
REAL
SUMMIT 2023
SAMSUNG SDS

기업의 숨겨진 보물,
비정형 데이터를 제대로 활용하는 법:
No More Hallucination

삼성SDS GenAI 아키텍처 담당 **백창현 상무**

AGENDA

I. LLM 시대의 의의

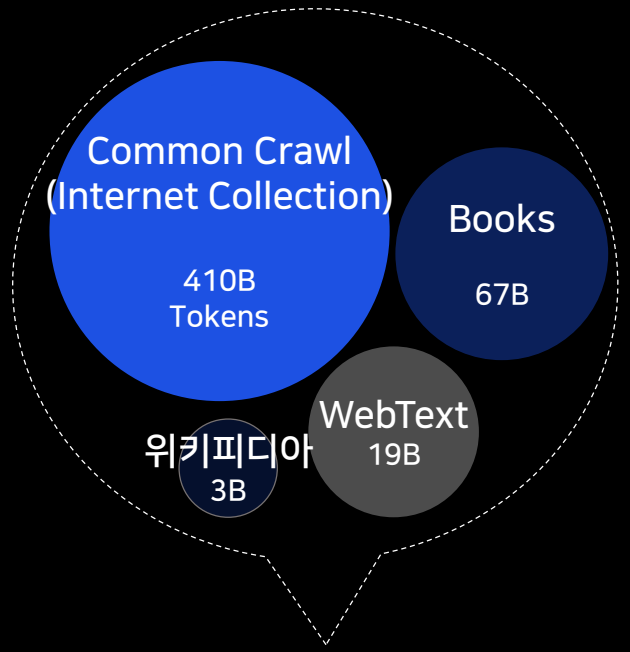
II. 비정형 데이터 활용 방안

III. 준비 필요 사항

Large Language Model 시대의 도래

LLM은 무엇이 다른가?

방대한 학습 데이터



GPT-3 학습 데이터 : 570 GB, 1.5조 토큰
(도서 일백 오십만 권 이상)

사전 학습

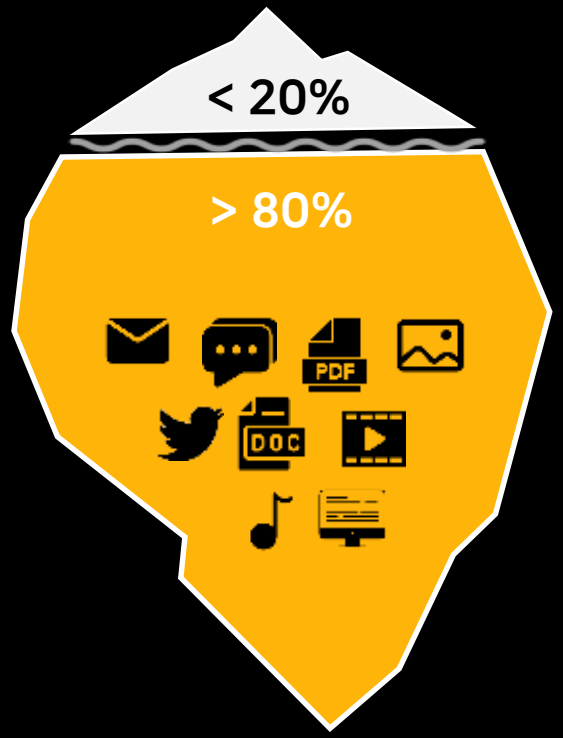
거대한 트랜스포머 모델

GPT-3 모델 파라메터 : 1750억 개

LLM의 차별점

- 인간의 언어를 이해하는 고도의 능력
- 자연어 기반의 다양한 업무 처리 능력

LLM과 비정형 데이터



정형 데이터

- 열과 행으로 표시
- 숫자, 날짜, 스트링
- 판매, 잔고, 체결 등
- 관계형 DB에 저장

XY	1	2
A	A1	A2
B	B1	B2
C	C1	C2
D	D1	D2

비정형 데이터

- 기업 데이터의 80% 이상을 차지
- 텍스트, 이미지, 오디오, 동영상
- 보고서, 기술문서, 회의록, 고객 대응 기록 등
- 파일로 저장

Large Language Model

잠자고 있던
기업 내 비정형 데이터를
쉽고 가치 있게 활용 가능

기업 경쟁력 제고를 위한
중요 요소로 부상

LLM 중심의 기업 정보 활용 체계

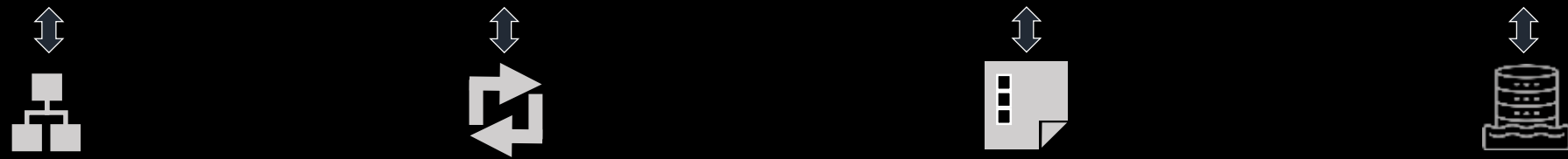
외부 검색 조합 | 대화형 웹 인터페이스 | 지식 검색 | 벡터 기반 검색 인덱스

AI 기반 사용자 인터페이스

LLM Powered Enterprise Data Access

- 맞춤학습
- 검색기반지식답변
- Vector DB

업무 시스템 인터페이스 (데이터 접근/수집/조회)



