

LLM 기반 실무형 AI 에이전트 개발 직무연수

한국과학기술정보연구원

이 홍 석
(hsyi@kisti.re.kr)

Day4-3

RAG (Retrieval-Augmented Generation) 이해하기

❖ RAG (검색 증강 생성) 정의

- ✓ 사용자가 질문을 입력하면 입력한 질문으로 연관된 문서를 검색하고, 검색 결과를 바탕으로 답변하는 기술
- ✓ Retrieval (검색)
 - “어디선가 가져오는 것, 집어 오는 것”으로 이해
- ✓ Augmented (증강되었다.)
 - “원래 것에 뭔가 덧붙이거나 보태여 더 충실하게 좋아졌다는 뜻”으로 이해
- ✓ Generation (생성)
 - “프롬프트라고 하는 사용자 질문/질의에 대한 응답을 텍스트로 표출하는 것”을 의미



❖ LLM의 한계

- ✓ LLM은 학습된 데이터에 기반하여 동작하기 때문에 최신 정보나 외부 데이터에 대해 접근할 수 없다.
- ✓ 도메인 지식이 필요한 복잡한 질문에서는 신뢰할 수 없는 답변을 제공하는 경우가 발생합니다.
 - **편향된 답변 (Hallucination - 할루시네이션 원인)**을 생성할 가능성이 있다.
- ✓ LLM의 문제점 중에서 환각 현상의 사례 : 세종대왕 맥북 던짐 사건

```
question = "세종대왕이 누구인지 설명해주세요"  
result = llm.invoke(question)  
print(result.content)
```

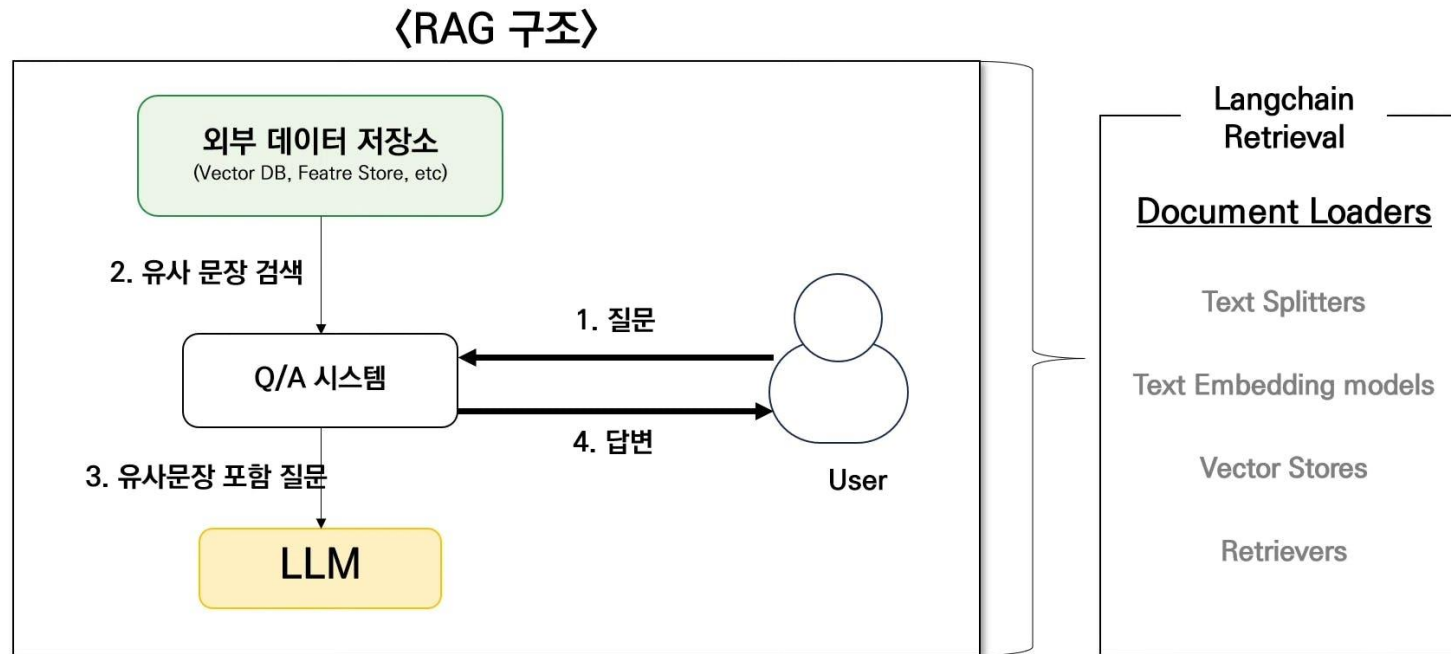
세종대왕은 조선시대 4대 왕 중 한 명으로, 조선시대의 제4대 군주로 알려져 있습니다. 그의 본명은 이도이며, 세종대왕은 1418년에 태어나 1450년부터 1455년까지 왕위에 오르게 되었습니다. 세종대왕은 조선시대를 대표하는 왕으로, 한글을 창제하고 과학기술, 문화, 예술 등 다양한 분야에서 발전을 이루게 했습니다. 또한, 세종대왕은 국내외 정치적인 사안에도 많은 관심을 기울였으며, 그의 통치는 조선시대의 전성기를 이루는 데 큰 역할을 하였습니다. 세종대왕은 한국 역사상 가장 위대한 왕 중 한 명으로 인정받고 있습니다.

```
question = "조선왕조 실록에 기록된 세종대왕의 맥북프로 던짐 사건에 대하여 알려줘"  
result = llm.invoke(question)  
print(result.content)
```

세종대왕의 맥북프로 던짐 사건은 조선왕조 실록에 기록된 사건으로, 세종대왕이 실록 작성 중 맥북프로를 사용하던 중 화가 나서 맥북프로를 던진 사건을 의미합니다. 이 사건은 세종대왕의 열정적인 성격과 업무에 대한 집중력을 보여주는 에피소드로 전해지고 있습니다. 이 사건은 세종대왕이 실록 작성을 위해 얼마나 열정적으로 노력했는지를 보여주는 사례로 전해지고 있습니다.

