Chain-of-Thought 기반 사실성 검증을 통한 Agentic RAG 농작물 QA 시스템의 성능 향상 연구

Performance Enhancement of an Agentic RAG-Based Agricultural QA System through Faithfulness Verification with Chain-of-Thought

요 약

기존 농작물 질의응답 시스템은 검색된 문서에 포함된 내용을 바탕으로 응답을 생성하지만, 문서와 질의 간 문맥적 불일치로 인해 실제와 다른 병명을 제시하거나 엉뚱한 작물 정보를 응답에 포함하는 등의 사실성 오류가 발생하고 있다. 본 논문에서는 이러한 문제를 해결하기 위해 Agentic RAG(Retrieval-Augmented Generation) 기반의 농작물 QA 시스템을 제안하고 응답의 사실성을 높이기 위한 두 가지 기법을 소개한다. 첫 번째는 별도의 학습 없이 추론 과정을 포함한 예시를 프롬프트에 삽입하여 모델의 사고 흐름을 유도하는 Chain-of-Thought(CoT) 기법, 두 번째는 생성된 응답이 검색 문서와 실제로 일치하는지를 평가하는 Fact-checker 기반 사실성 검증 기법이다. 실험 결과, 제안한 기법들은 기존 시스템 대비 병해 진단의 정확성과 문맥 기반 사실성(Contextual Faithfulness)을 효과적으로 개선함을 확인하였다. 본 연구는 농업 도메인에서의 신뢰도 높은 Large Language Model(LLM) 기반 질의응답 시스템 설계에 기여한다.

1. 서론

최근 인공지능 기반 질의응답(Question Answering, QA) 시스템은 다양한 도메인에서 사용자 질문에 대한 자동화된 응답을 제공하여 활용도가 증가하고 있다. 특히 농업 분야에서는 농작물의 병해충 진단, 생육 환경 개선, 재배 방법 제안 등의 실질적인 문제 해결을 위해 질의응답 시스템이 도입되었다. 그러나 기존의 질의응답 시스템은 주로 검색 기반 방식에 의존하며, 사전 정의된 규칙 또는 검색된 문서 내용을 직접 사용하는 방식에 머무르는 경우가 많다[1].

기존의 대표적인 농업 QA 시스템인 '이삭이'는 사용자의 질의에 대해 관련 문서를 검색한 후 해당 내용을 바탕으로 응답을 생성한다[2]. 하지만 이러한 방식은 다음과 같은 사실성 문제(factuality issue)를 가지고 있다. 첫째, 검색된 문서가 질의와 관련 있는 정보가 아니지만 시스템이 이를 응답 생성에 사용하여 질병을 오진(misdiagnosis)하는 등의 정확도 감소 문제가 있다. 둘째, 검색된 문서가 실제로는 다른 작물과 관련한 내용이지만 이를 사용자의 질의 대상 작물에 그대로 적용하는 문맥 오류(Contextual Faithfulness Error)가 발생한다.

이러한 문제는 단순한 문서 검색이나 keyword matching만으로는 해결이 어렵고, 생성된 응답이 문맥적으로 사용자의 질의에 부합하는지, 그리고 검색 문서에 기반한 사실인지를 판단할 수 있는 사실성 검증(faithfulness verification) 기법이 필요하다[3].

본 연구에서는 이러한 문제를 해결하기 위해 Agentic RAG(Retrieval-Augmented Generation) 기반의 농작물 질의응답 시스템을 제안하고, 두 가지 사실성 검증 기법을 사용하여 QA 시스템의 성능 향상을 도모한다[4]. 첫 번째는 Chain-of-Thought(CoT) 기법으로, 모델에 사전 학습 없이 사고 과정을 포함한 예제를 프롬프트에 삽입하여 질의에 대한 단계적 추론 과정을 유도한다[5]. 두 번째는 Fact-checker 알고리즘을 통해 생성된 응답이 실제 검색된 문서와 논리적으로 일치하는지를 판단하여 사실성이 낮은 응답을 사후 검증한다[6]. 본 연구는 두 기법을 실험적으로 적용하여 기존 시스템 대비 병해 진단의 정확성과 문맥 기반 사실성을 높일 수 있음을 입증하였다.

2. 관련 연구

최근 농업 분야에서는 병해충 진단 등 실질적인 문제 해결에 '이삭이'와 RAG 질의응답 시스템을 활용하고 있다. 하지만 이러한 시스템은 잘못된 병명 제시나 문맥 오류 등 사실성 문제(factuality issue)가 빈번하게 발생하며, 검색된 문서의 내용이 질의와 부정확하게 이어질 경우 문맥 오류가 발생한다. 이를 보완하기 위해 최근에는 응답의 사실성을 검증하거나 높이기 위한 다양한 연구가 진행되고 있다[7].