협업 도구, 메일, 게시 | 고객 대화/대응 기록 (CRM) | 보고서, 회의록, 기술문서 (파일 저장소) | 데이터 웨어하우스 /데이터 레이크

기업 Data를 LLM에 적용하는 방안

	맞춤 학습 (Finetuning)	검색 기반 지식 답변 (RAG)
활용 방식	기업 내부지식으로 추가 학습하여 기업 전용 모델을 생성함으로써 질문에 대한 기업에 특화된 답변 능력 향상	AI에 기반한 정교한 검색으로 문서 집합에서 필요한 정보를 찾아내고 이를 기반으로 의미 있는 답변을 LLM으로 생성하여 요청자에게 제공
활용 기술	• Domain Adaptation (기업 특화 지식 훈련) • Instruction Tuning (작업용 데이터 셋 훈련)	• 정확한 의미기반 검색 (Vector Search) • 답변 생성 (Answer Generation)
활용 예시	• 제조/국방/금융 업종 지식을 훈련 • 회의록 요약, 정리 등의 Task 중심 업무	• 사내 전문가 Chatbot, 고객 상담 Chatbot • AI 기반 업무 지식 검색/활용 시스템

맞춤 학습 (Finetuning)

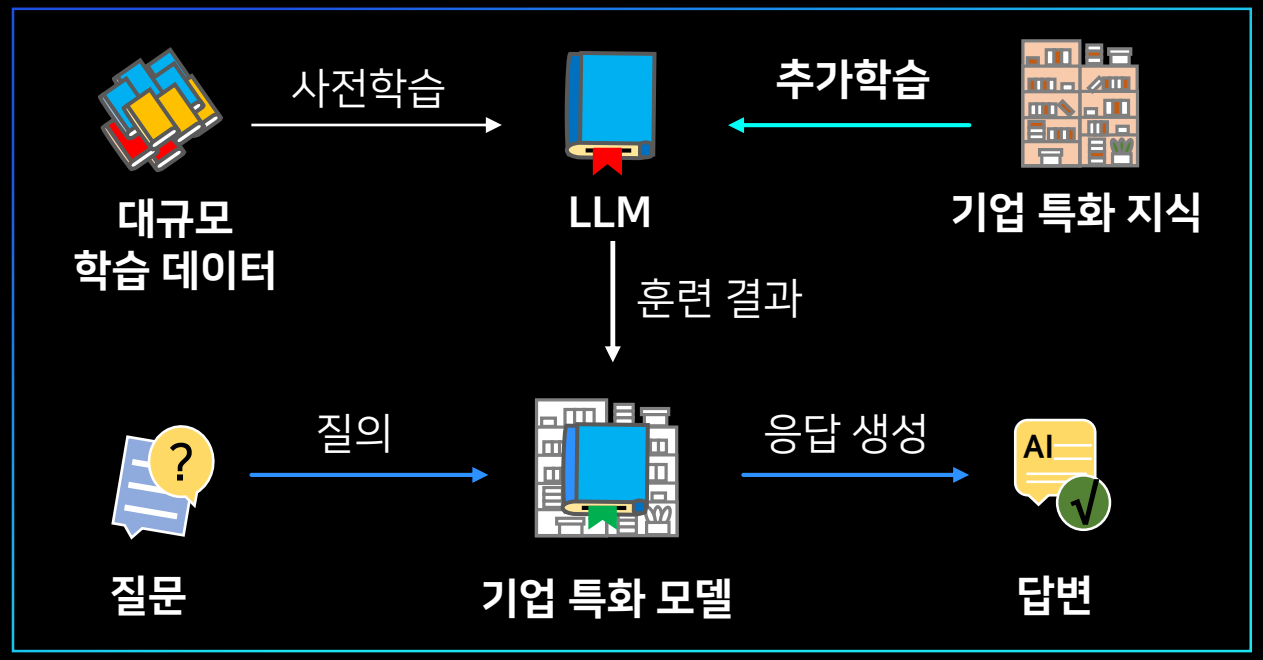
기업 내부 지식으로 LLM을 추가학습 하여 특화 LLM를 생성하여 답변을 생성하고 작업에 활용

① Domain Adaptation

전문 문서 기반의 문장/용어/지식 학습

② Instruction Tuning

프롬프트 예제에 기반한 Task 수행 학습



👍 **내부 지식 답변 가능**

학습에 사용된 문서에 대한 답변 가능

Task 실행력 강화

요약, 번역, 템플릿 활용 등 작업 능력↑

Domain 해석 능력 강화

전문/특화 용어 이해도↑, 문장 이해도↑

👎 **환각(Hallucination) 현상**

맞지 않는 답변, 거짓으로 지어낸 답변

보안/권한 이슈

접근 제어 불가

환각 현상 (Hallucination)

전체 또는 일부가 누락되어 훈련된 LLM에 해당 정보의 요청할 때 날조된 내용을 생성하는 현상

의학적 정의

외부적 자극이 없는데도 불구하고,
마치 외부에서 자극이 있는 것처럼
느끼는 현상

LLM

- **날조된 내용, 잘못된 정보,
  편향/차별적인 시각 등을 생성하는 동작**
- 문법적 오류가 없고 의미가 전달되는
  그럴듯한 답변을 생성해 내지만,
  제시된 프롬프트와 확률적으로 매칭된 답을
  만드는 데만 초점을 두기 때문에
  내용에 대한 문제를 인식하지 못함

사례

한국의 동화 **콩쥐 팥쥐전**에 대해서 알려주세요

'콩쥐 팥쥐'는 한국의 전래동화 중 하나로,
두 마리의 쥐가 서로 협력하여 어려움을 극복하는
이야기입니다.

이 이야기는 매우 가난한 삶을 살던 콩쥐와 팥쥐가
각각 콩과 팥을 모아 아껴서 겨울을 버티며 살아가던
중, 갑자기 온갖 재난이 찾아오면서 그들의 고통이
시작됩니다. (중략)

이 이야기는 작은 것이라도 노력과 협력을 통해 큰
것을 이룰 수 있다는 가치를 강조하며, 서로의 도움을
받아 어려움을 극복하는 중요성을 보여줍니다.
또한 이 이야기는 한국의 전통 문화와 가치관을
담고 있어, 한국어 교육과 한국 문화 이해에 중요한
역할을 하고 있습니다.

