Journal of KIIT. Vol. 24, No. 6, pp. 99-108, Jun. 30, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.6.99

스마트팜 RAG를 위한 질의 적응형 검색 채널 제어

김유석*[1], 김용운*[2], 변영철*[3]

Query-Adaptive Multi-Channel Retrieval Control for Smart Farming RAG

Yu-Seok Kim*[1], Yong-Woon Kim*[2], and Yung-Cheol Byun*[3]

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2024-00405278). 본 과제(결과물)는 2026년도 교육부 및 제주특별자치도의 재원으로 제주RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다(2026-RISE-17-001)

요 약

스마트팜 검색 증강 생성(RAG) 시스템에서는 작물 이상, 병해 원인, 환경 제어처럼 구조적 관계를 고려해야 하는 질의가 증가하고 있다. 그러나 의미 기반 및 키워드 기반 검색만으로는 이러한 관계를 충분히 반영하기 어렵고, 그래프 기반 검색을 단순 결합하면 단순 질의에서 검색 순위가 저하될 수 있다. 본 논문은 질의 유형별 프로파일을 조회하여 의미 기반, 키워드 기반, 그래프 기반 검색의 가중치를 조절하는 질의 적응형 다중 검색 채널 제어 구조를 제안한다. 실험 결과, 제안 구조는 농업 질의응답 데이터에서 고정 3채널 결합으로 인한 검색 순위 저하를 완화하고, 다중 홉 질의응답 데이터에서 검색 순위 품질 개선 가능성을 확인하였다. 또한 엣지 참조 환경에서 실행 시점 LLM 기반 가중치 산정보다 검색 지연 시간을 줄였다.

Abstract

Smart farming Retrieval-Augmented Generation (RAG) systems increasingly require queries that link crop abnormalities, disease causes, and environmental control actions. Semantic and keyword retrieval alone cannot fully capture these structural relations, whereas naive graph retrieval may degrade rankings for simple queries. This paper proposes a query-adaptive multi-channel retrieval control architecture that selects semantic, keyword, and graph weights from query-type profiles. Experiments show that the proposed architecture mitigates ranking degradation from fixed three-channel fusion on agricultural QA data, indicates potential ranking gains on multi-hop QA data, and reduces retrieval latency over runtime LLM-based weight estimation in an edge reference environment.

Keywords

retrieval-augmented generation, smart farming, hybrid retrieval, graph retrieval, query-adaptive control, edge computing

* 제주대학교 컴퓨터공학과(*[3] 교신저자)
- ORCID[1]: https://orcid.org/0009-0009-3065-7656
- ORCID[2]: https://orcid.org/0000-0002-4759-0138
- ORCID[3]: https://orcid.org/0000-0002-1579-5323

· Received: May 08, 2026, Revised: May 27, 2026, Accepted: May 30, 2026
· Corresponding Author: Yung-Cheol Byun
Dept. of Computer Engineering Jeju National University, 102 Jejudaehak-ro, Jeju-si, Jeju-do, Korea
Tel.: +82-064-754-3657, Email: ycb@jejunu.ac.kr

Ⅰ. 서 론

스마트팜은 온도, 습도, 이산화탄소, 양액 농도, 조도와 같은 환경 데이터를 기반으로 작물 생육 상태를 관리하고 생산성을 높이는 농업 정보화 기술이다. 시설 재배와 식물공장 환경에서는 센서와 제어 장치가 확산되면서 데이터 수집은 빠르게 자동화되고 있다. 그러나 현장의 의사결정은 단순 수치 확인에 그치지 않는다. 잎의 황화, 생장 지연, 뿌리 활력 저하와 같은 증상이 발생하면 농가는 현재 환경 수치, 과거 작업 기록, 품종 특성, 병해충 정보, 재배 매뉴얼을 함께 확인해야 한다. 따라서 스마트팜 현장 질의는 단순 사실 검색이 아니라 원인, 증상, 조치 사이의 관계를 해석하는 지식 기반 의사결정 문제에 가깝다.

대형 언어 모델은 자연어 질의에 대해 설명형 답변을 제공할 수 있어 스마트팜 의사결정 지원에 활용될 가능성이 크다. 그러나 모델 내부 지식만으로는 개별 농가의 최신 작업 기록이나 지역별 재배 조건을 즉시 반영하기 어렵다. 또한 경량화된 현장 실행 환경에서는 복잡한 원인 분석 질의에서 환각 답변이 발생할 수 있다. 이러한 이유로 본 논문은 외부 지식 검색 결과를 답변 생성에 함께 사용하는 검색 증강 생성(RAG) 방식에 주목한다[1][2].

RAG 시스템의 품질은 검색 단계에 상당 부분 의존한다. 의미 기반 검색은 문장 의미가 유사한 문서를 찾는 데 강점을 가지며[3], 키워드 기반 검색은 특정 용어, 수치, 품종명, 병해명처럼 정확한 단어 일치가 필요한 경우에 유리하다[4]. 그러나 스마트팜 현장에서 자주 발생하는 원인 추론형 질의는 의미 유사도나 키워드 일치만으로 충분하지 않다. 예를 들어 "잎이 노랗게 변하고 뿌리 활력이 낮을 때 가능한 원인과 조치는 무엇인가"와 같은 질의는 증상, 원인, 환경 조건, 조치가 여러 문서에 분산되어 있을 수 있으며, 이들 사이의 관계를 따라가야 한다.