대표적으로 Chain-of-Thought(CoT) 기법은 모델이 사고 과정을 따라가며 정답을 도출하게 하여 응답의 정확도를 높인다. 또한, Fact-checker 알고리즘은 생성된 응답이 검색 문서와 논리적으로 일치하는지를 판단하여 응답의 신뢰도를 높인다.

본 연구는 이러한 최신 기법들을 RAG 기법에 추가한 Agentic RAG 시스템을 통해 병해 진단의 정확성과 문맥 기반 사실성을 동시에 높이는 방법을 제안한다.

3. 방법론

3.1 연구 목적 및 구성

본 연구는 사실성 향상을 목표로 제안한 Agentic RAG 기반 QA 시스템의 성능을 검증하기 위해 사전에 구축한 평가셋과 다

양한 임베딩 및 리랭커 모델 조합을 기반으로 실험을 수행하였다. 실험은 Chain-of-Thought(CoT) 기법과 Fact-checker 알고리즘의 개별 및 병행 효과로 구성되었다. 각 기법이 응답 생성 과정에서 의미 해석력 및 문서 기반 사실성 확보에 미치는 영향을 중심으로 분석을 진행하였다.

3.2 실험 준비

데이터셋 농작물 질의응답 시스템 평가를 위해 실제 농업 현장에서 사용하는 다양한 문서 및 커뮤니티 데이터를 기반으로 하여 QA 데이터셋을 만들었다. 활용된 문서는 농촌진흥청의 농업과학도서관에서 제공하는 PDF 자료이며, 커뮤니티 데이터는 '농사로'의 현장 기술 상담 결과 및 병해충 상담 결과를 이용하였다[8]. 문서에서 텍스트를 추출하여 Markdown 형식으로 변경 후 문서 구조 기반으로 Chunk화 하였으며, 이후 metadata를 추가하여 데이터셋을 구성하였다.

평가셋 평가셋은 농작물 문헌 기반 QA 270개와 hallucination 확인을 위한 농작물과 관련이 없는 QA 30개, 커뮤니티 데이터를 기반으로 한 QA 50개로 총 350개의 QA로 구성하였고 이를 GitHub에 게시하였다[9]. PDF 문서를 기반으로 만든 QA 270개는 병충해 정보 및 재배 정보와 같은 일반적인 지식을 포함하여 제작하였다. 실제 평가에는 350개 중 무작위로 추출한 100개의 QA 세트를 사용하였고, 모든 실험에서 동일한 QA 세트를 사용하였다. 무작위로 추출된 100개의 QA 세트에는 농작물 문헌 기반 QA 78개, 커뮤니티 데이터 기반 QA 13개, 농산물과 관련 없는 질문 9개가 포함되었다.

시스템 구성 CoT 기법과 Fact-checker 기법의 효과를 확인하기 위해 Baseline이 될 RAG 기반 농작물 QA 시스템을 제작하였다. Baseline은 FAISS DB 기반으로 제작하였으며 문서 검색을 위한 임베딩 모델로 BAAI/bge-m3을 선택하였다. 이후 Baseline에 CoT와 Fact-checker를 적용 후 두 기능을 선택 가능하도록 제작하여 차이점을 비교하였다. CoT 구현에서는 OpenAI의 gpt-4-turbo를 사용하였으며 Fact-checker 구현에서는 Upstage의 Solar-1-Mini를 사용하였다. 이후 Streamlit을 이용해 사용자 인터페이스를 구성하였다.

평가지표 기법 적용 후 생성된 응답을 평가하기 위해 5가지 지표를 이용하였다. RAGAS[10]의 신뢰성(Faithfulness) 지표를 통해 생성된 답변이 문서에 기반하는지를 평가하고, 응답 연관성(Answer Relevancy) 지표를 통해 질문과 생성된 답변의 연관성을 평가하였다. 본 연구에서는 자체적으로 제작한 평가 프롬프트를 이용하여 LLM-as-a-judge 방식으로 생성된 응답이 사용자의 의도를 올바르게 이해하였는지(Intent Understanding, 의도 이해도), 생성된 응답이 평가셋의 모범답안과 의미 측면에서 유사한지(Semantic Similarity, 의미 유사도)를 평가하였고, 평가 프롬프트로 평가한 점수의 신뢰도(Score Reliability)를 평가하여 응답의 사실성과 적절성을 정량적으로 검증하였다. LLM-as-a-judge 방식을 적용하기 위해 사용한 모델은 OpenAI의 gpt-4o이다.

