

LangChain을 활용한 뉴스 QA RAG 설계 및 프롬프트 기법 분석

12191735 김선우
지도: 안남혁 교수님

CONTENTS

01 주제 선정 배경

02 프로젝트 목표

03 진행 프로세스

04 기술스택 및
구현

05 결과 및 결론

연구 주제 및 선정 배경

- 기술 생태계의 빠른 변화

매일 수많은 정보와 기술이 새롭게 등장하고 빠르게 소비
정보를 빠르게 습득하고 싶어도 그 양이 방대하여 시간 소요가 크다

- LLM을 활용한 요약·분석 자동화 가능성

대규모 언어 모델(LLM)은 비정형 데이터를 요약·분석하고
정리하는데 뛰어난 성능을 보여 기술 트렌드 정리에 유용한 도구



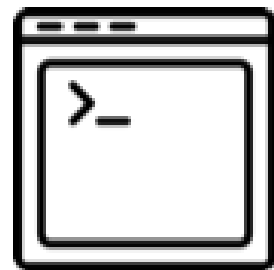
수많은 기술 뉴스 중 핵심 정보를 자동으로 추출하여 전달하는 기술 설계

설계 목표

- 실제 뉴스 데이터와 LLM을 결합(RAG)
- 프롬프트 설계를 통해 응답을 정형화



기술 동향 분석 자동화,
요약 및 질의응답 시스템 구축



진행 프로세스 - 전처리(Pre-processing)

2025-1 전기전자융합설계

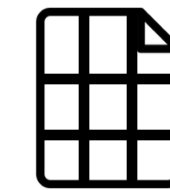
1 데이터 크롤링

2021년부터 2025년까지 작성된 기술 뉴스
정보 데이터를 분야별로 수집



2 로드 및 분할

수집한 데이터를 읽어와서
chunk 단위로 분할



3 임베딩

데이터를 벡터 표현으로 변환

[0.3, 0.4, 0.1, 1.8, 1.1...]

[0.7, 1.4, 2.1, 4.8, 4.1...]

4 벡터DB

변환된 데이터를 DB에 저장



진행 프로세스 - 수행(RunTime)

2025-1 전기전자융합설계

5

검색(Retrieval)

DB에서 검색하여 결과를 가져오기 위한
리트리버(검색 알고리즘) 정의



6

프롬프트(Prompt)

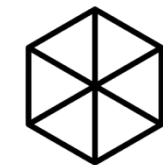
RAG 수행을 위한 프롬프트 작성



7

모델(LLM)