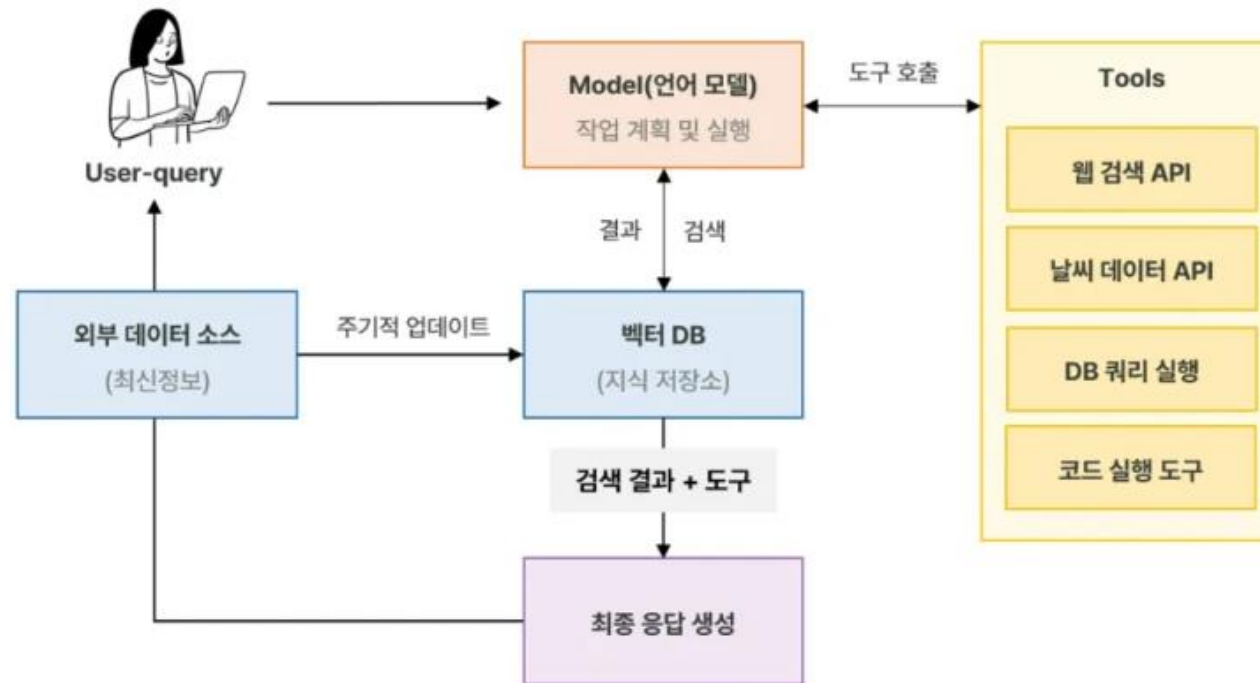
❖ RAG 모델

- ✓ 방대한 외부 데이터베이스에서 질의와 관련된 정보를 실시간으로 검색하고
- ✓ 이를 바탕으로 답변을 생성하는 강력한 자연어 생성 기술



❖ RAG 동작 방식

- ✓ 쿼리 인코더 : 사용자의 질문을 지식 검색기가 이해할 수 있는 벡터 형태로 인코딩
- ✓ 지식 검색기 : 인코딩 된 질문을 바탕으로 외부 데이터에서 관련된 데이터를 검색
- ✓ 지식 증강 생성기 : 검색된 데이터를 활용해서 질문에 대한 답변을 생성하는 언어모델



<https://uracle.blog/2025/05/16/ai-agent>

- ❖ 다수의 가정통신문 PDF 데이터를 활용하여 작성 팁, 표현, 주의사항
 - ✓ RAG로 제공하는 자동화 실습
 - ✓ 한글 파일을 pdf로 읽어보기
- ❖ 교사 및 행정 직원이 가정통신문 작성 시 필요한 자료를 빠르게 검색 및 생성해 활용하도록 지원
- ❖ 검색 기반 생성(RAG)을 통해 근거 기반 자동 응답 및 작성 보조 기능을 체험
- ❖ LangChain을 활용하여 PDF 로드, Chunk 처리, 임베딩, 벡터 검색, LLM 연동까지 실전형 RAG 파이프라인 학습

- ❖ LangChain 프레임워크
 - ✓ 문서 로드, Chunk 처리, Embedding, Vector DB 구축, Retriever 구성
- ❖ RAG 기법
 - ✓ 검색과 LLM 생성을 결합하여 답변 생성
- ❖ OpenAI API 활용
 - ✓ 임베딩 생성 및 GPT 기반 LLM 연동
- ❖ FAISS 또는 Chroma
 - ✓ 고속 벡터 검색용 벡터 데이터베이스
- ❖ Python 및 가상환경 관리
 - ✓ 실습 환경 구성 및 코드 실행
- ❖ PDF 처리 기술
 - ✓ PyPDFLoader 등으로 PDF 문서 데이터 전처리 및 학습 데이터화