LLM-as-a-judge 점수의 신뢰성 평가 본 연구의 프롬프트를 이용한 LLM-as-a-judge를 통해 산출된 점수가 사람의 평가와 얼마나 일치하는지를 평가하기 위해 아래 그림 1을 이용하였다.

$$\text{Consistency (\%)} = \left(1 - \frac{|\text{Score}_{\text{LLM}} - \text{Score}_{\text{Human Avg}}|}{\text{MaxScore}}\right) \times 100$$

그림 1. Consistency 계산식

LLM-as-a-judge 점수와 두 명의 평가자가 부여한 점수의 평균 간 절대 차이를 계산하였고, 이를 최대 점수로 정규화한 후 퍼센트로 환산하였다. 이와 동일한 방식으로 두 명의 평가자 간 점수 차이도 계산하여 사람 간 유사성을 측정하였다. 표 1을 보면 LLM-as-a-judge는 사람 평가자 간 일치도보다 높은 수준의 일치도를 보이며, 이는 LLM-as-a-judge 방식으로 측정된 의도 이해도 및 의미 유사도의 신뢰성을 뒷받침한다.

표 1. LLM과 사람 간 평가 일치도 및 평가자 간 비교

평가지표	LLM-Human Consistency (%)	Human-Human Consistency (%)
Intent Understanding	90.7	86.6
Semantic Similarity	89.8	89.6
Average	90.25	88.1

4. 실험 결과 및 분석

4.1 실험 결과

표 2. 모델 조합별 평가지표

모델 조합	Intent Understanding	Semantic Similarity	Score Reliability
Baseline	0.784	0.758	1.000
Fact Checker	0.782	0.762	1.000
CoT	0.868	0.832	1.000
CoT + Fact Checker	0.868	0.828	1.000

모델 조합	Faithfulness	Answer Relevancy
Baseline	0.836	0.723
Fact Checker	0.842	0.725
CoT	0.812	0.674
CoT + Fact Checker	0.832	0.649

표 3. LLM-as-a-Judge 평가 항목별 Score Reliability 반영 점수

모델 조합	Adjusted Intent Understanding	Adjusted Semantic Similarity
Baseline	0.784	0.758
Fact Checker	0.782	0.762
CoT	0.868	0.832
CoT + Fact Checker	0.868	0.828

Adjusted Intent Understanding = Intent Understanding × Score Reliability
Adjusted Semantic Similarity = Semantic Similarity × Score Reliability

Chain-of-Thought CoT 기법을 적용한 시스템은 의도 이해도(0.868)와 의미 유사도(0.832)에서 높은 성능을 보였으나, 신뢰성(0.812)은 상대적으로 낮다.

Fact-checker Fact-checker 알고리즘을 적용한 시스템은 신뢰성(0.842)에서 높은 성능을 보였다. 그러나 의도 이해도(0.782), 의미 유사도(0.762)가 낮게 나와 CoT 기법을 적용한 실험에 비해 내용 구성이나 질문 이해의 정밀도에서는 낮은 평가를 받았다.

CoT + Fact-checker CoT와 Fact-checker를 병행한 실험에서는 신뢰성(0.832)이 높아졌다. 그러나 의미 유사도(0.828)는 CoT 기법 하나만 적용한 실험에 비해 감소하여, 사고 흐름 유도(CoT)에 의한 문맥 파악 강화와 Fact-checker 기반 사실성 보정 사이의 trade-off가 관찰되었다.

4.2 실험 결과 분석

본 연구에서 사용한 5가지 지표 중 의미 유사도는 사실성이 보장된 모범답안과의 의미 유사도를 평가하고, 신뢰성은 검색된 문서와의 연관성을 평가한다. 그러므로 이 2가지 지표를 통해 사실성을 확인할 수 있다. 또한 의도 이해도와 의미 유사도의 경우 본 연구의 평가 프롬프트를 이용하여 평가하였기에 점수 신뢰도를 곱하여 2가지 지표를 보정하였다.

CoT 기법을 적용하면 신뢰성이 감소하는 것으로 보아 LLM이 추론 과정에서 문서에 포함되지 않은 정보를 사용한 빈도수가 높아졌다고 판단할 수 있으나, 추론 과정 중 필요한 보편적 사실이 포함되어 신뢰성이 감소했다고도 분석할 수 있다. CoT 기법을 적용한 후 의미 유사도가 대폭 증가하였다는 결과는 이러한 분석을 뒷받침한다. 또한 의도 이해도도 대폭 증가하였는데, 이를 통해 논리적인 추론 과정(CoT)이 사용자의 의도를 파악하고 적절한 응답을 생성하는 데에 효과적임을 알 수 있다. 응답 연관성의 경우 감소하지만, 응답을 살펴보았을 때 농산물과 관련이 없는 질문에 대해 답변하지 않았을 경우 응답 연관성이 0으로 나타나는 것을 확인할 수 있다. 이는 RAGAS가 특정한 목적을 가진 시스템의 평가지표로는 부적합하다는 것을 시사한다.

