

일반논문 (Regular Paper)

방송공학회논문지 제31권 제5호, 2026년 9월 (JBE Vol.31, No.5, September 2026)

<https://doi.org/10.5909/JBE.2026.31.5.1101>

ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

배수펌프장 안전대응 온프레미스 RAG 챗봇의 결정적 오케스트레이션 설계와 다중 LLM 평가

장 일 식^{a)}, 박 구 만^{b)†}

Deterministic Orchestration Design and Multi-LLM Evaluation of an On-Premise RAG Chatbot for Drainage Pump Station Safety Response

Ilsik Chang^{a)} and Gooman Park^{b)†}

요 약

집중호우 시 배수펌프장 운영자는 수위, 예측, 펌프 가동, 대응 절차를 동시에 판단해야 한다. 본 연구는 이를 지원하는 온프레미스 RAG 안전대응 챗봇을 제안한다. 챗봇은 자연어 질의를 매뉴얼 검색, 실시간 데이터 호출, 일반 답변으로 라우팅하고 단일 응답 계약 (ChatResponse)으로 정규화하며, 데이터와 절차 근거를 한 답변에 결합해야 하는 복합 질의는 멀티라운드 탐색과 결정적 매뉴얼 보강으로 처리한다. 500문항을 4개 오픈 웨이트 LLM으로 답변한 2,000건을 RAGAS로 평가한 결과는 다음과 같다. 첫째, 복합 질의 라우팅 정확도는 모델별 0.55~0.87로 편차가 컸고, 실제 실험 결과 결정적 보강의 기여는 평균 정확도 상승이 아니라 매뉴얼 검색 누락 사례를 차단하는 안전 보장임을 확인하였다. 둘째, 빈 답변과 멈춤을 방지하는 생성 안정성이 품질의 핵심 변수였다. 셋째, 검색 성능은 모델과 무관했고(Context Precision 0.51~0.70), 변별력은 라우팅과 생성에서 발생하였다. 또한 에이전트 프레임워크 없이 결정적 오케스트레이션을 구현한 근거를 정리하여 안전 중요 RAG의 설계 및 모델 선택 기준을 제시한다.

Abstract

During heavy rainfall, drainage pump station operators must simultaneously assess water levels, forecasts, pump operation, and response procedures. We propose an on-premise RAG safety-response chatbot that routes natural-language queries to manual retrieval, real-time data calls, or general answers, normalizing all of them into a single response contract (ChatResponse). For compound queries requiring both live data and procedural grounding in one answer, the system applies multi-round exploration with deterministic manual augmentation. We evaluated 2,000 answers, produced by answering 500 questions with four open-weight LLMs, using RAGAS. Three findings emerge. First, compound-query routing accuracy varied widely across models (0.55 - 0.87); an ablation toggling deterministic augmentation on the same evaluation set showed that its contribution is not higher average accuracy but a deterministic safeguard against the residual cases in which a model omits manual retrieval. Second, generation stability, preventing empty answers and stalling, is a critical determinant of safety chatbot quality. Third, retrieval performance was largely model-independent (Context Precision 0.51 - 0.70), indicating that discriminating power arises in routing and generation rather than retrieval. Finally, we document the rationale for implementing deterministic orchestration without an agent framework, offering design and model-selection criteria for safety-critical RAG systems.

Keyword : Drainage Pump Station, On-premise RAG, Compound-query Routing, Deterministic Orchestration, RAGAS Evaluation

Copyright © 2026 Korean Institute of Broadcast and Media Engineers. All rights reserved.

“This is an Open-Access article distributed under the terms of the Creative Commons BY-NC-ND (<http://creativecommons.org/licenses/by-nc-nd/3.0>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited and not altered.”

1. 서론

1. 연구 배경 및 동기

도시 배수펌프장은 집중호우 시 내수 침수를 방지하는 핵심 방재 시설로, 운영자는 짧은 시간 내에 수위, 예측, 펌프 가동 상태, 대응 절차를 동시에 판단해야 한다. 그러나 이러한 판단에 필요한 정보는 첫째, 실시간 운영 데이터(수위, 예측, 펌프 상태, 출입구 영상 분석 결과), 둘째, 안전 대응 매뉴얼 및 행동 요령(절차, 기준, 법령), 셋째, 일반 도메인 지식 등 서로 이질적으로 분산되어 있다. 운영자가 자연어로 질문하면 이 세 가지 정보를 근거(출처)와 함께 제공하는 보조 도구가 있다면, 야간이나 집중호우 상황에서 인지 부하를 크게 줄일 수 있을 것이다.

본 연구의 대상은 이러한 요구에 맞춰 구축된 온프레미스 RAG(Retrieval-Augmented Generation) 기반 안전 대응 챗봇이다. 이 챗봇은 기존 5분 주기 예측·의사결정 시스템 위에 구축되어 매뉴얼 검색, 실시간 데이터 함수 호출, 일반 답변을 의도에 따라 라우팅하며, 답변을 단일 응답 계약(ChatResponse)으로 정규화한다. 본 논문은 단순 정확도 벤치마크를 넘어, “복합 질의에 효과적으로 응대하는 챗봇이 되기 위해 어떤 구조와 설계가 필요한가”를 다중 모델 실험을 통해 규명하는 데 초점을 둔다.

2. 최근 동향

공공 인프라 및 방재 분야에서는 운영 데이터의 외부 유출이 엄격히 제한되는 경우가 많아, 상용 클라우드 LLM 대신 로컬 GPU에서 구동되는 경량 및 중형 오픈 웨이트

모델(Llama, Qwen, Gemma, Granite 계열)을 Ollama나 vLLM으로 서빙하는 사례가 증가하고 있다^{[1]-[3]}. 본 시스템 역시 모든 구성 요소가 온프레미스 환경에 위치하며 외부 호출을 전혀 수행하지 않는다. 이러한 폐쇄망 환경에서 사실 기반 응답을 확보하기 위한 표준 접근법으로 자리 잡은 것이 RAG이다. RAG는 LLM의 환각(hallucination) 및 지식 시효 문제^[4]를 검색으로 보완함으로써^[5] 사실 기반 응답의 표준이 되었으나^[6], 동시에 검색 품질, 근거 정합성, 라우팅이 전체 품질을 좌우한다는 점이 부각되면서 단순 밀집 검색을 넘어 하이브리드 검색, 리랭킹, 그리고 근거가 부족할 때의 거부^[7]가 함께 강조되고 있다. 이러한 한계 인식은 질의를 적절한 도구로 연결하는 라우팅 설계로 이어진다. 자연어를 SQL로 변환하는 Text2SQL 대신, 사전에 정의된 읽기 전용 함수들을 도구로 제공하고 모델이 어떤 도구를 호출할지 선택하게 하는 함수 호출 방식이 안전성과 신뢰성 측면에서 선호된다^{[8][9]}. 이때 모델의 도구 선택 자체가 곧 의도 라우팅이 된다. 검색 단계에서는 밀집(의미) 신호^[10]와 희소(정확 토큰) 신호^[11]를 RRF^[12]로 융합한 뒤, cross-encoder 리랭커로 재정렬하는 하이브리드 검색·리랭킹 구성이 보고되었는데, 이는 표, 번호 체계, 전문 용어(모델 코드·약어)가 빈번한 한국어 검색에서 특히 강건한 성능을 보인다^[13].

평가 측면에서는 생성된 답변의 충실도, 적합성, 검색 적합성을 참조 답안 없이 LLM 심판으로 채점하는 RAGAS 유형의 접근법이 확산되고 있으며^{[14][15]}, 본 연구 역시 LLM-as-judge를 활용하여 전수 채점을 수행한다. 한편, 한국어 안전·방재 문서는 표, 번호 체계, 전문 용어가 많아 다국어 임베딩(BGE-M3^[13]), 한국어 특화 생성 모델, 구조 기반 청킹이 품질에 직접적인 영향을 미친다. 무엇보다 방재 및 산업 인프라와 같은 안전 중요 도메인에서는 틀린 답변을 그럴듯하게 제시하는 것이 답을 찾지 못하는 것보다 더 위험하므로^[4], 근거가 불충분할 경우 답변 생성을 거부하고^[7], 안전 임계값 변경 권한을 LLM에 부여하지 않는 보수적인 설계가 요구된다.

3. 연구의 타당성과 기여

기존 RAG 평가 연구의 대부분은 개방형 QA 및 위키 도

a) 서울과학기술대학교 IoT융복합기술연구소(IoT Convergence Technology Lab., Seoul National University of Science and Technology)

b) 서울과학기술대학교 스마트ICT융합공학과(Dept. of Smart ICT Convergence Engineering, Seoul National University of Science and Technology)

✉ Corresponding Author : 박구만(Gooman Park)

E-mail: gmpark@seoultech.ac.kr

Tel: +82-2-970-6430

ORCID: <https://orcid.org/0000-0002-7055-5568>

· Manuscript received August 20, 2026; Revised September 8, 2026;

Accepted September 8, 2026.

메인에서 단일 모델과 소규모 질의로 수행되어 왔다^{[6][14]}. 본 연구는 이와 달리 다음과 같은 현장 제약을 모두 만족하는 조건에서 대규모 평가를 수행한다는 점에서 차별화된다. 첫째, 안전 중요 폐쇄망 환경으로 외부 API 호출이 금지되며, LLM에 대한 쓰기 및 임계값 변경 권한이 부여되지 않고, 근거가 부족할 경우 답변을 거부한다. 둘째, 매뉴얼(검색), 실시간 데이터(함수), 일반 지식, 복합 질의가 혼재하는 이질적인 정보의 동시 응대를 다룬다. 셋째, “지금 위험한 데 대응 절차는?”, “무단 침입인데 어떻게 대응?”과 같이 데이터 조회와 절차 근거를 한 답변에 결합해야 하는 복합 질의를 단순 질의와 분리하여 집중 분석한다. 넷째, 동일한 500문항을 4개 오픈 웨이트 LLM(qwen3.5:9b, gemma4:12b, gemma4:e4b, granite4.1:8b)으로 측정하는 다중 모델 비교를 통해 온프레미스 모델 선택에 대한 실증적 근거를 제공한다. 다섯째, 표본이 아닌 2,000개(500×4) 답변 전체를 동일한 루브릭으로 채점하는 전수 RAGAS 평가를 수행한다.

