GongsiMate

Financial Disclosure Research Agent

이동경

AI Engineer / LLM Systems Engineer

dhtmwjsdyd95@gmail.com

susincamp.com



목차

1 **제품 설명**
질문으로 공시 근거까지, 한 번에
1 2 3 4 5 지원 기업 5개 · 색인 공시 29건 02

2 **시스템 아키텍처**
원본부터 인용까지, 책임을 나눈 5단계 구조
1 2 3 4 5 retrieval unit 65,684개 03

3 **수집·파싱·표현**
공시 특성에서 출발한 수집·파싱·검색 설계
1 2 3 4 5 2,869 검색 유닛 · 길이 제한 초과 0 04

4 **평가셋 설계**
근거 기반 평가셋과 편향 통제
1 2 3 4 5 120문항 · 판정 일치율 0.923 05

5 **검색·가설 흐름**
가설 검증과 실패 분석을 통한 검색 시스템 개선
1 2 3 4 5 dev Hit@10 42 → 43 / 45 06

6 **에이전트·생성**
사람 판정과 사전 등록 실험으로 개선안의 채택 여부 판단
1 2 3 4 5 P01 정답 2/3 · 오답 1/3 → 보류 07

질문으로 공시 근거까지, 한 번에

GongsiMate는 기업에 대해 궁금한 점을 묻고, 실제 금융 공시를 바탕으로 답을 찾는 리서치 에이전트입니다. 답변에 사용된 근거와 공시 원문을 함께 확인할 수 있으며, 현재 제품 색인은 다섯 기업의 정기·정정공시를 대상으로 구축했습니다.

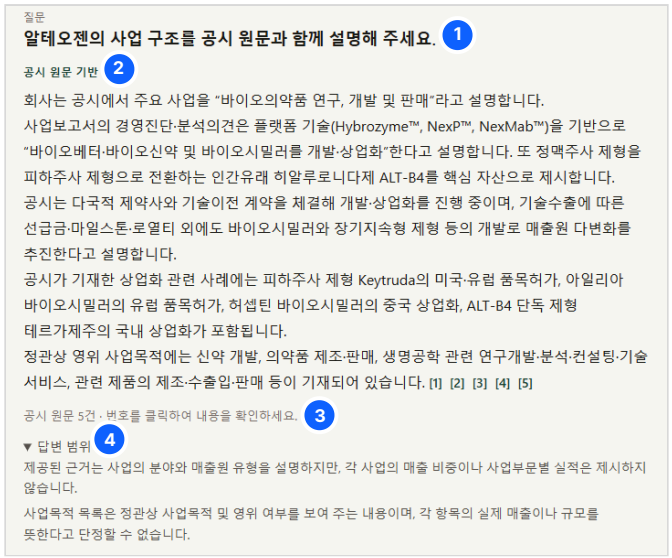
GongsiMate, 왜 필요할까요?

기업을 조사하려면 공시를 봐야 하지만, 필요한 정보를 확인하는 일은 어렵습니다.

1. **긴 공시에서 답이 되는 근거를 찾기 어렵습니다.**
   사업보고서에는 본문과 표, 주석, 감사보고서가 함께 들어 있고, 사용자에게 필요한 것은 문서가 아니라 질문에 답이 되는 문단이나 표 한 칸입니다.
2. **어떤 공시를 기준으로 봐야 하는지도 어렵습니다.**
   같은 기업에도 사업보고서·분기보고서·정정공시가 계속 쌓입니다. 원하는 시점의 정보를 확인하려면 보고 기간과 공개 시점, 정정 여부를 함께 따져야 합니다.
3. **답변이 그럴듯해도 사용자는 원문을 다시 봅니다.**
   금융 정보에서는 답변 자체보다 검증 가능성이 중요하므로, 답변에서 근거를 거쳐 실제 공시 원문까지 이어져야 합니다.

신뢰를 위한 5가지 제품 원칙

1. **기업의 재무·사업 사실**은 공시 검색 근거 없이 단정하지 않습니다.
2. 사용법이나 일반적인 질문은 검색 없이 답하고, **기업이나 시점이 모호하면** 먼저 되묻습니다.
3. **사건 연도와 공시 연도**를 구분합니다. FY2024 사건의 설명이 FY2025 사업보고서에 있을 수 있기 때문입니다.
4. **근거가 부족하면** 부분 답변하거나 보류하고, 인용은 실제 공시 요소에 연결합니다.
5. **지원 범위와 데이터 상태를 사용자에게 숨기지 않습니다.** 수집되지 않은 공시, 검색 결과 없음, 처리 실패를 구분해서 답변합니다.





실제 제품 화면

1. **자연어 질문** 기업명과 질문하고 싶은 내용을 자연어로 입력합니다.
2. **공시 근거 기반 답변** 검색한 공시 근거만으로 답하고, 문장 끝에 인용 번호를 붙입니다.
3. **근거 목록** 답변이 사용한 공시 원문 건수를 밝히고 번호를 누르도록 안내합니다.
4. **답변 범위** 찾은 공시로 확인되지 않는 부분을 답변과 분리해 따로 밝힙니다.
5. **원문 확인** 인용 번호를 누르면 공시 원문 표와 함께 공시일·보고 기간·섹션·접수번호를 보여 주어, DART에서 같은 위치를 다시 찾을 수 있습니다.