❖ OpenAI Key 발급 및 설정

- ✓ <https://platform.openai.com/signup> 에서 가입 후 API Key 발급

❖ 가상환경 생성 및 라이브러리 설치

- ✓ `'python -m venv venv'` 혹은 `'source venv/bin/activate'`
- ✓ `'pip install langchain langchain-community langchain-openai faiss-cpu'`

❖ 실습 목표 요약 (정리)

- ✓ 다수의 가정통신문 PDF를 불러와 Chunk → 임베딩 → 벡터 DB 구성
- ✓ LangChain RAG로 질의 시 가정통신문 작성 팁/표현/주의사항 제공
- ✓ 출력으로 로드, 분할, 임베딩 완료 상태 확인 및 RAG 응답 검증


❖ 코드 실행


- ✓ 멀티 PDF 경로 설정 후 실행
- ✓ 질문 예시로 "여름 폭염 대응 가정통신문 작성 팁 알려줘" 등 사용
- ✓ 응답 및 소스 문서 수로 검증

❖ LangChain 영역: 여러 PDF 로드 및 처리

LangChain 영역: 여러 PDF 로드 및 처리

```
pdf_files = ["/data/가정통신문_여름방학영어캠프참여안내.pdf", "/data/가정통신문_폭염발생_시_행동요령.pdf"]
```

	가정통신문	스스로 공부하는 학생 근지와 보람을 느끼는 교사 개성과 능력이 발현되는 학교																										
대전시 유성구 노은서로 62(노은1동) ☎ 교무실 : 828-5300																												
2025학년도 여름방학 영어캠프 참여 안내																												
<p>안녕하십니까? 영어에 대한 흥미를 불러일으키고 자신감을 키워주고자 여름방학 영어캠프에 대해 아래와 같이 안내해 드리오니 신청을 희망하는 학생은 7월 4일(금)까지 신청서를 제출해 주시기 바랍니다.</p> <p>1. 일 시 : - 3~4학년(1기) : 2025. 7. 28.(월) ~ 7. 31.(목) 08:50 ~ 12:10 (점심식사 미제공) [4일간] - 5~6학년(2기) : 2025. 8. 4.(월) ~ 8. 7.(목) 08:50 ~ 12:10 (점심식사 미제공) [4일간]</p> <p>2. 장 소 : 본교 영어실</p> <p>3. 대 상 : 본교 3~6학년 학생 중 희망자(참가비 없음/기수별 최대 15명)</p> <p>4. 강 사 : 원어민 보조교사, 본교 영어전담교사</p> <p>5. 운영일정</p> <table border="1"> <thead> <tr> <th>운영 일자</th> <th>학습 내용</th> </tr> </thead> <tbody> <tr> <td>Day1</td> <td>- Science</td> </tr> <tr> <td>Day2</td> <td>- Music</td> </tr> <tr> <td>Day3</td> <td>- Social</td> </tr> <tr> <td>Day4</td> <td>- Movie</td> </tr> </tbody> </table> <p>※ 기수별 신청 학생이 15명을 초과할 경우 추첨 실시 예정</p> <p>2025. 6. 30.</p> <p>대전수정초등학교장</p> <p>2025학년도 여름방학 영어캠프 참가신청 및 보호자 동의서</p> <table border="1"> <thead> <tr> <th>학년-반</th> <th>학생 성명</th> <th>성별</th> <th>보호자 성명</th> <th>학생과의 관계</th> </tr> </thead> <tbody> <tr> <td>-</td> <td>-</td> <td>-</td> <td>(인)</td> <td>-</td> </tr> </tbody> </table> <p>연락처 본인: _____ 집: _____ 학부모: _____</p> <p>□ 개인정보 수집·이용 동의</p> <table border="1"> <thead> <tr> <th>항 목</th> <th>수집목적</th> <th>보유기간</th> </tr> </thead> <tbody> <tr> <td>학생(학년, 반, 번호, 성명, 연락처) 보호자(성명)</td> <td>영어캠프 참가 신청</td> <td>1년(학년 말까지)</td> </tr> </tbody> </table> <p>※ 개인정보 수집·이용에 대한 동의를 거부할 권리가 있습니다. 그러나 동의를 거부할 경우 영어캠프참가 신청에 제한을 받을 수 있습니다.</p> <p>개인정보 수집·이용 동의 <input type="checkbox"/>예 <input type="checkbox"/>아니요</p> <p>위와 같이 개인정보를 수집·이용·제공하는데 동의합니다. ※ 만 14세 미만 학생의 경우 보호자(법정대리인)의 동의가 필요합니다. 보호자(법정대리인) : _____ (서명)</p> <p>대전수정초등학교 귀하</p>			운영 일자	학습 내용	Day1	- Science	Day2	- Music	Day3	- Social	Day4	- Movie	학년-반	학생 성명	성별	보호자 성명	학생과의 관계	-	-	-	(인)	-	항 목	수집목적	보유기간	학생(학년, 반, 번호, 성명, 연락처) 보호자(성명)	영어캠프 참가 신청	1년(학년 말까지)
운영 일자	학습 내용																											
Day1	- Science																											
Day2	- Music																											
Day3	- Social																											
Day4	- Movie																											
학년-반	학생 성명	성별	보호자 성명	학생과의 관계																								
-	-	-	(인)	-																								
항 목	수집목적	보유기간																										
학생(학년, 반, 번호, 성명, 연락처) 보호자(성명)	영어캠프 참가 신청	1년(학년 말까지)																										