이를 통한 본 연구의 기여는 다음과 같다. (1) 안전 중요 온프레미스 RAG 챗봇의 복합 질의 응대 구조(멀티라운드 탐색 및 결정적 매뉴얼 보강)를 제시하고, 동일 평가셋에서의 보강 on/off 절제 실험으로 그 작동 특성(결정적 안전 보강)을 실증한다. (2) 난이도, 종류, 영상, 복합 요소가 균형 있게 구성된 500문항 한국어 평가셋과 다중 모델 RAGAS 결과를 공개 가능한 형태로 산출한다. (3) 챗봇 품질에 영향을 미치는 요인들을 검색, 라우팅, 복합 처리, 안전의 관점에서 통찰하여 정리한다.

4. 관련 연구 및 배경

4.1 RAG와 검색 증강

RAG^[5]는 질의와 관련된 문서를 검색하여 생성 컨텍스트로 제공함으로써, 파라미터 지식만으로 답변할 때 발생하는 환각 및 시효 문제를 완화한다^{[6][4]}. 그 품질은 검색과 생성 두 단계 모두에 의존하며, 검색이 부정확할 경우 생성 결과도 부정확해진다. 따라서 본 시스템은 검색 단계에 하이브리드 및 리랭킹 기법을 적용하고, 생성 단계에는 근거 제한 프롬프트와 판정 게이트를 둔다^[7].

4.2 검색: 밀집·희소·하이브리드·리랭킹

밀집 검색^[10]은 문장 임베딩 간 의미 유사도에 기반하므로 페러프레이즈나 동의 표현에 강하지만, 정확한 토큰(모델 코드, 수치 등)에는 약할 수 있다. 반면, 희소 검색(BM25 계열)^[11]은 정확한 토큰 매칭에 의존하여 전문 용어, 약어, 코드에 강한 대신 의미 일반화에는 약하다. 이 두 신호를 RRF^[12]로 융합하는 하이브리드 방식은 점수 스케일 차이에 강건하게 양자를 상호 보완하며, 본 시스템은 밀집과 희소를 동시에 산출하는 BGE-M3^[13] 임베딩을 사용한다. 마지막 단계에서는 1차 후보를 질의-문서 교차 인코딩으로 재정렬하여 정밀도를 높이고, 그 최고 점수를 근거로 충분성 판정에도 함께 활용한다.

4.3 라우팅과 도구 호출

질의의 의도가 매뉴얼 근거를 요구하는지, 실시간 수치를 요구하는지, 아니면 일반 지식을 요구하는지를 판별하는 라우팅은 별도의 분류기 대신 LLM의 도구 호출(function calling)^[8]로 수행할 수 있다. 모델이 도구를 호출하면 그 종류가 곧 경로가 되기 때문이다^[9]. 이 방식은 별도의 분류기 학습이 불필요하며, 도구 스키마와 설명을 통해 의도 신호를 풍부하게 전달하고, 복합 질의에서 여러 도구를 함께 호출하는 자연스러운 표현을 가능하게 한다는 이점을 가진다.

4.4 평가 지표: RAGAS와 LLM-as-judge

RAGAS^[14]는 참조 답안 없이 LLM 심판을 활용하여 RAG 품질을 평가하는 프레임워크로, 주요 지표는 세 가지이다. Faithfulness(충실도)는 답변의 주장이 검색된 컨텍스트나 도구 결과에 의해 뒷받침되는지, 즉 환각 여부를 측정한다. Answer Relevancy(답변 적합성)는 질문에 대해 직접적이고 완전하게 답변하는지를 평가하며, Context Precision(검색 적합성)은 검색된 컨텍스트가 질문에 적합한지를 판단한다. LLM-as-judge는 사람의 채점을 근사하는 자동화되고 확장 가능한 방법이지만, 심판 모델의 편향(위치, 상황, 자기 선호)과 일관성 한계가 존재한다^[15]. 본 연구에서는 이를 인지하여 단일 심판(Claude), 동일 루브릭, 기준 점 척도를 적용하여 일관성을 확보하였다.

4.5 에이전트 패턴과 오케스트레이션 프레임워크

도구를 사용하는 LLM의 제어 흐름에는 몇 가지 대표적인 패턴이 있다. Toolformer^[8]는 모델이 외부 API(도구)를 언제, 어떻게 호출할지 자기지도 학습으로 익히는 접근법으로, 도구 사용 LLM의 출발점에 해당한다. ReAct^[9]는 추론과 행동을 교대로 수행하는 패턴으로, 모델이 도구를 호출하고 그 결과를 관찰하여 다시 입력받아 다음 행동을 결정하는 관찰 → 행동 반복 루프를 형성하며, 사전 계획 없이 단계별로 동적으로 도구를 선택한다. Plan-and-Solve/Plan-and-Execute^[16]는 전체 과제를 하위 단계로 분해한 뒤 그 계획에 따라 실행하는 방식으로, 누락 단계 오류를 줄이지만 사전 분해 단계가 추가된다는 비용이 따른다. Reflexion^[17]은 시도 결과(피드백)를 언어적 자기비판으로 요약하여 에피소드 메모리에 축적하고, 이를 다음 시도에 반영함으로써 궤적을 개선한다. 한편 LangChain^[18]과 LangGraph^[19]는 이러한 패턴의 체인·그래프·에이전트·메모리·도구 추상화를 제공하는 오케스트레이션 라이브러리로, 개발 생산성을 높이는 대신 의존성, 추상화 계층, 비결정성을 함께 도입한다.

II. 본 론

1. 전체 시스템 구성

챗봇은 펌프 예측 및 의사결정 시스템과 같은 Docker Compose 프로젝트에 RAG 프로파일로 통합되며, 코드와 자원은 격리되어 있다. 모든 구성 요소는 온프레미스 환경에 위치하며 외부 API를 호출하지 않는다. 그림 1은 시스템 아키텍처를 보여준다.

라우팅 4경로를 요약하면 다음과 같다. 첫째, **manual: search_manual** → 근거 생성 및 가드레일 적용. 둘째, **data: 운영 함수(read-only)** → DB 및 플랫폼 값 활용. 셋째, **composite: 데이터 함수와 search_manual 결합**. 넷째, 멀티라운드 및 결정적 보강 수행. 마지막으로 **general: 도구 없이 일반 답변 제공 및 면책 처리**. 모든 답변은 항상 ChatResponse 형식으로 정규화한다.

2. 질문 → 답변 파이프라인(단계별)

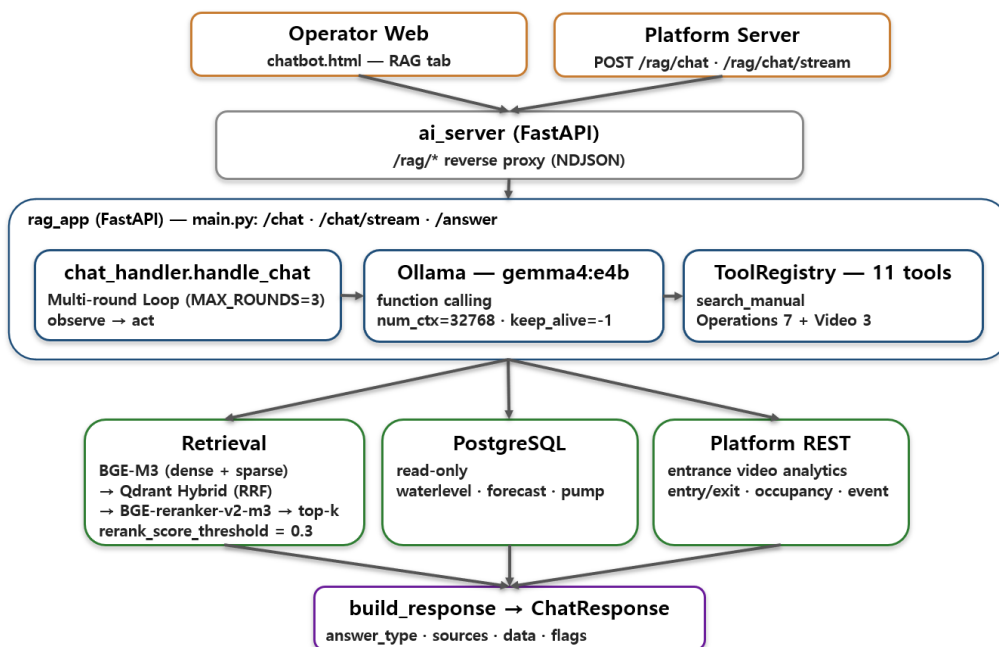


그림 1. 시스템 아키텍처

Fig. 1. System Architecture

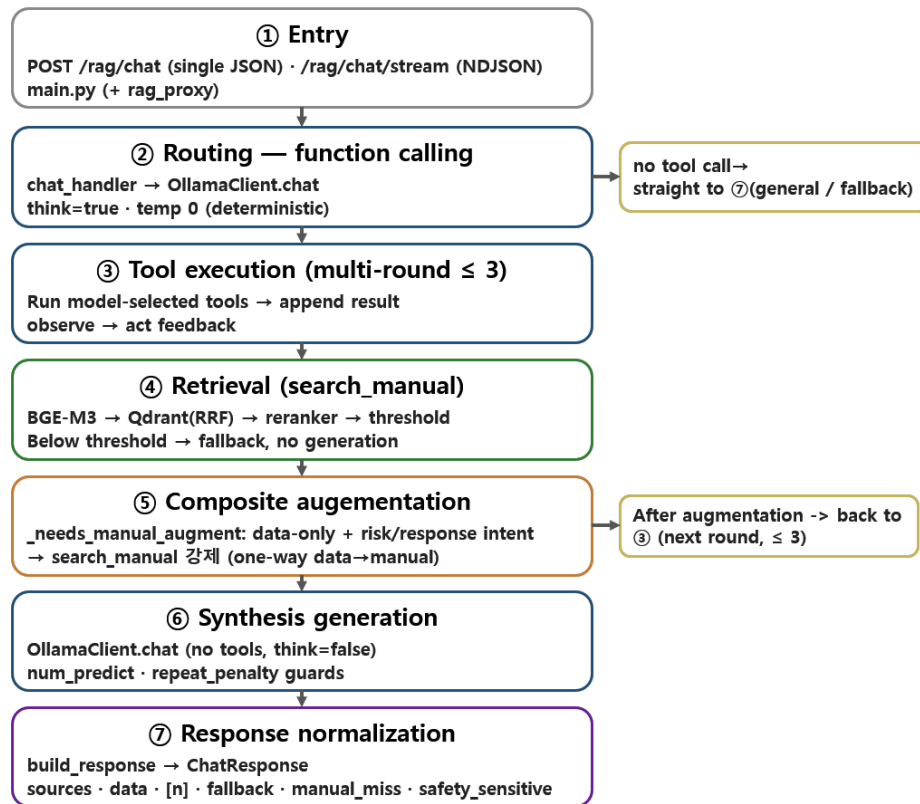


그림 2. 질문 → 답변 파이프라인
Fig. 2. Question → Answer Pipeline

질의는 진입부터 응답까지 총 7단계를 거친다. 각 단계는 독립된 모듈로 분리되어 있어 교체와 검증이 용이하다. 그림 2는 질문에서 답변으로 이어지는 파이프라인을

보여준다.