표 1은 스마트팜 검색 질의가 단순 사실 확인부터 관계 추론까지 다양한 형태로 나타남을 보여준다. 단순 사실형 질의는 작물명, 수치, 관리 기준과 같은 명시적 단서가 중요하므로 의미 기반 검색과 키워드 기반 검색이 우선된다. 반면 병해·원인 분석형 질의는 증상과 환경 조건 사이의 연결을 함께 확인해야 하므로 그래프 채널이 보조 근거를 제공할 수 있다. 따라서 모든 질의에 동일한 검색 채널 조합을 적용하기보다, 질의의 목적과 관계 추론 필요성에 따라 채널 비중을 조절하는 것이 실용적이다.

그래프 기반 관계 검색은 이러한 구조적 관계를 명시적으로 표현할 수 있다는 점에서 스마트팜 RAG 시스템의 중요한 보완 채널이다. 그러나 그래프 검색 채널을 단순히 추가한다고 해서 항상 성능이 향상되는 것은 아니다. 예를 들어 "딸기 재배의 적정 pH 범위는 얼마인가"라는 질의는 수치와 작물명이 핵심인 단순 질의이다. 이때 그래프 검색이 주변 병해 노드, 인접 양분 관계, 과거 관수 이력 노드를 함께 가져오면 상위 순위에 노이즈가 섞일 수 있다.

표 1. 스마트팜 질의 유형과 선호 검색 채널 예시
Table 1. Smart farming query types and preferred retrieval channels

Query type	Example query	Preferred retrieval channels
Simple fact	"What is the appropriate EC range for tomato nutrient solution?"	Semantic + keyword
Environment control	"How can nighttime humidity be reduced in a strawberry greenhouse during hot periods?"	Semantic + keyword
Disease and cause analysis	"What causes cucumber leaf yellowing and reduced root vigor to appear together?"	Semantic + graph
Relational reasoning	"How are calcium deficiency, rapid water fluctuation, and blossom-end rot related?"	Semantic + graph
Cultivation management	"How should irrigation and nutrient solution management be adjusted two weeks after tomato transplanting?"	Semantic + keyword + graph

특히 단순 질의에서는 정답 후보가 소수의 매뉴얼 문서에 집중되는 경우가 많기 때문에, 관계 이웃을 과도하게 확장하면 기존 상위 문서의 상대 순위가 낮아질 수 있다.

본 논문은 이 문제를 해결하기 위해 의미 기반, 키워드 기반, 그래프 기반 검색을 병렬로 수행하되, 질의 유형별 프로파일을 조회하여 검색 채널 가중치를 결정하는 질의 적응형 다중 검색 채널 제어 구조를 제안한다. 특히 가중치 산정을 위한 실행 시점 LLM 호출에 의존하지 않고 오프라인 프로파일링 기반 질의 가중치 선택 구조를 적용함으로써, 제한된 엣지 컴퓨팅 환경에서 검색 품질과 응답 지연 시간 간의 균형을 확보하고자 하였다.

본 논문의 기여는 다음과 같다. 첫째, 스마트팜 현장 질의를 구조적 관계 필요성에 따라 구분하고 세 검색 채널의 역할을 정리하였다. 둘째, 오프라인 검증으로 생성한 질의 유형별 가중치 프로파일을 실행 시점에 조회하는 검색 채널 제어 구조를 제안하였다. 셋째, 공개 질의응답 데이터를 활용하여 균등 3채널 결합 대비 순위 저하 완화와 조건부 그래프 활용 가능성을 확인하였다. 넷째, 엣지 참조 환경에서 프로파일 조회 방식이 실행 시점 LLM 기반 가중치 산정보다 검색 지연 시간을 줄일 수 있음을 보였다.

II. 관련 연구

2.1 RAG와 하이브리드 검색

RAG는 외부 지식 검색과 언어 모델 생성을 결합하여 모델의 환각을 줄이고 근거 기반 답변을 제공하는 접근이다[1]. 최근 연구는 단순 검색 후 생성 구조를 넘어 검색기 구성, 질의 변환, 재순위화, 모듈형 파이프라인을 포함하는 다양한 구조로 확장되고 있다[2]. 이러한 연구는 RAG가 도메인 지식 갱신과 사실성 보완에 유용함을 보이지만, 대부분 의미 기반 검색 또는 2채널 검색을 중심으로 한다.

하이브리드 검색은 의미 기반 검색과 키워드 기반 검색을 결합하여 단일 검색 방식의 한계를 보완한다. 밀집 검색은 의미 유사도에 강하지만 수치와 고유명사에 약하고, 희소 검색은 정확한 용어 일치에 강하지만 표현 변형에 약하다[3][4]. 상호 순위 융합과 같은 순위 기반 결합 방식은 서로 다른 점수 체계를 단순하게 통합할 수 있어 실용적이다[5]. 다만 기존 하이브리드 검색은 문서 간 구조적 관계를 명시적으로 추적하기 어렵다는 한계가 있다.

2.2 그래프 기반 RAG와 스마트팜 적용

그래프 기반 RAG는 문서 간 개체와 관계를 명시적으로 표현하여 다중 단계 질의에 필요한 구조적 관계 정보를 제공한다. 선행 연구는 그래프 구조를 활용한 질의 중심 요약과 관계 경로 선택이 그래프 기반 검색 품질에 중요함을 보였다[6][7]. 그러나 그래프 기반 검색은 다중 단계 질의에서는 도움이 될 수 있으나, 단순 질의에서는 불필요한 구조적 관계 정보가 검색 순위에 영향을 줄 수 있으므로 그래프 채널의 활용 여부와 비중을 질의 조건에 따라 조절할 필요가 있다.