PROJECT

대상	이수페타시스 · 삼성전자 · 삼양식품 · 대상 · 알테오젠의 정기·정정공시
규모	지원 기업 5개 · 색인 공시 29건 · retrieval unit 65,684개 · BGE-M3 + BM25
구성	OpenDART 수집 → DART4 XML 파서 → 검색 → LLM Agent(도구 호출) → SSE 답변 → 원문 인용 UI
스택	Python · FastAPI · SSE · BGE-M3 · BM25 · OpenAI · Gemini
상태	제품 런타임 구현 · 실제 공시 데이터 기반 질의응답과 원문 인용 smoke 검증

원본부터 인용까지, 책임을 나눈 5단계 구조

각 단계는 앞 단계의 산출물을 식별자로 연결합니다. 에이전트는 검색 필요 여부를 판단하고, 서버는 검색 범위와 도구 인자·인용의 유효성을 검증합니다.

1 데이터 수집 · 원본 관리
SOURCE & PROVENANCE

공시 원본을 재현 가능한 버전으로 관리합니다.

공시 발견
기업·접수 기간을 기준으로 정기공시와 관련 정정공시를 수집합니다

원본 재검증
로컬에 저장된 공시도 원격 원본과 다시 비교해 변경·손상 여부를 확인합니다

버전 보존
원본의 해시와 취득 이력을 함께 남기고, 변경된 공시는 이전 버전을 덮어쓰지 않습니다

정정 관계 추적
정정공시와 기존 공시의 관계를 기록하고, 확인되지 않은 관계는 추정 상태로 구분합니다

격리와 복구
손상된 로컬 ZIP은 별도 폴더로 옮기고 원격에서 다시 받습니다

주기적 동기화
신규·정정공시를 주기적으로 확인해 반영합니다

원본 공시 26건
정정 관계 5건

2 파싱 · 검색 표현
PARSING & REPRESENTATION

사용자에게 보여줄 문서 형태와 검색기가 색인할 데이터 형태를 분리합니다.

구조 파싱
DART4 XML의 제목·문단·표를 요소로 뽑고, 병합 셀과 중첩 표의 관계를 보존합니다

원문 위치 추적
요소마다 원문 위치 정보를 남겨 검색 결과와 인용에서 다시 해당 문단·표로 돌아갈 수 있게 합니다

구조 기반 청킹
문단·표 등 문서 구조를 우선 보존하고, 토큰 길이는 목표 512 · 상한 1,024로 제한합니다

검색용 표현 분리
검색에는 절 경로와 표 헤더를 포함한 텍스트를 사용하고, 인용에는 원문 요소를 그대로 연결합니다

품질 게이트
내부 검증을 통과한 공시만 색인에 넣습니다

청크 길이 검증
전체 청크의 길이 분포와 상한 초과 사례를 확인해 비정상적인 청크를 따로 점검합니다

retrieval unit 65,684개
partial 2건 색인 제외

3 평가 데이터셋
EVALUATION DATASET

검색기를 비교하기 전에 비교의 기준을 먼저 만듭니다.

평가 분포 사전 정의
검색 전 회사·주제·근거 유형별 테스트 문제 구성을 먼저 정의합니다

근거 기반 질문 생성
근거를 유형별로 균형 있게 선별한 뒤 질문과 요구사항을 만듭니다

다양한 검색 후보 수집
키워드 검색과 의미 검색의 상위 결과와 다양한 순위대의 후보를 샘플링해 특정 검색기에 치우치지 않게 합니다

독립 판정 · 불일치 조정
두 모델이 독립적으로 관련성을 판정하고, 의견이 다른 사례만 별도 모델이 다시 조정합니다

평가셋 버전 고정
Retrieval Unit 기준의 정답 관계와 DEV 45 · TEST 55 split을 버전으로 고정합니다

요구사항 분해
복합 질문은 여러 개의 요구사항으로 분리해 평가합니다

120문항
판정 2,414건 · 일치율 0.923

4 검색
RETRIEVAL

공시 범위를 먼저 정한 뒤 근거 후보들의 순위를 매깁니다.

공시 범위 결정
정정으로 대체된 공시와 질문 기준 시점 이후 공개된 공시를 검색 대상에서 제외합니다

검색 정책
BM25 상위 결과를 우선 유지하고, Dense의 최상위 후보만 제한적으로 보완합니다

검색 후보와 컨텍스트 분리
검색 단계의 후보 순위와 LLM에 실제 전달되는 근거 수를 분리하고, 최종 컨텍스트는 토큰 예산 안에서 구성합니다

오프라인 검증에서 제품 기본 검색으로 연결

확인	Dense · BM25 상보성
기각	전체 순위 RRF
부분	Top-20 후보 RRF
진단	남은 실패의 순위 원인 분석
적용	BM25-primary asymmetric

dev Hit@10 — Dense 33 · BM25 42 · BM25-primary 43 / 45
test 1회 — BM25 47 · BM25-primary 48 / 55

5 에이전트 · 생성
AGENT & GENERATION

판단은 LLM이 하고, 사실의 범위는 서버가 제한합니다.