표 1은 질문 → 답변 파이프라인 단계별 처리 내용이다.

표 1. 질문 → 답변 파이프라인 단계별 처리
Table 1. Question → Answer Pipeline Step-by-Step Processing

Stage	Processing	Key points
① Entry	POST /rag/chat(single JSON) · /rag/chat/stream(NDJSON)	input {query}; streaming is token-by-token
② Routing	tool selection via function calling	think=true · temp 0 (deterministic); no tool call = general
③ Tool execution	run model-selected tool, append result to messages	multi-round (≤3), observe → actfeedback
④ Retrieval	search_manual: hybrid + rerank + threshold	below threshold → fallback without generation
⑤ Composite augmentation	data-only call + risk/response intent → force search_manual	one-waydata → manual, one extra round
⑥ Synthesis generation	synthesize final answer from tool results	think=false · num_predict · repeat_penalty (block hang/repeat)
⑦ Response normalization	Configuration of sources, data, citations, and flags	sources · data · [n] · adjustment · fallback · manual_miss · safety_sensitive

3. 핵심 설계

3.1 하이브리드 검색 + RRF + 리랭킹

단순한 밀집 검색만으로는 충분하지 않다. 안전 및 설비 매뉴얼에는 모델 코드, 약어, 수치 기준이 많아 밀집 임베딩만으로는 정확한 토큰을 놓치기 쉽기 때문이다. 회소 신호를 함께 사용하면 이러한 토큰을 포착할 수 있고, 밀집 신호는 자연어 패러프레이즈를 잡아낸다. 두 순위를 RRF로 융합하여 스케일 차이에 강건하게 만든 뒤, 리랭커로 상위 후보를 재정렬해 정밀도를 높인다. 이때 리랭커 최고 점수가 임계값에 미치지 못하면 답변을 생성하지 않고 대체 응답하는데, 이는 근거가 약할 때 답을 지어내는 것을 원천 차단하기 위함이다. 검색기는 모델과 무관하게 동작하여 모든 LLM이 동일한 하이브리드 + 리랭크 결과를 받는다. 따라서 모델 간 차이는 검색이 아니라 라우팅 및 생성에서 발생하며, 본 실험의 Context Precision이 동일한 질의에 대해 모델 간에 동일하게 나타나는 것이 정상이다.

3.2 함수 호출 라우팅 vs Text2SQL

본 시스템은 Text2SQL 방식을 채택하지 않는다. 운영 DB에 LLM이 생성한 SQL을 직접 실행할 경우, SQL 인젝션, 과도한 데이터 노출, 실수로 인한 쓰기 작업 및 스키마 변경 등의 위험이 있기 때문이다. 대신, 파라미터화된 사전 정의된 읽기 전용 함수들을 도구로 제공한다. 이 함수들은 순위, 예측, 펄프, 이력, 임계, 권고에 출입구 영상 분석 결과인 카운팅, 재실, 이벤트를 더한 총 11종으로 구성되어 있다. 각 함수는 결정적으로 동작하며, 화이트리스트(테이블 및 펄프장), 입력 검증, 조회 기간 상한을 엄격히 적용한다. 또한, 실패 시 예외로 종료되지 않고 명확한 오류 객체를 반환한다. 모델은 어떤 함수를 호출할지 선택하고, 그 결과를 문장으로 정리하는 역할만 수행하며, 숫자와 상태 값은 데이터 출처의 값을 그대로 사용한다.

3.3 안전 우선: 매뉴얼 미검출·안전 게이트·대체 응답

안전이 중요한 도메인에서 가장 위험한 실패는 두 가지다. 첫째, 위험 신호가 담긴 질의를 매뉴얼 검색 없이 흘려보내는 누락이고, 둘째, 근거 없이 그럴듯한 답을 만들어내는 환각이다. 본 시스템은 “근거가 충분할 때만 답변을

생성하고, 부족하면 생성을 거부한다”는 원칙으로 두 실패를 함께 차단하며, 이는 모델이 스스로 검색 필요성과 근거를 점검하는 자기 성찰 기반 RAG^[7]와 같은 문제의식에서 출발한다. 이 원칙 아래 시스템 프롬프트는 위험·이상 신호가 조금이라도 감지되면 다른 무엇보다 먼저 매뉴얼을 검색하도록 지시한다(누락 방지).

구현 수준에서 이 원칙은 세 가지 장치로 강제된다. 첫째, 근거 게이트: 리랭커 최고 점수가 임계값에 미달하면 LLM이 생성한 본문을 폐기하고 고정된 대체 문구(“해당 내용을 매뉴얼에서 찾지 못했습니다. 담당자에게 확인하세요.”)로 응답하며 fallback 플래그를 세운다(환각 차단). 둘째, 매뉴얼 미검출(manual_miss) 전환: 매뉴얼을 검색했음에도 근거가 부족한 경우, 답변 없이 종료하는 막다른 길 대신 일반 지식 기반 답변으로 전환하되, 그 사실을 구조화된 플래그(manual_miss)와 안전 주제 여부 분류(safety_sensitive)로 함께 표시한다. 분류가 실패하면 보수적으로 안전 주제로 간주한다. 웹 인터페이스는 이 구조화 신호에 따라 안전 주제 답변을 기본 보류(접기) 상태로 두고 운영자가 직접 펼칠 때만 노출하여, 근거가 약한 답변이 그대로 노출되지 않게 한다. 셋째, 빌드 게이트: 매뉴얼 밖 질의에 대해 폴백 대신 답변을 생성하면 빌드를 실패 처리하는 자동 검사를 두어, 이후 코드 변경으로 안전 동작이 회귀하는 것을 방지한다. 이 검사는 LLM·검색 서버가 내려간 오프라인 상태에서도 항상 실행된다.

3.4 복합 질의 처리: 멀티라운드 탐색 + 결정적 매뉴얼 보강

복합 질의는 도구를 한 번만 호출해서는 데이터와 절차 근거를 모두 가져오지 못하는 경우가 많다. 예를 들어, “지금 위험한데 대응 절차는?”이라는 질문에 답하려면 현재 순위·위험 등급과 같은 데이터와 위기 단계별 대응 절차 매뉴얼을 한 답변에 함께 답아야 한다. 본 시스템은 두 가지 장치로 이러한 복합 응대를 보강한다.

첫째, 멀티라운드 탐색은 첫 라운드에서 도구를 호출한 뒤에도 추가 라운드를 허용하여, 데이터 두 개 또는 데이터와 매뉴얼을 함께 가져오도록 한다. 라우팅의 적합성 평가도 ‘경로 하나가 맞았는지’가 아니라 ‘필요한 도구를 모두 호출했는지’로 진행한다.

둘째, 결정적(‘결정적’이란 모델 판단에 맡기지 않고, 조

건이 충족되면 코드가 항상 동일하게 실행된다는 의미) 매뉴얼 보강은 위험 상황(침수·위급·비상·대파·범람·무단진입 등)을 묻는 질문에 모델이 데이터 도구만 호출하고 매뉴얼 검색을 누락할 경우, 코드가 정해진 규칙에 따라 매뉴얼 검색을 한 번 더 자동으로 호출하여 대응 절차 근거를 함께 제공한다. 단, 이미 매뉴얼을 검색했거나 단순히 상태만 묻는 질문(대응 의도가 없는 경우)에는 작동하지 않는다(데이터 → 매뉴얼 단방향).

결정적 보강의 트리거 조건은 코드 수준에서 다음과 같이 엄격히 정의된다. 두 한국어 키워드 목록 — 위급 상황 어휘 9종(‘침수’, ‘위급’, ‘비상’, ‘대파’, ‘범람’, ‘차오’, ‘무단진입’, ‘무단 진입’, ‘침입’)과 대응 의도 어휘 13종(‘대응’, ‘절차’, ‘조치’, ‘대처’, ‘행동요령’, ‘어떻게 해’, ‘어떻게 대응’, ‘어떻게 운영’, ‘어떻게 가동’, ‘뭘 해야’, ‘뭘 해야’, ‘무엇을 해야’, ‘어찌’) — 중 하나 이상이 질의에 부분 문자열로 포함되고(조건 ①), 해당 턴에서 데이터 도구가 1회 이상 호출되었으며(조건 ②), 매뉴얼 검색(search_man-

ual)이 아직 호출되지 않았을 때(조건 ③), 세 조건이 모두 충족되는 경우에만 발동한다. 아울러 오프리저(과탐)를 방지하는 예외 규칙도 코드로 강제된다. 첫째, 단순 상태 질문 필터링: 상태 조회 표현에 흔히 나타나는 ‘위험’은 위급 어휘 목록에서 의도적으로 제외하였다(강한 위급 신호인 ‘위급’만 포함). 따라서 “현재 위험 등급이 뭐야?”처럼 대응 의도가 없는 상태 질의에는 트리거 어휘 자체가 존재하지 않아 보강이 개입하지 않는다. 둘째, 중복 방지: 모델이 이미 search_manual을 호출한 경우에는 발동하지 않아 같은 검색의 중복 수행과 보강의 반복 발동을 동시에 차단한다. 셋째, 단방향성: 데이터 도구를 전혀 호출하지 않은 질의(순수 일반 지식·무도구 질의)에는 발동하지 않으며, 보강은 데이터 → 매뉴얼 한 방향으로만 작동한다. 이 트리거-예외 규칙 일체는 단위 테스트로 고정되어 있어 코드 변경 시 회귀가 자동 검출된다.

