RAG를 활용한 사용자 맞춤 AI 뉴스레터

목차

1. 프로젝트 소개
2. 워크플로우
3. 실험설계
4. 데이터 크롤링
5. RAG
   - 뉴스 크롤링 DB에서 쿼리 문장 유사도 기반 문서 Retrieve
   - Retrieve된 뉴스 문서 텍스트와 쿼리 문장을 조합한 프롬프트
   - 프롬프트를 이용해 LLM에서 출력된 뉴스 요약문으로 레터 구성
6. 모델 성능 측정
7. 프롬프트 엔지니어링
   - LLM을 이용한 키워드 추출
   - 요약문 출력을 위한 프롬프트
8. 마무리

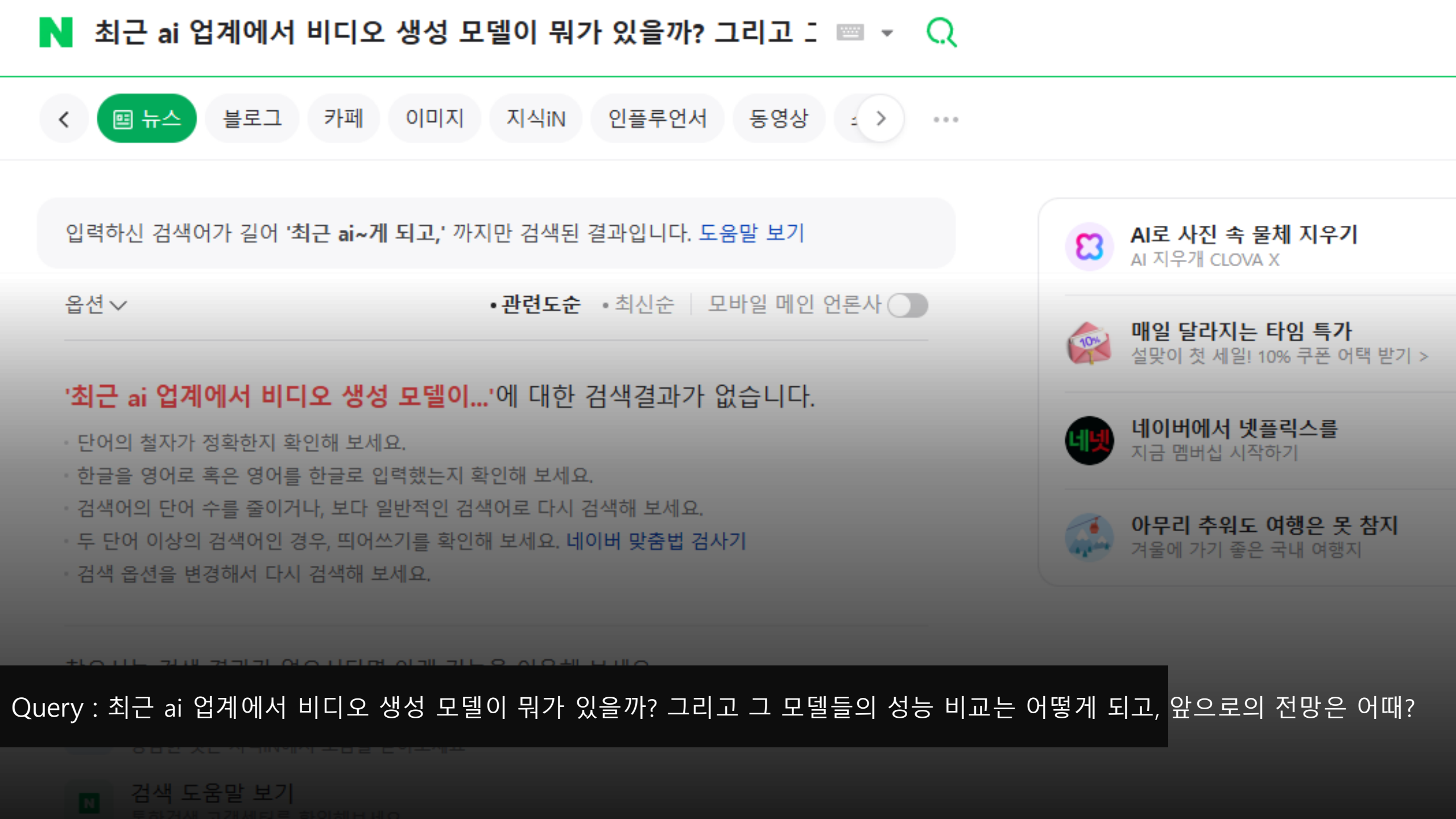
프로젝트 소개

문제 정의

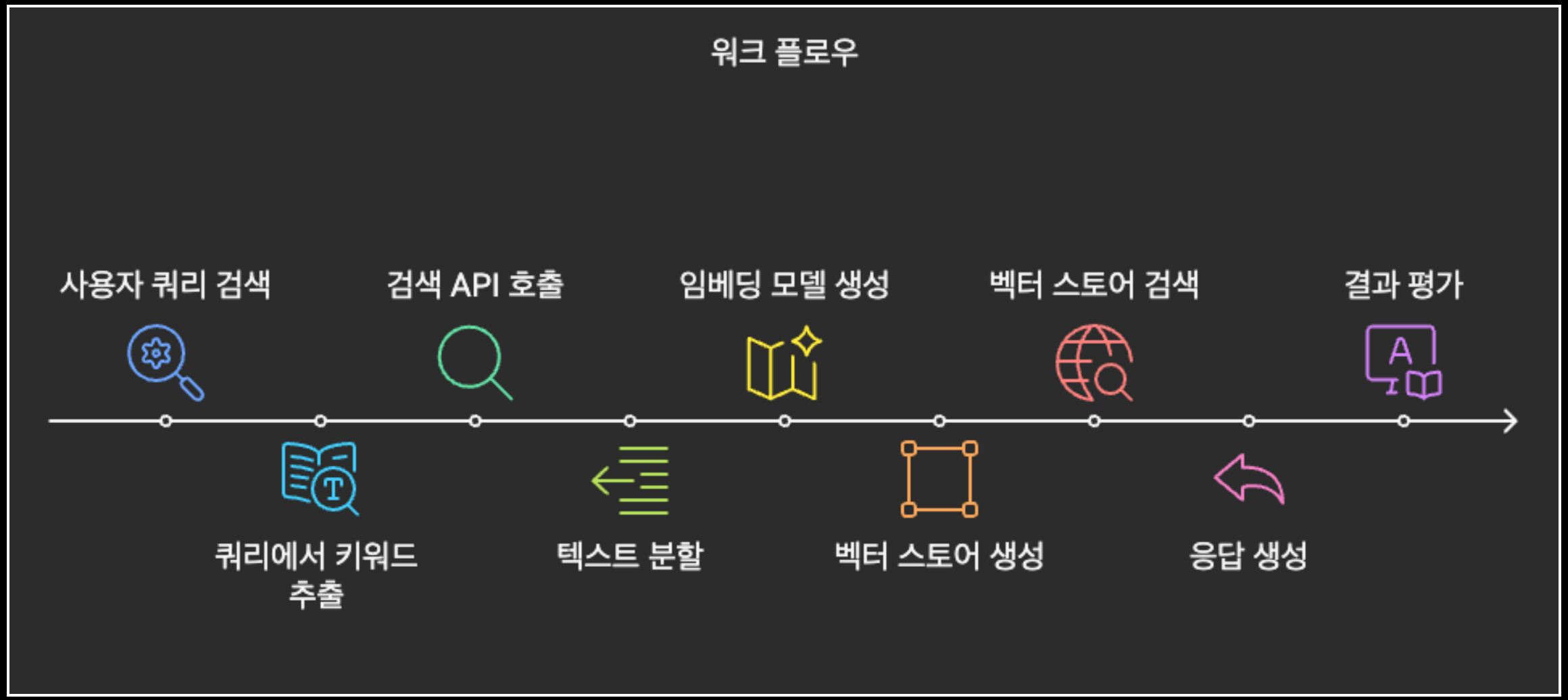
- 기존 검색 엔진을 활용하여 특정 도메인(AI)에 대해 키워드 나열이나 질문형 쿼리로 검색을 시도했을 때, 관련된 정확한 답변이나 유용한 정보를 얻는 데 어려움 있음

- 검색 결과는 사용자 의도를 충분히 반영하지 못하거나, 필요한 정보를 정확히 제공하지 못하는 경우가 빈번하게 발생 이는 사용자가 원하는 정보를 얻기 위해 추가적인 탐색 시간을 필요로 함.

- 이런한 문제를 해결해보고자 **사용자 맞춤형 RAG(Retrieval-Augmented Generation) 기반 뉴스레터 시스템**을 설계 및 구현하고자 함



Query : 최근 ai 업계에서 비디오 생성 모델이 뭐가 있을까? 그리고 그 모델들의 성능 비교는 어떻게 되고, 앞으로의 전망은 어때?


워크 플로우
사용자 쿼리 검색
검색 API 호출
임베딩 모델 생성
벡터 스토어 검색
결과 평가
쿼리에서 키워드 추출
텍스트 분할
벡터 스토어 생성
응답 생성

워크플로우

Brainstorming - 실험설계

확장성 : 더욱 넓은 범위의 도메인은
RAG가 검색(Retrieval)할 수 있는 DB의
크기 영역

신뢰도 : 검색 자체의 신뢰도는 DB를
구축할 내용물의 영역