LLM planner
직접 답변 · 재질문 · `search_disclosures` 중에서 고르고 검색 인자를 만듭니다

서버 검증
도구 인자·회사 범위·호출 한도를 검사하고, 범위를 벗어난 결과는 실패로 처리합니다

근거 구성
검색 결과를 문자열로 붙이지 않고 근거 객체로 만들어 4,096토큰 예산까지 담습니다

답변 · 인용 검증
보류가 아니면 인용을 필수로 두고, 현재 실행에서 검색한 근거만 허용합니다

원문 보기
인용을 element locator로 풀어 원문 문단·표를 그대로 보여 줍니다

한도
검색 2회 · LLM 호출 4회 · 요청 180초. 금액 기준 비용 한도는 현재 없습니다

생성 요건 충족 42/58
지시문 실험 P01 2/3 · S4 FAIL → HOLD

각 단계를 잇는 식별자

검색 결과와 인용에서 사용한 공시 원본까지 추적할 수 있습니다

`receipt_no` — 공시 접수번호
→ `source SHA-256` — 원본 버전
→ `parser version · config hash` — 파싱 버전
→ `element_id · locator` — 원문 요소 · 위치
→ `retrieval_unit_id` — 검색 · 평가 단위
→ `evidence_id` — 평가 근거 · 답변 인용

원본은 보존하고, 검색 표현과 원문을 분리

금융 공시 원본은 바이트가 바뀌어도 이전 버전을 덮어쓰지 않습니다. DART4 XML의 제목·문단·표와 병합 셀은 파서가 유지하고, 검색용 Retrieval Unit과 사용자가 보는 Source Element는 같은 위치를 가리키되 표현은 다릅니다. 인용은 그 위치로 원문과 원본 버전까지 돌아갑니다.

SOURCE VERSION

바뀐 원본은 덮어쓰지 않음

접수번호	원격 ZIP	SHA-256	별도 보존
`created`	최초본 등록	`unchanged`	동일 해시 확인
`updated`	새 버전, 이전본 유지	`repaired`	원격 원본으로 복구
`failed`	직전 정상본 유지		

파일 해시 변경이 실제 본문 정정을 의미하지는 않으므로(단순 재압축 가능성 등), 도메인 수준의 검증을 거치지 않은 변경점은 정정 공시로 섣불리 확정하지 않고 추정 상태로 분리 관리합니다.

DART4 PARSING

문서 구조를 요소로 보존

제목 heading · **문단** paragraph · **표** rowspan · colspan · **중첩 표** · **원문 위치**

- 문단과 표를 각각 하나의 요소로 분리해 파싱하고, 요소마다 원문 위치를 남깁니다.
- 셀의 행·열 범위를 보존하고, 헤더 행에서 열 라벨을 따로 만들어 본문 셀에 붙입니다.
- 단위(예: 백만 원)의 적용 범위가 불명확하면 원문만 보존하고 임의로 전파하지 않습니다.

CHUNKING · TOKEN AUDIT

문서 구조를 우선하고, 보조적으로 길이 제한

- 문단은 목표 800자·상한 4,000자. 설명·캡션은 같은 단위에 붙입니다.
- 표는 8행 이상이면 4행씩 나누고, **단위 행과 헤더를 반복**합니다.
- 토큰 목표 512·상한 1,024. 청크의 토큰 길이 분포를 근거로 정했습니다.

분할 전 청크 · 2,803개

51	558	42	5	2
p50	p99	>512	>1024	>8192

초과는 포함 관계. 8,192 초과 2건은 1×1 셀, 최대 12,711토큰.

분할 이후 **2,869 검색 유닛** · **> 1024 · 0** · **길이 제한 초과 · 0**

같은 표의 세 가지 형태 — 원문 구조, 검색 표현, 인용 화면

이수페타시스 반기보고서 (2026.06) · II. 사업의 내용 > 1. 사업의 개요 · 접수번호 20260814002283

1 · Source Element 보여줄 원문 요소

문단

당사는 전자제품의 핵심부품인 인쇄회로기판(PCB)을 전문으로 생산하고 있으며, … 해외에는 생산공장(중국) 및 판매법인(미국,태국), 영업사무소(대만, 상해)을 보유하고있으며, 총 4개의 자회사와 2개의 손자회사를 두고 사업을 운영하고 있습니다. [사업부문 정보]

표 (rowspan 유지)

사업부문	연결 회사명	제품 유형
PCB 생산과 판매	한국 ㈜ 이수페타시스	고다층PCB(MLB)의 양산과 판매
	나머지 법인 행. 병합 셀은 반복하지 않음	

병합 셀이 원문입니다. 절 경로와 헤더 반복은 검색 표현에만 있습니다.