무단진입과 같은 영상 보안 신호도 위험 신호로 간주하여, 단순 출입 통계는 그대로 답변하고 이상 이벤트는 대응

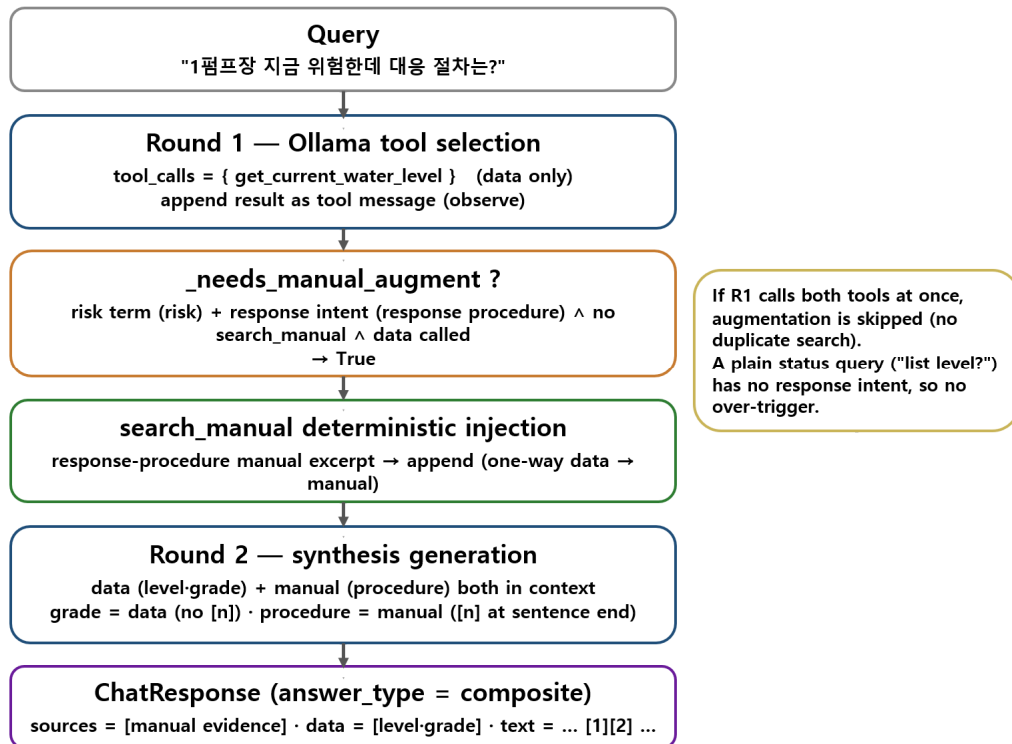


그림 3. 복합 질의의 라운드별 답변 경로
Fig. 3. Round-by-round answer paths for complex queries

절차와 함께 복합적이고 안전 우선적으로 처리한다.

이로부터 다음과 같은 설계 통찰(가설)을 도출할 수 있다. 복합 질의를 모델이 스스로 여러 도구를 호출하는 능력에만 의존하면, 모델과 프롬프트에 따라 결과가 불안정해져 취약해진다. 따라서 모델의 능력에 규칙 기반의 결정적 보강을 더한 하이브리드 방식이 복합 응답을 보다 안정적으로 만든다. 본 연구는 이를 모델별 복합 정확도와 보강 효과를 통해 검증한다. 그림 3은 결정적 보강이 개입하는 복합 질의의 라운드별 답변 경로를 예시(예: “1편프장 지금 위험한데 대응 절차는?”)와 함께 살펴본다. 여기서 핵심은, 이 결정적 보강이 어떤 모델을 사용하든 동일하게 작동하여, 모델이 절차 검색을 누락하는 실수를 자동으로 바로잡아 준다는 점이다. 반대로 모델이 처음(1라운드)부터 두 도구를 함께 호출했다면 보강이 개입하지 않아 같은 검색이 중복되지 않는다. 또한 “위험 등급이 뭐야?”처럼 단순히 상태만 묻는 질문에는 ‘대응·조치’를 의미하는 표현이 없어 보강이 잘못 개입하지 않는다.

멀티라운드 루프의 종료와 예외 처리는 다음과 같이 보장된다. 루프는 고정 상한(최대 3라운드)의 유한 반복으로 구현되어 구조적으로 무한 루프가 불가능하며, 결정적 보강 역시 발동 즉시 도구 호출 이력에 `search_manual`이 기록되어 다음 판정에서 트리거 조건(조건 ③)이 거짓이 되므로 최대 1회만 개입한다(보강 자체의 반복 발동 차단). 3라운드를 소진할 때까지 필요한 정보를 모두 얻지 못한 경우에도 시스템은 예외로 종료하지 않는다. 대신 도구 정의를 제거한 최종 합성 호출을 1회 수행하는데, 이 호출에는 도구가 제공되지 않으므로 모델은 더 이상 도구를 호출할 수 없고, 그때까지 수집된 도구 결과만으로 텍스트 답변을 생성하도록 구조적으로 강제된다. 개별 도구의 실행 실패(미등록 도구, 인자 오류, 실행 예외)도 예외를 전파하지 않고 구조화된 오류 객체로 반환되어 루프가 계속되며, 오류만 담긴 도구 결과는 원시 오류 내용 대신 “실시간 데이터 없음” 류의 명시적 한 줄 안내로 답변에 반영된다. 생성 단계에는 토큰 상한(`num_predict=1,536`)과 반복 억제(`repeat_penalty=1.3`)를 두어 무한 생성·반복으로 인한 멈춤(`hang`)을 차단하고, LLM 호출 자체에는 타임아웃(180초)과 2회 재시도를 둔다. 이렇게 어떤 경로로 종료하든 최종 텍스트는 동일한 정규화 함수를 거쳐 단일 응답 계약(`Chat-`

`Response`)으로 수렴한다. 정규화 단계는 리inker 점수가 임계값에 미달하면 본문을 고정 대체 문구로 교체하고(3.3절의 근거 게이트), 근거 목록 범위를 벗어난 인용 번호를 제거하며, `sources·data`·플래그 필드를 실제 도구 결과와 일치하도록 구성한다. 유일하게 예외가 상위로 전파되는 경우는 Ollama 추론 서버 자체가 응답 불가능일 때이며, 이때는 그럴듯한 빈 답변으로 위장하는 대신 HTTP 503으로 실패를 명확히 알린다.

3.5 에이전트 오케스트레이션 설계: 프레임워크 미채택 근거

본 챗봇의 제어 흐름은 ReAct 계열(관찰 → 행동 멀티라운드)에 해당하지만, 에이전트 프레임워크(LangChain·LangGraph 등)는 사용하지 않는다(의존성 및 임포트 전수 확인). 표준 웹·DB·검증 라이브러리만으로 Ollama를 직접 HTTP로 호출하며, 멀티라운드 루프(최대 3라운드)는 소규모 커스텀 코드로 구현했다. 도구 결과를 다음 라운드 입력으로 되돌리는 관찰 → 행동 피드백은 ReAct와 같지만, 사전 계획 분해(Plan-and-Execute) 및 언어적 자기비판/계획 재시도(Reflexion)는 포함하지 않았다.

프레임워크와 고차 에이전트 패턴을 채택하지 않은 설계 근거는 다음과 같다.

- (1) 결정성·감사 가능성: 안전 중요 도메인에서는 “왜 이렇게 답변했는가”에 대한 추적이 필수적이다. 본 시스템은 모든 도구 호출을 ‘`ChatResponse.tool_calls`’에 그대로 노출하며, 보강은 키워드 기반의 결정적 규칙을 통해 재현 가능하도록 설계되었다. 학습형 및 자기 성찰형 루프는 비결정성을 증가시켜 감사를 어렵게 만든다.
- (2) 안전 불변식의 코드화: “위험 신호가 있을 경우 매뉴얼 검색을 누락하지 않는다”는 불변식을 단순히 프롬프트 유도에만 의존하지 않고, 코드로 명확히 강제하였다(단방향 데이터 → 매뉴얼). 프레임워크의 자율 계획에 위임할 경우 이 불변식을 보장하기 어렵다.
- (3) 지연 및 자원: 상주 단일 모델(`gemma4:c4b`)에 라운드 상한(≤ 3)을 설정하여 지연을 예측 가능하게 했다. Plan 후 다수의 Execute를 수행하거나 Reflexion으로 재시도하는 방식은 호출 횟수와 지연을 증가시킨다.

- (4) 의존성 및 라이선스: 온프레미스 폐쇄망 배포 환경에서 무거운 의존성 트리가 보안, 라이선스, 유지보수에 부담을 준다. 최소한의 의존성은 공급망 공격 표면을 줄이는 데 기여한다.
- (5) 충분성: 본 도메인의 질의는 대부분 “데이터 + 절차”의 1~2홉 결합으로 이루어져 있어, 사전 분해(Plan)나 자기 성찰(Reflexion)의 이점이 크지 않다. 따라서 ReAct 방식의 관찰 → 행동에 결정적인 보강을 더한 접근법으로 충분하다. 표 2는 본 시스템과 에이전트 패턴(ReAct·Plan-and-Execute·Reflexion)의 비교를 나타낸다.

요약하자면, 본 시스템은 ReAct 방식의 관찰 → 행동 루프에 도메인 안전 불변식을 결정적으로 추가한 최소한의 오케스트레이션이다. ‘더 똑똑한 에이전트’가 아니라 ‘더

검증 가능한 에이전트’를 선택한 점이 안전 중요 및 온프레미스 설계의 핵심이다.

3.6 영상 분석 통합(플랫폼 REST 소비)

영상 분석(진입·진출 카운팅, 자전거·킥보드, 무단 진입)은 영상 분석 AI에서 플랫폼 서버를 거쳐 데이터베이스(DB)에 적재되며, 챗봇은 영상 DB에 직접 접근하지 않고 플랫폼 REST API를 통해서만 조회하는 읽기 전용 소비자로 동작한다. 즉, 챗봇은 실시간 CCTV나 원본 영상을 직접 처리하지 않고, 영상 분석 AI가 산출한 분석 결과(메타데이터)만을 소비한다. 평가 시에는 외부 플랫폼 의존성을 제거하기 위해 합성 데이터(결정적)를 사용하여 영상 함수를 구동한다.