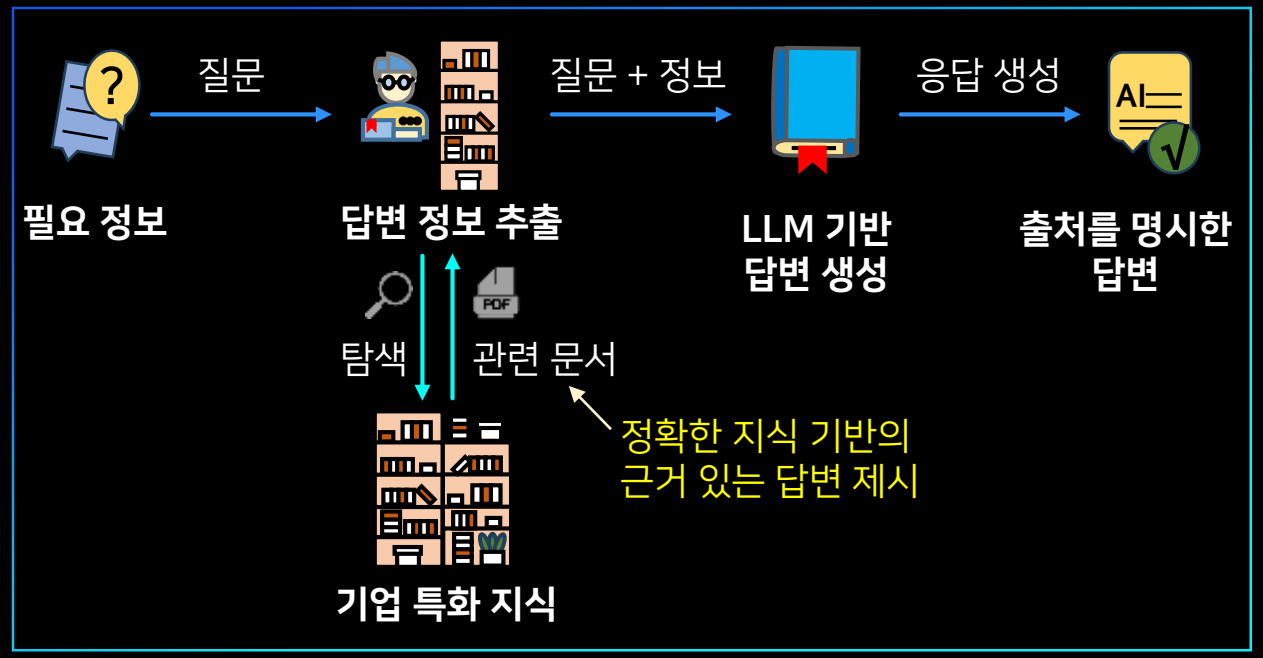
검색 기반 지식 답변 (RAG)

요청에 해당되는 정보를 추출하여 후보 중에서 정확한 내용을 선정하여 답변 생성

① 후보 정보 추출 (Retrieval)
- 지식 검색을 통하여 답변 후보 추출
- 유사성이 높은 지원이 필요함

② 답변 정보 선정 (Augmentation)
- 검색된 지식 중에서
  적절한 답변 및 근거자료 선택

③ 대화형 응답 생성 (Generation)
- 지식 기반 대화형 답변 생성
- 사용자와의 후속 질의 응답



[사내지식 검색 예시]

Q) 입사 2년차 신입사원입니다.
우리 회사의 전세 대출 지원 금액은
얼마나 되나요?

A) 당사는 입사 후 만 1년이 지난 사원들에게
최대 연 5천만원까지 지원하고 있습니다.
[Doc Link] https://guide.sds.co.kr/companylife/doc32.pdf

의미 정보 벡터화 (Vector Embedding)

단어, 문장, 단락의 의미를 벡터(Vector) 수치 값으로 표현

Male-Female

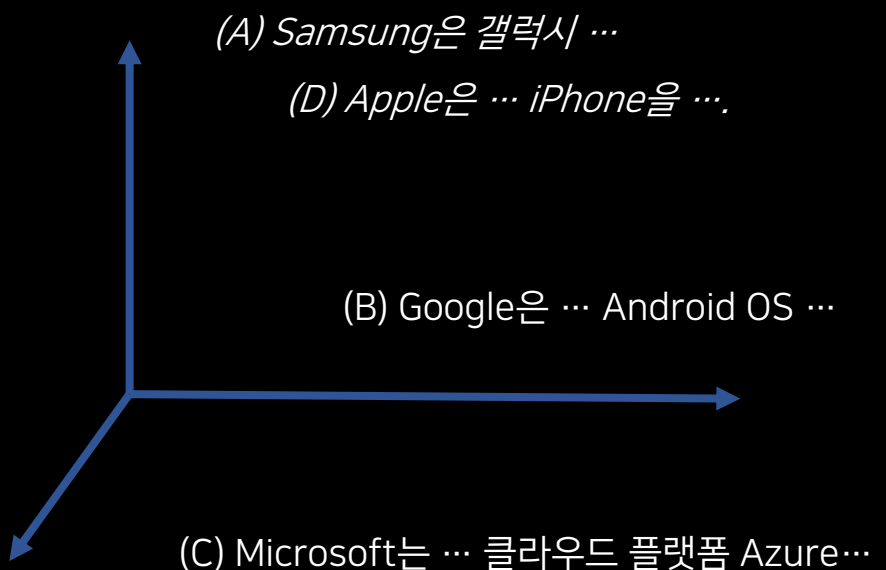
Verb Tense

Country-Capital

King - Queen = Man – ☐

문장 의미 벡터화 (Sentence Embedding)

의미공간 (Semantic Space)



문장 임베딩 (Sentence Embedding)

1. 오픈소스 임베딩 모델 : Sentence-BERT (SBERT)
Sentence-BERT는 문장 간의 유사성을 더 잘 캡처하기 위해 훈련된 모델

2. 예제 문장:
(A) "Samsung은 갤럭시 스마트폰의 새로운 모델을 8월11일에 공개했습니다. "
(B) "Google은 새로운 Android OS 버전을 발표했습니다."
(C) "Microsoft는 클라우드 플랫폼 Azure를 업그레이드했습니다. "
(D) "Apple은 9월13일에 새로운 iPhone을 공개합니다."