스마트팜 연구는 플랫폼 설계, 생산 최적화, 시스템 구현과 같이 현장 적용 가능한 시스템 구조를 중요하게 다루어 왔다[8]. 또한 최근 하이브리드 검색에서는 DAT(Dynamic Alpha Tuning)과 같이 질의 입력마다 검색 채널 가중치를 동적으로 산정하는 방식이 제안되었으나, 실행 시점 LLM 호출은 엣지 환경에서 추가 지연을 유발할 수 있다[9]. 따라서 스마트팜 질의에서 그래프 채널을 언제 억제하고 언제 활용할지, 그리고 이를 엣지 환경에서 어떻게 가볍게 적용할지에 대한 질의 적응형 검색 제어 구조가 필요하다.

III. 시스템 설계 및 구현

3.1 질의 적응형 검색기 구조

그림 1은 질의 적응형 검색 채널 제어 구조의 오프라인-온라인 처리 흐름을 나타낸다. 제안 시스템은 그림 1의 ①-⑥과 같이 오프라인 프로파일 생성 단계와 실행 시점 검색 제어 단계로 구성된다. 오프라인 단계에서는 ① 스마트팜 문서, 재배 매뉴얼, 연구 논문, 작업 기록과 같은 도메인 문서를 처리하고, ② 문서 조각, 벡터 인덱스, 키워드 인덱스, 지식 그래프를 구성한다.

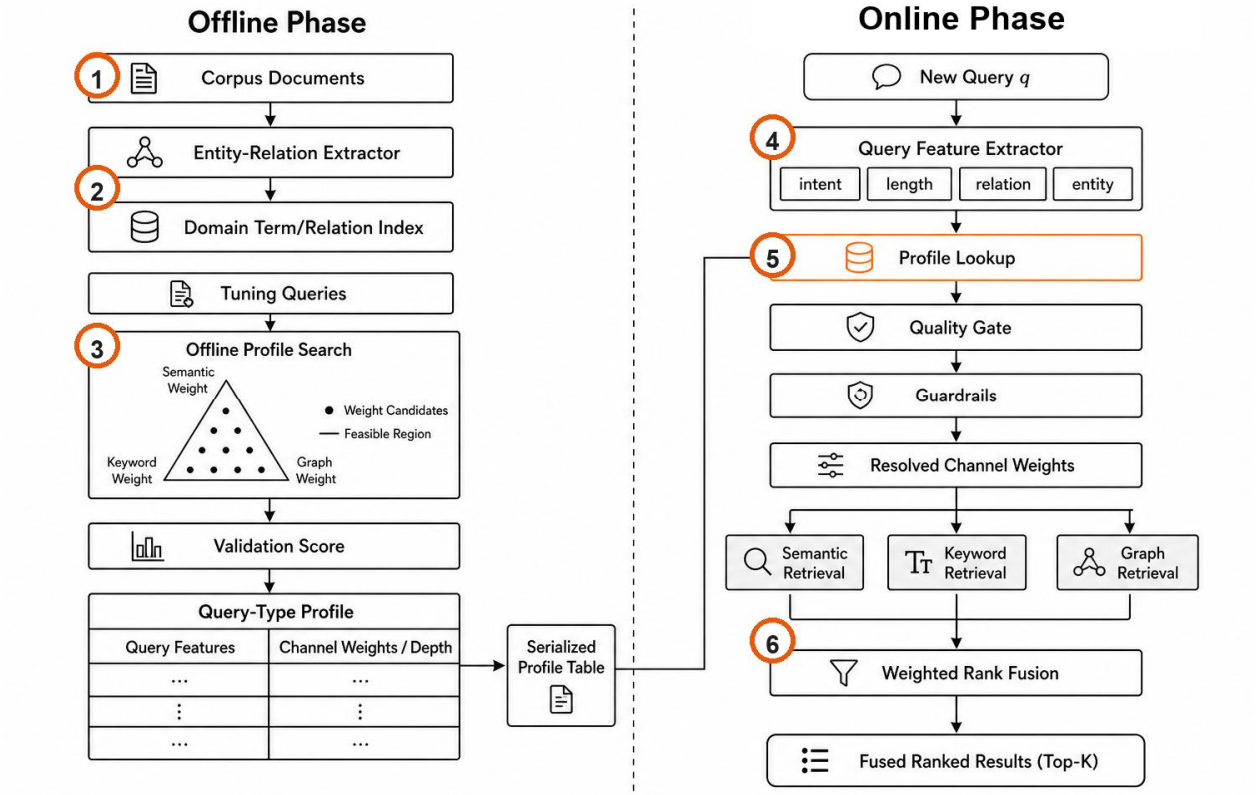

그림 1. 질의 적응형 검색 채널 제어 구조의 오프라인-온라인 처리 흐름
Fig. 1. Offline-online workflow of query-adaptive retrieval channel control

문서 조각은 답변 생성에 직접 사용되는 근거 단위이며, 그래프는 범주, 상태·조건, 환경, 관리 작업과 같은 도메인 개체를 노드로 구성하고 문서 조각과 개체 사이의 언급 관계 및 개체 간 관계를 엣지로 연결한 보조 색인이다. 그래프 검색은 질의 토큰과 일치하는 개체 주변 이웃을 찾는 저수준 검색과 1-3홉 경로를 탐색하는 고수준 검색을 함께 사용하며, 경로 신뢰도와 홉 길이를 반영해 후보 문서 점수를 계산한다.