	가정통신문	스스로 공부하는 학생 근지와 보람을 느끼는 교사 개성과 능력이 발현되는 학교												
대전시 유성구 노은서로 62(노은1동) ☎ 교무실 : 828-5300														
폭염 발생 시 행동요령 안내														
<p>안녕하십니까?</p> <p>요즈음 이른 더위로 온열질환자 발생이 예상됨에 따라 한여름 폭염 발생과 관련하여 행동요령을 안내하여 드리니 가정에서도 참고하시어 여름을 건강하고 안전하게 나시기 바랍니다.</p> <p>< 폭염의 정의 ></p> <p>☉ 일 최고온도 33℃ 이상의 불볕더위가 계속되는 현상</p> <table border="1"> <thead> <tr> <th>폭염주의보</th> <th>폭염경보</th> </tr> </thead> <tbody> <tr> <td>일 최고온도가 33℃ 이상인 상태가 2일 이상 지속될 것으로 예상될 때</td> <td>일 최고온도가 35℃ 이상인 상태가 2일 이상 지속될 것으로 예상될 때</td> </tr> </tbody> </table> <p>< 폭염 시 학생 행동 요령 ></p> <table border="1"> <thead> <tr> <th>등교 전</th> <th>등교 시</th> <th>학교에서</th> <th>가정에서</th> </tr> </thead> <tbody> <tr> <td> <ul style="list-style-type: none"> ○ 방송 매체(TV, 라디오) 및 인터넷(기상청)을 통해 기상 상황을 확인한다. </td> <td> <ul style="list-style-type: none"> ○ 최대한 햇볕을 피해 그늘로 걷는다. ○ 가볍고 얇은 옷을 입고, 모자나 양산 등으로 햇볕을 가린다. ○ 자외선 차단제로 피부부를 보호한다. </td> <td> <ul style="list-style-type: none"> ○ 학교 지시에 따라 안전한 학교생활을 준수한다. ▶ 쉬는 시간과 점심시간의 체육활동 등 실외 및 야외 활동을 자제한다. ※ 폭염경보 발령 시에는 체육활동 등 모든 실외 및 야외 활동을 금지한다. ▶ 손 씻기 등 개인위생을 철저히 한다. ○ 깨끗한 물을 규칙적으로 섭취한다. </td> <td> <ul style="list-style-type: none"> ○ 밀폐된 차 안에 혼자 있지 않는다. ○ 균형 있는 식사 및 식품 안전을 철저히 한다. ▶ 식사는 균형 있게 신선한 채소와 과일 등을 골고루 섭취한다. ▶ 물은 끓여 마시고, 날 음식은 삼가며 유통기한을 확인하여 변질이 의심 되면 버린다. ○ 냉방병 예방을 위해 적정온도를 유지한다. ▶ 에어컨, 선풍기는 잠들기 전에 끄거나 일정 시간 가동 후 꺼지도록 예약한다. ▶ 냉방기기 사용 시 실내외 온도를 5℃ 내외로 유지하여 냉방병을 예방한다. ○ 창문을 커튼이나 천 등으로 가려 직사광선을 최대한 차단한다. ○ 집에서 가까운 병원의 연락처를 확인하고, 자신과 가족의 건강 상태를 체크한다. ○ 준비운동 없이 물에 들어가거나 갑작스러운 찬물 샤워를 자제한다. (심장마비 위험) </td> </tr> </tbody> </table> <p>2025. 6. 18.</p> <p>대전수정초등학교장</p>			폭염주의보	폭염경보	일 최고온도가 33℃ 이상인 상태가 2일 이상 지속될 것으로 예상될 때	일 최고온도가 35℃ 이상인 상태가 2일 이상 지속될 것으로 예상될 때	등교 전	등교 시	학교에서	가정에서	<ul style="list-style-type: none"> ○ 방송 매체(TV, 라디오) 및 인터넷(기상청)을 통해 기상 상황을 확인한다. 	<ul style="list-style-type: none"> ○ 최대한 햇볕을 피해 그늘로 걷는다. ○ 가볍고 얇은 옷을 입고, 모자나 양산 등으로 햇볕을 가린다. ○ 자외선 차단제로 피부부를 보호한다. 	<ul style="list-style-type: none"> ○ 학교 지시에 따라 안전한 학교생활을 준수한다. ▶ 쉬는 시간과 점심시간의 체육활동 등 실외 및 야외 활동을 자제한다. ※ 폭염경보 발령 시에는 체육활동 등 모든 실외 및 야외 활동을 금지한다. ▶ 손 씻기 등 개인위생을 철저히 한다. ○ 깨끗한 물을 규칙적으로 섭취한다. 	<ul style="list-style-type: none"> ○ 밀폐된 차 안에 혼자 있지 않는다. ○ 균형 있는 식사 및 식품 안전을 철저히 한다. ▶ 식사는 균형 있게 신선한 채소와 과일 등을 골고루 섭취한다. ▶ 물은 끓여 마시고, 날 음식은 삼가며 유통기한을 확인하여 변질이 의심 되면 버린다. ○ 냉방병 예방을 위해 적정온도를 유지한다. ▶ 에어컨, 선풍기는 잠들기 전에 끄거나 일정 시간 가동 후 꺼지도록 예약한다. ▶ 냉방기기 사용 시 실내외 온도를 5℃ 내외로 유지하여 냉방병을 예방한다. ○ 창문을 커튼이나 천 등으로 가려 직사광선을 최대한 차단한다. ○ 집에서 가까운 병원의 연락처를 확인하고, 자신과 가족의 건강 상태를 체크한다. ○ 준비운동 없이 물에 들어가거나 갑작스러운 찬물 샤워를 자제한다. (심장마비 위험)
폭염주의보	폭염경보													
일 최고온도가 33℃ 이상인 상태가 2일 이상 지속될 것으로 예상될 때	일 최고온도가 35℃ 이상인 상태가 2일 이상 지속될 것으로 예상될 때													
등교 전	등교 시	학교에서	가정에서											
<ul style="list-style-type: none"> ○ 방송 매체(TV, 라디오) 및 인터넷(기상청)을 통해 기상 상황을 확인한다. 	<ul style="list-style-type: none"> ○ 최대한 햇볕을 피해 그늘로 걷는다. ○ 가볍고 얇은 옷을 입고, 모자나 양산 등으로 햇볕을 가린다. ○ 자외선 차단제로 피부부를 보호한다. 	<ul style="list-style-type: none"> ○ 학교 지시에 따라 안전한 학교생활을 준수한다. ▶ 쉬는 시간과 점심시간의 체육활동 등 실외 및 야외 활동을 자제한다. ※ 폭염경보 발령 시에는 체육활동 등 모든 실외 및 야외 활동을 금지한다. ▶ 손 씻기 등 개인위생을 철저히 한다. ○ 깨끗한 물을 규칙적으로 섭취한다. 	<ul style="list-style-type: none"> ○ 밀폐된 차 안에 혼자 있지 않는다. ○ 균형 있는 식사 및 식품 안전을 철저히 한다. ▶ 식사는 균형 있게 신선한 채소와 과일 등을 골고루 섭취한다. ▶ 물은 끓여 마시고, 날 음식은 삼가며 유통기한을 확인하여 변질이 의심 되면 버린다. ○ 냉방병 예방을 위해 적정온도를 유지한다. ▶ 에어컨, 선풍기는 잠들기 전에 끄거나 일정 시간 가동 후 꺼지도록 예약한다. ▶ 냉방기기 사용 시 실내외 온도를 5℃ 내외로 유지하여 냉방병을 예방한다. ○ 창문을 커튼이나 천 등으로 가려 직사광선을 최대한 차단한다. ○ 집에서 가까운 병원의 연락처를 확인하고, 자신과 가족의 건강 상태를 체크한다. ○ 준비운동 없이 물에 들어가거나 갑작스러운 찬물 샤워를 자제한다. (심장마비 위험) 											