Fact-checker를 baseline에 적용하였을 때 신뢰성이 근소하게 상승한 것으로 Fact-checker의 문서 선별로 응답의 hallucination이 줄어 문서와 연관된 신뢰성이 높아졌음을 알 수 있다. 그러나 응답 연관성의 경우 baseline과 CoT에 적용했을 때 경향성이 다르게 나타난다. 응답을 살펴보았을 경우 제공된 문서에는 정보가 포함되어 있지 않다는 답변을 한 경우 점수가 응답 연관성이 0점이 나온 것을 확인할 수 있다. 또한 추정 등의 단어가 CoT+fact-checker 시스템에서 더 빈번히 발생하였으며 응답 연관성 또한 근소하게 감소하는 것을 확인할 수 있었다. 그러나 이러한 경향성을 제외하더라도 완전히 동일한 내용임에도 점수가 달라지는 경우 또한 존재했으며, 결과적으로 RAGAS의 응답 연관성 지표만으로 시스템을 평가하기에는 부적합함을 확인할 수 있었다.

5. 결론 및 향후 연구

본 연구는 Agentic RAG 기반 농작물 QA 시스템에 CoT와 Fact-checker 기반 사실성 검증 기법을 적용하여 병해 진단의 정확성과 문맥 기반 사실성을 높였다. 그러나 CoT 기법은 모델의 추론 과정을 그럴듯하게 설명하는 데에는 유용하나, 해당 설명이 실제 예측의 근거를 정확히 반영하는지는 불분명하다[11]. 즉, CoT가 생성한 사고 흐름은 설득력 있어 보이지만 모델이 응답을 도출한 실제 결정 요인과는 불일치할 가능성이 있다[12]. 그러므로 CoT가 도출한 설명은 오히려 응답의 신뢰성을 해칠 수 있다.

또한 Agentic RAG 방식은 프롬프트에 의존하기 때문에 새로운 유형의 질의나 복잡한 검색 상황에서 일반화 성능이 제한적이며 검색 도구 활용 방식의 최적화에도 한계가 있다. 이러한 문제점을 해결하기 위해 LLM이 검색 쿼리를 스스로 생성하고, 다단계 추론 과정을 강화학습으로 최적화하는 방식을 제안할 예정이다.

향후 연구에는 강화학습 기반 검색 추론 기법을 도입하여 LLM이 질의에 따라 능동적으로 검색과 추론을 반복하며 응답의 정확도와 사실성을 더욱 높일 수 있는 구조를 설계할 예정이다. 이를 통해 Agentic RAG 기반 시스템을 더 지능적인 QA 시스템으로 발전시켜 실제 농업 현장에서 사용할 수 있는 신뢰도 높은 질의응답 서비스를 구현하는 것이 궁극적인 목표이다.

6. 참고 문헌

[1] Patrick Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”, in NeurIPS, 2020.

[2] 농촌진흥청, “AI 이삭이”, in https://www.nongsaro.go.kr/chatbot/aichatbot.html, 2025.

[3] Tobias Schimanski et al., “Towards Faithful and Robust LLM Specialists for Evidence-Based Question-Answering”, in arXiv preprint arXiv:2402.08277, 2024.

[4] Aditi Singh et al., “Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG”, in arXiv preprint arXiv:2501.09136, 2025.

[5] Jason Wei et al., “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”, in NeurIPS, 2022.

[6] Liyan Tang et al., “MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents”, in EMNLP, 2024.

[7] Bowen Jin et al., “Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning”, in arXiv preprint arXiv:2503.09516, 2025.

[8] 농촌진흥청, “농업과학도서관” in https://lib.rda.go.kr/main.do, 2025.

[9] “Evaluation Set”, in https://github.com/kirasma53/Evaluation_Set, 2025.

[10] Shahul Es et al., “RAGAS: Automated Evaluation of Retrieval Augmented Generation”, in arXiv preprint arXiv:2309.15217, 2023.

[11] Alan Jacovi et al., “Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness?”, in ACL, 2020.

[12] Miles Turpin et al., “Language Models Don’t Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting”, in NeurIPS, 2023.