동시에 ③ 튜닝 질의를 이용하여 질의 유형별 검색 채널 가중치와 검색 깊이를 탐색하고, 그 결과를 작은 프로파일 테이블로 저장한다. 오프라인 프로파일링 단계에서는 LLM 분석과 벤치마크 결과를 이용해 튜닝 질의를 세그먼트로 나누고, 각 세그먼트별 후보 채널 가중치와 후보 문서 깊이를 평가하여 검색 가중치를 사전 구성한다. 프로파일은 기본 가중치, 세그먼트별 매칭 조건, 채널 가중치, 후보 깊이, 가드레일, 품질 메타데이터를 저장한 경량 JSON 테이블이다. 이 방식은 실행 시점에 매 질의마다 가중치 산정 모델을 호출하는 부담을 줄이기 위한 설계이다.

실행 시점에는 사용자 질의가 입력되면 ④ 경량 질의 특성 추출기를 통해 질의 유형을 판별하고, ⑤ 해당 프로파일을 조회하여 의미 기반 검색, 키워드 기반 검색, 그래프 기반 관계 검색의 가중치를 결정한다. 질의 특성 벡터는 x(q)=[intent, query_length_bucket, entity_hint_presence, relation_hint]로 구성하며, 각 항목은 질의 목적, 질의 길이 구간, 숫자 기반 개체 단서, 도메인 관계 단서 포함 여부를 나타낸다. 이후 ⑥ 세 검색 채널의 결과를 가중 순위 결합 방식으로 통합하여 답변 생성 모델에 전달할 근거 문서를 선정한다.

표 2는 질의 유형별 특성 벡터와 검색 채널 가중치, 검색 깊이의 프로파일 예시를 나타낸다. 최종 반환 문서 수는 모든 프로파일에서 평가 지표와 동일하게 Top-10으로 고정하였다.

알고리즘 1은 오프라인 프로파일 생성과 실행 시점 검색 제어의 전체 절차를 나타낸다.

표 2. 질의 유형별 프로파일 예시
Table 2. Example of query-type retrieval profiles

Query type	Feature vector example	Dense	Sparse	Graph	Retrieval depth
Simple fact	[fact, short, true, false]	0.50	0.40	0.10	10
Disease and cause analysis	[cause_analysis, medium, false, true]	0.45	0.15	0.40	15
Cultivation management	[management, medium, true, true]	0.45	0.35	0.20	15

알고리즘 1. 오프라인 프로파일 생성 및 실행 시점 검색 제어
Algorithm 1. Offline profile generation and runtime retrieval control

```
Input : Domain documents D, tuning queries Q_t, runtime query q
Output: Fused top-K documents R_K
1: I_dense, I_sparse, G <- BuildIndexes(D)
2: for each q_i in Q_t do
3:       x_i <- OfflineLLMFeatureAnalysis(q_i)
4:       for each candidate pair (w, h) in Ω do
5:            m_i(w, h) <- CrossValidatedNDCG@10(q_i, I_dense, I_sparse, G)
6:       end for
7: end for
8: Profile <- SelectBySegment({x_i}, {m_i}, variance)
9: x <- RuntimeFeatureExtract(q)
10: w, h <- Lookup(Profile, g(x)); w <- ApplyGateAndGuardrails(w)
11: R_dense, R_sparse, R_graph <- Retrieve(q, I_dense, I_sparse, G, h)
12: R_K <- WeightedRankFusion({R_dense, R_sparse, R_graph}, w, K)
```

3.2 질의 적응형 채널 제어

본 논문에서 제안하는 질의 적응형 채널 제어는 세 검색 채널의 순위 정보를 가중 결합하는 구조이다. 각 채널에서 문서 d가 질의 q에 대해 가지는 순위를 $r_c(d,q)$라고 할 때, 최종 점수는 식 (1)과 같이 정의한다.

$$S(d|q) = \sum_{c \in C} \frac{w_c(q)}{k + r_c(d,q)} \quad (1)$$

여기서 S(d|q)는 질의 q에 대한 문서 d의 최종 결합 점수이고, C는 의미 기반 검색, 키워드 기반 검색, 그래프 기반 관계 검색의 채널 집합이다. 참조 구현에서는 이를 각각 vector_dense_text, vector_sparse, graph 채널로 매핑한다. $r_c(d,q)$는 채널 c에서 문서 d가 가지는 순위이고, $w_c(q)$는 프로파일에서 조회한 채널 가중치로서 $\sum_{c \in C} w_c(q) = 1$을 만족한다. k는 상위 순위 문서에 대한 과도한 점수 집중을 완화하기 위한 상수이다. 특정 문서가 어떤 채널의 상위 결과에 포함되지 않으면 해당 채널의 기여도는 0으로 처리한다. 이 수식은 채널별 원점수 보정 없이 순위 정보만으로 결과를 결합할 수 있으며, 질의 유형에 따라 채널별 기여도를 조정할 수 있도록 확장한 것이다.

각 질의 세그먼트에 대해서는 튜닝 질의 집합에서 후보 가중치와 후보 검색 깊이를 탐색한다. 여기서 검색 깊이는 채널별 후보 문서 수와 그래프 채널의 제한 홉 범위를 포함한다. 5겹 교차 검증으로 nDCG@10 평균과 순위 변동성을 함께 고려하여 최적 프로파일을 선택한다. 이때 한 질의에만 맞는 값을 선택하지 않고, 같은 세그먼트에 속한 질의들의 평균 성능과 변동성을 함께 확인한다. 선택된 프로파일은 정적 파일로 저장되고, 실행 시점에는 특성 벡터에서 만든 키를 해시 테이블로 조회한다. 즉, LLM 기반 질의 분석과 가중치 탐색은 오프라인 단계에서 수행되고, 실행 시점에는 사전 프로파일링된 가중치를 참조하는 경량 구조로 동작한다.