2 · Retrieval Unit 검색기가 읽는 단위

```
[section] 목차 > II. 사업의 내용 > 1. 사업의 개요
당사는 전자제품의 핵심부품인 인쇄회로기판(PCB)을 전문으로 생산하고 있으며, … 해외에는 생산공장(중국) 및 판매법인(미국,태국),영업사무소(대만, 상해)을 보유하고있으며, 총 4개의 자회사와 2개의 손자회사를 두고 사업을 운영하고 있습니다.

[사업부문 정보]
사업부문|연결 회사명|주요 생산 및 판매제품 유형
--- | --- | ---
PCB 생산과 판매|한국 ㈜이수페타시스|고다층PCB(MLB)의 양산과 판매
(병합)|태국 ISU-APEX CO.,LTD.|중·고다층PCB(MLB)의 판매
```

병합 셀 문구는 첫 행에만 있습니다. 파이프와 [section]은 원문이 아닙니다.

3 · Citation 인용 표현 · 사용자 제공 화면

원문 보기 이수페타시스 · 반기보고서 2026.06

(대만, 상해)을 보유하고있으며,총 4개의 자회사와 2개의 손자회사를 두고 사업을 운영하고 있습니다.

【사업부문 정보】

사업부문	연결 회사명	주요 생산 및 판매제품 유형
PCB 생산과 판매	한국 ㈜이수페타시스	고다층PCB(MLB:Multi Layer Board)의 양산과 판매
	미국 ISU Petasys Corp.	해외 고객 대상 영업 및 마케팅
	중국 ISU Petasys Asia Limited	중·고다층PCB(MLB:Multi Layer Board)의 판매
	중국 ISU Petasys Int't Ltd	중·고다층PCB(MLB:Multi Layer Board)의 판매
	중국 ISU Petasys Hunan Ltd.	중·고다층PCB(MLB:Multi Layer Board)의 양산과 판매
	태국 ISU-APEX CO.,LTD.	중·고다층PCB(MLB:Multi Layer Board)의 판매

기록된 위치 정보로 병합된 표를 엽니다. 접수번호로 같은 위치를 다시 찾습니다.

Representation Experiment 단일 공시 · 사업보고서 1건(위 반기보고서와 다른 문서) · claim recall / 27

표현	무엇을 바꿨는가	@1	@5	@10	@20
D0	분할 전 기준	6	15	20	25
D1	장황한 표	4	9	19	25
D2b	행 단위 분할 · 단위도 함께 변경	3	8	11	17
D2c	단위는 D2a, 표 표현만 변경	4	17	22	23
D2a	**큰 단위만 분할 · 채택**	**6**	**15**	**20**	**25**

최고 점수를 고른 것이 아닙니다. D2a는 길이 제한 초과 2건을 없앴습니다. D2c의 @10은 22지만 @1·@20은 낮습니다. D2b는 unit이 5,296개로 늘어 표현과 단위가 함께 바뀌어 기각했습니다.

검색 표현이 달라져도 원문과 연결됩니다. 채점 단위는 검색 유닛(Retrieval Unit), Source Element는 원문 위치입니다. 다음 페이지는 이 단위로 평가셋을 구성합니다. 또한 이후 페이지의 Dense는 D2a를 의미합니다.

근거 기반 평가셋과 편향 통제

전문가 검수 초기 31문항(v1)에서 복합 표·생산 법인 등 실패 유형을 찾았으나, 기업당 5–7문항으로 검색기 비교에는 표본이 부족했습니다. 사전 정의한 샘플링 기준과 공시 근거로 120문항을 구축했습니다. LLM 자동 벤치마크의 한계와 편향 통제 장치를 함께 명시합니다.

평가셋 구축 파이프라인

평가 분포 사전 정의
검색 편향 방지를 위해 기업당 답변 가능(20)·부분 가능(2)·답변 불가(2) 비율 고정

근거 층화 샘플링
주제 5종(사업·개요·경영진단·재무·지배구조) 및 문서 형태 4종(문단·표·여러 요소·기간 비교)

근거 기반 질의 생성
근거 1개(기간 비교는 2개)당 질문 1개. 도메인 용어·일반 사용자 표현·바꿔 쓴 표현

요구사항 분해
복합 질문은 원자 단위로 분해하여 정밀 채점 기준 수립. 단일 사실은 질문 자체가 요구사항

통합 평가 후보 풀 구성
Dense Top-10 + BM25 Top-10 + 하위 랭크(21–50, 51–200, 201+) 계층 샘플 및 Seed 문서

이중 블라인드 판정
이종 모델(GPT·Gemini 계열) 교차 평가, 편향 방지를 위해 검색기 메타데이터(순위·점수) 마스킹

불일치 조정
두 모델의 판정이 엇갈리는 경계 케이스에 한해 상위 티어 모델이 최종 판정

데이터셋 동결 및 분할
해시 기반 무결성 검증 후 Dev(45) / Test(55) 분리 (v2.1)