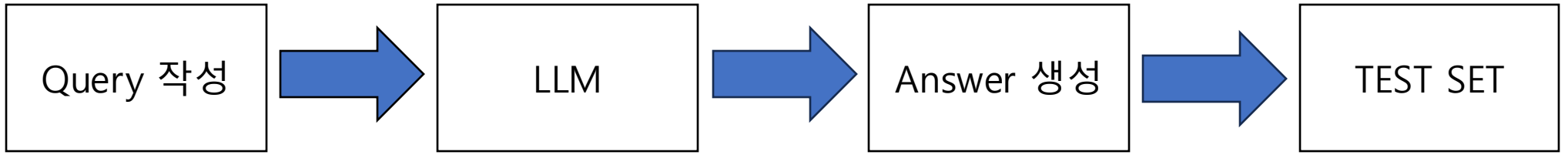
정확도 : 사용자 query에 대한 RAG의
출력물 평가는 사용자의 영역

사용자의 측면에서 정확도를 측정할 때, 자원의 한계를 극복할 정성평가의 대안의 필요성

실험설계 - 어떻게 평가할 것인가

Query 역생성을 통한 TEST SET 구축

기존



커스텀한 RAG의 벡터 데이터베이스와 LLM이 활용하는 데이터베이스 간의 극심한 격차

-> 벡터 데이터베이스 내부의 Retrieval 능력을 측정하는데 어려움

ZDNET Korea 구독 PICK

IBM "양자컴퓨터 상용화 핵심은 오류 수정…2029년 완성 목표"

입력 2025.01.22. 오후 10:56 기사원문

[redacted]

1 4 가가

| 현재 양자컴퓨터·슈퍼컴퓨터 융합 방식 활발…"3년 내 '양자 우월성' 발생할 것"

"양자컴퓨터가 상용화하려면 데이터 계산을 오류 없이 빠르게 처리해야 합니다. 이에 기업이 양자컴퓨터 오류 현상 방지를 위해 연구하고 있습니다. IBM은 2029년 오류 수정(error correction) 기능을 완벽히 갖춘 양자컴퓨터 개발을 목표로 뒀습니다."

한국IBM 표창희 아시아·태평양 지역 퀀텀 엔터프라이즈 영업 총괄상무는 최근 여의도 한국IBM 사옥에서 진행한 기술 세션에서 양자컴퓨터 개발 현황과 자사 목표를 이같이 밝혔다. 빠른 연산 처리를 오류 없이 할 수 있는 오류 수정 기술이 절실하다는 설명이다.

양자컴퓨터는 양자물리학 기반으로 대규모 연산 처리를 할 수 있는 새로운 형태의 컴퓨터다. 일반 슈퍼컴퓨터는 연산 처리를 '비트' 단위로 계산한다. 0 또는 1로만 사용해 출력값을 내놓는 이진법 형태다. 양자컴퓨터는 '큐비트' 단위로 처리한다. 큐비트는 0과 1 사이 무한한 값을 동시에 처리할 수 있다. 이를 '양자얽힘' 또는 '중첩'이라 부른다.

Query 생성

양자컴퓨터가 상용화되기 위해서는 어떤 기술들이 필요해? 장기적인 전망이나 실현 가능성은?