3. 문장 의미 수치화 수행:
(A) → [0.45, -0.34, 0.87, …]
(B) → [0.50, -0.32, 0.88, …]
(C) → [-0.21, 0.67, 0.54, …]
(D) → [0.48, -0.35, 0.85, …]

의미 정보 저장 및 검색을 위한 Vector DB

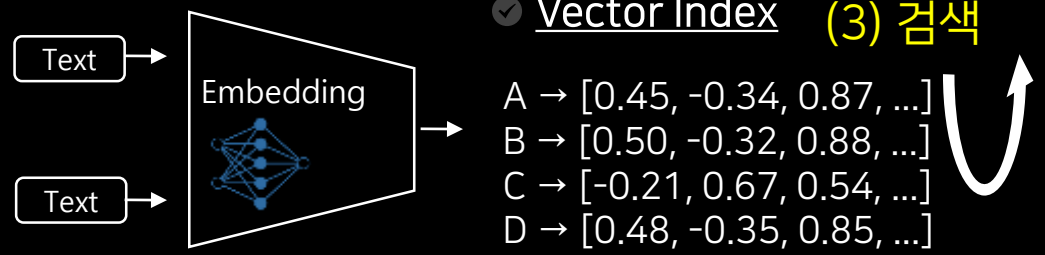
Q) 갤럭시 폰 신제품 출시일?

[0.48, -0.36, 0.84, ...]

A) Samsung은 갤럭시 ….
… 8월11일에 공개했습니다.

(2) Vector 값 생성

(4) 결과 제공



Vector Index

(3) 검색

A → [0.45, -0.34, 0.87, ...]
B → [0.50, -0.32, 0.88, ...]
C → [-0.21, 0.67, 0.54, ...]
D → [0.48, -0.35, 0.85, ...]

(1) Vector Index 생성

A: "Samsung은 갤럭시 스마트폰의 새로운 모델을 8월 11일에 공개했습니다."
B: "Google은 새로운 Android OS 버전을 발표했습니다."
C: "Microsoft는 클라우드 플랫폼 Azure를 업그레이드했습니다. "
D: "Apple은 9월 13일에 새로운 iPhone을 공개합니다."

주요 특징

Vector Search

- 고성능 대용량 Vector Search Engine
- HNSW 알고리즘 기반 고성능 Vector Search
  * Hierarchical Navigable Small World Graphs

주요 기능

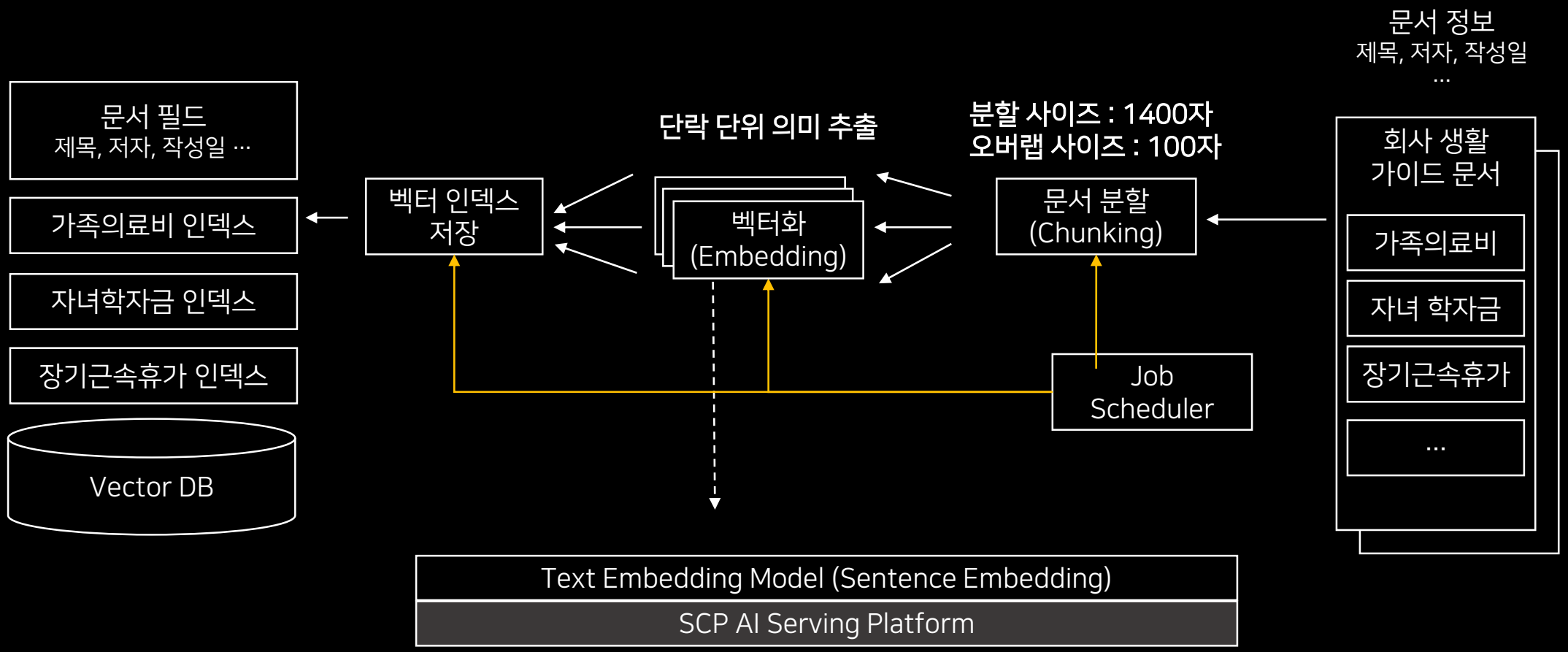
- 대규모 벡터 처리 - 1조 스케일의 벡터를 인덱싱
- 클라우드 기반의 확장성과 신뢰성
- Keyword Search 대비 검색 정확도 ▲