❖ 질문과 답변

```
# LangChain 영역: LLM 설정
temperature_setting = 0.3
model="gpt-3.5-turbo"
llm = ChatOpenAI(model=model, temperature=temperature_setting)
print(f"LLM 설정 완료 (model={model}, temperature={temperature_setting})")
```

LLM 설정 완료 (model=gpt-3.5-turbo, temperature=0.3)

```
# 테스트 질문 (검증 및 학습용)
```

```
queries = [
    "여름방학 캠프 안내 가정통신문 작성 팁 알려줘.",
    "폭염 예방을 위한 가정통신문 작성 시 주의사항은?",
    "가정통신문 작성에 필요한 기본 구성 요소는 무엇인가요?",
    "가정통신문 마무리 문구 예시 알려줘."
]
```

```
# RAG 영역: 각 질문별 검색+생성 QA 수행 및 출력
```

```
for idx, query in enumerate(queries, 1):
    print(f"\n{idx}. [질문]: {query}")
    result = qa_chain.invoke(query)
    print("[응답]:")
    print(result['result']) # 생성된 답변 출력
    print(f"참고한 소스 문서 수: {len(result['source_documents'])}")
```

❖ 질문1

1. [질문]: 여름방학 캠프 안내 가정통신문 작성 팁 알려줘.

[응답]:

캠프 안내를 위한 가정통신문을 작성할 때에는 다음과 같은 팁을 활용할 수 있습니다:

1. 목적 설명: 캠프의 목적과 이유를 간단명료하게 설명해 주세요.
2. 일정 안내: 캠프 일정, 장소, 시간 등을 구체적으로 안내해 주세요.
3. 참가 신청 방법: 참가를 원하는 학생들이 어떻게 신청할 수 있는지 자세히 안내해 주세요.
4. 안전 수칙: 캠프 참가자들이 지켜야 할 안전 수칙에 대해 명확히 안내해 주세요.
5. 문의처 안내: 부모님들이 궁금한 사항이나 문의할 사항이 있을 경우 연락할 수 있는 방법을 안내해 주세요.
6. 마무리: 참가를 기대하며 마무리 인사와 함께 긍정적인 마무리를 지어주세요.

이러한 요소들을 고려하여 가정통신문을 작성하면 학부모님들이 보다 이해하기 쉽고 참여하기 편리할 것입니다.
참고한 소스 문서 수: 4

2. [질문]: 폭염 예방을 위한 가정통신문 작성 시 주의사항은?

[응답]:

폭염 예방을 위한 가정통신문을 작성할 때 주의해야 할 사항은 다음과 같습니다:

1. 폭염의 정의와 폭염이 발생할 때의 주요 특징을 설명한다.
2. 폭염 발생 시 시민들이 취해야 할 행동요령을 상세히 안내한다.
3. 집안에서의 예방 조치 및 안전 수칙을 제시한다. (예: 직사광선 차단, 물 섭취 등)
4. 폭염 시 학생들의 행동요령을 명확히 안내한다. (예: 햇볕 피하기, 가벼운 옷 입기 등)
5. 필요 시 응급상황 대처 방법이나 연락처를 제공한다.
6. 가정에서의 안전한 환경 조성을 위한 조언을 포함한다.

참고한 소스 문서 수: 4

❖ 학교 적용 사례 등 본인 직무에 적합한 AI 에이저트를 생성하시오.

