

## 교육과정 기반 RAG 학습 챗봇 설계 및 구현: 근거기반 설명형과 질문유도형 이중 응답 전략을 중심으로

하 오 선\*

동국대학교 교무처 교수법연구초빙교수

# A Curriculum-Based RAG Learning Chatbot with Dual Response Strategies: Evidence-Based Explanations and Question-Prompting Responses

Oh-Sun Ha\*

Research Professor, Office of Academic Affairs, Dongguk University, Seoul 04620, Korea

### [요 약]

본 연구는 중·고등학교 교과 데이터를 활용하여 학습자의 발달 수준을 반영하는 교육과정 기반 검색증강생성(RAG) 학습 챗봇을 설계·구현하였다. 제안 시스템은 범용 생성형 AI에서 발생할 수 있는 환각 문제를 완화하고, 구조화된 상호작용을 통해 학습자의 사고를 촉진하는 것을 목적으로 한다. 이를 위해 학년 및 교과 정보를 고려한 문서 검색 구조와 근거기반 설명형 응답과 질문유도형 응답으로 구성된 이중 응답 전략을 적용하였다. 평가 결과, 제안 시스템은 교과 범위 내·외 및 교과 무관 질의 전반에서 근거 없는 정보 생성을 효과적으로 억제하는 것으로 나타났다. 또한 정량적 평가 결과, 범용 챗봇과 비교하여 낮은 환각률과 높은 BERTScore를 보여 응답의 사실적 신뢰성과 적절성이 향상되었음을 확인하였다. 이러한 결과는 제안 시스템이 근거 제시 중심의 설명과 학습 과정 지향적 상호작용을 지원하는 신뢰 가능한 교육용 AI 도구로 활용될 수 있음을 시사한다.

### [Abstract]

This study designs and implements a curriculum-based retrieval-augmented generation (RAG) learning chatbot that utilizes middle and high school curriculum data to reflect learners' developmental levels and support structured thinking. The proposed system addresses hallucination issues in general-purpose generative AI by incorporating grade- and subject-aware document retrieval alongside a dual response strategy. This strategy consists of two components: evidence-based explanatory responses and question-prompting responses that encourage learners' cognitive engagement. Evaluation results show that the system effectively suppresses unsupported information across curriculum-aligned, out-of-curriculum, and curriculum-irrelevant queries. Quantitative analysis further demonstrates a lower hallucination rate and a higher BERTScore compared with a general chatbot, indicating enhanced factual reliability and response adequacy. Overall, the findings suggest that the proposed system can function as a reliable educational AI tool that supports evidence-grounded explanations and learning-oriented interactive responses.

**색인어** : 교육과정 기반, 검색증강생성, 학습 챗봇, 근거기반 설명, 질문유도형 응답

**Keyword** : Curriculum-Based Learning, Retrieval-Augmented Generation, Learning Chatbot, Evidence-Based Explanation, Question-Prompting Response

<http://dx.doi.org/10.9728/dcs.2026.27.3.821>



This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Received** 08 February 2026; **Revised** 06 March 2026

**Accepted** 10 March 2026

\*Corresponding Author; Oh-Sun Ha

**Tel:** +82-2-2260-3666

**E-mail:** wffwff@nate.com

## I. 서론

생성형 인공지능(Generative AI)의 확산은 학습 환경에 빠른 변화를 가져오고 있으며, 학교 현장에서도 AI 기반 대화형 시스템을 활용한 학습지원 시도가 증가하고 있다. 특히 학생의 질문에 즉각적으로 응답할 수 있는 챗봇 기반 시스템은 자기주도 학습을 촉진할 수 있다는 점에서 주목받고 있다. 그러나 기존의 범용 생성형 AI는 교육과정 및 학습자의 발달 수준을 충분히 반영하지 못한 채 정답 중심의 응답을 제공하거나, 근거가 불명확한 내용을 생성하는 한계를 갖는다[1],[2]. 이는 학습 정확성뿐 아니라 학생의 사고 과정 형성에 부정적 영향을 미칠 가능성이 있으며, 교육적 활용을 위해서는 교육과정 기반의 신뢰 가능한 응답 구조가 요구된다. 이러한 문제를 보완하기 위한 대안으로 검색증강생성(RAG; Retrieval-Augmented Generation)은 외부 지식 기반 검색 결과를 바탕으로 답변을 생성함으로써 사실성 및 근거 제시 가능성을 높이는 접근으로 제시되고 있다. 특히 대학 LMS 질의응답 챗봇에 RAG 기반 거대언어모델을 적용하여, 질의 임베딩을 활용한 검색 및 유사도 기준 기반 응답 생성 구조를 제안한 연구는 RAG의 실질적 적용 가능성을 보여준다[3]. 또한 한국어 교육 분야에서도 생성형 AI의 환각(Hallucination) 문제를 보완하기 위해 검증된 자료 기반 RAG 적용 필요성이 강조되며, 문법 교정 챗봇에 RAG 구조를 적용하는 방안이 논의되고 있다[4]. 다만 기존 RAG 기반 학습 챗봇 연구는 대체로 “질문-정답 제공” 중심의 기능 구현에 머무르는 경우가 많아, 학생의 사고를 확장시키고 질문을 정교화하도록 돕는 교육적 피드백 기능이 상대적으로 부족하다. 즉, 학습자가 단순히 답을 얻는 것이 아니라 스스로 사고하도록 유도하는 상호작용 구조를 갖춘 RAG 설계가 필요하다. 이에 본 연구는 ‘교과단계별 교과 데이터(개방데이터)’를 기반으로 학년·교과에 적합한 근거 문서를 검색하고, 이를 바탕으로 학생에게 질문유도형 피드백을 제공하는 교육과정 기반 RAG 시스템을 설계·구현하고자 한다. 본 연구는 범용 생성형 AI의 한계를 교육과정 기반 데이터와 RAG 구조로 보완하고, 단순 정답 제공을 넘어 학습 과정 중심의 상호작용을 지원하는 AI 튜터 설계 방향을 제시한다는 점에서 의의가 있다.