활용 분야

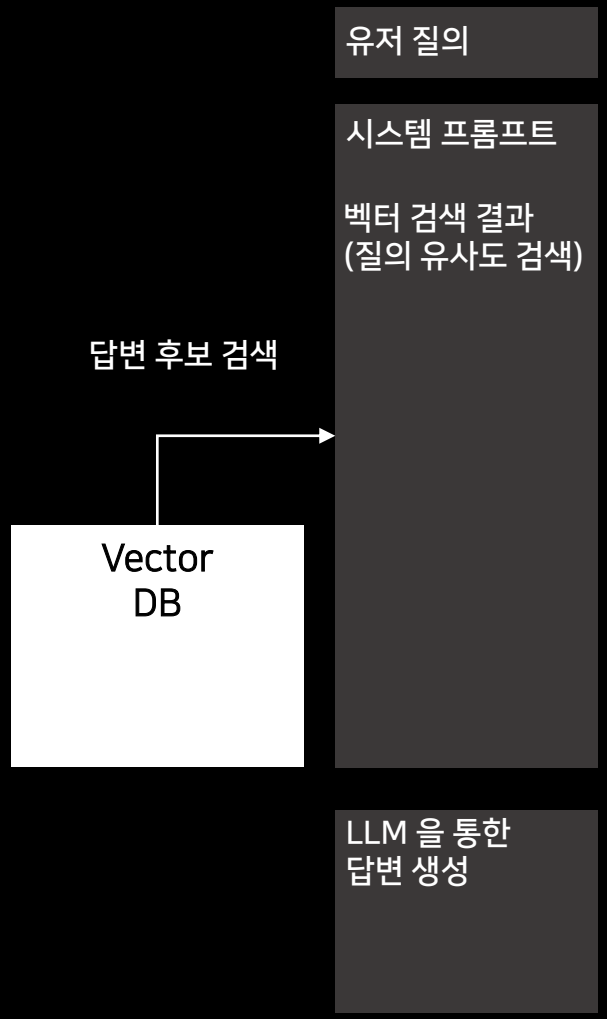
- 문서 검색 등
- 질의 응답 (Auto Q&A w/LLM)
- 상품 추천
- Video 검색
- 이미지 검색

단락 단위 정밀 문서 검색

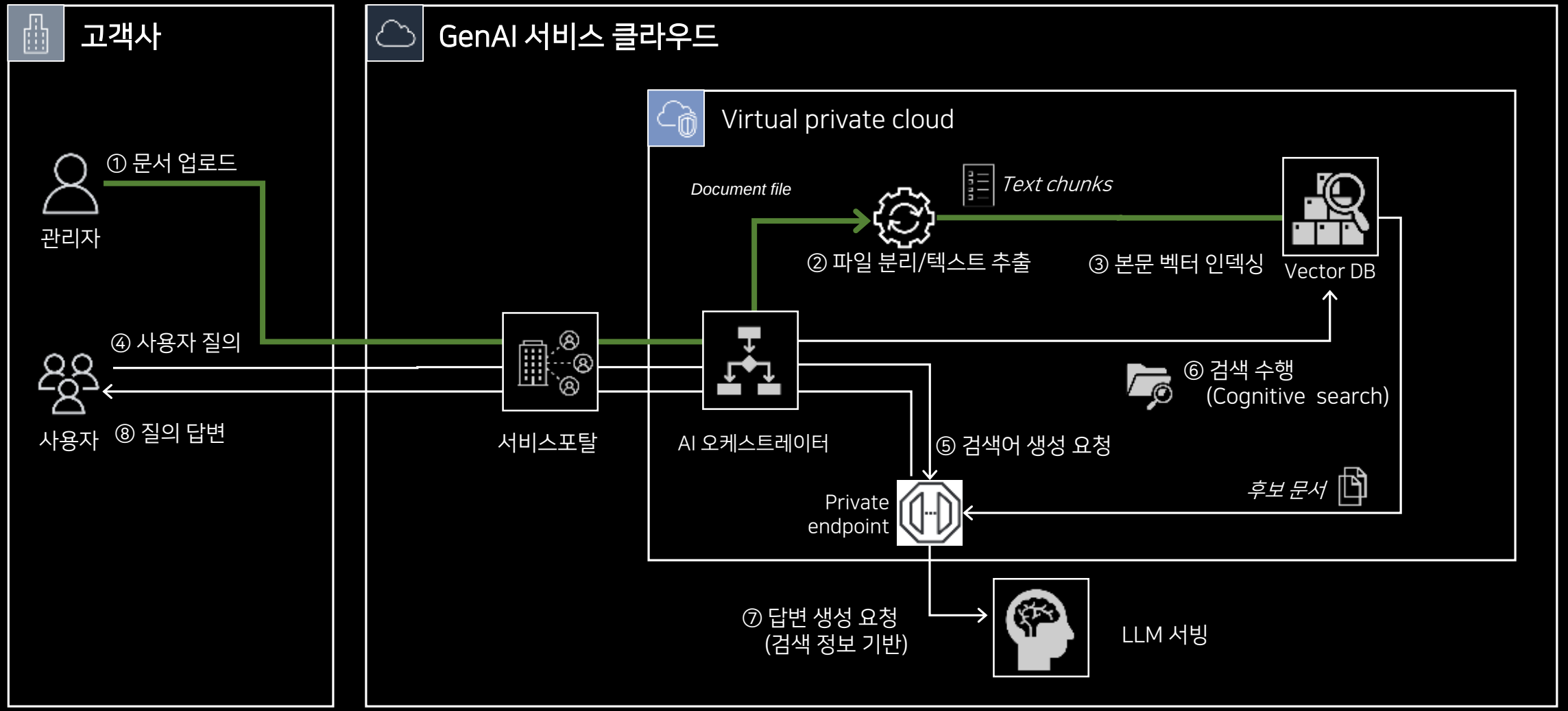
문서 내의 정보를 정확하게 추출하기 위해 단락별로 Vector로 변환하여 검색