4. 실험 설계

표 2. 본 시스템과 에이전트 패턴(ReAct·Plan-and-Execute·Reflexion) 비교

Table 2. Comparison of this system and agent patterns (React, Plan-and-Execute, Reflexion)

Aspect	This system	ReAct ^[9]	Plan-and-Execute ^[16]	Reflexion ^[17]
Framework	none (direct HTTP)	pattern	pattern	pattern
Tool orchestration	custom loop (≤ 3)	dynamic	plan then execute	dynamic + retry
Observe → act feedback	yes	yes	yes	yes
Upfront plan decomposition	no	no	yes	no
Self-critique / trajectory revision	no	no	no	yes
Deterministic safety augmentation	yes(keyword rule)	no	no	no
Fallback gate (threshold)	yes(rerank 0.3)	no	no	no
Nondeterminism · latency	low (predictable)	medium	medium - high	high

표 3. 평가 데이터셋 구성(유형·난이도별 문항 수)

Table 3. Evaluation dataset composition (number of items by type and difficulty)

Type (route)	Total	easy	medium	hard	Notes
manual · operations	75	25	25	25	facilities · inspection · operations
manual · safety	65	22	22	21	disaster · crisis levels · reporting chain
manual · guideline	65	22	22	21	flood action guidelines · criteria
manual · video	30	10	10	10	entrance video operating guideline (counting · intrusion)
data	85	29	28	28	level · forecast · pump · history · threshold · recommendation
data · video	40	16	14	10	entry/exit · occupancy · events (REST)
general	80	27	27	26	domain-adjacent · out-of-domain (trap)
composite	60	20	20	20	data+manual · data+data · video+response
total	500	171	168	161	manual 47%·data 25%·general 16%·composite 12%

4.1 평가 데이터셋(500문항)

평가에는 프로그램으로 생성한 500개의 질의 데이터셋을 사용한다. 각 질의는 정답 라우팅 경로, 기대 범주, 기대 함수, 난이도, 함정 여부, RAGAS 표본 포함 여부 등의 라벨이 생성 시점에 함께 부여되어 정답이 설계 단계에서 보장된다. 데이터셋은 난이도(easy/medium/hard), 유형(manual/data/general/composite), 그리고 영상 및 복합 차원에서 균형 있게 배치된다. 표 3은 평가 데이터셋 구성(유형 및 난이도별 문항 수)을 보여준다.

구체적으로 영상 분석과 관련된 70문항(매뉴얼 30문항 + 데이터 40문항)과 복합 내 영상 11문항을 포함하며, 함정 문항 57문항(일반 오프리거 방지 25문항, 데이터/영상 펌프장 미지정 26문항, 복합 6문항)을 두어 라우터의 과탐(false trigger) 및 명확화(list stations) 처리를 평가한다. 난이도는 약 1:1:1 비율로 구성하였으며, ragas_sample(난이도 × 종류 2/cell, 42문항)은 분석용 증화 플러그로서 전수 채점에 도움을 준다.

4.2 비교 모델(4종)

표 4는 평가 대상인 4종의 오픈 웨이트 LLM에 대한 설명이다.

gemma4:12b는 Ollama 0.30.7 버전을 요구하지만, 이 버전은 qwen 라우팅을 손상시킨다. 따라서 수집 오케스트레이션은 0.24.0 버전에서 qwen, e4b, granite를 측정 한 후, 0.30.7 버전으로 일시 전환하여 12b를 측정하고, 종료 시 다시 0.24.0 버전으로 복원한다.

gemma4:12b는 구동에 Ollama 0.30.5 이상을 요구하여 다른 세 모델의 측정 버전(0.24.0)에서는 실행 자체가 불가능하다. 따라서 각 모델을 구동 가능한 버전에서 측정하되, 실행 버전이 결과에 미치는 영향은 별도의 통제 측정으로 분리하였다. 구체적으로 gemma4:e4b를 gemma4:12b와 동

일한 0.30.7에서 동일한 500문항으로 재측정하여, 실행 버전 자체가 생성 안정성을 좌우하는지를 직접 검정하였다(5.4절). 그 결과 gemma4:e4b의 빈 답변은 0.24.0에서 3건, 0.30.7에서 4건으로 두 버전 모두 낮게 유지된 반면 같은 0.30.7의 gemma4:12b는 78건이었다. 즉 0.30.7이라는 실행 버전은 생성 불안정을 유발하지 않으며, gemma4:12b의 빈 답변은 모델 특성에 기인한다.

4.3 평가 코퍼스 및 데이터 출처

검색 코퍼스로는 Qdrant 벡터 컬렉션 safety_manuals(약 3,471개 청크)를 사용한다. 이는 실제 매뉴얼 PDF(시설물 안전·유지관리 세부지침, 풍수해 현장조치 행동매뉴얼, 배수펌프장 운영관리지침, 침수 대비 국민행동요령 등)에서 추출한 operations 2,332개, guideline 605개, safety 512개 청크에, 본 실험용으로 별도 생성·적재한 「배수펌프장 + 출입구 영상 분석 운영지침」 video 22개 청크를 더한 것이다. 검색은 모델과 무관하게 동일한 하이브리드 + 리랭크 방식으로 수행된다. 운영 데이터는 평가의 결정성 및 재현성을 위해 합성 데이터(수위·예측·펌프·임계 + 영상 카운팅·재실·이벤트)로 대체한다. 즉, 수치 자체는 합성이지만, 라우팅·도구 선택·근거 결합·문장화라는 챗봇의 핵심 행위는 동일한 조건에서 측정된다.

4.4 평가 지표

평가는 결정적 자동 채점과 RAGAS(Claude judge, 전수 2,000건)를 병행하여 진행한다. 자동 채점은 네 가지 지표로 구성된다. 라우팅 정확도는 기대 경로 대비 일치율로, manual과 composite, data와 composite 간의 전환을 허용한다. 함정 방지는 일반 질의에서 오프리거 사용을 금지하며, 펌프장이 지정되지 않은 data 및 video 질의에서 후보 확인 같은 명확화 요구 여부를 평가한다. 데이터 함수 매칭은 기

표 4. 평가 대상 오픈 웨이트 LLM 4종

Table 4. Open Weight LLM Types Subject to Evaluation

Model	Scale	Ollama	think	Notes
qwen3.5:9b	~9.65B	0.24.0	ON	strong Korean reasoning; pre-consolidation candidate
gemma4:e4b	~4.5B (efficient)	0.24.0	toggle	current single production model
gemma4:12b	~12B	0.30.7	ON	large; requires newer Ollama
granite4.1:8b	~8B	0.24.0	OFF(Not supported)	IBM Granite, --no-think

대 함수 집합을 호출한 비율(0~1)이며, 로컬 충실도(local faithfulness)는 토큰 포함 휴리스틱을 활용해 LLM과 무관하게 근거 정합성을 교차 검증한다. RAGAS는 정답 라벨 없이(reference-free) 0~1 척도(기준점 0, 0.25, 0.5, 0.75, 1.0)에서 세 가지 지표를 채점한다. Faithfulness는 근거 정합성으로, 컨텍스트나 도구 결과에 없는 내용과 환각을 감점하며 근거가 없으면 점수를 부여하지 않는다. Answer Relevancy는 질문에 대한 적당성으로, 빈 답변이나 대체 응답을 감점하되 적절한 명확화는 0.5~0.7 범위로 평가한다. Context Precision은 검색 적합성으로, manual과 composite 질의에만 적용하며 data와 general 질의에는 산정하지 않는다. 모든 배치에 동일한 루브릭과 기준점을 적용하여 일관성을 확보한다.

4.5 실험 절차 및 재현

평가는 다음 다섯 단계로 수행한다. (1) 500개의 질의 데이터셋을 프로그램으로 생성한다. (2) 영상 분석 운영지침 코퍼스(영상 분석 운영지침, category=video)를 검색 컬렉션에 적재한다. (3) 4개 모델로 모든 질의에 답변하게 하여 2,000개의 답변을 수집한다. 이때 gemma4:12b는 상위 Ollama(0.30.7)를 요구하므로 수집 중에 버전을 일시 전환하며, granite4.1:8b는 think를 지원하지 않아 비활성화한다. 수집은 백그라운드에서 약 8~12시간 소요된다. (4) 수집한 2,000개의 답변을 50개 단위 배치로 나누어 동일한 루브릭으로 RAGAS 채점(Claude judge)을 수행한 뒤 결과를 병합

한다. (5) 라우팅, RAGAS, 문항별 결과를 표로 집계하여 비교 리포트를 생성한다.

5. 결과

500문항×4모델=2,000개의 답변을 수집하여 전수 RAGAS 채점(Claude judge, 동일 루브릭)을 수행하였다. 라우팅, 합정, 데이터 함수 매칭은 코드 자동 채점(전수)으로 진행되었으며, RAGAS의 3가지 지표는 전수 정성 채점으로 평가되었다.

5.1 라우팅 정확도(전체 · 카테고리 · 난이도)

표 5는 모델별 라우팅 정확도(전체·카테고리·난이도)를 보여준다.

검색기는 모델과 무관하게 동일한 하이브리드 + 리랭크 방식을 사용하므로, 모델 간 변별은 라우팅과 생성 단계에서 발생한다. 난이도에 따라 모든 모델에서 easy에서 hard로 갈수록 정확도가 점진적으로 하락하지만, 모델 간 순위는 유지된다.

5.2 RAGAS 지표(모델별, 전수 500×4)

표 6은 모델별 RAGAS 지표(전수 500×4)를 보여준다. 핵심 결과는 다음과 같다. granite4.1:8b는 전체 지표에서 최하위를 기록했으며(Faithfulness 0.405, Relevancy 0.481로 다른 모델의 약 절반 수준), gemma4:12b는 Faithfulness

표 5. 모델별 라우팅 정확도(전체 · 카테고리 · 난이도)

Table 5. Routing accuracy by model (All · Category · Difficulty)

Metric	qwen3.5:9b	gemma4:e4b	gemma4:12b	granite4.1:8b
Overall routing accuracy	0.88	0.88	0.77	0.69
operations	73/75 (0.97)	64/75 (0.85)	75/75 (1.00)	32/75 (0.43)
safety	64/65 (0.98)	57/65 (0.88)	65/65 (1.00)	34/65 (0.52)
guideline	64/65 (0.98)	53/65 (0.82)	65/65 (1.00)	24/65 (0.37)
data	79/85 (0.93)	81/85 (0.95)	57/85 (0.67)	81/85 (0.95)
general	55/80 (0.69)	75/80 (0.94)	33/80 (0.41)	79/80 (0.99)
composite	51/60 (0.85)	52/60 (0.87)	33/60 (0.55)	43/60 (0.72)
video (combined 70)	54/70 (0.77)	56/70 (0.80)	59/70 (0.84)	49/69 (0.71)
Trap over-trigger avoidance	44/57 (0.77)	50/57 (0.88)	28/57 (0.49)	42/57 (0.74)
Data-function match (avg)	0.435	0.762	0.473	0.714
Difficulty easy / medium / hard	0.90 / 0.87 / 0.88	0.86 / 0.91 / 0.86	0.83 / 0.75 / 0.74	0.67 / 0.76 / 0.62

표 6. 모델별 RAGAS 지표(전수 500×4)

Table 6. RAGAS metrics by model (total 500×4)

Metric	qwen3.5:9b	gemma4:e4b	gemma4:12b	granite4.1:8b
Faithfulness	0.787	0.759	0.811	0.405
Answer Relevancy	0.691	0.755	0.635	0.481
Context Precision	0.697	0.622	0.642	0.509
(local token faithfulness cross-check)	0.951	0.965	0.950	0.885
Number of blank responses / 500	48	3	78	12

가 가장 높았으나(0.811) 빈 답변이 78건에 달해 Answer Relevancy가 0.635까지 떨어졌다. 반면, gemma4:e4b는 Answer Relevancy가 가장 높았고(0.755), 빈 답변이 최소(3건)로 가장 안정적인 모습을 보였다.