- ✓ 팀별 (2인)
- ✓ 알림장 및 가정통신문 요약을 자동으로 작성하여 배포합니다.
- ✓ 학교/학급 일정 Q&A 챗봇으로 학부모 및 학생 질의응답 자동화합니다.
- ✓ 수업 시간에 교과 질문응답 AI 챗봇을 도입해 즉시 피드백 제공합니다.
- ✓ 상담 기록 요약 및 회의 자료 자동 정리로 업무 효율을 높입니다.
- ✓ 교사 업무 경감과 학생 맞춤형 학습 지원에 활용할 수 있습니다.

Day4-4

AI 에이전트 직무 활용을 위한 실습

❖ 대규모 언어 모델(LLM) 기반 애플리케이션을 쉽게 만들 수 있도록 돕는 프레임워크.

✓ 역할:

- LLM을 데이터베이스, API, 파일, 검색 시스템 등과 연결하고, 대화 흐름·프롬프트 관리·메모리 관리 등을 구조화.

✓ 특징:

- 다양한 LLM(OpenAI, Anthropic, Hugging Face 등) 지원
- 검색 시스템, DB, 벡터스토어 연결 기능 내장

✓ 체인(Chain)과 에이전트(Agent) 개념으로 복잡한 워크플로우 구성

✓ 비유:

- "LLM 기반 앱을 만들 때 쓰는 레고 조립 키트"

- ❖ LLM이 답변하기 전에 외부 지식 소스에서 관련 문서를 검색(Retrieval)하고, 그 내용을 바탕으로 답변을 생성(Generation)하는 패턴/아키텍처.
 - ✓ 역할:
 - 모델이 훈련 데이터에 없는 정보나 최신 정보도 검색해서 활용할 수 있게 함.
 - ✓ 특징:
 - 검색 단계: 벡터 데이터베이스(예: Pinecone, FAISS)나 검색 엔진에서 관련 문서 검색
 - 생성 단계: 검색 결과를 프롬프트에 넣어 LLM이 응답 생성
 - 환각(hallucination) 줄이고, 최신성 및 도메인 특화 지식 제공
 - ✓ 비유:
 - "대답하기 전에 사전·검색엔진을 먼저 찾아보고 말하는 방식"

❖ RAG는

✓ "어떻게 검색하고 생성할지"에 대한 방법론이고,

❖ LangChain은

✓ 그 방법론(RAG 포함)을 구현할 수 있는 도구 상자입니다.

항목	LangChain	RAG
성격	개발 프레임워크	LLM 활용 아키텍처 패턴
목적	LLM 기반 앱을 쉽게 구현	외부 지식 검색 후 답변 정확도 향상
구현 여부	RAG 구현 가능	LangChain 없이도 가능
범위	LLM 연결, 체인, 에이전트, 메모리 등	검색 + 생성 단계에 집중
예시	대화형 챗봇, 자동 보고서 생성기	사내 문서 Q&A, 최신 뉴스 기반 챗봇

❖ 오픈소스 벡터 데이터베이스로 Python 친화적이며, LangChain과 RAG 프로젝트에서 많이 사용.

✓ 특징: 개발 편의성·메타데이터 관리 중시, RAG 실험·중소규모 앱에 적합

- 소규모·중규모 프로젝트에 적합
- 임베딩 생성은 외부 모델(OpenAI, Hugging Face 등) 사용

✓ 역할: 벡터와 함께 메타데이터(문서 내용, 출처, 태그 등) 저장 및 검색.

✓ 장점:

- 빠른 프로토타입 제작 가능
- RAG 예제와 튜토리얼이 풍부

✓ 단점:

- 대규모 데이터셋 처리 속도·확장성은 Faiss나 Milvus보다 약함
- GPU 최적화 부족

✓ 비유:

- "주소록 있는 벡터 저장소"
- 전화번호(벡터)뿐 아니라 이름·주소(메타데이터)도 함께 저장

- ❖ Meta(구 Facebook) AI Research에서 개발한 고성능 벡터 검색 라이브러리.
 - ✓ 특징: 대규모, 초고속 검색용. 메타데이터는 직접 관리해야 함.
 - ✓ 역할: 대규모 임베딩 벡터(예: 문장, 이미지 특징)를 빠르게 유사도 검색할 수 있도록 지원.
 - ✓ 특징: C++로 작성, Python 바인딩 제공 → 속도 빠름
 - 자체적으로는 메타데이터 관리 기능 없음 → 검색된 벡터에 연결된 문서나 추가 정보 관리하려면 따로 구현 필요
 - ✓ 장점:
 - 검색 속도 빠름, 메모리·성능 최적화 잘 돼 있음
 - ✓ 단점:
 - 메타데이터/필터링 기능 부족
 - ✓ DB 기능이 아니라 검색 엔진 라이브러리에 가까움
 - ✓ 비유:
 - "초고속 벡터 검색 엔진" — 단, 주소록 없이 전화번호만 저장하는 느낌

항목	Faiss	Chroma
개발사	Meta AI Research	Chroma 팀 (오픈소스)
성격	벡터 검색 라이브러리	벡터 데이터베이스
메타데이터 저장	없음 (직접 구현 필요)	내장
속도·성능	매우 빠름, 대규모에 적합	빠르지만 대규모에 한계
GPU 지원	강력	미지원
사용 난이도	비교적 어려움	쉬움
확장성	매우 큼	중간
주 사용처	대규모 검색 시스템	RAG 프로토타입, 소규모 서비스

❖ 목표:

- ✓ RAG(Retrieval-Augmented Generation) 개념 이해
- ✓ 교사 생활기록부 예시 문서를 바탕으로 질문
 - 세특 생성 실습
- ✓ LangChain + OpenAI API 연동 경험

❖ 실습 준비:

