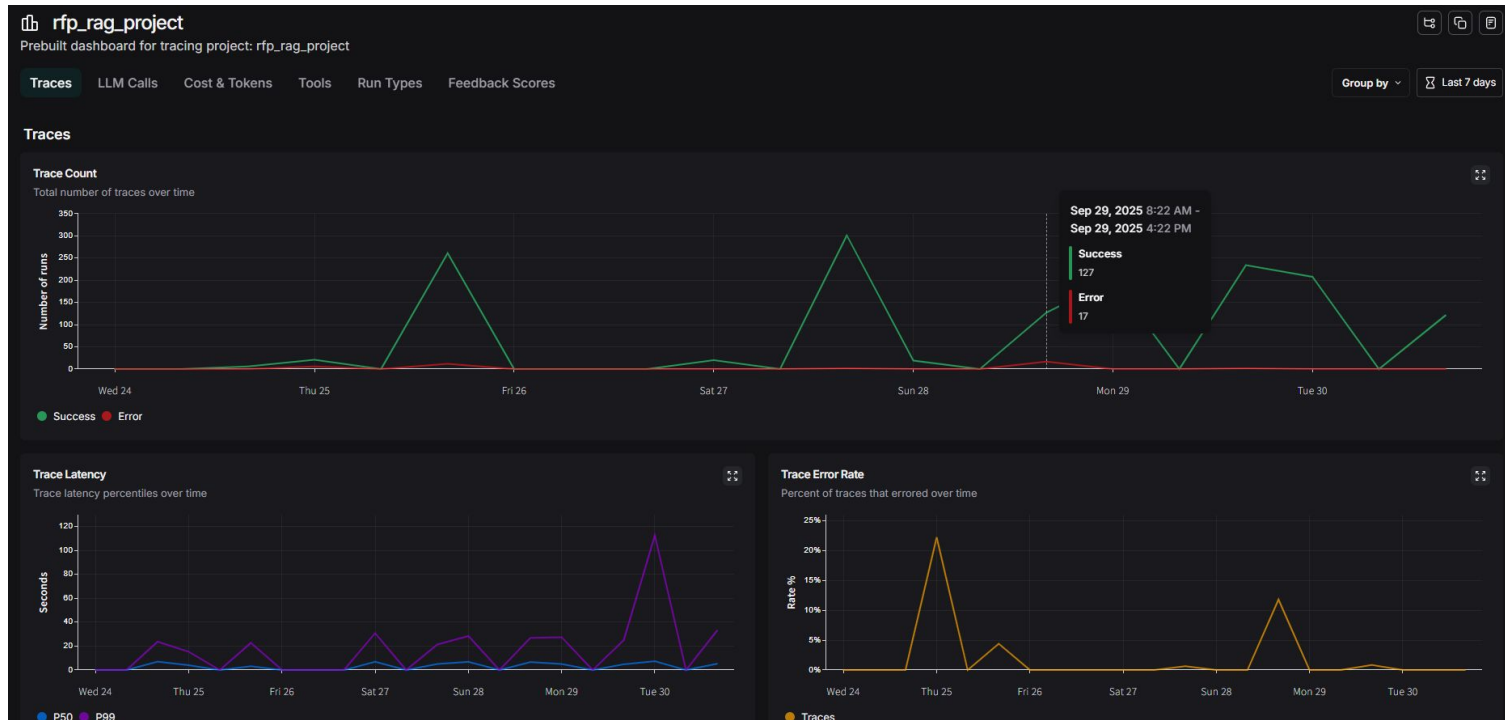
- **Generation:** *Faithfulness, AnswerCorrectness*
- **Retrieval:** *ContextRelevance, ContextRecall, ContextPrecision*
- 자체 데이터셋 (**25문항**): 5개 문서 × 5가지 카테고리