LLM 모델 호출



8

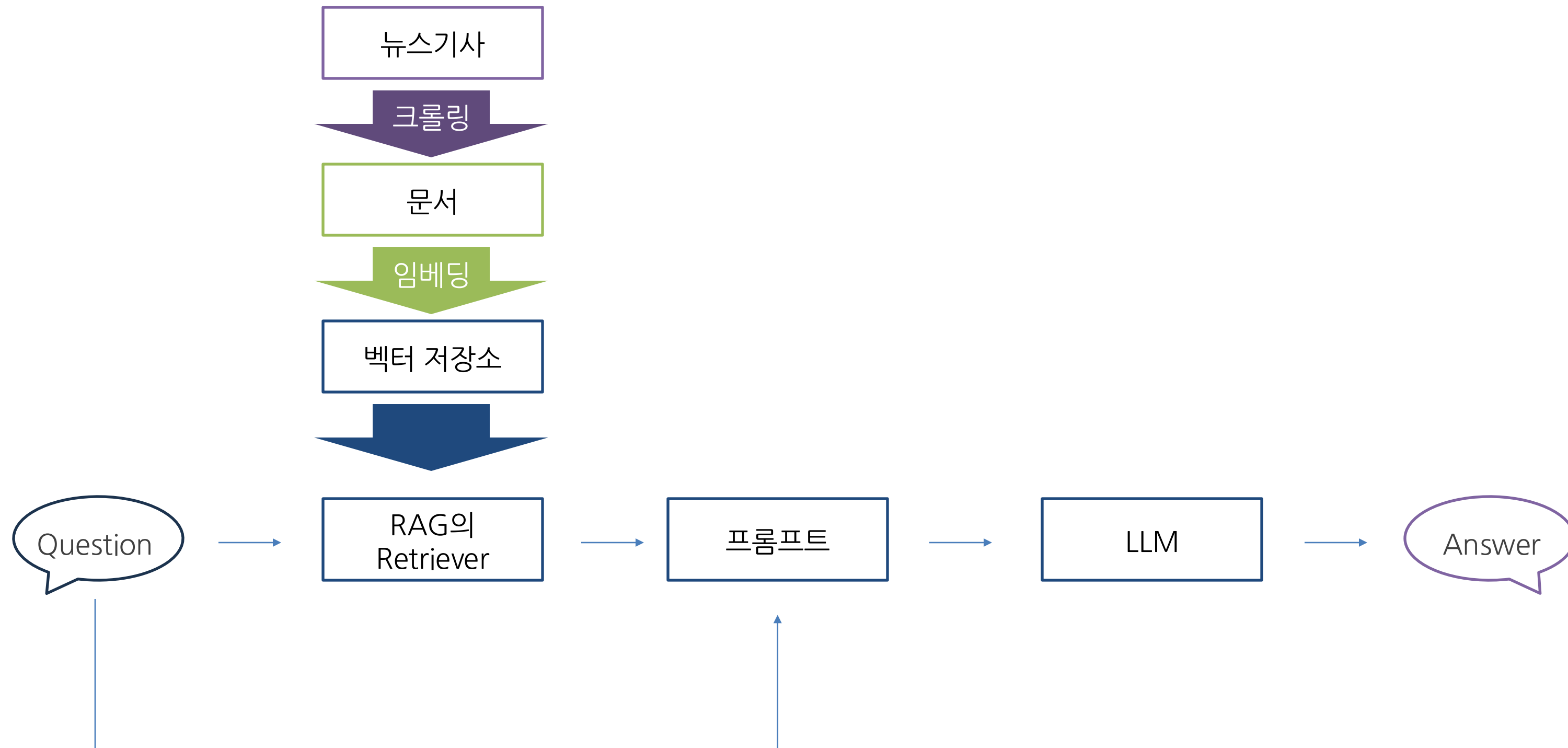
결과(Output)

텍스트, JSON, 마크다운 형태로 출력



진행 프로세스 - 요약

2025-1 전기전자융합설계



- LangChain

- LLM 모델을 활용한 애플리케이션 개발에 특화된 오픈소스 프레임워크
- 외부 데이터 소스와 통합하여 복잡한 애플리케이션을 만들 수 있다.



- Solar LLM

- Upstage에서 출시한 **한국어 특화 LLM**
- RAG 및 LangChain 연계 최적(Upstage 자체 임베딩 모델 제공)



- ChromaDB
 - 벡터 데이터베이스
 - 임베딩된 벡터 정보를 효율적으로 저장하고 유사도 검색 수행에 최적화
 - RAG의 검색기 역할
- LangSmith
 - 응답 품질을 정성적/정량적으로 평가할 수 있는 도구
 - LLM 기반 응답 흐름의 실행 기록 수집



Chroma



구현 - 크롤링, 임베딩

- 뉴스 플랫폼: ZDNet, Itworld
- 카테고리 : 클라우드 컴퓨팅, 인공지능, 소프트웨어 개발, 인터넷, 컴퓨팅, 방송
- 메타데이터: 제목, 요약(소제목), 날짜, 작성자, URL, 플랫폼

	구성
기사	3,385개
차원(벡터 길이)	1024차원
Chunk 사이즈	최대 1000토큰(오버랩 100)