## II. 선행연구

교육 환경에서 챗봇은 학습자의 질문에 즉각적인 피드백을 제공함으로써 자기주도 학습을 지원할 수 있는 도구로 주목받아 왔다. 초기 연구에서는 형태소 분석, 문서 분류, 정답 후보 추출 등 자연어 처리 기반 기법을 통해 특정 도메인에서의 질의응답 정확도를 높이기 위한 다양한 접근이 제안되었으며[5], 학교 현장에서 교과 질의응답 챗봇을 운영하고 학생 활용 양상을 분석한 사례도 보고되었다[6]. 최근에는 거대언어 모델(LLM)을 활용한 챗봇 개발이 확산되면서 규정·상담·과학·의료 등 여러 분야에서 적용 가능성이 확인되고 있다

[7]-[10]. 그러나 범용 LLM은 학습 시점 이후 정보 반영의 한계와 근거 없는 응답 생성 가능성이 존재하며, 교육적 맥락에서는 이러한 부정확성이 학습 오류로 이어질 수 있다는 점에서 신뢰성 확보가 중요한 과제로 제기된다[11]. 이러한 문제를 보완하기 위한 대안으로 검색증강생성(RAG)이 제시된다. RAG는 외부 문서에서 검색된 정보를 근거로 응답을 생성하는 구조로, 답변의 사실성과 근거 제시 가능성을 높이고 환각을 완화하는 접근으로 논의되어 왔다. 특히 질의 임베딩을 활용해 관련 문서를 검색하고, 검색 결과를 기반으로 응답을 생성하는 RAG 기반 질의응답 시스템은 특정 도메인에서의 적용 가능성을 보여준다[12]. 최근에는 교육 환경에서 RAG 기반 챗봇을 활용하여 학습자의 참여와 학습 효과를 향상시키려는 연구가 증가하고 있다. 예를 들어 온라인 학습 환경에서 RAG 기반 챗봇이 학습 자료를 기반으로 설명과 예시를 제공하여 학습자의 이해도와 만족도를 높일 수 있음이 보고되었으며[13], 대학 교육 환경에서 강의 자료를 기반으로 질의응답을 제공하는 RAG 학습 챗봇 역시 학습 지원 도구로서의 가능성을 보여주었다[14]. 또한 RAG 챗봇의 교육적 활용 사례를 종합적으로 분석한 연구에서는 다양한 교육 영역에서 RAG 기반 챗봇이 학습 지원 도구로 활용되고 있음을 확인하였다[15]. 또한 교육 분야에서도 검증된 자료를 기반으로 설명을 제공하기 위해 RAG 구조의 활용이 강조되고 있으며[16], 이는 교육과정에 정합적인 데이터 기반 응답 체계를 설계하는 데 중요한 기반이 된다. 다만 기존 RAG 기반 챗봇 연구는 주로 “정확한 답 제공” 중심으로 설계되는 경향이 있어, 학습자의 사고 확장과 질문 정교화를 촉진하는 상호작용 전략이 상대적으로 부족하다는 한계가 있다[17]. 이에 본 연구는 교과 단계별 교과 데이터(개방데이터)를 기반으로 학년·교과 맥락에 맞는 근거 문서를 검색하고, 근거 중심 설명형 응답과 함께 소크라테스식 질문유도형 응답을 제공하는 교육과정 기반 RAG 학습 챗봇을 설계·구현하고자 한다.

## III. 연구방법

### 3-1 전처리

본 연구에서는 AI허브의 교과 단계별 교과 데이터를 사용하였다. 포함된 JSON 형식의 원천 데이터를 수집하여 RAG 학습용 텍스트 청크 chunks\_train.jsonl을 생성하였다. 전처리의 목적은 검색 효율을 높이기 위한 노이즈 제거와 구조화, 그리고 근거 제시율 및 환각률 평가가 용이하도록 메타데이터를 정리하는 데 있다. 먼저 텍스트 데이터에 대해 공백 및 줄바꿈을 정규화하고 제어 문자를 제거하였다. 또한 페이지 번호나 캡션 등 검색에 의미가 없는 짧은 문장을 제거하기 위해 15자 미만의 텍스트는 제외하였다. 교과서 데이터의 반복성을 고려하여 교과·학교급·학년·텍스트 내용을 기준으로 중복 청크를 제거하였다. 청크는 RAG 입력의 일관성을 확보하기 위해 설명(text\_description), 질문(text\_qa), 답변