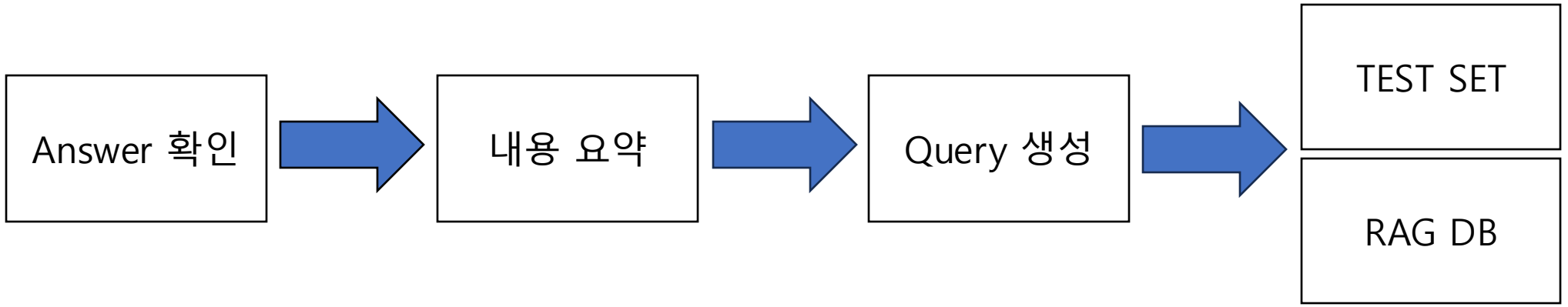
양자컴퓨터가 기존의 컴퓨터보다 훨씬 빠른 속도로 데이터를 처리할 수 있는 것은 이해가 되는데, 그래서 왜 오류 수정 기술이 왜 상용화의 핵심이라는거야?

그래서, 양자컴퓨터의 상용화를 위해서, IBM이나 다른 기업들이 이 오류 수정 기술을 어떻게 연구하고 있어?

실험설계 - 어떻게 평가할 것인가

Query 역생성을 통한 TEST SET 구축

아이디어



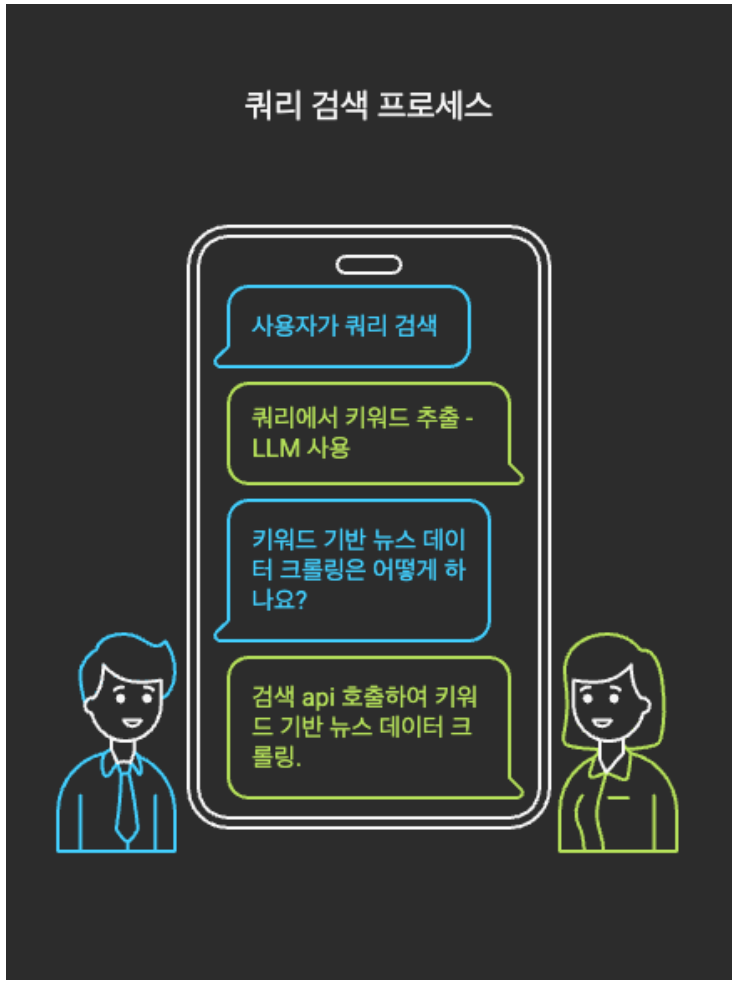
정량적인 지표를 통해 정성적인 평가를 간접적으로 파악 - 주관적 요소 감소

사용자 query 대해 RAG 알고리즘이 유의미한 정보를 제공할 수 있는지 평가하는데 유효

기계적인 평가가 가능

-> 정성 평가를 전부 대체할 순 없다는 한계점 존재

데이터 크롤링



Title	URL	Content
제목 1	https://n.news...	본문 내용 1
제목 2	https://n.news...	본문 내용 2

총 1,012개 데이터 확보

	Title	URL	Content
0	한발 늦게 'AI 원주민' 되겠다는 카카오...성능 발표는 '소심하게'	https://n.news.naver.com/mnews/article/138/000...	[디지털데일리 이건한 기자] 지난 22일 자체 개발한 통합 인공지능(AI) 플랫...
1	엔비디아가 포토샵을 '생성 AI 혁명'으로 이끈다	https://n.news.naver.com/mnews/article/262/000...	[박원익의 유익한 IT] "인공지능 '아이폰 모멘트' 시작됐다" ● 거대언어모델이...
2	오픈AI, 생산성 관리 '프로젝트' 기능 공개...향후 '쉽마스'서 AI 에이전트 등장 관심	https://n.news.naver.com/mnews/article/092/000...	생산성·자율성 증가 기대감...업계 "남은 5일간 차세대 모델·오퍼레이터 선보일 듯" ...
3	구글·삼성전자 'AI 동맹군' XR 전쟁서 위력 발휘할 것	https://n.news.naver.com/mnews/article/009/000...	산제이 굽타 구글 아시아·태평양 총괄사장 올해 'AI 에이전트' 경쟁 치열 내년 빅...
4	국방 AI 혁신...합성데이터에 주목 [각스]	https://n.news.naver.com/mnews/article/015/000...	KIDA 국방데이터연구단, 국방 데이터 활용 논의 젠젠에이아이, 합성데이터로 전시 ...