- ✓ OpenAI API Key 발급
- ✓ 실습 파일: 학생 활동 모의 데이터 (5명 정도)
 - student_activity_records.txt
- ✓ Python 환경
 - langchain, langchain-openai, openai, faiss-cpu 등 설치

텍스트 파일 불러오기

```
loader = TextLoader("./student_activity_records.txt", encoding="utf-8")  
documents = loader.load()
```

문서를 작은 조각(chunk)으로 분할

```
splitter = RecursiveCharacterTextSplitter(chunk_size=300, chunk_overlap=50)  
docs = splitter.split_documents(documents)
```

벡터 DB(Faiss)에 문서 저장

```
vectorstore = FAISS.from_documents(docs, embeddings)  
vectorstore.save_local("faiss_seotok")
```

저장한 벡터 DB 다시 로드 (주의: pickle 포함)

```
vectorstore = FAISS.load_local("faiss_seotok", embeddings, allow_dangerous_deserialization=True)  
retriever = vectorstore.as_retriever()
```

RAG 체인 구성 (앞서 만든 LLM 사용 가능)

```
rag_chain = RetrievalQA.from_chain_type(llm=llm, retriever=retriever)
```

텍스트 파일 불러오기

```
loader = TextLoader("./student_activity_records.txt", encoding="utf-8")  
documents = loader.load()
```

--- 문서 1 ---

홍길동: 과학 시간에 환경 문제에 대한 토론에 주도적으로 참여하였고, 자료 조사 내용을 발표하며 친구들의 이해를 도왔음.

홍길동: 과학 실험 보고서를 작성하며 실험 과정을 꼼꼼하게 기록하고 분석하여 논리적으로 정리하였음.

홍길동: 팀 프로젝트에서 의견을 조율하며 협력적으로 과제를 완수하였음.

김민지: 영어 독서 발표 시간에 자신만의 관점으로 내용을 요약하여 자신감 있게 발표하였음.

김민지: 친구들의 질문에 성실히 답변하며 영어 표현력 향상에 기여하였음.

김민지: 조별 활동에서 친구들과의 의사소통을 통해 협력적으로 문제를 해결하였음.

```
query1 = "홍길동 학생, 과학탐구활동 세특 초안 작성"  
print(rag_chain.invoke(query1))
```



{'query': '홍길동 학생, 과학탐구활동 세특 초안 작성', 'result': '홍길동 학생은 과학 시간에 환경 문제에 대한 토론에 주도적으로 참여하였고, 자료 조사 내용을 발표하며 친구들의 이해를 도왔습니다. 또한, 과학 실험 보고서를 작성하며 실험 과정을 꼼꼼하게 기록하고 분석하여 논리적으로 정리하였으며, 팀 프로젝트에서 의견을 조율하며 협력적으로 과제를 완수하였습니다. 이러한 내용을 바탕으로 홍길동 학생의 과학탐구활동 세부 내용을 세심하게 작성할 수 있을 것입니다.'}

❖ 쿼리의 차이는 RAG의 차이를 발생한다.

✓ '김민지 학생 '영어탐구활동' vs '영어 독서 발표 활동'

```
# 질의 예시: 세특 자동 생성
query0 = "김민지 학생, 영어탐구활동 세특 초안 작성"
result = rag_chain.invoke(query0)
print(result)
```

'영어탐구활동'

{'query': '김민지 학생, 영어탐구활동 세특 초안 작성', 'result': '죄송합니다. 영어탐구활동에 대한 김민지 학생의 내용은 제공되지 않았습니다.'}

```
query5 = "김민지 학생, 영어 독서 발표 활동 세특 초안 작성"
print(rag_chain.invoke(query5))
```

'영어 독서 발표 활동'

{'query': '김민지 학생, 영어 독서 발표 활동 세특 초안 작성', 'result': '김민지 학생은 영어 독서 발표 시간에 자신만의 관점으로 내용을 요약하여 자신감 있게 발표하고, 친구들의 질문에 성실히 답변하며 영어 표현력 향상에 기여했습니다. 따라서, 김민지 학생의 영어 독서 발표 활동 세부 내용을 포함한 초안을 작성할 때, 이러한 점을 강조하여 기술할 수 있습니다. 또한, 김민지 학생이 친구들과의 의사소통을 통해 협력적으로 문제를 해결한 경험도 함께 언급하면 좋을 것입니다.'}

❖ 활동별로 세특 자동 분류 실습

✓ 질의 예시

- "김민지 학생의 발표 활동에 해당하는 세특만 생성해줘"

✓ Prompt나 retriever 필터링 활용

❖ RAG 성능 비교 실험

✓ 동일 문서에 대해 직접 작성한 세특 vs. RAG 생성 세특 비교

✓ 학생 2~3명 예시로 작성, 교사들의 피드백 수집

Day4-5

실습 정리 및 마무리 토론

- ❖ 웹기반 Flowise를 이용한 AI 자동화
 - ✓ Colab 환경에서 구현은 복잡함
- ❖ 수준은 중급/고급 과정
 - ✓ 개인 노트북 환경에서 설치 및 AI Agent 자동화