단순 수치 확인 질의에서는 의미 기반 및 키워드 기반 검색의 비중을 높이고 그래프 채널은 억제한다. 반대로 원인, 증상, 조치 사이의 구조적 관계 단서가 강한 질의에서는 그래프 채널을 활성화하여 구조적 관계 정보를 보완한다.

제안 구조는 운영 적용 시 고려할 수 있는 세 가지 안전 장치를 포함한다. 첫째, 채널 가중치에는 최소값, 최대값, 연속 질의 간 변화율 제한을 둔다. 둘째, 프로파일의 표본 수와 유효 기간을 확인하는 품질 게이트를 둔다. 셋째, 특정 채널이 유효한 검색 결과를 반환하지 못하면 해당 채널의 가중치를 최소값 또는 0으로 설정한 뒤 나머지 채널 가중치를 재정규화한다.

3.3 구현 환경

참조 구현 측면에서 오프라인 지식 구축은 문서 파싱, 문서 청크 구성, 도메인 용어 기반 개체-관계 추출로 구성된다. 검색 저장소는 벡터 검색과 그래프 저장소를 함께 사용하고, 엣지 실행 환경에서는 경량 추론 엔진 기반 양자화 언어 모델을 사용한다[10]. 농가별 작업 기록과 생육 데이터는 민감한 지식일 수 있으므로, 참조 구현에서는 공개 지식과 농가별 사유 지식을 논리적으로 분리하는 저장 구조를 고려하였다.

Ⅳ. 실험 결과 및 성능 평가

4.1 실험 설정

제안 구조의 검색 성능을 평가하기 위해 세 가지 공개 질의응답 데이터를 사용하였다. AgXQA는 농업 영역 질의응답 데이터로, 단일 홉 성격의 질의에서 그래프 채널을 항상 추가할 때 발생할 수 있는 순위 저하와 그래프 채널 억제 효과를 확인하는 데 사용하였다[11]. MuSiQue와 2WikiMHQA는 여러 근거를 연결해야 하는 다중 홉 관계 질의를 포함하므로, 구조적 관계 탐색이 필요한 질의에서 그래프 채널의 조건부 활용 효과를 확인하는 데 사용하였다[12][13]. 이와 같이 데이터셋별 역할을 구분하여 단순 농업 질의와 다중 단계 관계 질의에서 제안 구조의 동작을 함께 평가하였다. 각 데이터셋에서는 모든 비교군에 동일한 근거 문서 집합, 의미 기반 색인, 키워드 기반 색인, 그래프 기반 관계 색인을 사용하였다. 그래프 기반 관계 색인은 문서 제목, 질의 내 핵심 명사구, 명명 개체 표현과 같은 텍스트 단서를 중심으로 개체와 관계 표현을 추출하고, 질의에 포함된 주요 개체를 시작점으로 동일한 제한 홉 조건에서 관련 문서를 탐색하는 방식으로 사용하였다.

비교 대상은 세 가지 검색 구조로 설정하였다. Baseline은 의미 기반 검색과 키워드 기반 검색만 결합한 2채널 구조로, 그래프 채널을 사용하지 않는다. Uniform은 의미 기반, 키워드 기반, 그래프 기반 검색을 1:1:1로 결합한 고정 3채널 구조이다. 이 비교군은 그래프 채널을 항상 같은 비중으로 추가했을 때 단순 질의에서 검색 순위가 저하될 수 있는지를 확인하기 위해 사용하였다. Ours는 질의 유형별 프로파일을 조회하여 세 검색 채널의 가중치를 조절하는 제안 구조이다. Uniform과 Ours는 동일한 그래프 탐색 후보 조건을 사용하며, 차이는 채널 가중치 적용 방식에만 두었다. 단순 질의에서는 그래프 채널을 억제하고, 다중 단계 질의에서는 구조적 관계 정보를 제한적으로 활용하도록 설정하였다. 질의 유형별 프로파일은 오프라인 LLM 분석과 표 1에서 정리한 질의 특성을 함께 사용하여 구성하였다.

가중치 산정 조건도 동일하게 통제하였다. Baseline은 dense/sparse만 사용하고 graph 가중치를 0으로 두었으며, Uniform은 세 채널을 1/3로 고정하였다. Ours는 x(q) 기반 프로파일을 조회하고 무효 채널 보정 후 재정규화하였다. Runtime LLM Baseline은 Qwen3-4B-Q4_K_M GGUF에 질의, 길이, 작물·병해·수치, 원인·비교·동반 증상 단서를 입력하여 ‘dense’, ‘sparse’, ‘graph’ JSON 가중치를 얻고, 결측·음수 제거 후 합이 1이 되도록 정규화하였다.

검색 성능은 정규화 할인 누적 이득(nDCG@10), 평균 역순위, 상위 10개 검색 회수율로 평가하였으며, 상위 검색 결과의 순위 품질을 반영하는 nDCG@10을 주 지표로 해석하였다. 공개 데이터셋별 질의 중 20%는 가중치 프로파일 튜닝에 사용하고, 나머지 80%는 최종 평가에 사용하여 튜닝 질의

와 평가 질의가 중복되지 않도록 분리하였다. 20% 튜닝 질의 내부에서는 5겹 교차 검증으로 후보 가중치와 검색 깊이를 선택하였다.