# 평가 체계

id	question	ground truth
1	예약발매시스템 개량 ISMP 용역 사업 발주 기관은 어디인가요?	예약발매시스템 개량 ISMP 용역 사업 발주 기관은 한국철도공사입니다.
2	한국철도공사 예약발매시스템 개량 ISMP 용역 사업의 주요 요구사항을 요약해줄 수 ...	본 과업은 예약발매시스템의 개량을 통해 ▲안정적 운영 ▲성능 개선 ▲장애 대응...
3	한국철도공사 예약발매시스템 개량 ISMP 용역에서 요구되는 PM 자격 조건과 일반 ...	PM은 정보처리기사 이상 자격증 보유 및 10년 이상 경력, 유사 프로젝트 관리 경...
4	한국철도공사 예약발매시스템 개량 ISMP 용역과 유사한 사업을 추천해줄 수 있나요?	본 사업과 유사한 사례로는 다른 공공기관의 승차권 발매 시스템 개선 용역, 교통·철...
5	한국철도공사 예약발매시스템 개량 ISMP 용역에서 제안사가 제출해야 하는 보안 준수...	제안사는 「정보보호 관리지침」을 준수해야 하며, 보안계획서 제출, 개인정보 보호대책...
6	대한장애인체육회 2025년 전국장애인체육대회 전산 및 시스템, 홈페이지 유지·보수 ...	배정예산은 220,000,000원(부가세 포함) 예정임
7	대한장애인체육회 2025년 전국장애인체육대회 전산 및 시스템, 홈페이지 유지·보수 ...	전국장애인체육대회 운영 안정성 확보, 외부 정보 제공 대응 강화, 데이터 관리 편의...
8	대한장애인체육회 2025년 전국장애인체육대회 전산 및 시스템, 홈페이지 유지·보수 ...	참가신청 시스템은 대회별 참가요강을 적용해 선수·임원 신청 등록을 지원하며, 경기기...
9	대한장애인체육회 2025년 전국장애인체육대회 전산 및 시스템, 홈페이지 유지·보수 ...	유사사업 범위로는 정보시스템 유지·운영 및 보수 관련 사업이 있으며, 특히 통합정보...
10	대한장애인체육회 2025년 전국장애인체육대회 전산 및 시스템, 홈페이지 유지·보수 ...	입찰방식은 제한경쟁입찰이며, 사업자 선정은 협상에 의한 계약으로 진행됨
11	부산관광공사 경영정보시스템 기능개선 사업의 총 사업금액은 얼마입니까?	사업금액: 금109,000,000원 (일억구백만원, 부가가치세 포함)
12	부산관광공사 경영정보시스템 기능개선 사업의 내용을 요약해 주세요.	사업목적은 업무환경 변화에 맞춘 기능개선으로 효율적 업무처리 지원. 범위는 현행 경...
13	부산관광공사 경영정보시스템 기능개선 사업에서 요구되는 주요 기능과 기존 시스템의 차...	기존 시스템: 그룹웨어(e-Gate EIP v5.0), 업무프로그램(NETRA R2...
14	부산관광공사 경영정보시스템 기능개선 사업과 유사한 다른 정보시스템 개선 사업을 추천...	문서에서 직접 다른 사업을 제시하지는 않으나, 유사 참고 시스템으로 공공기관 경영정...
15	부산관광공사 경영정보시스템 기능개선 사업의 제안서 평가 기준은 어떻게 되나요?	평가방식: 기술능력평가(90%) + 가격평가(10%). - 정량평가(20점): 수행...
16	국민연금공단 <2024년 이러닝시스템 운영 용역> 사업의 총 사업금액은 얼마입니까?	사업예산은 773,801천원이며, 교육과정 운영 620,801천원, 콘텐츠 개발·관...
17	국민연금공단 <2024년 이러닝시스템 운영 용역> 사업 내용을 요약해 주세요.	목적은 비대면 교육 수요 충족과 직원 역량 강화이며, 범위는 직무·자기개발 콘텐츠 ...
18	국민연금공단 <2024년 이러닝시스템 운영 용역>에서 내부 콘텐츠와 외부 콘텐츠의 ...	내부 콘텐츠는 공단 소유 콘텐츠(예: 직무교육, 개인정보보호, 정보보안 등)로 관리...
19	국민연금공단 <2024년 이러닝시스템 운영 용역>과 비슷한 사업 추천해 주세요.	문서 내 직접 추천은 없으나, 4차 산업혁명 관련 콘텐츠(인공지능, 빅데이터, VR...
20	국민연금공단 <2024년 이러닝시스템 운영 용역>의 기능 요구사항은 어떤 것들이 있나요?	주요 기능 요구사항에는 사이버(모바일) 연수원 구축(SFR-002), 큐레이션 서비...
21	국민연금공단 「사업장 사회보험료 지원 고시 개정에 따른 정보시스템 보완」 사업의 총...	본 사업의 사업예산은 **금 109,000,000원(금일억구백만원, 부가세 포함)*...
22	국민연금공단이 추진하는 「사업장 사회보험료 지원 고시 개정에 따른 정보시스템 보완」 ...	주요 과업은 ▲「사회보험료 지원 고시 개정」 반영을 위한 시스템 기능 보완 ▲지원 ...
23	국민연금공단의 「사업장 사회보험료 지원 고시 개정에 따른 정보시스템 보완」 사업과 ...	기존 시스템은 사회보험료 지원 관리 전반을 다루었으나, 본 사업은 고시 개정사항 반...
24	국민연금공단의 「사업장 사회보험료 지원 고시 개정에 따른 정보시스템 보완」과 유사한...	유사 사업으로는 타 공공기관 사회보장·복지 정보시스템 기능 개선 사업(예: 건강보험...
25	국민연금공단의 「사업장 사회보험료 지원 고시 개정에 따른 정보시스템 보완」 사업에서...	제안 자격은 「국가를 당사자로 하는 계약에 관한 법률 시행령」 제12조 요건을 충족...