실제 데이터 예시

기본 모델 구축

RAG_DB

임베딩 모델

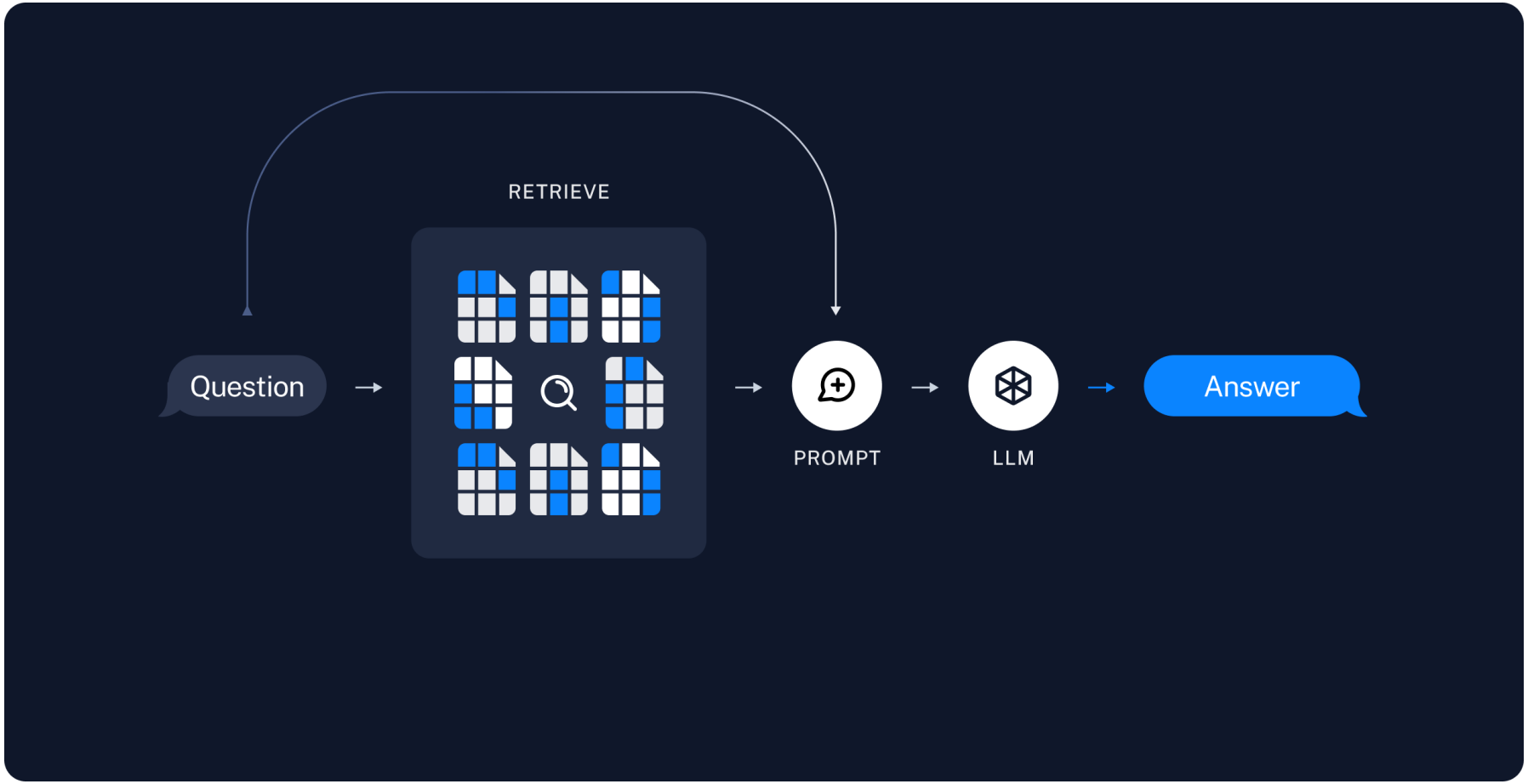
Adapter Models for

intfloat/multilingual-e5-large-instruct

huggingface.co/models

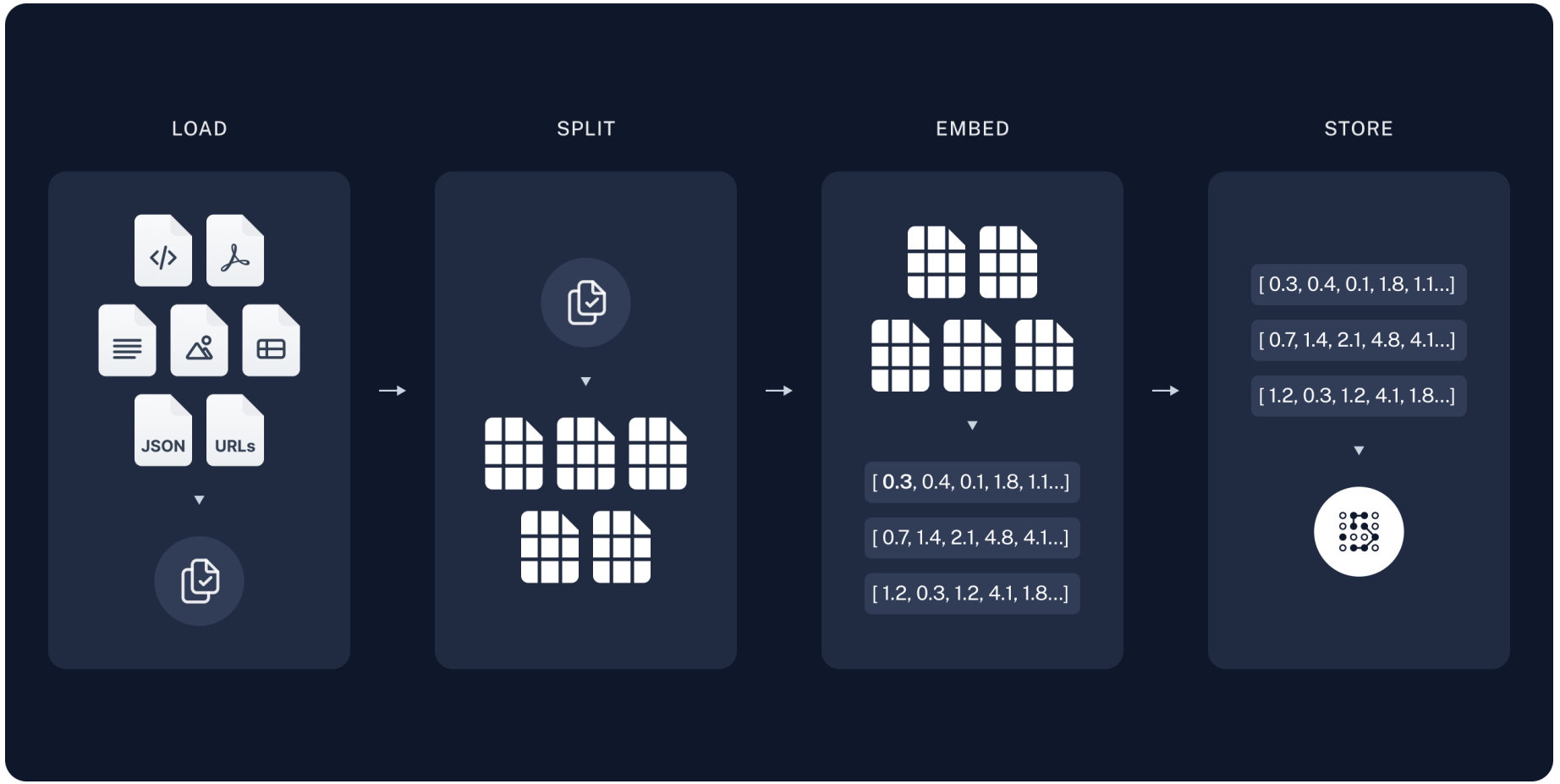
생성 모델

RAG (Retrieval-Augmented Generation)



https://wikidocs.net/2333410

RAG (Retrieval-Augmented Generation)



https://wikidocs.net/2333410

모델 성능 측정에 사용한 평가지표

평가항목	이름	설명
생성 (Generation)	Faithfulness	생성된 응답이 검색된 컨텍스트에 기반했는지를 평가
생성 (Generation)	Answer Relevancy	생성된 응답이 사용자 질문(프롬프트)과 얼마나 관련이 있는지를 평가
검색 (Retrieval)	Context Recall	검색된 문서 중 얼마나 많은 문서가 실제로 관련 있는지를 평가 (검색된 정보의 정확도)
검색 (Retrieval)	Context Precision	검색된 문서 중 실제로 관련 문서가 얼마나 포함되었는지를 평가 (필요한 정보를 얼마나 잘 찾는지)

모델 성능 측정에 사용한 평가지표