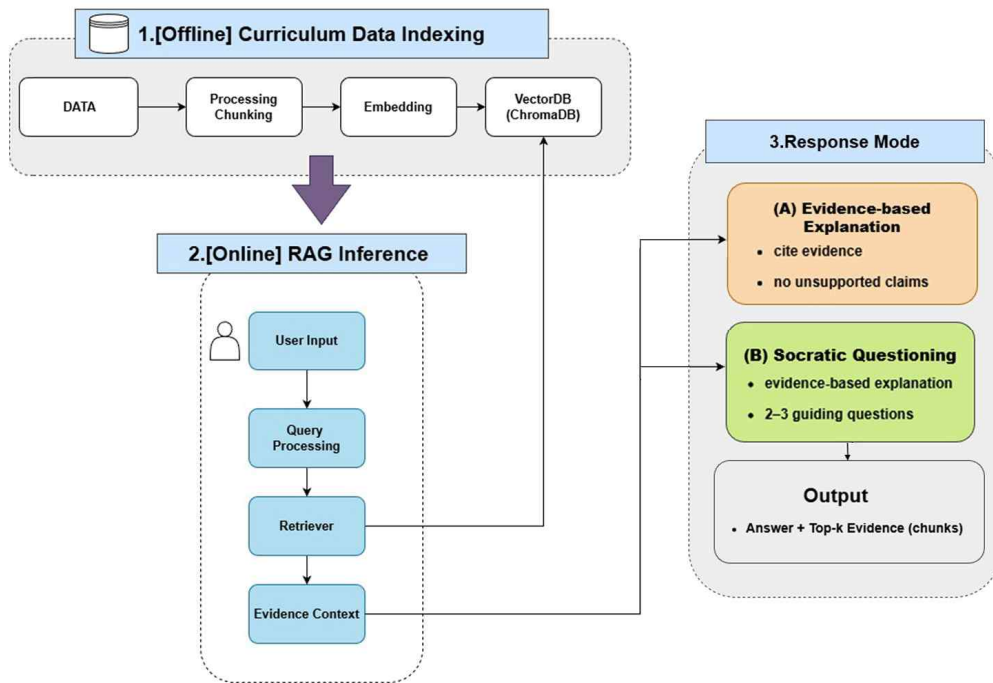


그림 1. 연구 흐름도  
Fig. 1. Overall research process

(text\_an)을 하나의 텍스트로 통합하는 방식으로 수행하였다. 이때 ‘설명-질문-답변’ 형식으로 구조를 통일하였으며, QA 정보가 없는 경우에는 설명 텍스트만을 사용하였다. 메타데이터는 학교급, 학년, 교과 정보를 표준화하여 저장하였고, 학년 정보는 숫자형 변수로 변환하였다. 또한 분석 대상의 일관성을 위해 class\_name이 “텍스트”가 아닌 데이터는 제외하였다. 이러한 전처리를 통해 검색 정확도와 평가 가능성을 동시에 고려한 RAG 학습 데이터셋을 구축하였다.

### 3-2 임베딩 및 ChromaDB 구축

처리된 텍스트 청크(chunks\_train.jsonl)를 대상으로 문장 임베딩을 수행하고, 벡터 검색을 위한 ChromaDB를 구축하였다. 임베딩 모델로는 한국어를 포함한 다국어 문장 의미 표현에 적합한 SentenceTransformer 기반의 paraphrase-multilingual-MiniLM-L12-v2 모델을 사용하였다. 각 청크는 임베딩 벡터로 변환된 후, 고유 식별자와 원문 텍스트, 그리고 학년·교과 정보를 포함한 메타데이터와 함께 ChromaDB에 저장하였다. 메타데이터는 이후 검색 단계에서 조건 필터링을 가능하게 하여 검색 정확도를 높이기 위한 목적으로 활용된다. 임베딩 적재는 배치 단위로 수행하였으며, 벡터 간 유사도 계산은 코사인 유사도를 기준으로 하였다. 구축된 ChromaDB는 질의 문장을 임베딩한 뒤 의미적 유사도를 기준으로 관련 텍스트 청크를 검색하는 데 사용되며, RAG 파이프라인에서 근거 텍스트로 활용된다.

### 3-3 Retriever

구축된 ChromaDB를 기반으로 사용자의 질의에 대해 관련 텍스트 청크를 검색하는 Retriever를 구성하였다. 질의 문장은 SentenceTransformer 기반 임베딩 모델을 통해 벡터로 변환되며, 코사인 유사도를 기준으로 의미적으로 유사한 상위 k개의 텍스트 청크(top-k 검색)를 반환한다. 검색 과정에서는 학년과 교과 정보를 메타데이터 필터로 활용하여 질의와 무관한 범위의 텍스트가 포함되는 것을 최소화하였다. 이를 통해 동일한 개념이라도 학년이나 교과가 다른 텍스트가 혼입되는 문제를 완화하고, 검색 결과의 적합성을 높였다. Retriever는 각 검색 결과에 대해 텍스트 내용과 함께 유사도 점수, 학년·교과·class 정보 등의 메타데이터를 함께 반환하도록 설계하였다. 이러한 구조는 이후 생성 단계에서 근거 텍스트를 구성하고 응답의 적합성을 평가하는데 활용된다.