플랫폼	카테고리	기사 수
ZDNet	인터넷	720
	컴퓨팅	785
	방송/통신	602
ITWorld	클라우드 컴퓨팅	590
	인공지능	578
	소프트웨어 개발	110
	합계	3,385

결론 - 프롬프트 (Zero-shot)

```
# === 프롬프팅 기법 정의 ===  
# 1. Zero-shot Prompt (기본)  
zero_shot_prompt_template = '''  
당신은 뉴스 기사 기반 Q&A 어시스턴트입니다.  
아래 context(근거 기사)에서 질문과 관련된 정보가 있으면, 그 내용을  
바탕으로 답변하세요.  
context에 전혀 관련 정보가 없으면 "관련 뉴스 기사 근거를 찾을 수  
없습니다."라고만 답하세요.  
당신의 지식이나 상식, 추측, 맥락 확장은 하지 마세요.  
답변은 3문장 이내로 간결하게 작성하세요.  
  
{context}  
'''
```

- 별도의 예시나 구체적인 추론 경로를 제공하지 않고
지시사항과 맥락만 전달
- 모델 자체 지식과 일반화 능력으로 작업을 수행
=> RAG 기본적인 성능을 측정하는 기준으로 활용

결론 - 프롬프트 (Few-shot)

```
# 2. Few-shot Prompt
few_shot_prompt_template = '''
당신은 뉴스 기사 기반 Q&A 어시스턴트입니다.
주어진 context(근거 기사)를 바탕으로 질문에 답변해야 합니다. 다음 예시와
같이 답변을 생성하세요.
```

```
---
```

```
**[예시]**
```

```
**context:** "Upstage는 최근 'Solar 10.7B'라는 새로운 대규모
언어 모델을 공개했습니다. 이 모델은 DUS(Depth Up-Scaling)라는
독자적인 기술을 통해 107억개의 파라미터를 가지며, 기존의 Llama 2나
Mistral 7B와 같은 모델들보다 뛰어난 성능을 보인다고 밝혔습니다."
```

```
**질문:** "Upstage의 Solar 모델에 대해 알려줘."
```

```
**답변:** Upstage의 Solar 10.7B는 107억개의 파라미터를 가진 대규모
언어 모델입니다. 이 모델은 DUS(Depth Up-Scaling)라는 효율적인 확장
기술을 사용하여 개발되었습니다. Solar 10.7B는 기존의 다른 오픈소스
모델들보다 뛰어난 성능을 보입니다.
```

```
---
```

이제 아래의 context와 질문에 대해 답변하세요.

만약 context에서 질문과 관련된 정보를 전혀 찾을 수 없다면, "관련 뉴스 기사 근거를 찾을 수 없습니다."라고만 답하세요.

```
{context}
'''
```

- 모델에게 완성도 높은 질의응답 예시(shot)을 제공
 - 모델을 직접 미세 조정하지 않고 원하는 형식으로 유도
 - 답변 스타일, 어조, 길이 등 구체적인 출력 형식을 일관성 있게 제어
- => 기사 내용을 자연스러운 문장으로 요약하고 종합하는지 학습하도록 유도

결론 - 프롬프트 (CoT)

```
# 3. Chain-of-Thought (CoT) Prompt
```

```
cot_prompt_template = '''
```

당신은 뉴스 기사 기반 Q&A 어시스턴트입니다.

아래 context(근거 기사)를 바탕으로 질문에 답변해야 합니다. 다음 단계를 따라서 답변을 생성하세요.

1. ****사실 확인 단계****: 질문에 답변하기 위해 context에서 필요한 핵심 사실들을 찾아서 한 문장씩 나열하세요.

2. ****종합 단계****: 위에서 찾은 사실들을 바탕으로, 질문에 대한 최종 답변을 3문장 이내로 간결하게 종합하여 생성하세요.