평가 신뢰도 검증 및 운영 비용



총 2,414개 판정 (일치도 0.923) 불일치 187건: 상위 모델 중재로 전량 해소

평가셋 규모	120문항 (답변 가능 100 · 부분 10 · 답변 불가 10)
API 호출량	질의 생성 123 · 분해 45 · 판정 4,818회 · 중재 187회
평가 예산	약 $8.9 (경량 모델 이중 배치로 평가 비용 최적화)
v2.1	추가 API 호출 없이 채점 단위 재조정만으로 벤치마크 정밀도 개선

매칭과 답변 가능성(Answerability)의 분리 계측

Hit@10 0.582
정답 유닛이 상위 10에 1개 이상 포함되는가

요구사항 Recall@10 0.561
요구사항별로 근거를 찾았는가

요구사항 완전 충족@10 0.527
한 질문의 요구사항을 모두 찾았는가

Recall@10 macro 0.392
정답 유닛을 얼마나 많이 찾았는가

왜 나누는가 "주요 제품과 각각의 용도는?" 같은 복합 질의에서 제품 이름만 찾으면 Hit은 성공이지만 모델은 제품 용도를 모른 채 답변하게 됩니다. 검색 결과의 완전성을 함께 측정하지 않으면 실제 서비스 답변 품질을 보장할 수 없습니다.

평가 방식의 한계와 보완 장치

예상되는 문제점	어떻게 보완했는가	여전히 남는 한계
질문을 임의로 만들면 문서 안에 정답이 없을 수 있다	공시 문서에서 근거 문장을 먼저 고른 뒤, 그 근거를 기반으로 질문을 만들어 문서 안에 정답이 있도록 설계했습니다	근거 문장을 기반으로 만들다 보니, 질문의 주제가 수집된 공시 문서 범위에만 머무릅니다
특정 검색기가 찾아온 결과만 판정하면 검색기가 놓친 정답은 평가 대상에서 제외된다	의미 기반(Dense) 검색과 키워드(BM25) 검색 상위 결과뿐만 아니라 하위 순위와 기준 정답까지 섞어, **특정 검색기에 치우치지 않는 후보군**을 만들었습니다	10위권 안쪽은 모두 판정됐지만 20위권까지 넓혀 보면 판정 비율은 0.536입니다
판정 LLM이 검색 순위를 알면 높은 순위의 문서를 유리하게 판정할 수 있다	판정 모델의 입력창에서 **검색기 이름·순위·점수를 모두 제거**하고, 두 모델이 서로의 결과를 모른 채 독립적으로 판정하게 했습니다	잘라낸 문단 조각만 보고 채점하므로, 문서 전체의 앞뒤 맥락까지는 고려하지 못합니다
LLM의 판정이 사람이 직접 검수한 정답과 완전히 같을 수 없다	두 모델의 의견이 갈린 187건은 **상위 모델이 다시 판단**해 결론을 냈으며, 결과를 'LLM 기준 벤치마크'로 규정했습니다	두 모델의 채점 일치율(92.3%)이 높다고 해서 정답이라는 보장은 없으며, 두 모델이 동시에 틀렸을 가능성도 있습니다
질문을 만든 모델과 채점하는 모델의 계열이 같으면 점수를 후하게 줄 수 있다	서로 다른 계열(GPT, Gemini)의 모델로 판정하게 하고, 판정이 엇갈릴 때 중재하는 모델은 상위 티어 모델로 분리했습니다	모델 패밀리 고유의 문체 선호 성향을 100% 없애기는 어렵습니다
인공적으로 만든 질문이고 대상 기업이 5개뿐이라 실제 사용 환경과 차이가 날 수 있다	전문 용어, 일상 표현, 다르게 바꿔 쓴 표현 등 9가지 질문 유형을 개발용(dev)과 최종 테스트용(test)에 모두 포함되도록 구성했습니다	실제 사용자의 질문 패턴과는 차이가 있으며, 기업당 질문이 10개 안팎이라 기업별 순위 우열은 비교하지 않습니다

운영 원칙 평가 기준 파일은 고유값(해시)으로 기록하고, 수정이 필요하면 새 버전으로 만듭니다. 검색 설정 비교는 dev 45문항에서만 하고 test 55문항은 설정 확정 후 단 한 번만 확인합니다. 평가 코드도 실제 서비스와 똑같은 공시 선택 및 검색 함수를 그대로 호출해 검증합니다.

가설 검증과 실패 분석을 통한 검색 품질 최적화

평가 기준과 검색 대상을 고정한 채 실험 변수를 하나씩 통제하여 각 검색 요소의 실질적인 기여도를 측정했습니다. 실험 과정의 데이터 누수와 과적합을 방지하고자 개발셋(45문항) 검증을 거친 평가 기준을 동결한 후, 테스트셋(55문항)에 대해 단 한 번의 평가로 일반화 성능을 최종 검증했습니다.