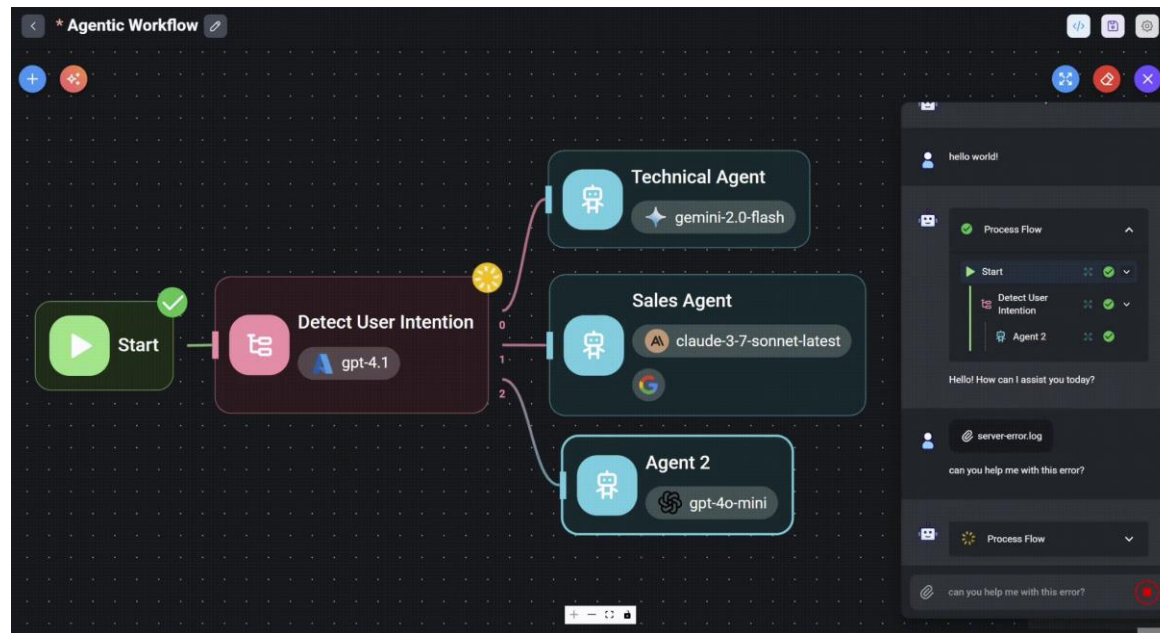
❖ 플로워즈는 코딩 없이 특화된 나만의 GPT를 만드는 무료 웹 기반 도구

✓ LangChain의 시각화 보완

- LangChain은 강력하지만 Python 코딩 필요
- 초보자 및 비전공자도 시각적 노드 연결 방식으로 쉽게 Agent 구축 가능

✓ 활용분야 및 예시

- 챗봇 개발
- 데이터 검색 시스템
- AI 자동화 에이전트
- 추천 시스템



- ❖ Flowise = LangChain을 기반으로 한 시각적 워크플로우 빌더
 - ✓ 브라우저에서 드래그 앤 드롭으로 LLM 파이프라인(프롬프트, 검색, API 호출 등)을 구성할 수 있음
 - ✓ GitHub에서 무료로 다운로드 가능 (MIT 라이선스)
- ❖ 장점
 - ✓ 코딩 경험이 없어도 LLM 앱 개발 가능
 - ✓ 시각적으로 구조를 파악하고 조정 가능
 - ✓ RAG 구현 속도가 매우 빠름 (몇 분이면 동작하는 프로토타입 제작 가능)
 - ✓ 오픈소스라서 자유롭게 수정·배포 가능
- ❖ 단점
 - ✓ 대규모 서비스용 최적화는 부족 → 프로덕션 환경에서는 별도 확장 필요
 - ✓ 커스터마이징이 복잡한 경우 여전히 코드 작성 필요
 - ✓ LangChain 업데이트 주기와 동기화가 필요
- ❖ Flowise = “LangChain의 레고 블록을 시각적으로 끼워 맞추는 도구”
 - ✓ 코드 없이도 블록(노드)을 연결해서 RAG, 챗봇, 문서 검색 시스템을 바로 만들 수 있음.

구조오프라인: 문서 임베딩 & 색인 파이프라인

문서를 쪼개고(Chunk), 임베딩으로 변환한 뒤 벡터 DB (Chroma/Pinecone/FAISS 등)에 저장

1. Document Loader

- PDF/HTML/웹/폴더/Notion/S3 등에서 문서 로드

2. Text Splitter

- 길이·의미 기반으로 문서 분할
- 추천 시작값: `chunk_size=800~1200`, `chunk_overlap=100~200`

3. Embeddings

- 예: OpenAI text-embedding-3-large, BGE/KorNLI 기반 한국어 임베딩 등

4. Vector Store (Upsert)

- Chroma(개발 편의) / Pinecone(클라우드 확장) / FAISS(고성능 로컬)
- 메타데이터(제목, 출처 URL, 페이지, 섹션 등) 함께 저장

2) 온라인: 질의응답(RAG) 파이프라인

1. Input (Chatflow API In)

- 사용자 질문이 들어오는 시작점

2. Retriever (Vector Store → Similarity Search)

- `topK=3~5` 추천, `similarity_threshold` 사용 가능
- 필터(메타데이터)로 문서 종류/날짜/태그 제한 가능

3. Prompt Template

- 시스템 프롬프트에 “규칙”과 “컨텍스트 삽입 자리” 정의

4. LLM

- GPT-4o/4.1-mini, Llama, Claude, Ko-전용 모델 등 선택
- 토큰 한도 고려해 `context window` 내로 유지

5. Output Parser (선택)

- JSON 스키마로 정답/근거/요약을 꺼내 쓰고 싶을 때

6. Memory (선택)

- 대화 맥락 유지 시 `ConversationBuffer` / `Summary` 메모리
- RAG 컨텍스트와 대화기억을 분리 관리하는 게 안전

7. Output

- 최종 답변 + 사용한 출처 링크(메타데이터) 포함해 반환

감사합니다.

Korea Institute of Science
and Technology Information

TRUST
KISTIL