만약 context에서 질문과 관련된 정보를 전혀 찾을 수 없다면, "관련 뉴스 기사 근거를 찾을 수 없습니다."라고만 답하세요.

```
{context}
```

```
'''
```

- LLM이 답변을 내놓기 전에, 추론 과정을 단계별로 명시적으로 생성하도록 유도하는 기법
- 여러 단계의 추론이나 계산의 정확도를 향상시키는 효과
=> 저장소 데이터를 맹목적으로 인용하는 것이 아니라 어떤 정보가 질문과 관련 있는지 분석하여 환각 현상 감소

결론 - 프롬프트(ReAct)

```
# 4. ReAct (Simplified) Prompt
react_prompt_template = '''
당신은 뉴스 기사 기반 Q&A 어시스턴트입니다.
주어진 context(근거 기사)와 질문을 바탕으로 다음 절차에 따라 답변을
생성하세요.

**Thought(생각):**
질문에 답하기 위해 context에서 어떤 정보가 필요한지 분석하고, 관련된
핵심 내용을 요약합니다.

**Action(행동):**
위 'Thought' 단계의 내용을 바탕으로, 질문에 대한 최종 답변을 3문장
이내로 생성합니다. 만약 필요한 정보가 context에 없다면 "관련 뉴스 기사
근거를 찾을 수 없습니다."라고 답합니다.

{context}
'''
```

- 명시적인 생각 단계와 행동 단계를 나누어 문제에 체계적으로 접근하도록 유도
- 모델이 무엇을 해야하는지 정확히 인지하도록 함
=> 검색된 기사를 어떻게 분석하고 요약할지 계획을 세우도록 하여 집중도와 관련성을 높이는 효과

결과 - 프롬프트(챗봇 UI)

IT Tech News RAG Chatbot

프롬프팅 기법을 선택하여 답변의 차이를 비교해보세요.

☒ RAG(검색 증강 생성) 활성화

프롬프팅 기법 선택:

기본 (Zero-shot)

기본 (Zero-shot)


소수 예시 학습 (Few-shot)

사고의 연쇄 (Chain-of-Thought)


생각-행동 (ReAct)

- RAG On/Off 기능을 통해 저장소 검색 없이 순수 LLM 모델 자체 답변 확인 가능
- 프롬프팅 기법을 선택하여 프롬프트에 따른 응답 차이 확인 가능

결과 - 프롬프트(Zero-Shot 문답 예시)



오픈소스 라이선스 관리가 중요한 이유는 무엇인가요?



오픈소스 소프트웨어를 사용할 때 소스코드가 공개돼 누구나 자유롭게 사용할 수 있지만, 오픈소스 라이선스의 의무사항을 준수하지 않을 경우 저작권 침해 등의 문제가 발생할 수 있기 때문에 오픈소스 라이선스 관리가 중요합니다.

[근거 문서]

1. 엘에스웨어, ISO/IEC 5230:2020 오픈소스 국제 표준 인증 획득

- URL:
<https://www.itworld.co.kr/article/3553816/%ec%97%98%ec%97%90%ec%8a%a4%ec%9b%a8%ec%98%a4%ed%94%88%ec%86%8c%ec%8a%a4-%ea%b5%ad%ec%a0%9c-%ed%91%9c%ec%a4%80-%ec%9d%b8%ec%a6%9d-%ed%9a%8d%eb%93%9d.html>
- 본문 일부: ISO/IEC 5230:2020은 오픈소스 라이선스 준수를 위한 핵심 사항을 정의한 최초의 오픈소스 관리 국제 표준이다. 기업의 오픈소스 소프트웨어 체계 및 컴플라이언스 역량을 평 ...

ISO/IEC 5230:2020은 오픈소스 라이선스 준수를 위한 핵심 사항을 정의한 최초의 오픈소스 관리 국제 표준이다. 기업의 오픈소스 소프트웨어 체계 및 컴플라이언스 역량을 평가해 국제 인증을 부여한다. 오픈소스 소프트웨어는 소스코드가 공개돼 누구나 자유롭게 사용할 수 있지만 오픈소스 라이선스의 의무사항을 준수하지 않을 경우 저작권 침해 등의 문제가 발생할 수 있으므로 기업의 지속적인 관심과 관리가 필요하다.