# LangSmith 모니터링





# LangSmith 모니터링

The screenshot displays the LangSmith monitoring interface for a project named **rfp\_rag\_project**. The interface is divided into several sections:

- Runs List:** A table on the left showing a list of runs. The selected run is **RunnableSequence**, which is highlighted in green and has a status of **Success**. The table includes columns for Name, Input, and various metrics.
- Trace View:** A central pane showing the execution trace of the selected run. It details the sequence of operations, including **map:key:rephrased\_question**, **ChatOpenAI** calls, **log\_and\_pass\_through**, **get\_context**, **ContextualCompressionRetriever**, **VectorStoreRetriever**, **SelfQueryRetriever**, **query\_constructor**, **FewShotPromptTemplate**, **ChatOpenAI**, and **StructuredQueryOutputParser**. Each step shows its duration and output.
- RunnableSequence Details:** A right-hand pane providing more information about the selected run. It includes the **Input** (a question about a university), the **MESSAGES** (human and AI messages), and the **Output** (a list of relevant information about the university).
- Metadata:** A section on the far right showing the run's **START TIME** (10/01/2025, 09:17:49 AM), **END TIME** (10/01/2025, 09:18:02 AM), **STATUS** (Success), **TOTAL TOKENS** (9,588 tokens / \$0.00147015), **LATENCY** (13.31s), and **TYPE** (Sequence).

# 평가 결과

---

## 강점

- **Retrieval 단계 성능 양호**

- **Context Relevance 0.85, Precision 0.83** → 검색된 문서와 질문의 관련성이 높음.
- 불필요한 정보는 줄이고, 필요한 문서를 대부분 찾아냄.
- 즉, 검색 파이프라인 안정성 확보.