$$\text{Faithfulness score} = \frac{|\text{Number of claims in the generated answer that can be inferred from given context}|}{|\text{Total number of claims in the generated answer}|}$$

$$\text{answer relevancy} = \frac{1}{N}\sum_{i=1}^{N} cos(E_{g_i}, E_o)$$

$$\text{context recall} = \frac{|\text{GT sentences that can be attributed to context}|}{|\text{Number of sentences in GT}|}$$

$$\text{answer relevancy} = \frac{1}{N}\sum_{i=1}^{N} \frac{E_{g_i} \cdot E_o}{\|E_{g_i}\|\|E_o\|}$$

$$\text{Context Precision@K} = \frac{\sum_{k=1}^{K} (\text{Precision@k} \times v_k)}{\text{Total number of relevant items in the top } K \text{ results}}$$

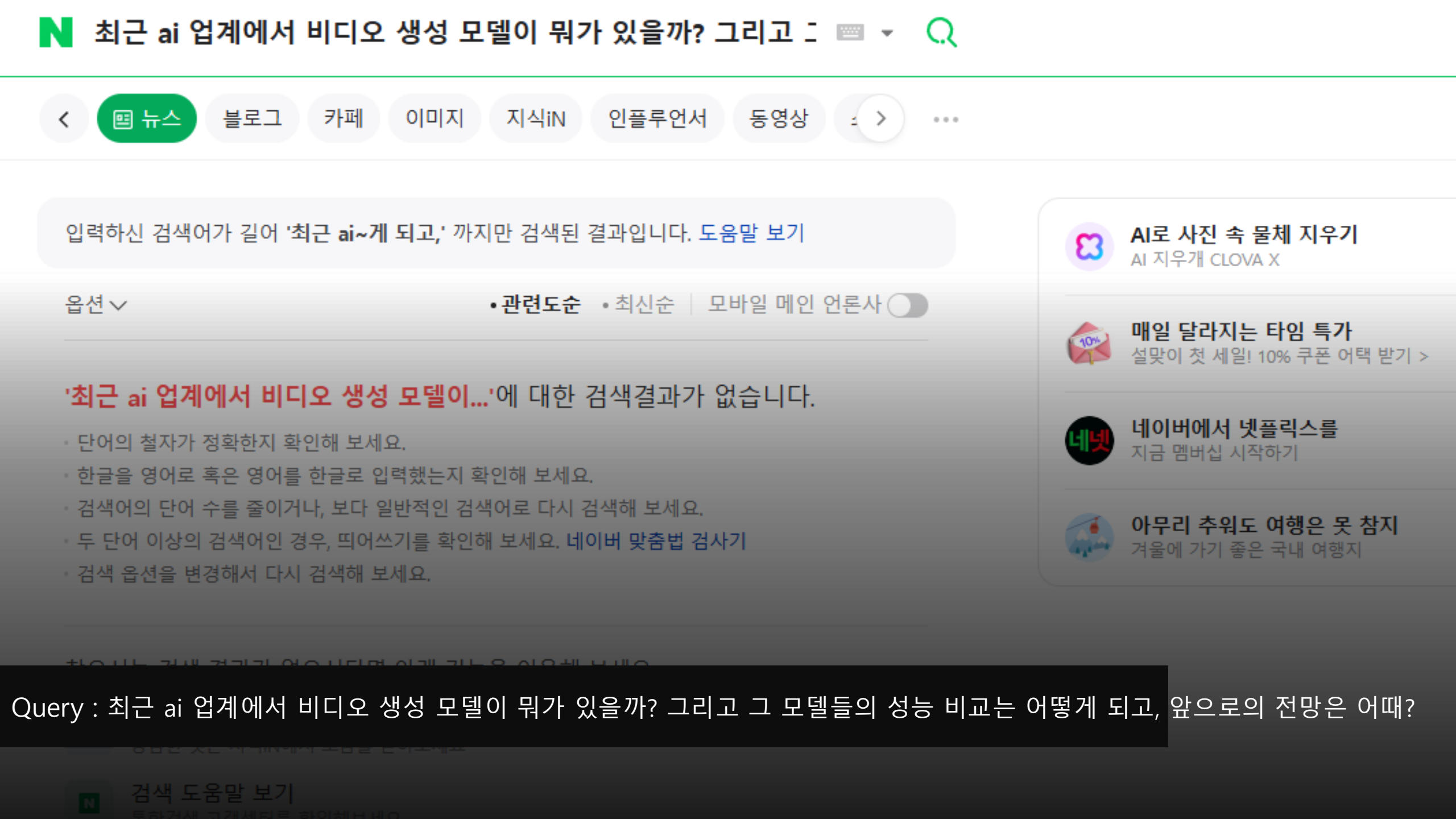
$$\text{Precision@k} = \frac{\text{true positives@k}}{(\text{true positives@k} + \text{false positives@k})}$$