↑ 실제 기사 내용

```
[DEBUG] MMR을 위한 후보군 검색 결과 (점수가 낮을수록 유사):
- 후보 1: score=0.5220, title='엘에스웨어, ISO/IEC 5230:2020 오픈소스 국제
- 후보 2: score=0.5784, title='글로벌 칼럼 | 오픈소스의 미래는 이념이 아
- 후보 3: score=0.5821, title='KMS테크놀로지, 오픈소스 거버넌스 관리 포털
- 후보 4: score=0.5913, title='KMS테크놀로지, 오픈소스 관리 포털 시스템
- 후보 5: score=0.5916, title='글로벌 칼럼 | 개방성을 다시 생각하게 만드
```


결과 - 프롬프트(Few-Shot 문답 예시)

프롬프팅 기법 선택:

소수 예시 학습 (Few-shot)



오픈소스 라이선스 관리가 중요한 이유는 무엇인가요?



오픈소스 라이선스 관리가 중요한 이유는 오픈소스 소프트웨어를 사용할 때 발생할 수 있는 법적 문제를 예방하기 위해서이다. 오픈소스 소프트웨어는 소스코드가 공개돼 누구나 자유롭게 사용할 수 있지만, 오픈소스 라이선스의 의무사항을 준수하지 않을 경우 저작권 침해 등의 문제가 발생할 수 있다. 따라서 기업의 지속적인 관심과 관리가 필요하다.

[근거 문서]

1. 엘에스웨어, ISO/IEC 5230:2020 오픈소스 국제 표준 인증 획득

- URL:
<https://www.itworld.co.kr/article/3553816/%ec%97%98%ec%97%90%ec%8a%a4%ec%9b%a8%ec%9c%bc%ec%9d%bd%ec%9f%9c%ec%9d%ac%ec%9c%ba%ec%9c%bb%ec%9c%be%ec%9c%bf%ec%9c%c0%ec%9c%c1%ec%9c%c2%ec%9c%c3%ec%9c%c4%ec%9c%c5%ec%9c%c6%ec%9c%c7%ec%9c%c8%ec%9c%c9%ec%9c%ca%ec%9c%cb%ec%9c%cc%ec%9c%cd%ec%9c%ce%ec%9c%cf%ec%9c%d0%ec%9c%d1%ec%9c%d2%ec%9c%d3%ec%9c%d4%ec%9c%d5%ec%9c%d6%ec%9c%d7%ec%9c%d8%ec%9c%d9%ec%9c%da%ec%9c%db%ec%9c%dc%ec%9c%dd%ec%9c%de%ec%9c%df%ec%9c%e0%ec%9c%e1%ec%9c%e2%ec%9c%e3%ec%9c%e4%ec%9c%e5%ec%9c%e6%ec%9c%e7%ec%9c%e8%ec%9c%e9%ec%9c%ea%ec%9c%eb%ec%9c%ec%ec%9c%ed%ec%9c%ee%ec%9c%ef%ec%9c%f0%ec%9c%f1%ec%9c%f2%ec%9c%f3%ec%9c%f4%ec%9c%f5%ec%9c%f6%ec%9c%f7%ec%9c%f8%ec%9c%f9%ec%9c%fa%ec%9c%fb%ec%9c%fc%ec%9c%fd%ec%9c%fe%ec%9c%ff>