가설	목적과 아이디어	통제 조건	성공 판정 조건	실험 결과	판정	실패 분석 · 인사이트	후속 가설 및 개선 방향
H1 실패 상보성	단독 Hit@10은 33/45에 불과. 결합 전 두 검색기의 **실패 영역이 상보적인지** 선제 검증	동일 검색 단위·평가 기준·문서 풀 고정. BM25 하이퍼파라미터 기본값 고정	Dense·BM25 각각의 단독 정답 문항이 **모두 존재**할 것	Hit@10 — 공통 **32** · Dense 단독 **1** · BM25 단독 **10** · 공통 실패 **2** (이론상 상한 **43**/45)	확인	BM25 고유 계정과목·표 헤더·날짜 일치에 강세. Dense는 어휘 불일치("구매" ↔ "매입")를 해소하여 하이브리드 결합 타당성 확보	두 순위를 RRF로 결합. 상한 여유가 1문항에 불과하므로 **순위 지표 (MRR/NDCG)**를 필수 추적 지표로 지정
H2 전역 순위 RRF	동일 가중 RRF(k=60) 적용 시, BM25 커버리지를 보존하면서 의미 검색의 이점을 살릴 수 있다	공시 내 전체 문서를 두 검색기로 순위화 후 전역 RRF 합산. 키워드 매칭이 없는 문서도 1/(60+rank) 점수를 부여받음	BM25 Hit@10 유지 및 Dense 단독 정답 복구, 순위 지표 비하락 (BM25 단독 정답 2건 이상 상실 시 기각)	Hit@10 **36**/45 (BM25 42 대비 하락), MRR **.686** (.742 대비 하락). BM25 단독 정답 10건 중 **8건 상실** (Dense 단독은 복구)	기각	하위 순위 점수 누적에 의한 순위 왜곡. BM25 1위인 정답이 Dense 705위 점수 희석으로 27위로 밀려나고, Dense 1위인 무관 주석이 BM25 32위 점수를 더해 1위로 부상	런타임 검색 계약(Top-20 반환)에 맞춰 **Top-20 외부 문서의 RRF 점수 기여를 0으로 차단**
H3 후보군 풀링 RRF	각 검색기의 **Top-20 반환 문서만으로 후보 풀을 제한**하면 BM25 커버리지를 지키며 Dense 신호도 남길 수 있다	Dense Top-20 ∪ BM25 Top-20만 평가(외부 순위 기여도 0). 계약 규격(N=20), k=60 고정	Hit@10 ≥ 41 달성, Dense 단독 정답 유지, H2 상실의 8건 과반 회복, 순위·요구사항 충족 지표 방어	Hit@10 **40**/45 달성, 상실 8건 중 **5건 회복**, Dense 단독 정답 유지. 상위 10 판정 비율 .787 → **.969** 개선	부분 지지	하위 순위 노이즈는 차단했으나 BM25 단독 성능에 미달. 양쪽 Top-20 교집합의 최소 점수(0.250)가 단독 문서의 최대 점수(0.164)를 상회하여 **품질과 무관하게 교집합 문서가 단독 정답을 구조적으로 밀어냄**	미회복 3건(BM25 상위권에서 10위 밖으로 밀린 사례)에 대해 Top-20 내부 **순위 압력 요인 분해** 진행
H4 순위 역전 진단	미회복 3건의 순위 하락이 **교집합 가산점 압력**과 **Dense 단독 문서 유입** 중 어느 요인에 기인하는지 규명한다	상위 랭크된 오답 문서를 3개 유형(교집합, Dense 단독, BM25 단독)으로 세분화한 뒤 제거 시뮬레이션 수행	Top-10 이탈 원인이 내부 경쟁 기전으로 설명될 것 (단일 기전 확인 시 강지지)	사례별 주 원인이 상이함. 대상 지분증권: Dense 단독 유입 7건 압력 · 대상 특수관계자: 교집합 8건 및 Dense 3건 복합 압력 · 알테오젠: 교집합 8건 압력 (교집합 제거 시 **4위 회복**)	부분 지지	내부 경쟁이 원인임을 밝혔으나 건별 하락 기전은 서로 달랐음. **대칭적 RRF 자체가 BM25의 명확한 키워드 단서를 필연적으로 희석**시킨다는 구조적 한계 도출	대칭 결합을 배제하고, **BM25 상위 순위를 보호**하며 Dense 상위 순위만 제한적으로 주입하는 비대칭 결합 설계
H5 BM25-primary	BM25 상위 슬롯을 고정시키고 **Dense 1위만 마지막 슬롯**에 넣으면, BM25 손실 없이 의미 검색 보완을 달성할 수 있다	BM25 1–9위 순위 보존. Dense 1위가 BM25 Top-10 밖일 때만 10위에 조건부 삽입 (규칙 동결 후 test **1회 평가**)	BM25 Hit@10 유지 이상, BM25 단독 정답 상실 ≤ 1, Dense 단독 정답 유지, 순위·요구사항 지표 방어	dev Hit@10 42 → **43**/45, test 47 → **48**/55. 양측 모두 BM25 정답 손실 **0**건 (test Recall@10 **.701 → .693**, NDCG@10 **.640 → .635**)	dev 강지지 test 일반화 검증	Dense 모델의 자체 정확도를 올린 것이 아니라 Dense 오류 시 **부정적 영향 반경을 10위 슬롯 1개로 제한**한 것이 성공 요인	기본 검색 파이프라인으로 프로덕션 반영 및 스모크 테스트 통과 완료