본 논문은 검색 채널 제어 구조가 검색 순위 품질, WASSABI 스마트팜 코퍼스 기반 생성 품질 보조 지표, 엣지 실행 성능에 미치는 영향을 함께 평가하였다. WASSABI 평가는 626개 문서와 11개 유형 54개 질의로 구성하고, 유형별 4-5개 질의를 균형 배치하였다. 생성 품질 보조 평가는 동일 검색 근거와 동일 Qwen3-4B-Q4_K_M GGUF 답변 생성 조건에서 RAGAS Faithfulness를 10회 측정한 평균으로 보고하였으며, 평가모델은 Ollama 기반 gpt-oss:120b로 고정하였다. 질의별 통계 검정은 향후 확장 검증 항목으로 남겨 두었다.

엣지 실행 성능 비교에는 Dynamic Alpha Tuning(DAT) 계열의 실행 시점 LLM 기반 가중치 산정 관점[9]을 단순화한 구조를 추가 비교 대상으로 사용하였다. 해당 방식은 질의 입력마다 Qwen3-4B-Q4_K_M GGUF 모델을 호출하여 dense, sparse, graph 채널 가중치를 계산하는 구조이다. 반면 제안 구조는 오프라인에서 생성한 프로파일을 실행 시점에 조회하므로, 가중치 산정을 위한 추가 모델 호출을 제거한다.

엣지 참조 환경은 NVIDIA Jetson AGX Orin Developer Kit로 구성하였다. 운영체제는 Ubuntu 22.04.5 LTS와 JetPack 6.2.1/L4T 36.4.7을 사용하였고, 서빙 스택은 FastAPI 기반 API 서버, CUDA 기반 llama.cpp 추론 엔진, 벡터 검색 저장소, 그래프 저장소로 구성하였다[10]. 답변 생성에는 Qwen3-4B-Q4_K_M GGUF 양자화 모델을 사용하였다. 지연 시간은 동일 질의 집합에 대해 1회 예열 실행 후 5회 반복 측정한 결과를 집계하였으며, 전체 요청 지연 시간, 검색 단계 평균 지연 시간, 생성 단계 중앙값 지연 시간, 초당 질의 처리량을 비교하였다.

4.2 검색 성능

표 3은 세 공개 데이터에서 Baseline, Uniform, Ours의 검색 성능을 비교한 결과이다. AgXQA에서는 Baseline의 nDCG@10이 0.900이었으나, 그래프 채널을 균등하게 추가한 Uniform은 0.721로 낮아졌다. 이는 단일 단계 농업 질의에서 그래프 채널이 항상 같은 비중으로 필요한 것은 아니며, 관련성이 낮은 구조적 관계 정보가 순위를 흐릴 수 있음을 보여준다. Ours는 nDCG@10을 0.809까지 회복하여 Uniform 대비 0.088의 개선을 보였다. 다만 AgXQA는 단일 홉 농업 질의가 많아 의미 기반 및 키워드 기반 2채널만으로 정답 문서가 상위에 집중되는 경우가 있으므로, 그래프 채널의 보조 정보가 항상 Baseline 이상의 순위 개선으로 이어지지는 않았다. 이 결과는 농업 질의응답 데이터에서 질의 유형별 채널 제어가 고정 3채널 결합으로 인한 순위 저하를 완화할 수 있음을 보여준다.

다중 단계 질의응답 데이터에서는 Ours가 Baseline보다 높은 nDCG@10을 보였다. MuSiQue에서 Ours는 nDCG@10 0.614를 기록하여 Baseline의 0.568보다 0.046 높았다. 2WikiMHQA에서도 Ours는 0.788로, Baseline의 0.771보다 0.017 높았다.

표 3. 공개 데이터셋에 대한 검색 성능 비교
Table 3. Retrieval performance comparison on public datasets

Dataset	Method	nDCG@10	MRR	Recall@10
AgXQA	Baseline	0.900	0.872	0.981
	Uniform	0.721	0.655	0.929
	Ours	**0.809**	**0.751**	**0.981**
MuSiQue	Baseline	0.568	0.747	0.617
	Uniform	0.512	0.664	0.574
	Ours	**0.614**	**0.789**	**0.669**
2WikiMHQA	Baseline	0.771	0.951	0.781
	Uniform	0.700	0.849	0.750
	Ours	**0.788**	**0.955**	**0.805**

두 데이터 모두 여러 근거를 연결해야 하는 질의를 포함하므로, 그래프 기반 관계 검색이 제한적으로 활용될 때 구조적 관계 정보가 검색 순위 개선에 기여할 수 있음을 보여준다.

따라서 본 실험 결과는 질의 유형에 따라 그래프 검색 채널의 비중을 조절할 때 단순 질의의 검색 순위 저하를 완화하면서, 다중 단계 질의에서는 구조적 관계 정보를 활용할 수 있음을 뒷받침한다.

실제 스마트팜 적용 도메인에서의 근거 충실성을 보조적으로 확인하기 위해 WASSABI 코퍼스에 대한 RAGAS 평가를 수행하였다. 답변 생성에는 검색 근거와 질의만 제공하였고, 평가모델에는 RAGAS Faithfulness 설정에 따라 질의, 생성 답변, 검색 근거를 입력하였다.