### 3-4 Generation

검색 단계에서 반환된 상위 k개의 텍스트 청크는 파이프라인 모듈을 통해 하나의 근거 텍스트(evidence)로 결합된 후, 생성 단계의 입력으로 사용된다. 이 과정에서 각 청크는 유사도 점수와 메타데이터 정보를 함께 유지하여, 검색 결과가 응답의 근거로 활용될 수 있도록 구성하였다. 생성 단계에서는 Gemini 기반의 생성형 언어모델을 사용하여 최종 응답을 생

성하였다. 검색-칭킹-근거 구성 과정을 거친 텍스트는 사전에 정의된 프롬프트와 함께 언어모델에 입력되며, 이를 통해 검색 결과와 생성 모델이 분리된 파이프라인 구조를 유지하였다. 이러한 파이프라인 설계는 검색 결과가 생성 단계에 명확히 반영되도록 하여, 근거 없는 응답 생성을 최소화하기 위한 목적을 가진다. 아래에 제시한 프롬프트는 생성 단계에서 사용된 입력 형식으로, 제공된 근거 텍스트에 기반하여 응답을 생성하도록 언어모델의 행동을 제어한다. 설명형 응답과 질문 유도형 응답을 구분하여 구성하였으며, 근거에 포함되지 않은 내용은 생성하지 않도록 명시적인 제약을 포함하였다. 이를 통해 교과 범위를 벗어난 환각 응답을 줄이고, 근거 제시가 가능한 응답 생성을 유도하였다.

표 1. 근거기반 응답 생성을 위한 프롬프트 설계

Table 1. Prompt design for evidence-based generation

mode	System Instruction	User Prompt Structure
MODE A	<ul style="list-style-type: none"> <li>- Answer strictly based on the provided evidence.</li> <li>- Do not fabricate any information.</li> <li>- Indicate evidence references in the answer.</li> </ul>	[Question] [Evidence Text] → Generate an evidence-based explanation.
MODE B	<ul style="list-style-type: none"> <li>- Guide student thinking using a Socratic approach.</li> <li>- Do not provide direct answers.</li> <li>- Include 2-3 follow-up questions.</li> </ul>	[Question] [Evidence Text] → Generate a guided response with follow-up questions.

3-5 Streamlit UI

본 연구에서 제안하는 RAG 시스템의 효용성과 접근성을 검증하기 위해, Python의 Streamlit 라이브러리를 사용하여 웹 기반의 프로토타입 UI를 구현하였다. 본 UI는 사용자가 자신의 학습 수준과 범위를 명확하게 설정하고, 목적에 따라 다른 유형의 피드백을 받을 수 있도록 설계되었다. 사용자는 드롭다운 메뉴를 통해 ‘학교급(초/중/고)’, ‘학년’, ‘교과’를 순차적으로 선택할 수 있다. 이러한 계층적 선택 구조는 사용자의 학습 맥락에 맞는 정확한 근거문서를 필터링하여 검색 정확도를 높이는 데 기여한다. 또한, 사용자는 ‘근거기반 설명형’과 ‘질문유도형’ 두 가지 응답 모드 중 하나를 라디오 버튼으로 선택하여, 동일한 질문에 대해서도 목적에 따라 다른 유형의 피드백을 받도록 구성하였다. 답변이 생성된 후에는 AI가 참고한 근거 자료의 원문과 출처 정보(학년, 교과, 유사도 점수 등)가 답변 하단에 함께 제시된다. 사용자는 확장 가능한(expander) 메뉴를 통해 각 근거의 상세 내용을 직접 확인할 수 있다. 이를 통해 생성된 답변의 신뢰도를 검증하고, 관련 내용에 대한 심화 학습의 단서로 활용할 수 있도록 지원한다.

표 2. 교과 단계별·학년별 교과 적용 범위

Table 2. Curriculum-based configuration by school level and grade in the Streamlit UI

School level	Grade range	Subjects
Elementary	Grades 3-4	Korean language, Ethics, Social studies, Mathematics, Science, English
Elementary	Grades 5-6	Korean language, Ethics, Social studies, Mathematics, Science, English, Information
Middle school	Grades 1-3	Korean language, Ethics, Social studies, Mathematics, Science, English, Information, Technology and Home Economics
High school	Grade 1	Korean language, Social studies, Mathematics, Science, English, Information, Technology and Home Economics, Sociology and Culture

IV. 연구결과

4-1 정성적 평가

본 섹션에서는 제안하는 RAG 시스템의 답변 품질과 특성을 다각적으로 분석하기 위해 정성적 평가를 수행한다. 평가는 질문의 유형을 ‘교과 범위 내’, ‘교과 범위 밖’, ‘교과 무관’ 세 가지 시나리오로 나누고, 각 시나리오에 대해 ‘근거기반 설명형’과 ‘질문유도형’ 두 가지 응답 모드가 어떻게 반응하는지 비교 분석하는 방식으로 진행되었다.

1) 시나리오별 응답 특성 비교 분석

먼저, 각 시나리오에 대한 두 응답 모드의 전반적인 응답 특성을 요약하면 표 3과 같다.

표 3. 시나리오 유형별 응답 모드 특성 요약

Table 3. Summary of response characteristics by scenario

Scenario type	Evidence-based explanatory	Question-Prompting Response
Within curriculum scope	Provides accurate and detailed explanations grounded in retrieved curriculum-aligned evidence	Encourages deeper understanding by guiding learners with follow-up questions based on the curriculum
Outside curriculum scope	Avoids unsupported explanations and explicitly indicates the absence of relevant evidence	Redirects the learner’s inquiry through exploratory or clarifying questions without fabricating information
Curriculum-irrelevant	Refrains from generating factual claims and responds conservatively to prevent misinformation	Prompts reflection or reformulation of the question to align with learnable content