지표 비교 Hit와 요구사항 모두 충족은 문항 수 · test는 규칙 고정 후 1회 실행

		Hit@1	Hit@5	Hit@10	Recall@5	MRR	NDCG@10	요구사항 Recall@10	요구사항 모두 충족	상위 10 판정 비율
dev 45	**Dense**	20	28	**33**	.387	.525	.478	41/58	32	1.000
	BM25	29	40	**42**	.572	.742	.708	51/58	41	1.000
	전체 순위 RRF 기각	28	33	**36**	.468	.686	.654	46/58	35	.787
	Top-20 RRF 부분 지지	29	37	**40**	.505	.737	.684	49/58	39	.969
	BM25-primary 선정	**29**	**40**	**43**	**.572**	**.744**	**.713**	**52/58**	**42**	**1.000**
test 55 · 1회	**Dense** 과거 기준선	19	27	**32**	.246	.411	.325	37/66	29	1.000
	BM25	36	45	**47**	.495	.717	.640	56/66	45	1.000
	BM25-primary 선정	**36**	**45**	**48**	**.495**	**.718**	**.635**	**57/66**	**46**	**1.000**

RRF는 상위 10에 판정되지 않은 unit이 있어 판정된 unit만으로 NDCG를 계산했습니다(미판정을 0으로 두면 .615 · .675).

최종 파이프라인 선정 결과

- **프로덕션 기본 검색기:** BM25 상위 순위를 보존하고 10위 슬롯에만 Dense 1위 결과를 선별 주입하는 **비대칭 하이브리드 정책** 확정 (키워드 매칭 안정성과 의미 검색 커버리지 동시 확보)
- **테스트셋 일반화 검증:** 독립 테스트셋(55문항)에서 BM25 단독 정답의 이탈 없이 의미 누락 정답을 추가 확보 (순위 지표의 미세 변동 대비 정답 회수 실익 우위)
- **Cross-Encoder Re-ranker 미도입 결정:** 서빙 비용 및 응답 지연 증가 대비 기대 효과가 미미하여 도입 보류

계산 가능한 질문의 답변 회피는 줄였으나, 새로운 연산 오류가 발생하여 프로덕션 적용을 보류했습니다

실제 서비스 답변을 정밀 검수한 결과, 본문의 숫자를 더하기만 하면 되는 단순 합산 질의임에도 모델이 "원문에 합계 행이 없다"며 답변을 회피하는 문제가 있었습니다. 검색 조건은 고정한 채 프롬프트 지시문만 수정하여 비교 실험을 거쳤으나, 사전에 정의한 품질 안전 기준을 충족하지 못해 배포를 보류했습니다.

01 올바른 근거 검색과 정확한 답변 생성은 다른 문제였습니다

검색 설정(H10)과 컨텍스트 주입 규칙을 동일하게 유지한 상태에서, 개발셋 45문항에 대한 서비스 응답 품질을 측정했습니다. 아래 수치는 수작업 보정을 거치지 않은 LLM 심판 모델의 순수 판정 결과입니다.

42/58	50/51	51/53	37 · 8
질의 요건 충족	근거 충실도	출처 인용 정확도	답변 생성 · 회피

정량 점수만으로는 실제 실패 원인 파악이 어려워 표본 8건을 선별해 직접 검수를 진행했습니다. 그 결과, 검색 시스템이 올바른 근거를 찾아도 답변 생성 단계에서 왜곡이 발생하는 사례가 확인되었습니다.

정밀 검수 8건에서 식별된 오류 유형 · 전체 오류율을 대변하지는 않습니다

검색 관련성	P05 전체 계정과 일부 항목 혼동 · P06 타 유형 보고서 조회
근거 충분성	P02 표 제목 누락 · P03 비중 계산의 분모·범위 누락
답변 정확성	**P01 합산 가능한 항목의 답변 회피** · P04 시점 모호
인용 완전성	P07 최종 답변은 맞으나 인용 출처만으로 확인 불가

개선 대상 P01 · 특수관계자 채권·채무 합계 질의

원문에 기준·단위가 일치하는 내역이 있어 단순 합산으로 **채권 24,528백만 원, 채무 20,005백만 원**을 도출할 수 있었으나, 모델은 "합계 행이 없다"며 답변을 회피했습니다. 기존 프롬프트가 수치 왜곡 방지만 강조하고 기초 연산 허용 기준을 명시하지 않은 것이 원인이었습니다.