검색 정보 기반 답변 생성

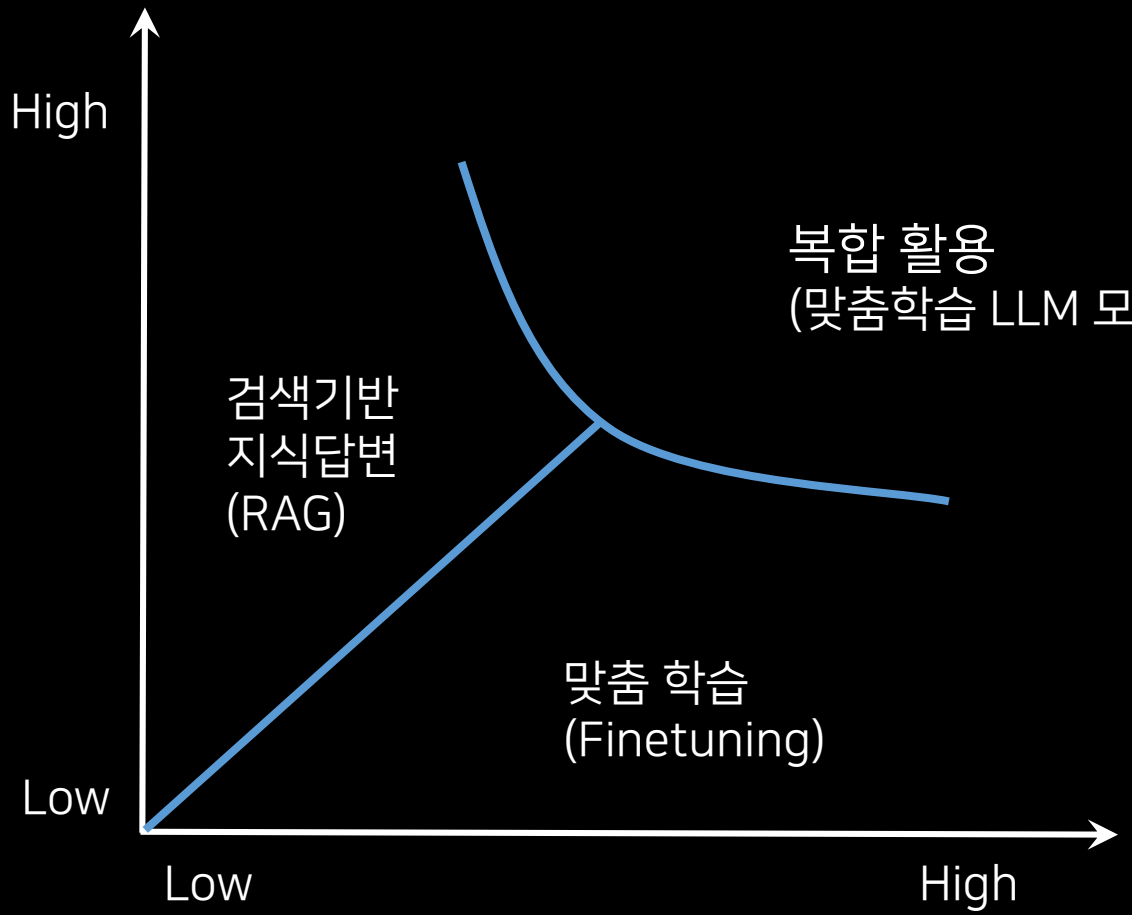


검색 기반 지식 답변 아키텍처



맞춤 학습 vs 검색 기반 지식 답변

기업 지식 활용 (근거 제시, 접근 권한, 최신 정보 답변 등)



LLM 모델 최적화
(문장 이해력, 대화 능력, 자연어 작업처리 능력,
스타일 추가 지원 등)

유형별 문서 저장 구성

데이터 권한과 보관 주기 고려하여, 공유 데이터와 개인 데이터를 관리할 수 있는 환경 필요

데이터 유형	주요 특징	구성 방안
공유 문서 (Restricted, Shared)	- 여러 사람이 접근 가능 - 팀 별 협업에 이용됨 - 기밀자료, 공개 자료 저장	- 권한 기반 접근 제어 시스템 적용 - 변경 이력 관리 및 백업 시스템 도입 - 통합 문서 저장소 필요
개인 자료 (Private, Personal)	- 개인만 접근 가능 - 이메일, 메모, 대화기록 - 개인 업무 자료 보관	- 개인 데이터 접근 제어 적용 - 주기적인 백업 필요 - 개인 문서 저장소 필요