## 보완 필요

- **Faithfulness (0.46), Answer Correctness (0.44) 저조**

- 모델이 문서 기반 근거를 충분히 반영하지 못함.
- 생성된 답변이 원문과 정확히 일치하지 않거나, 추론이 개입된 경우 발생.
- 특히 요약(summarization)·추천(recommendation) 응답에서 문서 외부 추론이 섞여 사실과 불일치.

# 평가 결론

---

## 정리

- 현재 시스템은 검색 정확도는 높지만 답변 생성 신뢰성은 중간 수준.
- 문서를 잘 찾지만, 생성된 답변이 컨텍스트와 덜 일치하는 문제가 있어, 생성 단계 성능 개선이 필요.

## 개선 방향

### 1. 데이터 처리 측면

- 제공된 RFP 문서들의 내용 유사도가 높음 → 임베딩 벡터를 표준편차 기반 정규화 (**normalization**) 적용.
- 보다 현 작업(RFP 검색/추천)에 적합한 커스텀 유사도 함수(**similarity function**) 도입 필요.

### 2. 생성 단계 강화

- **Chain-of-Thought (CoT) 기반 Reasoning** 단계 추가.
- 답변 과정에서 근거 인용 → 추론 → 최종 답변의 단계적 구조화로 Faithfulness와 Correctness 개선.

Wrap Up

# 프로젝트 의의

---

- **데이터 중심 접근 → 시스템 중심 접근**
  - 초반: 전처리·칭킹 위주 개선
  - 전환: RAG 자체 성능 강화로 **불완전 데이터도 처리 가능한 Robust 시스템** 구축
- **라우팅 (Routing) 도입**
  - 사용자의 다양한 질문 유형을 5개 카테고리로 분류
  - 실제 컨설턴트가 자주 하는 질문을 유연하게 처리 가능
- **Baseline → 점진적 고도화**
  - 기본 RAG 시스템 구축 후 기능 확장
  - Hybrid Retrieval, Re-ranking, Summarization 등 적용
  - 지속적 **성능 모니터링** 을 통한 개선
- **프롬프트 엔지니어링**
  - 정밀·엄격한 설계로 LLM 환각 최소화
  - 답변 품질 및 신뢰성 향상
- **팀워크와 협업**
  - 전원 참여 + 역할 분담으로 효율 극대화
  - 모든 팀원이 코드/구조를 공유 이해 → **안정적 개발 기반 확보**

# 핵심 가치 실현

---

- **핵심 정보 신속 파악**

- 자동 탐색 및 요약: 시스템이 **RFP**를 자동으로 분류하고, 클릭 한 번에 핵심 구조를 구조화하여 요약 → 시간 낭비 **X**, 유망한 입찰 건 즉시 선별

- **의사결정 속도 향상**

- 다중 문서 비교 분석: 시스템이 다중 문서를 분석하고 차이점을 제시 → 고객의 강점과 약점을 고려한 전략적 의사결정 가능
- 고객 맞춤형 추천: 복합적인 조건의 질문에 대해 시스템이 내부 데이터를 라우팅하여 최적의 **RFP** 추천 → 컨설턴트가 놓칠 수 있는 새로운 사업 기회 발굴

- **제안서 작성 효율 극대화**

- 맥락 이해 기반 연속 답변(**Multi-turn**): 대화의 맥락을 유지하여 제안서의 논리와 완성도를 높이는 데 결정적

# 비즈니스 기대효과

---

- **컨설턴트 생산성 극대화**
  - 컨설턴트가 분석 가능한 RFP 수가 비약적으로 증가, 더 많은 고객사에게 맞춤형 컨설팅 가능 → 회사의 매출 증대
- **수주 성공률 증가**
  - 데이터 기반의 정교한 RFP 분석을 통해 제안서의 질을 향상 → 최종 낙찰 가능성 증가
- **고객 만족도 상승**
  - 쏟아지는 RFP 중 맞춤형 필터링 & 추천 가능 → 시간 및 인력 비용 절감

Thank You!