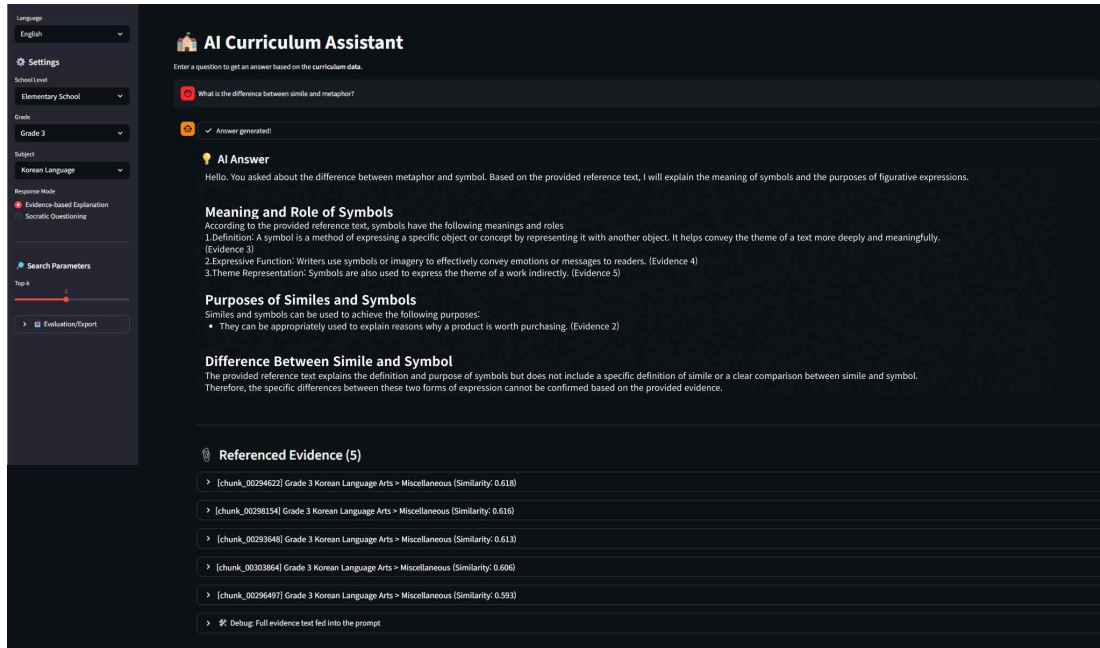


그림 2. 교육과정 기반 RAG 학습 챗봇의 사용자 인터페이스

Fig. 2. User interface of the curriculum-based RAG learning chatbot

분석 결과, ‘근거기반 설명형’은 정보의 정확한 전달에, ‘질문유도형’은 사용자의 능동적 학습을 유도하는 데 각각 강점이 있음을 확인했다. 이러한 응답 특성은 표 3에 제시된 시나리오별 비교 결과를 통해 명확히 드러난다. 특히, 근거가 없는 ‘교과 범위 밖’ 및 ‘교과 무관’ 질문에 대해서도 두 모드 모두 사실이 아닌 정보를 생성하지 않고 안정적으로 대응함으로써, 제안하는 RAG 시스템의 환각 억제 성능을 입증하였다.

2) 시나리오별 심층 분석 사례

응답 특성이 실제 답변에서 어떻게 나타나는지를 시나리오

별로 심층 분석한 사례는 표 4, 표 5, 표 6과 같다.

본 사례 분석은 교과 범위 내·외·무관 질문에 대해 제안한 RAG 기반 챗봇이 응답 전략을 어떻게 달리 적용하는지를 보여준다. 교과 범위 내 질문에서는 근거기반 설명을 제공하고 추가 사고를 유도하는 질문형 응답을 통해 학습 심화를 지원하였다. 반면 교과 범위 밖 및 교과 무관 질문에 대해서는 응답 범위를 명확히 제한하면서도, 교과 내용과의 연결 가능성을 제시하는 질문을 통해 학습 맥락을 유지하고자 하였다. 이는 제안 시스템이 단순한 응답 생성이 아니라 교육적 타당성과 범위 인식을 기반으로 상호작용을 조절함을 시사한다.

표 4. 사례 분석 (1) - 교과 범위 내 질문

Table 4. Curriculum-aligned query

Scenario	In-scope query (High/Korean)	
QUESTION	What's the difference between a metaphor and a symbol?	
ANSWER	Gemini	Although metaphors and symbols are both ways of representing one object by comparing it to another, there are several important differences. The metaphor is a temporary expression in which the original and the auxiliary concepts are connected 1:1 ("My heart is a lake"), while the symbol is a continuous expression that shows universal and multi-layered meaning only by the auxiliary idea without the original idea ("Pigeons are peace")
	Evidence-based explanatory mode	The main differences between metaphors and symbols are the presence or absence of 'original ideas' and 'temporary nature of relationships'. 1. Continuity of the relationship: The metaphor is a one-on-one relationship that is temporarily established only in a specific context, like "My heart is a lake," but the symbol is conventionally solidified like "Dove = Peace," continuing regardless of the context. It holds a one-to-many relationship. 2. Raw Idea: Although metaphors reveal both the original and the auxiliary concepts, symbols often have hidden concepts.
	Question-Prompting Response	Good question! Both metaphors and symbols have one thing in common: they express the original idea by comparing it to other objects. Before I know the difference between the two concepts, can you explain what the original and auxiliary concepts are, respectively, in the sentence "My heart is a lake"?