AI 친화적 문서 정책

PPT에서 Word로 변화

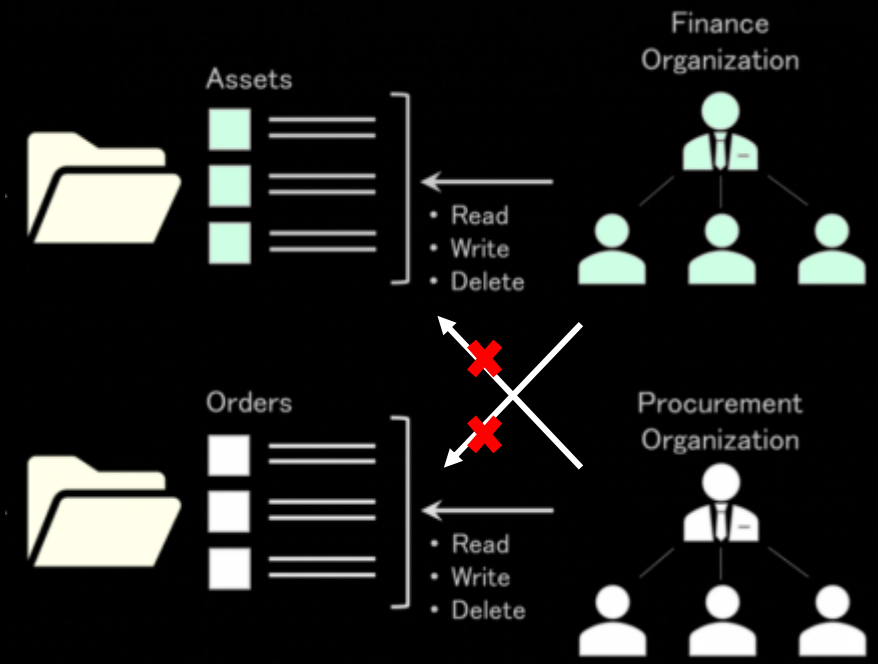
항목	PPT (**Presentation**)	Word (Report/**Documentation**)
사용 목적	발표 및 프레젠테이션	문서 작성 및 편집
파일 크기	이미지 및 멀티미디어 요소로 인해 상대적으로 큼	텍스트 기반으로 상대적으로 작음
정보 표현	함축적, 추상적, 정보량 불충분	명시적, 설명적, 정보량 상대적으로 많음
AI **활용성**	활용 가능성 낮음	LLM 활용 용이함

문서 작성 규칙

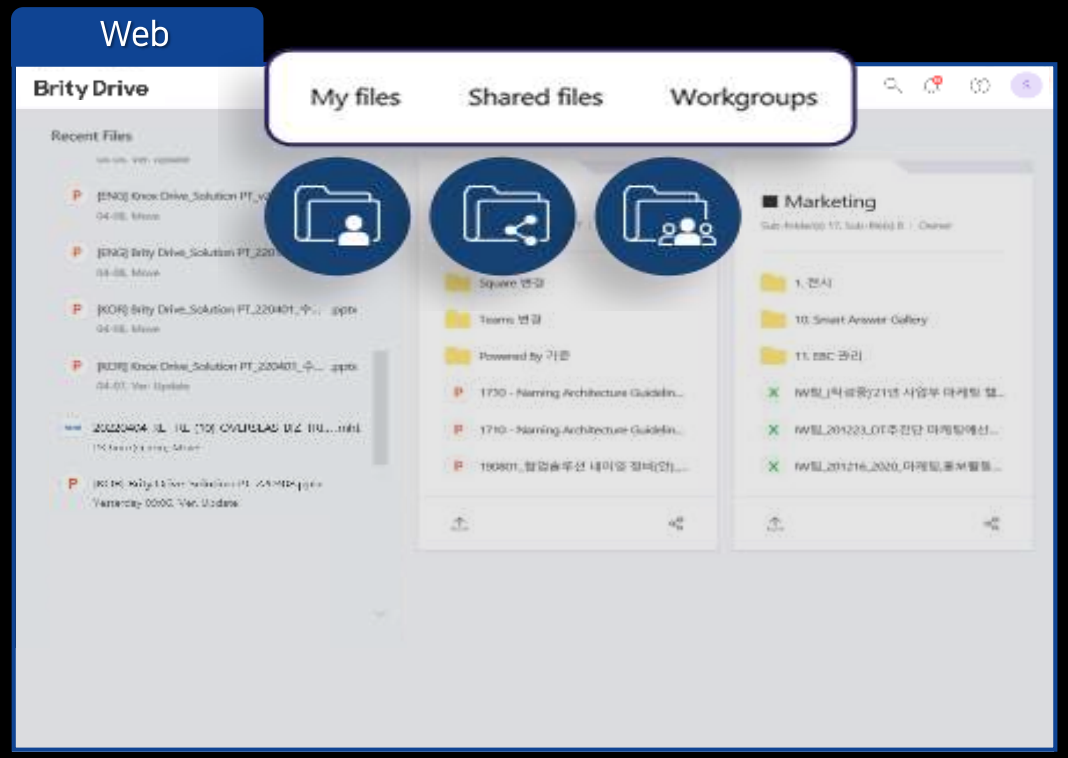
- 따로 설명이 필요 없이 이해될 수 있도록 텍스트 형태로 작성
- 문장의 주어, 동사를 명확히 작성
- 한글과 영어만 사용하고, 한자 사용 금지 등

Cloud Document Storage

역할 및 조직에 따른 데이터 접근 권한 관리
협업 목적으로 초대된 사용자만 접근



Cloud에 저장된 개인/부서/협업 문서를
Windows 탐색기와 Web을 통해서 접근



19개 기업, 18만 사용자가 하루 약 100만 건의 문서 기반 협업을 처리 중

한글 LLM 활용

Tokenizing 테스트 : “안녕하세요, 오늘은 날씨가 좋네요.”