비교 실험 설계 검색 조건은 통제하고 프롬프트 지시문만 변경했습니다

A0 기존안	A1 실험안
서비스 답변 지시문 그대로	계산 허용 조항 추가: 세부 항목·기간·회계 기준·단위가 일치할 때만 연산 허용, 단순 합계 행 부재만으로 답변 회피 금지, 반드시 산출식을 함께 명시

- 45문항 × 2개 안 × 3회 = **총 270회 생성**. 동일한 모델과 파라미터를 사용했으며, 해시 검증을 통해 270건 모두 모델에 전달된 프롬프트와 컨텍스트가 원본과 일치함을 입증했습니다.
- 서비스 런타임 코드는 보존한 채, 독립된 실험 환경에서 시스템 프롬프트만 교체하여 외부 간섭을 배제했습니다.
- 실험자의 사후 편향을 방지하기 위해 첫 API 호출 전 성공 기준과 기각 조건을 확정하고 사후 변경 없이 적용했습니다.

02 답변 회피는 줄었지만, 새로운 수치 왜곡이 발생했습니다

P01 질의 · 동일 조건 3회 반복 검증 결과

	r1	r2	r3	요약
A0 기존안	회피	회피	회피	정답 0/3 · 회피 3/3
A1 실험안	정답	정답	오답	정답 2/3 · 오답 1/3

실험안(A1)은 의도대로 3회 모두 연산을 수행하여 답변을 산출했고, 2회는 올바른 산출 근거를 명시하며 정답을 냈습니다. **그러나 1회(r3)에서 표의 원문 수치를 잘못 추출하여 오답을 생성했습니다.**

A1 3회차(r3) 상세 분석: 수치 오추출과 오차 상쇄로 인한 우연한 일치

항목 (백만 원)	원문	생성	차이
매출채권	22,800	22,800	0
기타채권	1,728	1,745	+17
채권 합계	**24,528**	**24,545**	**+17**
매입채무	513	508	−5
기타채무	19,492	19,497	+5
채무 합계	**20,005**	**20,005**	**0 · 상쇄**

채무 합계의 최종값은 원문과 일치했으나, 이는 두 개 세부 항목의 오차(−5, +5)가 우연히 상쇄된 결과였습니다. **최종 수치가 정답이라도 내부 추출 과정이 올바른 것은 아니었습니다.** 프롬프트에 산출식을 반드시 명시하도록 강제한 덕분에 어떤 항목의 수치가 왜곡되었는지 추적할 수 있었습니다.

18/18	1 vs 2	+4.1%
답변 보류 대상 6문항 × 3회 모두 회피 유지	기존 정상 답변 질의의 신규 회피 발생 횟수	답변 생성 비용 증가율
검수 결과 회피가 타당했던 4문항 및 미검수 2문항	실험안(A1) 1회 · 기존안(A0) 2회 (37문항 × 3회 중)	A0 $0.0873 → A1 $0.0909 (각 135회) · 응답 시간 차이는 오차 범위 내

03 사전 정의한 품질 안전 기준을 넘지 못한 개선안은 배포하지 않았습니다

사전 등록 기준 · S1–S5 조건을 모두 충족해야 프로덕션 채택

S1	P01 질의 3회 중 2회 이상 정답 (기존안 대비 개선)	PASS
S2	답변 보류 대상 질의에 임의 수치 생성 0건 (환각 방지)	PASS
S3	신규 답변 회피 발생 건수가 기존안 + 1회 이하	PASS
S4	산출 수치 왜곡(연산 및 추출 오류) 0건	FAIL
S5	출처 인용 무결성 오류 0건	PASS
R3	기존 연산 질의의 신규 산술 오류 발생 여부	검수 예정

HOLD

실험안(A1)은 필수 기준(S4)을 충족하지 못해 프로덕션 배포를 전면 보류했습니다. 대기 중인 R3 판정 결과와 관계없이, 신뢰성이 핵심인 금융 공시 서비스에서 오답 발생 가능성을 확인한 이상 기존 배포 상태를 유지하는 것이 타당하다고 판단했습니다.

차기 아키텍처 개선안

이번 수치 오류는 연산 허용 여부의 판단이 아니라 표 데이터의 값을 파싱하는 과정에서 발생했습니다. 따라서 연산 필요 여부만 LLM이 판별하고, 실제 수치 추출과 합산은 별도 도구나 런타임 시스템에 위임하는 아키텍처 분리 방식을 다음 개선 후보로 수립했습니다. 아직 구현·실험 전 단계입니다.

검색 품질 평가부터 생성 검증까지, 모든 프로덕션 의사결정은 사전 정의된 기준과 재현 가능한 실험 로그에 근거하여 내렸습니다.

직접 검수한 기록은 별도 자동화 검증 코드를 통해 원문 데이터와 정합성을 재검증했습니다(5개 패킷 기준 일치 226건 · 불일치 0건 · 판정 불가 26건). 이는 기록과 원문의 정합성을 확인한 것이며, 검수 판정 자체의 정확성을 보증하지는 않습니다.

실험 한계: 해당 실패 양상은 개발셋 45문항 중 1건(P01)에서 관측된 표본이며, 3회 반복 검증만으로 일반 오류율을 확정할 수는 없습니다. 또한 지시문 수정 후 별도 심판 모델을 호출하지 않아 변경된 응답 전체의 충실도와 인용 정확도를 재측정하지는 못했습니다.

감사합니다.

이동경

AI Engineer / LLM Systems Engineer

dhtmwjsdyd95@gmail.com

susincamp.com