표 5. 사례 분석 (2) - 교과 범위 밖 질문

Table 5. Out-of-curriculum query

Scenario	Out-of-scope query (Historical questions in higher/Korean settings)	
QUESTION	What was Admiral Yi Sun-shin's first victory in the Imjin War?	
ANSWER	Gemini	Admiral Yi Sun-shin's first victory in the Japanese invasion of Korea was in the Battle of Okpo. On May 7, 1592, he defeated the Japanese army's Dodo Takatora's fleet off the coast of Okpo.
	Evidence-based explanatory mode	I'm sorry, but I couldn't find the information related to Imjin War or Lee Soon-shin within the current 'High Korean Language' curriculum.
	Question-Prompting Response	That is an interesting historical question! However, we are currently preparing to answer the Korean language curriculum. Is it related to the content of literary works or archives such as Nanjung Ilgi?

표 6. 사례 분석 (3) - 교과 무관 질문

Table 6. Curriculum-irrelevant query

Scenario	Curriculum-irrelevant query (Routine questions in higher/Korean settings)	
QUESTION	Please recommend a menu for dinner	
ANSWER	Gemini	Of course! What kind of food do you like? Do you prefer Korean, Chinese, or Western cuisine? And are you okay with spicy food? If you tell me a bit more, I can recommend a better option. For something simple, how about kimchi stew or pasta?
	Evidence-based explanatory mode	I'm sorry, but I am an AI assistant specialized in the high school Korean language curriculum. Within the currently configured scope, I am unable to provide everyday information such as dinner menu recommendations.
	Question-Prompting Response	It seems like you're thinking about what to have for dinner today. However, since I am an AI assistant trained specifically on curriculum-based content, it's difficult for me to recommend menus based on personal preferences. Instead, would you like me to recommend an interesting novel related to your Korean language studies?

4-2 정량적 평가

1) 평가 데이터셋

본 평가는 chunks\_validation.jsonl에 포함된 질의-응답 쌍을 기반으로 구축된 평가 데이터셋을 사용하였다. Validation 데이터셋에서 질문과 정답이 모두 존재하는 50개의 질의-응답 쌍을 무작위로 추출하여 평가를 수행하였으며, 데이터셋은 국어, 수학, 사회, 과학 등 다양한 교과와 초·중·고 학년군을 포괄하도록 구성하였다. 다만 본 평가는 제한된 표본을 활용한 초기 정량 평가로서, 제안 시스템의 성능을 절대적으로 확정하기보다 응답 특성을 탐색적으로 비교하기 위한 목적에서 수행되었다.

2) 근거제시율과 환각률

본 절에서는 개발된 RAG(Retrieval-Augmented Generation) 모델의 응답 특성을 정량적으로 파악하기 위해 근거제시율(Grounding Rate)과 환각률(Hallucination Rate)을 분석하였다. 다만 본 연구에서 사용한 두 지표는 RAG 응답의 신뢰성을 절대적으로 판정하는 단일 지표라기보다, 근거기반 응답 특성을 탐색적으로 비교하기 위한 운영적 지표(operational proxy)로 활용하였다. 기존 연구에서도 groundedness와 hallucination은 RAG 평가의 핵심 진단 축으로 널리 활용되지만, 정답성, 완전성, 검색 품질, 근거 부족 시의 거절 능력 등을 포괄하는 단독 지표로 보기에는 한계가 있음이 지적되고 있다[18].

• 평가 방법

평가는 evaluation\_results.csv에 포함된 50개의 질의응답 쌍을 대상으로 수행되었다. 해당 데이터셋은 질문(Question), 표준 정답(Reference), RAG 모델의 응답(Rag\_answer), 그리고 의미 유사도 점수(Rag\_bert\_score) 등을 포함하고 있다. 각 항목의 정의는 다음과 같다.

- evaluation\_results.csv: 평가에 사용된 데이터 파일로, 질문(Question), 표준 정답(Reference), RAG 모델의 답변(Rag\_answer), 그리고 성능 점수(Rag\_bert\_score 등)를 포함한다.
- rag\_answer: 본 연구에서 개발한 RAG 모델이 생성한 최종 응답이다.
- rag\_bert\_score: RAG 모델 응답(Rag\_answer)과 표준 정답(Reference) 간의 의미적 유사도를 BERT 기반 임베딩으로 계산한 값으로, 값이 높을수록 두 문장 간 의미적 유사성이 높음을 의미한다[19].

이를 바탕으로, 본 논문에서는 근거제시율과 환각률을 다음과 같이 정의하여 측정하였다.

- 근거제시율(Grounding Rate)

본 연구의 RAG 시스템은 검색된 교과 데이터를 기반으로 답변이 생성될 경우 해당 정보를 명시적으로 설명하도록 응답 템플릿을 설계하였으며, 이때 “관련된 내용은 다음과 같습니다”와 같은 근거 설명 표현을 사용하도록 구성하였다. 반면 검색된 근거가 충분하지 않을 경우에는 “답변이 없습니다”와 같은 기본 응답을 반환하도록 설계하였다. 따라서 본 연구의 근거제시율은 응답 내 근거 설명 표현의 출현 여부를 기준으

로 산출한 운영적 지표로 정의하였다. 문헌에서 groundedness는 일반적으로 답변의 주장들이 제공된 컨텍스트에 의해 지지되는 정도로 정의되므로[20], 본 연구의 근거 제시율은 이러한 groundedness의 간접적·탐색적 대리치(proxy)로 해석할 필요가 있다.