Model	Tokens	Token 개수
영문 LLM (한글 전용 Tokenizer 미적용)	['_', '안', '<0xEB>', '<0x85>', '<0x95>', '하', '세', '요', ',', '_', '오', '<0xEB>', '<0x8A>', '<0x98>', '은', '_', '<0xEB>', '<0x82>', '<0xA0>', '씨', '가', '_', '<0xEC>', '<0xA2>', '<0x8B>', '<0xEB>', '<0x84>', '<0xA4>', '요']	29 (100%)
한글 LLM (한글 Tokenizer 적용)	['_안녕', '하세요', ',', '_오늘은', '_날', '씨가', '_좋네요']	7 ***(24.1%)**

[출처]: huggingface.co/beomi

한글 질의

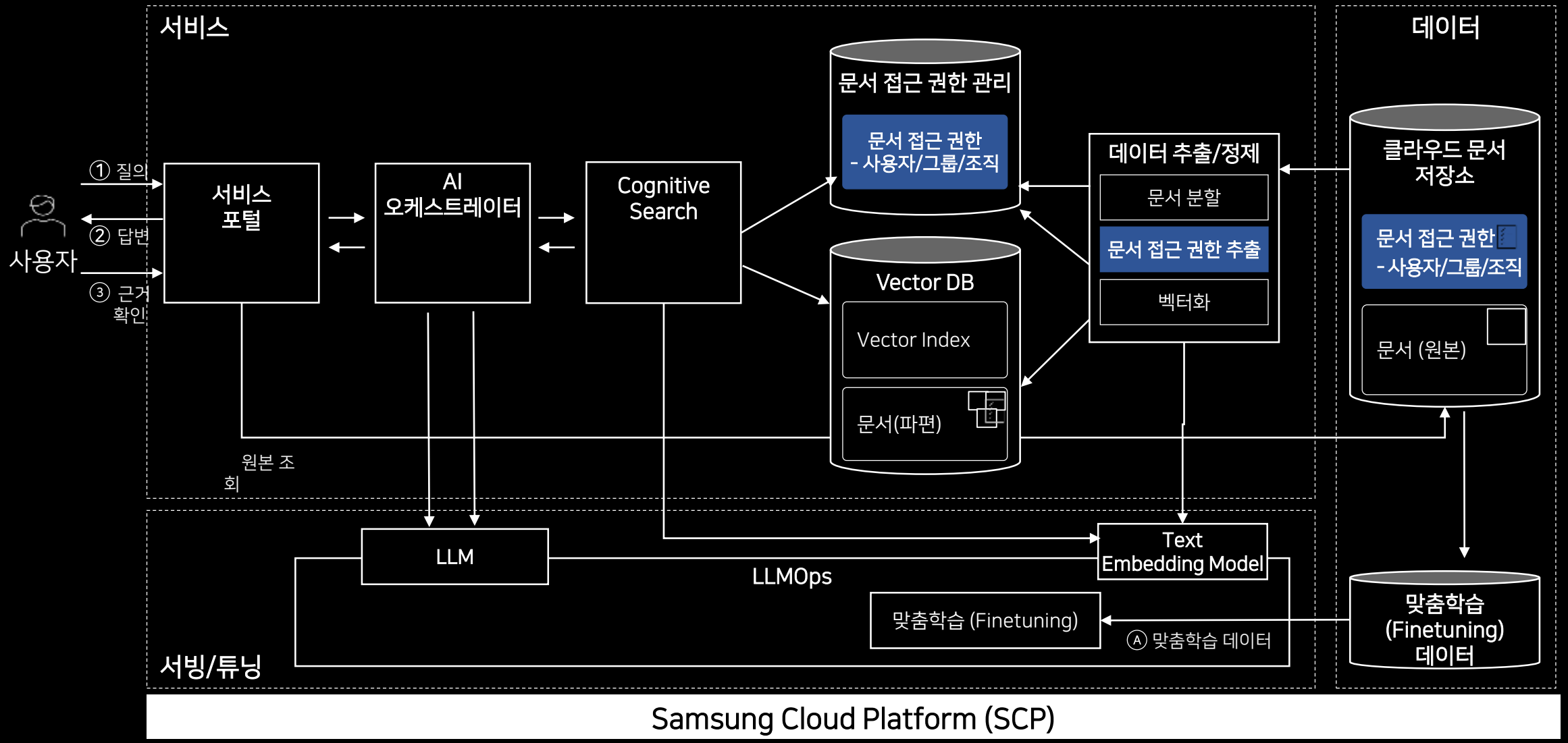
AI를 이용한 기업의
효율적 지식관리 방법을
설명해주세요.

한글 LLM

풍부한 답변

AI를 이용하면 기업의 내부정보,
전문 지식, 데이터 및 리소스를
효율적으로 조직화, 저장, 검색 및
공유하는데 중요한 역할을 할 수 있습니다.
문서자동분류, 질의응답 시스템,
추천 시스템, 지식업데이터 및 유지관리,
음성인식 및 텍스트 변환 등의
여러 업무에 활용할 수 있습니다.
...

Gen.AI 기반 비정형 데이터 활용 구조



정리

영역	필요사항	상품
① 맞춤 학습(Finetuning) 환경	LLM 학습 플랫폼	SCP AI & MLOps Platform
② 의미 추출, 벡터화	LLM 서빙 환경	SCP AI Serving Platform
③ 대량 Vector 검색	Vector DB Service	SCP Vector DB Service
④ 의미 중심 문서 검색	Cognitive Search	SCP Cognitive Search
⑤ 지식 문의 답변 Chat	생성형 AI 포탈	SCP Gen.AI Portal
⑥ 문서 정보 공유/축적	Cloud Document Storage	Brity Drive
⑦ GPU 포함 AI 인프라	AI 클라우드 인프라	SCP AI HPC

Thank you

삼성SDS 백창현 상무
architeam@samsung.com

SAMSUNG SDS