5.3 복합 질의(composite, 60문항) 전용 분석

표 7은 복합 질의(60문항)에 대한 전용 성능을 보여준다.

복합 질의에서 라우팅 정확도는 모델에 따라 0.55(gemma4:12b)에서 0.87(gemma4:e4b)까지 큰 편차를 보였다. 이 0.55~0.87은 서로 다른 모델 간의 편차 범위이며, 구조적 보강(멀티라운드 탐색·결정적 매뉴얼 보강) 자체의 전후 효과는 5.4절의 절제(ablation) 실험으로 분리하여 검증한다. 복합 60문항 중 매뉴얼 검색(search_manual)이 최종 도구 집합에 포함된 문항은 모델별 35~45건이었다(표 7 자발적 호출과 결정적 보강의 합산이므로 열 이름을 'search_manual included'로 표기한다). 보강 주입만의 기여는 5.4절의 절제 실험에서 별도로 계수하였다.

다만, 복합 답변의 품질(Answer Relevancy)은 라우팅보다 생성의 안정성에 크게 좌우된다. gemma4:12b는 라우팅과 충실도는 높지만, 긴 복합 답변을 생성하던 중 멈추거나 반복하여 빈 답변이 23건 발생하였고, 그 결과 Relevancy가 0.42로 크게 하락했다. 반면 gemma4:e4b는 빈 답변이 1건

에 불과하여 Relevancy 0.74로 가장 안정적이었다.

5.4 결정적 매뉴얼 보강 절제(ablation) 실험

구조적 보강의 기여를 모델 간 편차와 분리하기 위해, 동일 평가셋의 composite 60문항에 대해 결정적 매뉴얼 보강만을 켜고 끈(on/off) 두 조건으로 라우팅 정확도를 재측정하였다. 측정은 각 모델을 본 실험 수집과 동일한 Ollama 버전(gemma4:e4b=0.24.0, gemma4:12b=0.30.7)에서 수행하여 표 7의 절대값과 직접 비교할 수 있게 하였고, 추가로 최신 단일 버전(0.31.1)에서도 같은 절차를 반복하여 실행 환경 변화의 영향을 함께 관측하였다. 모든 조건에서 시스템 프롬프트·도구 정의·멀티라운드 프로브(최대 2라운드)·채점 규칙(경로 일치 또는 기대 함수 집합 완전 호출)은 동일하다. 보강 코드에는 실험 전용 토글(기본값 on)을 추가하였고, 강제 주입 횟수를 명시적으로 계수하였다. off 조건에서는 보강이 전혀 개입하지 않아 모델의 자발적 도구 선택만으로 라우팅이 결정된다. 표 8은 그 결과다.

결과의 해석은 세 가지다. 첫째, 본 실험 수집과 동일한 버전에서 절제 결과는 표 7의 모델 간 서열과 절대값을 근사 재현하면서(e4b 0.83 대 표 7의 0.87, 12b 0.47 대 0.55 - 시행 간 변동 범위), on/off 간 평균 정확도 차이가 전혀 없음을 보였다(두 모델 모두 동일 정확도). 최신 버전에서

표 7. 복합 질의(60문항) 전용 성능

Table 7. Dedicated performance for complex queries (60 items)

Model	Routing accuracy	RAGAS Faith	RAGAS Rel	Empty answers	search_manual included	Multi-tool (≥2)
qwen3.5:9b	0.85	0.76	0.64	10	35	52
gemma4:e4b	0.87	0.71	0.74	1	45	53
gemma4:12b	0.55	0.79	0.42	23	42	33
granite4.1:8b	0.72	0.57	0.39	0	36	54

표 8. 결정적 매뉴얼 보강 on/off 절제 실험(composite 60문항)

Table 8. Ablation of deterministic manual augmentation (60 composite items)

Model	Ollama	augment OFF	augment ON	Forced injection(ON)	Voluntary search_manual(OFF)
gemma4:e4b	0.24.0 (Same as this experiment)	0.83 (50/60)	0.83 (50/60)	1 time	42/60
gemma4:12b	0.30.7 (Same as this experiment)	0.47 (28/60)	0.47 (28/60)	1 time	40/60
gemma4:e4b	0.31.1 (Reference)	0.82 (49/60)	0.77 (46/60)	1 time	43/60
gemma4:12b	0.31.1 (Reference)	0.42 (25/60)	0.42 (25/60)	3 times	38/60

도 같았다(12b 동일, e4b의 -0.05는 강제 주입 1회와 무관한 시행 간 비결정성 범위 - 두 시행 사이에 정오가 뒤바뀐 문항이 양방향으로 13건 존재). 보강이 발동한 문항은 채점 구조상 데이터 경로가 복합 경로로 승급되어 정답으로만 전환될 수 있으므로, 정확도를 낮추는 부작용은 원리적으로 발생하지 않는다. 둘째, 강제 주입 자체가 드물었다(모든 조건에서 60문항 중 1~3회). 두 모델 모두 멀티라운드 프로브에서 복합 질의의 대부분에 search_manual을 자발적으로 호출하여(off 조건 기준 e4b 42~43/60, 12b 38~40/60) 보강의 트리거 조건(매뉴얼 미호출)이 성립하는 잔여 사례 자체가 적었기 때문이다. 즉, 표 7의 search_manual 포함 35~45건의 대부분은 모델의 자발적 호출이었고, 결정적 보강은 그 잔여를 메우는 역할이었다. 한편, 복합 라우팅 절대값은 실행 버전에 따라 관측 가능하게 이동했다(e4b 0.87→0.83→0.77~0.82, 12b 0.55→0.47→0.42). 이는 라우팅 행태가 실행 버전에 민감함을 실측으로 보여주며, 실행 버전 통제 측정(표 9)에서 함정 회피가 버전에 따라 이동한 것과도 부합한다. 셋째, 따라서 결정적 보강의 가치는 ‘평균 정확도를 끌어올리는 장치’가 아니라, 실행 환경에 따라 변동하는 모델의 자발적 호출 성향과 무관하게 ‘위험·대응 질의에서 매뉴얼 검색이 누락되는 잔여 사례를 항상 차단하는 결정적 보강(안전망)’으로 해석하는 것이 정확하다. 개입이 필요 없는 환경에서는 침묵하고(오프리거 없음), 필요한 환경에

서는 모델과 무관하게 결정적으로 작동한다는 성질이 안전 중요 도메인에서 이 장치가 갖는 설계 의의다.

아울러 실행 버전이 결과를 좌우했을 가능성(4.2절)을 검증하기 위해, gemma4:e4b를 gemma4:12b와 동일한 0.30.7에서 동일한 500문항으로 재측정하였다. 빈 답변은 응답 본문의 공백 여부로 판정되어 LLM 심판 없이 코드로 채점되므로, 이 비교는 RAGAS 재채점 없이 수행할 수 있다. 표 9는 그 결과다.

동일한 0.30.7에서 gemma4:e4b의 빈 답변은 4건인 반면 gemma4:12b는 78건으로 약 20배 차이가 났고, gemma4:e4b의 값은 0.24.0에서의 3건과 사실상 같았다. 따라서 0.30.7이라는 실행 버전은 생성 불안정의 원인이 아니며, gemma4:12b의 빈 답변은 모델 고유 특성이다. 라우팅 정확도(0.876→0.866)와 데이터 함수 매칭(0.762→0.749)도 두 버전에서 거의 동일했다. 다만 함정 회피는 0.877에서 0.702로 낮아져 라우팅 행태의 일부가 실행 버전에 민감함을 보여주는데, 이 경우에도 같은 버전의 gemma4:12b(0.491)보다는 높아 모델 간 서열은 유지되었다. 복합 질의에 한정된 절제 실험에서도 gemma4:12b의 라우팅 열세는 수집 당시 버전(0.30.7: 0.47)과 최신 버전(0.31.1: 0.42) 모두에서 재현되었다(gemma4:e4b는 각각 0.83, 0.77~0.82; 무도구 응답 13~16건이 주요 실패 양상).

본 절제 실험의 한계는 로컬 미보유 모델(qwen3.5:9b,

표 9. 실행 버전 통제 측정(동일 500문항, 코드 채점)

Table 9. Controlled measurement of the runtime-version effect (same 500 items, code-scored)

Model	Ollama	Empty Response	Routing Accuracy	Trap Avoidance	Data-Function Matching
gemma4:e4b	0.24.0 (Main Experiment)	3/500 (0.6%)	0.876	0.877	0.762
gemma4:e4b	0.30.7 (Control Measurement)	4/500 (0.8%)	0.866	0.702	0.749
gemma4:12b	0.30.7 (Main Experiment)	78/500 (15.6%)	0.774	0.491	0.473

granite4.1:8b)과 RAGAS 생성 품질 측을 포함하지 않았다는 점이다.

5.5 영상 분석 관련 질의(70문항) 분석

표 10은 영상 분석 관련 질의 70문항에 대한 성능을 보여준다.

영상 분석 운영 지침(video manual, 합성 적재)과 영상 분석 데이터(합성) 모두에서 gemma4 계열이 뛰어난 성능을 보였다. 반면, qwen3.5:9b는 video-data에서 빈 답변이 19건 발생하여 영상 분석 데이터 함수의 선택 및 합성에 취약한 모습을 보였으며, 이는 전체 데이터 함수 매칭 점수 0.435와 일관된 결과이다.