- 환각률(Hallucination Rate):

본 연구에서는 모델 응답의 의미적 불일치 가능성을 파악하기 위해 rag\_bert\_score가 0.85 미만인 경우를 환각 가능 응답으로 분류하였다. 생성된 응답(Rag\_answer)과 기준 정답(Reference) 간의 의미적 유사도를 BERTScore로 계산하여 분석하였다. BERTScore는 두 문장 간 의미적 유사도를 계산하는 자동 평가 지표로, 값이 높을수록 두 문장 간 의미적 일치도가 높음을 의미한다. BERTScore는 공식적인 임계값이 존재하지 않고 상대적 의미 유사도를 측정하는 지표이다[19]. 본 연구에서는 평가 데이터셋의 분포를 사전 탐색한 결과 대부분의 응답이 0.85 이상에 분포하는 점을 고려하여 상대적으로 의미 유사도가 낮은 응답을 탐지하기 위한 탐색적 기준으로 0.85를 설정하였다. 다만 BERTScore는 의미 유사도 기반 자동 평가 지표이므로 실제 사실 오류와 완전히 일치하지 않을 수 있으며, 본 연구에서는 의미적 불일치 가능성을 탐지하기 위한 보조적 지표로 활용하였다.

• 평가 결과

상기 정의에 따라 측정한 결과는 다음과 같다.

표 7. RAG 응답의 근거 제시율 및 환각률 평가 결과  
Table 7. Evaluation Results of Grounding Rate and Hallucination Rate in RAG Responses

Metric	Methodology	Result
Grounding Rate	Evidence citation pattern rate	22.00%
Hallucination Rate	Low semantic similarity rate (RAG-BERTScore <0.85)	8.00%

- 근거제시율: 22.00%

총 50개의 답변 중 11개에서 근거 제시 패턴이 발견

- 환각률: 8.00%

총 50개의 답변 중 4개의 rag\_bert\_score가 0.85 미만으로 측정

• 결과 분석

근거제시율은 22.00%로, 모델이 검색된 정보를 활용하여 답변을 생성하면서도 항상 그 출처를 명시적으로 밝히지는 않는 경향을 보였다. 이는 답변의 간결성을 우선하는 모델의 특성일 수 있으나, 사용자가 답변의 신뢰도를 판단할 근거를 제공한다는 측면에서 향후 응답 생성 템플릿의 개선을 통해 보완할 수 있는 지점이다[21]. 반면, 환각률은 8.00%로 상대적으로 낮은 수준으로 나타났다. 이는 모델이 생성한 답변의 92%가 실제 정답과 의미적으로 매우 유사하다는 것을 의미하며, RAG 모델이 사실에 기반한 신뢰도 높은 답변을 생성하

는 데 효과적임을 시사한다. 낮은 환각률은 본 시스템이 교육용 질의응답과 같이 정확성이 중요한 도메인에서 안정적으로 활용될 수 있는 가능성을 보여준다.

3) ROUGE

본 절에서는 생성된 답변 자체의 완성도와 정확성을 정량적으로 측정하기 위해 자연어 처리 분야의 표준 평가지표인 ROUGE(Recall-Oriented Understudy for Gisting Evaluation)와 BERTScore를 사용하여 시스템의 답변 생성 성능을 분석한다.

- ROUGE

생성된 답변과 정답(Reference) 간의 n-gram 중복을 기반으로 점수를 계산한다. ROUGE-1은 unigram, ROUGE-2는 bigram, ROUGE-L은 가장 긴 공통 부분 문자열(LCS)을 측정하여 각각 어휘, 구문, 문장 수준의 유사도를 평가한다[22].

- BERTScore

BERT 임베딩을 사용하여 생성된 답변과 정답 간의 의미적 유사도를 계산한다. 단순한 단어 중복을 넘어 문맥적 의미를 반영하므로, 표현이 다르더라도 의미가 같으면 높은 점수를 받을 수 있어 보다 정확한 성능 측정이 가능하다. 제안하는 RAG 시스템과 일반 챗봇(Gemini)을 동일한 평가 데이터셋으로 비교한 결과는 표 8과 같다.

표 8. 제안 RAG 시스템과 일반 챗봇(Gemini)의 정량적 성능 평가 결과 비교

Table 8. Quantitative performance comparison between the proposed RAG system and a general chatbot (Gemini)

Evaluation Metrics	Proposed RAG System	General-purpose Chatbot (Gemini)
ROUGE-1	0.1600	0.1419
ROUGE-2	0.1400	0.1185
ROUGE-L	0.1600	0.1419
BERTScore	0.9721	0.8453

평가 결과, 제안하는 RAG 시스템은 모든 지표에서 일반 챗봇보다 우수한 성능을 보였다. 어휘 및 구문적 유사도를 측정하는 ROUGE 점수에서 제안 시스템은 일반 챗봇 대비 소폭 높은 점수를 기록하여, 정답의 핵심 어휘를 더 잘 포함하고 있음을 보여주었다. 가장 주목할 점은, 문맥적 의미 유사도를 측정하는 BERTScore에서 제안 RAG 시스템이 0.9721점을 기록하여, 일반 챗봇의 0.8453점 대비 현저히 높은 성능을 달성했다는 점이다. 이는 제안 시스템이 단순히 키워드를 나열하는 것을 넘어, 제공된 교과과정 지식 내에서 질문의 의도에 맞는 핵심적인 내용을 의미적으로 훨씬 더 충실하게 생성하고 있음을 강력하게 시사한다. 이러한 성능 차이는 두 시스템의 작동 방식에서 기인한다. 일반 챗봇은 방대하고 일반적인 지식을 활용하여 유창한 답변을 생성하지만, 그 과정에서 질문의 핵심 범위를 벗어나는 부가 정보를 포함하거나 미묘한 환각을 일으킬 수 있다. 반면, 제안 RAG 시스템은 명확히