모델 성능 측정에 사용한 평가지표

평가항목	이름	TEST SET 평균 수치
생성 (Generation)	Faithfulness	**1.000000**
생성 (Generation)	Answer Relevancy	**0.894704**
검색 (Retrieval)	Context Recall	**0.914095**
검색 (Retrieval)	Context Precision	**0.417773**

모델 성능 측정에 사용한 평가지표 - 요약

Metrics	Score
BERTScore Recall	0.6523
BERTScore Precision	0.7093
BERTScore F1	0.6792
BLEU Score	0.0142
METEOR Score	0.0940
ROUGE-1 Recall	0.0816
ROUGE-1 Precision	0.3562
ROUGE-1 F1	0.1270
ROUGE-2 Recall	0.0306
ROUGE-2 Precision	0.1377
ROUGE-2 F1	0.0485
ROUGE-L Recall	0.0792
ROUGE-L Precision	0.3473
ROUGE-L F1	0.1234



Query : 최근 ai 업계에서 비디오 생성 모델이 뭐가 있을까? 그리고 그 모델들의 성능 비교는 어떻게 되고, 앞으로의 전망은 어때?

```
# 대화형 QA 시스템 실행
interactive_qa()
```

```
AI 질문 답변 시스템 (종료하려면 'quit' 입력)
--------------------------------------------------
질문을 입력하세요: 최근 AI 업계에서 비디오 생성 모델이 뭐가 있을까? 그리고 그 모델들의 성능 비교는 어떻게 되고, 앞으로의 전망은 어때?

답변 생성 중...

답변:
최근 AI 업계에서 주목받는 비디오 생성 모델로는 오픈AI의 Sora, 구글의 Imagen Video (비오2), 메타의 Make-A-Video (무비젠)이 있습니다.

**성능 비교:** 세 모델 모두 장단점이 있으며, 단순한 성능 비교는 어렵습니다. 기사에 따르면, 메타는 자사의 Make-A-Video가 오픈AI의 Sora보다 더 나은 결과물을 만들어냈다고 주장하지만, 정식 출시되지 않아 실제 성능(

**앞으로의 전망:** 동영상 콘텐츠 수요 증가와 AI를 통한 영상 제작 시간 및 비용 절감 효과로 인해 AI 비디오 생성 도구 시장은 계속 성장할 것으로 예상됩니다. 각 기업은 기술력을 더욱 발전시켜 해상도, 물리적 정확도,

참고 출처:
1. https://n.news.naver.com/mnews/article/009/0005429554?sid=105
2. https://n.news.naver.com/mnews/article/009/0005429554?sid=105
3. https://n.news.naver.com/mnews/article/009/0005429554?sid=105
4. https://n.news.naver.com/mnews/article/052/0002142109?sid=101

계속하려면 Enter를 누르세요...
```

최근 AI 업계에서 주목받는 비디오 생성 모델로는 오픈AI의 Sora, 구글의 Imagen Video (비오2), 메타의 Make-A-Video (무비젠)이 있습니다.

성능 비교 : 세 모델 모두 장단점이 있으며, 단순한 성능 비교는 어렵습니다. 기사에 따르면, 메타는 자사의 Make-A-Video가 오픈AI의 Sora보다 더 나은 결과물을 만들어냈다고 주장하지만, 정식 출시되지 않아 실제 성능에 대한 객관적인 평가는 부족합니다.

Sora는 일반 사용자를 대상으로, Imagen Video와 Make-A-Video는 전문가를 대상으로 하는 것으로 보이며, 각 모델의 강점이 다릅니다.

Sora는 텍스트 외 이미지 입력이 가능하고, Imagen Video는 높은 해상도를 자랑하며, Make-A-Video는 기존 영상 편집에 강점이 있다고 언급되어 있습니다. 물리 엔진 구현 능력 또한 모델마다 차이가 있을 것으로 예상됩니다.

앞으로의 전망 : 동영상 콘텐츠 수요 증가와 AI를 통한 영상 제작 시간 및 비용 절감 효과로 인해 AI 비디오 생성 도구 시장은 계속 성장할 것으로 예상됩니다.

각 기업은 기술력을 더욱 발전시켜 해상도, 물리적 정확도, 편집 기능 등을 개선하려 할 것이며, 사용자 경험을 향상시키는 데 집중할 것으로 보입니다.

하지만, 모델 훈련에 사용된 데이터의 투명성 부족과 저작권 문제 등 해결해야 할 과제도 남아 있습니다.

google

최근 AI 업계에서 비디오 생성 모델이 뭐가 있을까? 그리고 그 모델들의 성능 비교는 어

전체 이미지 뉴스 도서 동영상 지도 쇼핑 더보기

모두의연구소
https://modulabs.co.kr › blog › openaisora

OpenAI Sora, 비디오 생성 AI의 시대를 열다 | 블로그

2024. 2. 19. — OpenAI가 공개한 **비디오 생성 AI모델** Sora에 대해 알려드립니다. 텍스트 프롬프트를 통해 최대 1분 길이의 **비디오**를 **생성**할 수 있는 **AI 모델** Sora를 ...
누락된 검색어: 뭐가 있을까? 그리고 전망 어때?

hahahahr.com
https://hahahahr.com › business-ai

영상 생성 AI 비교 : runway VS Lumalab 성능을 비교해보자.

2024. 6. 18. — **LumaLabs AI의 Dream Machine**은 텍스트와 이미지에서 고품질의 현실적인 비디오를 생성하는 최신 AI 모델이다. 주요 기능으로는. 1) 사용자가 입력한 ...

AI 히어로즈
https://aiheroes.ai › community

구글 영상 생성 AI '비오2(Veo2)' 소라(Sora)와 비교하기

2024. 12. 27. — 오픈 **AI에서** 공개한 소라(Sora) **AI** 영상 **생성 모델이** 올해 많은 관심을 받았는데, 구글 딥마인드의 비오2 **모델**은 소라에 비해 해상도는 4배, **생성** ...
누락된 검색어: 최근 뭐가 있을까? 그리고 전망 어때?