5.6 라우팅 행태: 응답 유형 분포와 함정 회피

표 11은 모델별 응답 유형(answer_type) 분포와 빈 답변 수(500개당)를 나타낸다.

granite4.1:8b는 매뉴얼 질의를 일반 지식(자체 지식)으로 처리하는 경향이 강해(manual answer_type 38/500), 매뉴얼 근거화 실패가 RAGAS 붕괴로 직결되었다(operations Faith 0.32, safety 0.12, guideline 0.15). gemma4:12b는 매뉴얼 과트리거 경향을 보여(manual 192, general 함정 방지 28/57로 최저), 일반 지식 질의까지 매뉴얼로 끌어들이는 탓에 함정 방지가 약하고, 데이터 도구 사용도 84로 가장

적었다. 이에 비해 gemma4:e4b는 빈 답변 3건으로 생성 강건성이 압도적이었다.

6. 논의 및 통찰

6.1 챗봇이 잘 되기 위한 요인

챗봇이 성공하기 위한 요인은 세 가지로 나눌 수 있다. 첫째, 검색은 모델과 무관하며, 변별은 라우팅과 생성 과정에서 발생한다. Context Precision이 0.51~0.70으로 모델 간 차이가 크지 않은 이유는 검색기가 공통으로 사용되기 때문이며, 실제 성능 차이는 올바른 경로와 도구를 선택하는 라우팅과 컨텍스트를 충실히 활용하는 생성 과정에서 나타난다. 따라서 검색 파이프라인(하이브리드 + 리랭킹)을 고정하고 라우팅 및 생성 모델을 선별하는 것이 합리적이다. 둘째, 생성 강건성, 즉 빈 답변과 멈춤 현상을 회피하는 능력이 안전한 챗봇 품질의 핵심 변수이다. 빈 답변 수(e4b 3, qwen 48, granite 12, gemma4:12b 78)가 Answer Relevancy 순위를 사실상 결정했으며, 라우팅과 충실도가 높더라도(gemma4:12b) 생성이 불안정하면 운영 품질은 저하된다. 셋째, 안전 편향과 과도한 탐색의 균형이 중요하다. 위험 신호에 대해 매뉴얼을 적극적으로 검색하여 누락을 방지하되, 일반 지식까지 과도하게 트리거하면(gemma4:12b 함정 0.49) 신뢰도가 떨어진다. gemma4:e4b의 함정

표 10. 영상 분석 관련 질의(70문항) 성능

Table 10. Performance of video analysis related queries (70 items)

Model	video RAGAS Faith	video RAGAS Rel	video-data empty	video-manual empty
qwen3.5:9b	0.76	0.59	19	0
gemma4:e4b	0.82	0.71	0	0
gemma4:12b	0.89	0.73	10	1
granite4.1:8b	0.46	0.58	1	0

표 11. 모델별 응답 유형(answer_type) 분포 및 빈 답변 수(500개당)

Table 11. Distribution of response types (answer_type) by model and number of empty answers (per 500)

Model	manual	data	general	composite	Empty/500
qwen3.5:9b	159	144	179	18	48
gemma4:e4b	143	145	184	28	3
gemma4:12b	192	84	184	40	78
granite4.1:8b	38	176	284	2	12

0.88은 ‘필요할 때만 검색’이라는 적절한 균형을 보여준다.

6.2 복합 질의 구조·설계 제언

복합 질의는 모델의 자발적인 다중 도구 호출에만 의존할 경우 취약하다. 본 실험에서 복합 라우팅 정확도는 모델 간 0.55~0.87의 편차를 보였고, 동일 평가셋 절제 실험(5.4절)은 결정적 매뉴얼 보강이 평균 정확도를 끌어올리는 장치가 아니라, 모델의 자발적 매뉴얼 호출이 닿지 않는 잔여 누락 사례를 모델·실행 환경과 무관하게 차단하는 결정적 안전망임을 보였다. 여기서 복합 질의의 성공은 라우팅(탐색 + 보강)과 생성 안정성의 곱으로 볼 수 있으며, 이 두 축이 분리된다는 점이 본 연구의 핵심 발견이다. 결정적 보강은 “데이터만 보고 절차를 빠뜨리는” 라우팅 실패를 모델과 무관하게 교정하며(5.4절), 생성이 불안정한 모델(gemma4: 12b, 복합 빈 답변 23건)은 라우팅이 교정되어도 여전히 실패한다. 즉, 구조가 라우팅의 절반을, 모델이 생성의 절반을 책임진다. 또한 함수 집합 일치로 채점할 경우 한 답변에 데이터 함수와 매뉴얼 검색이 함께 호출되었는지를 확인하기 때문에, 단일 경로 일치만 보는 방식보다 복합 질의의 실패를 더 정확히 포착한다. 이로부터 세 가지 설계 제언을 도출한다. 첫째, 위험·보안 의도 어휘에 대한 결정적 매뉴얼 보강을 유지·확장한다(영상 무단 진입 등 포함). 둘째, 복합 답변 생성 시 생성 길이 상한과 반복 억제 가드(num_predict·repeat_penalty)를 두어 멈춤과 반복을 차단함으로써 빈 답변을 줄인다. 셋째, 라우터로는 과도한 탐색이 적고 생성이 안정적인 모델(본 실험의 gemma4:e4b)을 선택한다.

6.3 모델별 강약과 운영 선택

표 12는 모델별 강점, 약점 및 운영 적합성을 보여준다.

온프레미스 단일 모델 선택에 대한 실증적 근거: 운영 측면에서 의사결정, 보고서 작성, 챗봇, RAG가 단일 모델을 공유하는 것이 자원 효율성과 일관성 면에서 유리하다. 본 실험은 gemma4:e4b 모델이 RAG 라우터에서도 균형 잡힌 1위 성능을 보임으로써 단일 모델 통합의 타당성을 입증한다.

III. 결론 및 향후 연구

1. 향후 연구

본 연구에는 몇 가지 한계점이 있다. 첫째, RAGS 채점은 정답 라벨이 없는(reference-free) 단일 심판(Claude)에 의존하므로, 심판의 편향과 일관성 문제를 완전히 배제하기 어렵다. 따라서 향후 다중 심판 및 인간 평가 교차검증과 정답셋 기반의 컨텍스트 리콜 보강이 필요하다. 둘째, 실행 버전의 영향은 gemma4:e4b를 두 버전에서 측정해 분리하였으나(4.2절·5.4절), 이 통제는 gemma4:e4b·gemma4:12b 두 모델과 코드 채점 지표에 한정된다. 4모델 전체를 단일 버전에서 재수집하고 RAGS까지 재채점하는 전면 검증은 향후 과제로 남는다. 셋째, 운영 데이터는 합성값으로 대체하였다. 라우팅, 근거 결합, 문장화 조건은 동일하지만 수치 자체가 합성된 것이므로, 실제 플랫폼 연동 후 실행 데이터로 재측정할 가치가 있다. 넷째, 구조적 보강의 절제 실험

표 12. 모델별 강점 · 약점 · 운영 적합성

Table 12. Strengths, Weaknesses, and Operational Suitability by Model

Model	Strengths	Weaknesses	Operational fit
gemma4:e4b	top balance of routing · trap · data-function · Relevancy; empty 3; 0.24.0 · shared with pump	slightly lower Faithfulness	★optimal (cuts resident 2→1 model)
qwen3.5:9b	strong manual routing · Faithfulness · Context Precision	lowest data-function 0.435; empty 48 (video-data 19)	secondary
gemma4:12b	perfect manual routing · highest Faithfulness (0.811)	empty 78 · trap 0.49 · composite collapse · needs 0.30.7 (breaks qwen)	not recommended (unstable)
granite4.1:8b	Data · Pitfalls · Unit Conversion Good	Failure to justify Korean manual (manual 38/500) · RAGS Half	Unsuitable

(5.4절)은 로컬 보유 모델 2종(gemma4:e4b, gemma4:12b)의 라우팅 측에 한해 수행하였다. 4모델 전체와 RAGAS 생성 품질 측까지 포괄하는 전면 절제 실험은 향후 과제다. 향후 동일 평가셋에서 보강 on/off를 비교하면 그 기여도를 보다 엄밀히 분리할 수 있을 것이다. 다섯째, 본 시스템의 영상 분석 통합은 실시간 원본 영상이 아니라 영상 분석 AI가 플랫폼 DB에 적재한 메타데이터(진입·진출 카운팅, 재실 추정, 자전거·킥보드·무단진입 이벤트)를 플랫폼 REST API로만 조회하는 읽기 전용 구조다(3.6절). 이 구조의 장점은 명확하다. ① 챗봇이 원본 영상-RTSP 스트림에 접근하지 않으므로 개인정보·보안 노출 표면이 최소화되고, 플랫폼 응답에 포함된 스트림 주소 등 민감 필드는 수신 즉시 폐기된다. ② 챗봇 측에 영상 처리 연산(GPU)·저장 부담이 없다. ③ 영상 분석 AI와 챗봇이 REST 계약으로 느슨하게 결합되어 장애가 격리된다(조회 타임아웃 5초, 실패 시 예외 전파 대신 ‘데이터 없음’ 명시 응답). 반면 한계도 분명하다. ① 답변 품질이 상류 영상 분석 AI의 정확도와 적재 지연에 종속되며, 챗봇이 원본 영상으로 재검증할 수 없다. ② 소비 가능한 이벤트 유형이 사전 정의된 목록(화이트리스트)에 고정되어, 새로운 유형의 이상 상황을 즉시 다루지 못한다. ③ 메타데이터 적재 시각과 조회 시각 사이의 최신성 격차가 존재한다. 향후 실제 펌프장 현장에 연동할 때 예상되는 문제와 해결 방안은 다음과 같다. (a) 플랫폼 API 장애-지연: 구현된 타임아웃·요류 객체 경로로 답변 생성 자체는 유지되되 데이터 부재를 응답에 명시하고, 플랫폼 헬스 모니터링을 병행한다. (b) 인증 토큰 만료·권한 변경: 토큰 발급·갱신 절차를 운영 문서화하고 만료를 사전 감지한다. (c) API 응답 스키마의 버전 드리프트: 인터페이스 계약 문서에 기반한 계약 테스트(contract test)를 배포 게이트에 편입하여 스키마 변경을 배포 전에 검출한다. (d) 이벤트 유형 확장(예: 침입 외 신규 이상 행동): 이벤트 화이트리스트 확장과 함께 해당 대응 매뉴얼 청크를 동시 적재하고, 결정적 보강의 트리거 어휘도 함께 갱신한다. (e) 합성 데이터로 검증한 영상 관련 70문항을 실데이터 조건에서 재측정하여 평가의 외적 타당성을 확보한다.