표 4는 WASSABI 기반 보조평가에서 graph 채널 제어가 생성 근거 충실성에 미치는 영향을 확인한 결과이다. 10회 반복 평가의 평균 기준으로 Ours는 Faithfulness 0.972을 기록하여 Uniform의 0.902보다 높게 나타났으며, 이는 고정 3채널 결합보다 질의 특성에 따른 채널 가중치 조정이 현장형 스마트팜 질의의 근거 기반 답변 유지에 유리할 가능성을 시사한다. Baseline은 graph 채널을 사용하지 않는 검색 비교군이므로 graph 채널 제어 효과를 비교하는 표 4에서는 제외하였고, 해당 결과는 주 검색 성능 비교인 표 3에서 보고하였다. 다만 WASSABI 평가는 자체 구축 코퍼스 기반의 보조 검증이므로, 본 논문의 주 실험 결과를 대체하기보다는 스마트팜 도메인 적합성에 대한 추가 근거로 해석하였다.

표 4. WASSABI 코퍼스 기반 도메인 적합성 보조 평가
Table 4. Domain-fit auxiliary evaluation on the WASSABI corpus

Method	Faithfulness
Uniform	0.902
Ours	0.972

4.3 엣지 실행 성능

제안 구조의 엣지 실행 가능성을 확인하기 위해 앞서 정의한 Jetson AGX Orin 기반 참조 환경에서 Runtime LLM Baseline과 Ours를 비교하였다. Runtime LLM Baseline은 Dynamic Alpha Tuning(DAT) 계열의 실행 시점 LLM 기반 가중치 산정 관점을 단순화한 구조로, 질의마다 모델을 호출하여 검색 채널 가중치를 계산한다. Ours는 사전에 생성한 프로파일을 조회하여 가중치를 적용하는 구조이다. 두 방식은 동일한 검색 저장소, 동일한 생성 모델, 동일한 API 서빙 환경, 동일한 질의 집합에서 실행하였고, 차이는 검색 채널 가중치 산정 방식에만 두었다. 표 5는 엣지 참조 환경에서 Runtime LLM Baseline과 Ours의 실행 성능을 비교한 결과이다. Ours는 전체 지연 시간 중앙값을 11502.89밀리초에서 8545.77밀리초로 줄였으며, 검색 평균 지연 시간은 3821.75밀리초에서 957.71밀리초로 감소하였다. 이는 검색 단계에서 약 74.9%의 지연 시간 감소에 해당한다.

표 5. 엣지 참조 환경에서의 실행 성능 비교
Table 5. Runtime performance comparison in the edge reference environment

Metric	Runtime LLM Baseline	Ours
End-to-end latency p50 (ms)	11502.89	**8545.77**
End-to-end latency p95 (ms)	11585.87	**8682.35**
Mean retrieval latency (ms)	3821.75	**957.71**
Generation latency p50 (ms)	7636.51	7534.97
QPS	0.087	0.117

생성 지연 시간은 두 방식이 거의 유사하였다. 따라서 지연 시간 개선은 답변 생성 모델 자체의 변화가 아니라, 검색 단계에서 실행 시점 LLM 호출을 제거한 효과로 해석할 수 있다. 표 3의 검색 순위 품질 결과와 표 4의 도메인 보조 평가 결과를 함께 볼 때, 제안 구조는 Runtime LLM 방식의 실행 시점 가중치 산정을 오프라인 프로파일 조회로 대체하면서도 질의 유형별 적응성을 유지한다. 이는 엣지 환경에서 가중치 산정에 따른 추가 지연을 줄일 수 있음을 보여준다.

V. 결론 및 향후 과제

본 논문에서는 스마트팜 RAG를 위한 질의 적응형 다중 검색 채널 제어 구조를 제안하였다. 제안 구조는 의미 기반 검색, 키워드 기반 검색, 그래프 기반 관계 검색을 병렬로 구성하고, 질의 유형별 가중치 프로파일을 실행 시점에 조회하여 최종 검색 순위를 구성한다. 이를 통해 그래프 검색 채널을 무조건 강화하는 방식이 아니라, 질의 조건에 따라 억제하거나 제한적으로 활용하는 구조를 구현하였다.

실험 결과, AgXQA에서 Uniform은 Baseline의 nDCG@10 0.900보다 낮은 0.721을 기록하였으나, Ours는 0.809까지 회복하였다. WASSABI 코퍼스 기반 RAGAS 보조 평가에서도 Ours는 Uniform 대비 Faithfulness 개선 가능성을 보였다. MuSiQue와 2WikiMHQA에서는 Ours가 Baseline 대비 각각 0.046 및 0.017 높은 nDCG@10을 보여, 다중 단계 질의에서의 조건부 그래프 활용 가능성을 확인하였다. 또한 엣지 장치 기반 참조 환경에서는 Ours가 실행 시점 LLM 기반 Baseline보다 검색 평균 지연 시간을 줄였다. 이러한 결과는 제안 구조가 스마트팜 환경에서 그래프 검색 채널을 질의 조건에 따라 운용하기 위한 실용적 제어 방식이 될 수 있으며, 제한된 연산 자원을 갖는 스마트팜 환경에서의 적용 가능성을 보여준다.