네이버 프리미엄콘텐츠
https://contents.premium.naver.com › aipost › contents

"더 사실적인 영상 생성"...루마 AI, 새로운 영상 생성 모델 '레이 2 ...

5일 전 — 동영상 생성 인공지능(**AI**) 스타트업인 루마 **AI**(Luma **AI**)가 새로운 **비디오 생성 모델**인 레이 2(Ray 2)를 공개했다. 드림 머신 홈페이지와 모바일 앱을 ...

Chat GPT

최근 AI 업계에서는 다양한 **비디오 생성 모델**이 개발되고 있습니다. 주요 모델과 그들의 성능 및 전망을 살펴보겠습니다.

주요 비디오 생성 모델

1. **KlingAI**:
   - **특징**: 중국에서 개발된 이 모델은 텍스트를 기반으로 비디오를 생성하는 'T2V'와 기존 이미지를 활용한 'I2V' 기능을 제공합니다. 현재 1.5버전까지 출시되었으며, 무료 버전은 5초 길이의 비디오만 생성할 수 있습니다. 생성 속도가 다소 느리지만, 유료 버전으로 업그레이드하면 더 긴 비디오와 빠른 생성 속도를 경험할 수 있습니다. 스테이블 디퓨전의 모든것
2. **RunwayML의 Gen-3 Alpha Turbo**:
   - **특징**: 이 모델은 이미지에서 비디오로 변환하는 기능을 제공하며, 기존 모델보다 7배 더 빠른 속도로 작업을 수행할 수 있습니다. 가격이 절반으로 줄어들어 많은 사용자들에게 접근성이 높아졌습니다. [블로그광고] 수익내기
3. **Haiper AI**:
   - **특징**: 비교적 신생 플랫폼으로, 텍스트 기반 비디오 생성에 최적화되어 있습니다. 사용자가 입력한 간단한 텍스트를 기반으로 고품질의 비디오 콘텐츠를 생성할 수 있으며, 빠른 렌더링 속도와 사용자 친화적인 인터페이스를 제공합니다. 지식의 힘
4. **구글의 Veo 2**:
   - **특징**: 구글이 최근 발표한 동영상 생성 AI로, 텍스트 프
     질 동영상을 자동으로 생성합니다. 시각적 스토리텔링
     기대됩니다. 김쿠삼 블로그

지식의 힘
최신 AI 영상 생성 플랫폼 대결 : Haiper AI, Lumalabs.ai, RunwayML Gen-3 ...
2024년 10월 28일 — 최근 AI 영상 생성 플랫폼의 신흥 강자로 떠오르고 있는 Haiper AI, Lumalabs....

5. **메타의 Movie Gen**:
   - **특징**: 메타가 발표한 새로운 AI 모델로, 텍스트 프롬프트를 통해 현실감 있는 비디오와 오디오 클립을 생성할 수 있습니다. 이 ↓ 델은 비디오 편집, 사운드 매칭, 개인화된 콘텐츠 생성 등 다양한 기능을 제공합니다. 로이터

서비스 고도화를 위한 작업

서비스 고도화 - Prompt Engineering

	Phase 1: Prompt to Keyword Extraction	Phase 2: Chunk Data to Summary Extraction
모델 한계	• 문장 형식의 쿼리를 사용하는 경우, 답변이 없거나 매우 희소한 결과값을 얻을 확률이 높음 (네이버 뉴스는 0000 알고리즘을 사용하기 때문에....) • 사용자의 쿼리를 단순 활용하는 경우, 검색 키워드가 한정적으로 충분한 데이터를 얻지 못함 • 사용자의 쿼리가 모호하거나 불명확한 경우 원하는 정보값을 얻기 어려움	▪ 상용 모델을 사용하더라도 구체적 입력값이 주어지지 않았을 때 일반적으로 출력되는 요약문의 형식은 한정되어 있음. (개조식 서술, 근거-결론식 구조) ▪ 각각 다른 쿼리를 입력할 때마다 상이한 형식, 어투의 요약문이 출력될 수 있다. ▪ LangChain에 내장된 기본 Q&A 프롬프트 서식을 사용해 출력
개선 사항	• 문장 형식의 쿼리에서 문맥을 유지하는 핵심 키워드를 추출하여 0000 형식의 알고리즘에 적합한 포맷으로 변환함 • 핵심 키워드 추출 및 중복되거나 모호한 표현을 배제함으로써 더욱 명확한 쿼리를 생성하도록 지원함 • 동의어 및 유의어 범위까지 확장하여 유사한 의미의 키워드를 포괄하는 키워드 리스트셋을 생성함	▪ 서비스에서 제공하고자 하는 의도에 맞는 일관된 어투, 서로 다른 쿼리에 대해서도 동일한 형식의 요약문 출력 ▪ 뉴스 본문을 읽지 않아도 될 정도의 세부정보 유지, 불필요한 정보를 제거한 요약문. ▪ 키워드 추출 과정에서 얻은 증강된 쿼리 정보를 활용

활
- Prompt Engineering
좋은 Chatbot 기반으

Prompt Engineering: Prompt to Keyword Extraction

Base LLM Model : Llama-3.2-3B-Instruct (Quantized Model)

Step 1: Prompt for Key Phrases Extraction

```
""" You are an intelligent assistant specializing in generating search-friendly keyword phrases. Your task is to extract concise, contextually accurate keyword phrases from a given user query. These keyword phrases should be optimized for news search and adhere to the following guidelines:

1. Ensure each phrase is specific and clear.
2. Maintain the user's intended context.
3. Include relevant dates, time periods, or locations when applicable.
4. Use Named Entities (e.g., people, places, organizations, events) mentioned in the query.
5. Incorporate synonyms or related terms to broaden the search scope.
6. Keep phrases concise and avoid redundant words.
7. Construct keyword phrases suitable for Boolean operators (e.g., AND, OR).
8. Include regional or contextual relevance when specified.
9. Reflect actions or dynamic processes, not just static terms.
10. Ensure the phrases are structured for search engine optimization.

**Output Format**:
- Provide 3-5 keyword phrases as a numbered list.
- Each phrase should be a complete and logical search query.

**Example User Query**:
"2025년에 헬스케어 분야에서 AI 기술이 어떻게 발전하고 있나요?"

**Example Output**:
1. 2025년 AI 기술 헬스케어 발전
2. AI 헬스케어 트렌드 2025
3. 2025년 헬스케어 AI 최신 동향
4. AI 기반 헬스케어 혁신 2025

Now, extract keyword phrases from the following query:
"{USER_QUERY}"""
```