시스템 측면에서는 일부 카테고리(예: 가정 침수 대비)와 영상 분석 운영 지침의 코퍼스 깊이가 제한적이어서, 근거가 부족할 때 대체 응답(fallback)이 정상적으로 발생한다.

이는 시스템 결합이 아니라 코퍼스 범위의 한계이며, 코퍼스 확충을 통해 개선할 수 있다. 또한 결정적 매뉴얼 보강은 위험 및 대응 의도 어휘와 도구 호출에 기반한 휴리스틱이므로, 보다 일반적인 복합 의도까지 포괄하려면 학습형 플래너나 계획 기반(plan-based) 접근법을 검토할 필요가 있다.

2. 결론

안전이 중요한 배수펌프장 운영을 지원하는 온프레미스 RAG 안전 대응 챗봇을 대상으로, 난이도, 유형, 영상, 복합 요소를 균형 있게 반영한 500문항을 4개의 오픈 웨이트 LLM(qwen3.5:9b, gemma4:e4b, gemma4:12b, granite4.1:8b)으로 답변하게 하였으며, 총 2,000개의 답변을 RAGAS로 평가하였다.

본 연구의 핵심 발견은 복합 질의 응답이 모델 능력에만 의존할 경우 취약하며, 구조적 설계가 강건성을 결정한다는 점이다. 복합 라우팅 정확도는 모델 간 0.55~0.87의 편차를 보였고, 동일 평가셋에서 결정적 매뉴얼 보강을 켜고 끈 절제 실험(5.4절)은 보강의 기여가 평균 정확도 상승이 아니라, 실행 환경에 따라 변동하는 모델의 자발적 매뉴얼 호출 성향과 무관하게 매뉴얼 검색 누락의 잔여 사례를 항상 차단하는 결정적 보장임을 보였다. 한편 복합 질의의 성공은 라우팅(구조가 담당)과 생성 안정성(모델이 담당)의 곁으로 나타나 이 두 요소가 분리됨을 확인했다. 특히, 빈 답변과 멈춤 현상을 피하는 생성 안정성은 안전한 챗봇 품질의 숨은 1차 변수로 작용하며, 빈 답변 수가 사실상 Answer Relevancy 순위를 결정하였다(gemma4:e4b 3건 < qwen3.5:9b 48건 < gemma4:12b 78건). 반면, 검색 성능은 모델과 무관하게(Context Precision 0.51~0.70) 나타났으므로, 하이브리드 + 리랭킹 검색 파이프라인을 고정하고 라우팅 및 생성 모델을 선택하는 전략이 타당하다. 온프레미스 모델 선택에 대한 실증적 근거로, gemma4:e4b가 라우팅, 함정 회피, 데이터 함수, Answer Relevancy의 균형에서 1위를 차지하였다. gemma4:12b는 충실도와 매뉴얼 라우팅에서는 우수하나 생성 불안정(빈 답변 78건)과 과탐으로 운영에는 권장되지 않는다. 이 불안정이 실행 버전 차이가 아니라 모델 특성에서 비롯됨은 동일 버전 통제 측정으로 확

인하였다(5.4절). granite4.1:8b는 한국어 매뉴얼 근거화에 실패하여 본 도메인에 부적합하다. 끝으로, 프레임워크 없는 최소 결정적 오케스트레이션이 온프레미스 안전 중요 RAG에 적합함을 입증하였다. 본 시스템은 LangChain·LangGraph 등 에이전트 프레임워크 없이 ReAct 방식의 관찰 → 행동 루프에 결정적 안전 보강을 더한 형태로 구성된다. 안전 중요 도메인에서는 ‘더 똑똑한 에이전트’보다 감사 가능하고 결정적인 에이전트가 핵심 설계 요소이며, 절대 실험에서 확인한 ‘개입이 필요 없으면 침묵하고, 필요하면 모델과 무관하게 결정적으로 작동하는’ 보강의 성질이 그 방향을 뒷받침한다. 본 연구는 안전 중요·폐쇄망이라는 현장 제약 하에서 복합 질의 응답 구조와 다중 모델 RAGS 결과를 함께 제시하여, 유사 인프라 챗봇 설계 및 온프레미스 모델 선택에 실증적 기준을 제공한다.

참 고 문 헌 (References)

- [1] Qwen Team, Qwen3.5: Towards Native Multimodal Agents, Qwen Blog, February 2026, <https://qwen.ai/blog?id=qwen3.5> (model card: <https://huggingface.co/Qwen/Qwen3.5-9B>) (accessed Sep. 9, 2026).
- [2] Gemma Team, Google DeepMind, “Gemma 4 Technical Report,” arXiv preprint, arXiv:2607.02770, July 2026, <https://arxiv.org/abs/2607.02770>
doi: <https://doi.org/10.48550/arXiv.2607.02770>
- [3] IBM Research, Granite 4.1 Language Models, IBM, 2026, <https://huggingface.co/blog/ibm-granite/granite-4.1> (source code: <https://github.com/ibm-granite/granite-4.1-language-models>, Apache-2.0) (accessed Sep. 9, 2026).
- [4] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, “Survey of Hallucination in Natural Language Generation,” ACM Computing Surveys, Vol.55, No.12, Article 248, pp.1-38, March 2023.
doi: <https://doi.org/10.1145/3571730>
- [5] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Proceeding of Advances in Neural Information Processing Systems (NeurIPS), Online (Virtual), Vol.33, pp.9459-9474, 2020.
doi: <https://doi.org/10.48550/arXiv.2005.11401>
- [6] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, “Retrieval-Augmented Generation for Large Language Models: A Survey,” arXiv preprint, arXiv:2312.10997, 2024, <https://arxiv.org/abs/2312.10997>
doi: <https://doi.org/10.48550/arXiv.2312.10997>
- [7] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, “Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection,” Proceeding of the 12th International Conference on Learning Representations (ICLR), Vienna, Austria, 2024, <https://arxiv.org/abs/2310.11511>
doi: <https://doi.org/10.48550/arXiv.2310.11511>
- [8] T. Schick, J. Dwivedi-Yu, R. Dessi, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom, “Toolformer: Language Models Can Teach Themselves to Use Tools,” Proceeding of Advances in Neural Information Processing Systems (NeurIPS), New Orleans, USA, Vol.36, pp.68539-68551, 2023, https://proceedings.neurips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html
- [9] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafraan, K. Narasimhan, and Y. Cao, “ReAct: Synergizing Reasoning and Acting in Language Models,” Proceeding of the 11th International Conference on Learning Representations (ICLR), Kigali, Rwanda, 2023, <https://arxiv.org/abs/2210.03629>
doi: <https://doi.org/10.48550/arXiv.2210.03629>
- [10] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, “Dense Passage Retrieval for Open-Domain Question Answering,” Proceeding of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, pp.6769-6781, November 2020.
doi: <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- [11] S. Robertson and H. Zaragoza, “The Probabilistic Relevance Framework: BM25 and Beyond,” Foundations and Trends in Information Retrieval, Vol.3, No.4, pp.333-389, 2009.
doi: <https://doi.org/10.1561/15000000019>
- [12] G. V. Cormack, C. L. A. Clarke, and S. Büttcher, “Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods,” Proceeding of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, USA, pp.758-759, 2009.
doi: <https://doi.org/10.1145/1571941.1572114>
- [13] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, “M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation,” Proceeding of Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, pp.2318-2335, August 2024.
doi: <https://doi.org/10.18653/v1/2024.findings-acl.137>
- [14] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, “RAGAs: Automated Evaluation of Retrieval Augmented Generation,” Proceeding of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, St. Julians, Malta, pp.150-158, March 2024.
doi: <https://doi.org/10.18653/v1/2024.eacl-demo.16>
- [15] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena,” Proceeding of Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, New Orleans, USA,

Vol.36, pp.46595-46623, 2023, https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html

- [16] L. Wang, W. Xu, Y. Lan, Z. Hu, Y. Lan, R. K.-W. Lee, and E.-P. Lim, "Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models," Proceeding of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada, pp.2609-2634, July 2023. doi: <https://doi.org/10.18653/v1/2023.acl-long.147>

- [17] N. Shinn, F. Cassano, E. Berman, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language Agents with Verbal Reinforcement Learning," Proceeding of Advances in Neural Information Processing Systems (NeurIPS), New Orleans, USA, Vol.36, pp.8634-8652, 2023. doi: <https://doi.org/10.48550/arXiv.2303.11366>
- [18] H. Chase, LangChain, software, 2022, <https://github.com/langchain-ai/langchain> (accessed Sep. 9, 2026).
- [19] LangChain Inc., LangGraph, software, 2024, <https://github.com/langchain-ai/langgraph> (accessed Sep. 9, 2026).

저 자 소 개



장 일 식

- 2011년 2월 : 서울과학기술대학교 NID융합기술대학원 석사
- 2020년 3월 ~ 현재 : 서울과학기술대학교 IoT융복합기술연구소 책임연구원
- 2026년 2월 : 서울과학기술대학교 나노IT디자인융합대학원 정보통신미디어공학전공 박사
- ORCID : <https://orcid.org/0000-0003-0822-9857>
- 주관심분야 : 컴퓨터비전, 생성형 AI



박 구 만

- 1984년 2월 : 한국항공대학교 전자공학과 공학사
- 1986년 2월 : 연세대학교 전자공학과 공학석사
- 1991년 2월 : 연세대학교 전자공학과 공학박사
- 1991년 3월 ~ 1996년 9월 : 삼성전자 신호처리연구소 선임연구원
- 2016년 1월 ~ 2017년 12월 : 서울과학기술대학교 나노IT디자인융합대학원 원장
- 1999년 8월 ~ 현재 : 서울과학기술대학교 스마트ICT융합공학과 교수
- 2006년 1월 ~ 2007년 8월 : Georgia Institute of Technology Dept.of Electrical and Computer Engineering, Visiting Scholar
- ORCID : <https://orcid.org/0000-0002-7055-5568>
- 주관심분야 : 컴퓨터비전, 지능형실감미디어