- 법적인 문제에 대한 지적은 동일하나
훨씬 구체적인 답변

결과 - 응답 내용 평가(LangSmith)

input	output
최근 SKT 유심 문제가 많다는데, 피해를 방지하려면 어떻게 해야하나요?	유심보호서비스부터 가입하세요. 그리고 유심 교체를 원한다면 꼭 온라인 예약부터 하세요. 해외 출국자는 공항에서 유심을 교체하셔야 합니다.
오픈소스 라이선스 관리가 중요한 이유는 무엇인가요?	오픈소스 라이선스의 의무사항을 준수하지 않을 경우 저작권 침해 등의 문제가 발생할 수 있으므로 중요합니다.
어떤 언어들이 Wasm 지원으로 잘 알려져 있나요?	러스트, C/C++, 고(Go)가 아마도 가장 잘 알려진 Wasm 지원 언어일 것이다
줌 워크플레이스에서 제공하는 AI 기반 협업 플랫폼의 주요 기능은 무엇인가요?	업무 생산성과 효율성을 향상하도록 지원해 고객이 실제로 일상에서 겪는 문제를 해결하도록 돕는다
쿠버네티스 환경의 애드온 관리가 어려운 이유는 무엇인가?	쿠버네티스 환경의 애드온은 여러 층의 복잡성을 유발하고 다양한 전문 지식을 요구해 관리하기 어려운 것으로 악명이 높다
파이썬 3.12에서 새로 추가된 기능은 무엇인가요?	질문에 대한 답이 근거 기사에 없습니다.
클라이언트-서버 구조의 보안 강화 기술에는 무엇이 있나요?	질문에 대한 답이 근거 기사에 없습니다.
코드 리뷰 자동화에 활용되는 AI 도구는?	검색 된 근거 자료에서는 리뷰 자동화 AI 도구에 대한 내용은 없지만, AI 도구를 통한 자동화를 확대할 수록 QA 통제에 우선순위를 두는 것이 매우 중요하다고 설명합니다.
GPT-4의 성능은 이전 버전과 어떻게 다르나요?	업로드된 스크린 샷, 사진, 문서 또는 차트를 인식하고 이에 응답할 수 있으며, 얼굴 표정과 종이에 손으로 쓴 정보를 인식하기도 합니다.
AI 환각(hallucination) 현상은 왜 발생하나요?	AI가 학습 데이터에서 확률적으로 다음 단어나 구를 예측할 뿐, 의미나 단어가 현실에서 어떤 의미를 가지는지 이해를 못하기 때문이다.

it-news-rag-chatbot ID

Experiments Examples Evaluators Pairwise Experiments Feedback

Filters Select split

Inputs	Reference Outputs
<input type="checkbox"/> 오픈소스 라이선스 관리가 ...	오픈소스 라이선스의 의무사항...
<input type="checkbox"/> 파이썬 3.12에서 새로 추가...	질문에 대한 답이 근거 기사에 ...
<input type="checkbox"/> AI 환각(hallucination) 현...	AI가 학습 데이터에서 확률적으...
<input type="checkbox"/> GPT-4의 성능은 이전 버전...	업로드된 스크린 샷, 사진, 문서 ...
<input type="checkbox"/> 최근 SKT 유심 문제가 많...	유심보호서비스부터 가입하세요...
<input type="checkbox"/> 쿠버네티스 환경의 애드온 ...	쿠버네티스 환경의 애드온은 여...
<input type="checkbox"/> 코드 리뷰 자동화에 활용되...	검색 된 근거 자료에서는 리뷰 ...
<input type="checkbox"/> 줌 워크플레이스에서 제공...	업무 생산성과 효율성을 향상하...
<input type="checkbox"/> 클라이언트-서버 구조의 보...	질문에 대한 답이 근거 기사에 ...
<input type="checkbox"/> 어떤 언어들이 Wasm 지원...	러스트, C/C++, 고(Go)가 아...

- 예시 정답표(csv) 작성하여 LangSmith DataSet 등록 (30 SET)

결과 - 응답 내용 평가(LangSmith)

- LLM 평가자 등록 -> 정답, 키워드 포함률(단순어휘), LLM 기반 포함률(의미/문맥), 환각 발생률 비교
- 평가 모델은 OpenAI gpt-4.1-mini

it-news-rag-chatbot ID

Experiments Examples Evaluators Pairwise Experiments

Name	Type	Feedback Key
Conciseness	LLM-as-a-Judge	■ conciseness
Hallucination	LLM-as-a-Judge	■ hallucination
Correctness	LLM-as-a-Judge	■ correctness

Prompt & Model

Create your evaluator by defining the criteria you want to evaluate in your prompt. Map parts of your Example dynamically to your prompt using `{{variable}}`. Create your prompt below or use a saved prompt from Prompt Hub.