다만 본 연구의 결과는 3채널 검색이 모든 조건에서 더 낫다는 의미는 아니다. 이는 그래프 채널의 가치를 낮추는 것이 아니라, 그래프 기반 관계 검색을 질의 조건에 맞게 선택적으로 활용해야 함을 의미한다. 또한 WASSABI 결과는 자체 구축 코퍼스 기반 보조 평가이므로, 질의별 통계적 유의성 검정과 다중 시드 생성 품질 평가는 추가 검증이 필요하다. 향후 연구에서는 국내 스마트팜 현장 질의와 재배 작물별 문서를 확대하여 질의 유형별 가중치 프로파일의 일반성을 검증하고, 다양한 엣지 장치와 장기간 운영 환경에서 지연 시간과 검색 순위 품질을 함께 검증할 계획이다.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W. Yih, T. Rocktaschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", Advances in Neural Information Processing Systems, Online, Vol. 33, pp. 9459-9474, Dec. 2020.

[2] 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, pp. 1-21, Mar. 2024. https://doi.org/10.48550/arXiv.2312.10997.

[3] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense Passage Retrieval for Open-Domain Question Answering", Proc. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, pp. 6769-6781, Nov. 2020. https://doi.org/10.18653/v1/2020.emnlp-main.550.

[4] 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, Apr. 2009. https://doi.org/10.1561/1500000019.

[5] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods", Proc. 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, MA, USA, pp. 758-759, Jul. 2009. https://doi.org/10.1145/1571941.1572114.

[6] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, and J. Larson, "From Local to Global: A Graph RAG Approach to Query-Focused Summarization", arXiv preprint arXiv:2404.16130, pp. 1-26, Feb. 2025. https://doi.org/10.48550/arXiv.2404.16130.

[7] B. Chen, Z. Guo, Z. Yang, Y. Chen, J. Chen, Z. Liu, C. Shi, and C. Yang, "PathRAG: Pruning Graph-based Retrieval Augmented Generation with

Relational Paths", Proc. of the AAAI Conference on Artificial Intelligence, Singapore, Vol. 40, No. 36, pp. 30183-30191, Jan. 2026. https://doi.org/10.1609/aaai.v40i36.40268.

[8] J. Y. Oh, J. Lee, and E. Hong, "A Study on Research Trends in the Smart Farm Field using Topic Modeling and Semantic Network Analysis", Journal of Digital Convergence, Vol. 20, No. 2, pp. 203-215, Feb. 2022. https://doi.org/10.14400/JDC.2022.20.2.203.

[9] H.-L. Hsu and J. Tzeng, "DAT: Dynamic Alpha Tuning for Hybrid Retrieval in Retrieval-Augmented Generation", arXiv preprint arXiv:2503.23013, pp. 1-12, Mar. 2025. https://doi.org/10.48550/arXiv.2503.23013.

[10] G. Gerganov and contributors, "llama.cpp: LLM Inference in C/C++", GitHub repository. https://github.com/ggml-org/llama.cpp. [accessed: May 07, 2026]

[11] J. Kpodo, P. Kordjamshidi, and A. P. Nejadhashemi, "AgXQA: A benchmark for advanced Agricultural Extension question answering", Computers and Electronics in Agriculture, Vol. 225, Art. no. 109349, Oct. 2024. https://doi.org/10.1016/j.compag.2024.109349.

[12] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, "MuSiQue: Multihop Questions via Single-hop Question Composition", Transactions of the Association for Computational Linguistics, Vol. 10, pp. 539-554, May 2022. https://doi.org/10.1162/tacl_a_00475.

[13] X. Ho, A.-K. D. Nguyen, S. Sugawara, and A. Aizawa, "Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps", Proc. 28th International Conference on Computational Linguistics (COLING), Online, pp. 6609-6625, Dec. 2020. https://doi.org/10.18653/v1/2020.coling-main.580.

[14] S. Es, J. James, L. E. Anke, and S. Schockaert, "RAGAs: Automated Evaluation of Retrieval Augmented Generation", Proc. 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, St. Julians, Malta, pp. 150-158, Mar. 2024. https://doi.org/10.18653/v1/2024.eacl-demo.16.

저자소개

김 유 석 (Yu-Seok Kim)

2024년 3월 : 제주대학교 컴퓨터공학과(공학사)
2025년 3월 ~ 현재 : 제주대학교 컴퓨터공학과 석사과정
관심분야 : 머신러닝, LLM, 모델양자화 최적화, 이미지처리, RAG 시스템

김 용 운 (Yong-Woon Kim)

1994년 3월 : 연세대학교 컴퓨터과학과(공학사)
1997년 3월 : 연세대학교 컴퓨터과학과(공학석사)
2023년 8월 : CHRIST University 컴퓨터과학 및 공학과(공학박사)
2023년 8월 ~ 현재 : 제주대학교 컴퓨터공학과 교수
관심분야 : 컴퓨터 비전, 인공지능, 그린수소, 사물인터넷(IoT)시스템, 블록체인

변 영 철 (Yung-Cheol Byun)

1993년 2월 : 제주대학교 정보공학과(공학사)
1995년 8월 : 연세대학교 컴퓨터과학과(공학석사)
2001년 8월 : 연세대학교 컴퓨터과학과(공학박사)
1998년 3월 ~ 2001년 2월 : 삼성전자, SDS 전문강사
2001년 9월 ~ 2002년 11월 : 한국전자통신연구원 선임연구원
2012년 9월 ~ 2014년 2월 : University of Florida 컴퓨터공학과 방문 교수
2003년 12월 ~현재 : 제주대학교 컴퓨터공학과 교수
관심분야 : 딥러닝, 패턴인식, 시계열 데이터 처리, 추천 시스템, 지식발견, 딥러닝 기반 신재생에너지 시스템, 블록체인