- **LLM 역할 명시**
  검색 엔진 기반 포맷에 맞춰 사용자 쿼리 기반 핵심 구문을 파악하여 추출하는 역할 부여
- **가이드라인 제공**
  뉴스 검색에 적합한 키워드의 특징(i.e. 고유명사 포함, 키워드 관련 유의어 생성 등) 및 결과물에 필수적으로 포함되어야 하는 내용 명시
- **One-shot Learning**
  사용자의 쿼리 기반 핵심 구문 및 관련 동의어를 추출한 Key Phrases를 숫자 리스트로 제공할 수 있도록 예시 제공

<실행 결과>

- **User Query**:
  "최근 AI 업계에서 비디오 생성 모델이 뭐가 있을까? 그리고 그 모델들의 성능 비교는 어떻게 되고, 앞으로의 전망은 어때?"
- **Output:**
  1. AI 비디오 생성 모델 전망
  2. 최근 비디오 생성 모델 성능
  3. AI 비디오 생성 모델 최신 동향
  4. 비디오 생성 모델 성능 향상 전망

Step 2: Prompt for Keywords Extraction

```
You are an intelligent assistant specializing in keyword extraction.
Your task is to extract **Matched Keywords** and **Related Keywords** from a given document.
The output should be presented as two separate numbered lists for clarity.
Make sure that the list should be written in query language.

**Guidelines**:
1. **Matched Keywords**: Extract up to 5 keywords explicitly mentioned in the document.
2. **Related Keywords**: Extract up to 5 keywords contextually related to the document but not explicitly mentioned.

**Output Format**:
- Matched Keywords:
1. [Keyword 1] ₩n2. [Keyword 2] ₩n3. [Keyword 3]...
- Related Keywords:
1. [Keyword 1] ₩n2. [Keyword 2] ₩n3. [Keyword 3]...

**Example Input**:
Document: "트랜스포머 기법을 쓰지 않은 LLM모델의 트렌드"

**Example Output**:
- Matched Keywords:
1. 트랜스포머 ₩n2. 기법 ₩n3. LLM ₩n4. 모델

- Related Keywords:
1. 자연어 ₩n2. 언어모델 ₩n3. AI ₩n4. 챗봇

Now, extract keywords for the following document:
Document: {USER_QUERY}
```

- **LLM 역할 명시**
  사용자의 쿼리를 기반으로 매칭 키워드와 연관 키워드를 추출하는 LLM으로써의 역할 부여
- **가이드라인 제공**
  각 매칭 키워드와 연관 키워드 추출 방식에 대한 세부적인 가이드라인 제시
- **추출 포맷 및 예시 제공**
  리스트 형태의 결과물을 반환할 수 있도록 Output 형식 제공 및 One-shot Learning을 위한 예시 제공

<실행 결과>

- **User Query**:
  "최근 AI 업계에서 비디오 생성 모델이 뭐가 있을까? 그리고 그 모델들의 성능 비교는 어떻게 되고, 앞으로의 전망은 어때?"
- **Output:**
  ['비디오', '생성', '모델', '성능', '전망', 'AI', '컴퓨터', '화면', '영상', '인공지능']

Prompt Engineering: Prompt for Summarization

Base LLM Model : 000

```
from langchain.chains import create_retrieval_chain
from langchain.chains.combine_documents import create_stuff_documents_chain
from langchain_core.prompts import ChatPromptTemplate
from langchain_openai import ChatOpenAI



retriever = ...  # Your retriever
llm = ChatOpenAI()


system_prompt = (
    "Use the given context to answer the question. "
    "If you don't know the answer, say you don't know. "
    "Use three sentence maximum and keep the answer concise. "
    "Context: {context}"
)
prompt = ChatPromptTemplate.from_messages(
    [
        ("system", system_prompt),
        ("human", "{input}"),
    ]
)
question_answer_chain = create_stuff_documents_chain(llm, prompt)
chain = create_retrieval_chain(retriever, question_answer_chain)


chain.invoke({"input": query})
```

```
<SYSTEM>
 당신은 출판사의 담당 편집자입니다.
 아래 질문(query)을 바탕으로 주어진 문서(document)를 다섯 문장 이내로 요약(summary)해야 합니다. 키워드 문장(keywords)과 제목(title)을 참고해서 생성하세요.
<EXAMPLE 1>
 query:
 미국 정치권에 발생한 중요한 이슈에 따라서 첨단 산업에 어떤 변화가 있을까?
 keywords:
 미국 정치 이슈와 첨단 산업 트렌드
 title:
 "트럼프 2.0시대 개막 4대 테마는?" 에너지, AI·로봇,우주, 안보
 document:
 미국 제47대 대통령 선거뿐만 아니라 상·하원 모두 공화당이 승리함에 따라...
 summary:
 미국 정치 이슈와 첨단 산업의 트렌드에 대해 트럼프 2.0 시대의 미국 우선주의와...</s>
<EXAMPLE 2>:
 query:
 보험 산업에서 AI 기술을 활용하여 어떤 새로운 서비스와 사업이 등장하고 있나요?
 keywords:
 보험산업과 AI
 title:
 시니어·AI·해외시장...신사업 시동거는 보험사
 document:
 삼성생명 시니어리빙 TF →시니어 Biz 팀 격상 교보생명 '보장분석 AI 서포터'...
 summary:
 보험 산업에서 AI 기술을 활용한 새로운 서비스와 사업이 등장하고 있습니다...</s>
<ASSISTANT>:
 query:
 AI 에이전트가 기업의 생산성과 효율성을 향상시키는 데 어떻게 사용되고 있나요?
 keywords:
 에이전트 AI
 title:
 "AI와 AI가 협업하는 시대"...새해 핵심 트렌드 'AI 오케스트레이션'
 document:
 아마존웹서비스·마이크로소프트·카카오 등 주력 서비스로 제시 2025년 인공지능( AI ) 산업...
 summary:
```

Limitation & Future Work

- 이곳에 실제로 값을 넣으면 됩니다!
- 모델 측정 후 넣을 예정

- 부족한점
- -> 여러 모델(Gemini-1.5-pro or Lamma3.2 등)간 의 성능을 비교하려고 하였으나 시간 부족으로 인해 한 개의 모델만 평가하고 나머지는 수행하지 못하였음
- -> 추후 수행할 예정