Use hub prompt ▾ Preview ☐ OpenAI gpt-4.1-mini ⚙

<Show up questions or errors for more information.>

- Redundant information or restatements.
- Polite phrases like "hope this helps" or "let me know if you need anything else."

</Rubric>

<Instructions>

- Carefully read the input and output.
- Check for any unnecessary elements, particularly those mentioned in the <Rubric> above.
- The score should reflect how close the response comes to containing only the essential information requested based on the rubric above.

</Instructions>

<Reminder>

The goal is to reward responses that provide complete answers with absolutely no extraneous information.

</Reminder>

<input>

`{{input}}` ▾

</input>

<output>

`{{output}}` ▾

</output>

결과 - 응답 내용 평가(LangSmith)

2025-1 전기전자융합설계

프롬프트	정답률(%) (Correctness)	키워드 포함률(%) (Inclusion)	LLM기반 포함률(%) (Inclusion)	평균 응답 길이	환각 발생률(%) (Hallucination)
기본(Zero-shot)	83.3(25/30)	23.3(7/30)	53.3(16/30)	1.8문장	46.7(14/30)
소수 예시 학습(Few-shot)	73.3(22/30)	43.3(13/30)	83.3(25/30)	2.9 문장	73.3(22/30)
사고의 연쇄(Chain-of-Thought)	90(27/30)	30.0(9/30)	46.7(14/30)	2.9 문장	83.3(25/30)
생각-행동(ReAct)	70(21/30)	40(12/30)	73.3(22/30)	2.2문장	50(16/30)

결론 - 최고의 프롬프트는 없다

- 정확성과 신뢰성의 상충 : **CoT** 프롬프트는 가장 높은 정답률을 보였으나 동시에 높은 환각 발생률을 기록 => 답이 정확하더라도 문서 내용을 벗어난 정보 생성 위험
- 균형 잡힌 접근 : ReAct, Zero-shot 프롬프트는 가장 낮은 환각 발생률 => 신뢰성(낮은 환각)과 답변의 질(의미적 유사도)를 함께 고려하는 균형 잡힌 설계도 중요하다

목표에 따른 최적의 프롬프트 선택

프롬프트 기법	지표 요약	활용 분야
Zero-shot / ReAct	높은 신뢰성 지표 - <u>사실에 기반한 답변</u> 에 초점	금융, 의료, 법률
Few-shot	높은 의미/문맥 정확도 지표 - <u>자세하고 풍부한 답변</u> 에 초점	에세이, 마케팅
CoT	높은 정답률 & 환각 - <u>정답을 맞히는 것이 목표인 경우</u>	퀴즈, 논리 기반 추론

**RAG 시스템이 특정 도메인에 맞춰 LLM의 답변 능력을 향상시키는 것은 명백
=> 만들고자 하는 서비스의 목표에 맞게 ‘정확성’과 ‘신뢰성’의 균형점을 찾아
적절한 프롬프트 기법을 선택하여야 한다**

참고문헌

- [1] Chandra Irugalbandara. “**Meaning Typed Prompting: A Technique for Efficient, Reliable Structured Output Generation**”
- [2] Pranab Sahoo, Asyush Kumar Singh, Sriparna Saha ··· “**A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications** “
- [3] Shubham Vastal, Harsh Dubey “**A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks**
- [4] Patrick Lewis, Ethan Perez. **Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks**
- [5] LangChain Documents : https://python.langchain.com/api_reference/core/documents/langchain_core.documents.base.Document.html
- [6] Upsatge Blog - Solar Pro: <https://www.upstage.ai/blog/en/solar-pro>