정의된 근거(교과 데이터)에 기반하여 보다 간결하고 핵심적인 답변을 생성한다. 결론적으로, 이번 정량 평가는 제안 RAG 시스템이 사실 기반 정보 전달의 정확성 및 충실성 측면에서 일반 챗봇보다 월등한 강점을 가짐을 수치적으로 명확히 입증한다. 다만 이러한 결과는 제한된 평가 표본과 자동 평가지표에 기반한 것이므로, 이를 시스템의 절대적 우수성으로 단정하기보다는 제안 시스템이 교육과정 기반 질의응답 맥락에서 상대적으로 더 안정적인 응답 특성을 보였음을 시사하는 탐색적 결과로 해석할 필요가 있다.

## V. 결 론

본 연구는 범용 생성형 AI가 지니는 교육적 맥락의 부재와 환각 가능성의 한계를 보완하고자, 교과 단계별 교과 데이터를 외부 지식으로 활용하는 교육과정 기반 RAG 시스템을 설계하고 구현하였다. 제안 시스템은 학년 및 교과 필터링을 통해 학습자 수준에 적합한 정보를 검색하고, 사실에 기반한 설명형 응답과 학습자의 사고를 촉진하는 질문유도형 응답을 선택적으로 제공하는 이중 응답 전략을 채택하였다.

정성 및 정량 평가 결과, 제안 시스템은 동일한 평가 조건에서 일반 챗봇 대비 더 높은 BERTScore와 ROUGE 점수를 보였으며, 교육과정 기반 질의응답 맥락에서 상대적으로 높은 응답 일치도를 나타냈다. 또한 근거제시율과 환각률을 통해 RAG 응답의 근거 활용 및 의미적 불일치 가능성을 탐색적으로 살펴보았다.

본 연구의 의의는 다음과 같다. 첫째, 공교육 체계에 맞는 학년별·교과별 지식만을 선별적으로 제공하는 RAG 구조를 구현함으로써 교육 현장에서 요구되는 정보의 적합성과 활용 가능성을 제시하였다. 둘째, 기존의 정답 제공 중심의 질의응답 시스템에서 벗어나, 질문유도형 상호작용을 도입함으로써 AI가 학습자의 지식 탐색 과정에 직접 개입하고 사고를 유도하는 튜터(Tutor)의 역할을 수행할 수 있음을 보여주었다. 이는 AI 챗봇이 단순한 정보 검색 도구를 넘어, 자기주도적 학습을 지원하는 능동적인 학습 파트너가 될 수 있음을 시사한다[23].

그러나 본 연구는 몇 가지 한계를 가진다. 첫째, 정량 평가는 50개의 질의응답 쌍을 기반으로 수행된 초기 탐색 수준의 평가이므로 결과의 일반화에는 제한이 있다. 둘째, 근거제시율은 응답 내 근거 참조 표현의 출현 여부를 기준으로 산정한 값이 지표로서, claim-level groundedness를 직접 측정할 결과와는 구별된다. 셋째, 환각률 역시 BERTScore 임계값을 활용하여 의미적 괴리가 큰 응답을 식별한 운영적 추정치이므로, 사실 오류율과 동일하게 해석하기 어렵다. 따라서 본 연구의 정량 결과는 시스템의 절대적 성능 관정보다 상대 비교와 경향성 확인의 차원에서 이해될 필요가 있다.

향후 연구에서는 실제 학습자를 대상으로 한 사용성 평가를 통해 교육적 효과를 검증하고, claim-level human

annotation, entailment 기반 자동평가, citation precision/recall, 근거 부족 시 응답 보류 또는 거절 능력에 대한 평가를 포함함으로써 RAG 시스템의 신뢰성을 보다 정교하게 검증할 필요가 있다. 또한 멀티턴 대화와 멀티미디어 교과 자료를 통합하는 방향으로 시스템을 확장하는 연구도 의미 있는 후속 과제가 될 것이다.

## 참고문헌

- [1] J. Min and M. R. Kim, "A Qualitative Study on Generative AI-Based Teaching and Learning : Focusing on the Experience of Teachers in School," *The Journal of Korean Association of Computer Education*, Vol. 28, No. 8, pp. 25-40, 2025. <https://doi.org/10.32431/kace.2025.28.8.003>
- [2] S. Sonkar, K. Ni, S. Chaudhary, and R. Baraniuk, "Pedagogical Alignment of Large Language Models," in *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami: FL, pp. 13641-13650, 2024. <https://doi.org/10.18653/v1/2024.findings-emnlp.797>
- [3] J. Lee, C. Lee, and J. Rew, "A Student Service Chatbot for Learning Management System at University Using Retrieval-Augmented Generation-Based Large Language Model," *Journal of Broadcast Engineering*, Vol. 29, No. 5, pp. 581-595, 2024. <https://doi.org/10.5909/JBE.2024.29.5.581>
- [4] S.-H. Lee and C. K. Lee, "Implementation of a Korean Language Education Chatbot Using Retrieval-Augmented Generation (RAG): Focusing on Grammar Correction for Korean Learners," *Journal of AI Humanities*, Vol. 19, pp. 35-65, 2025. <https://doi.org/10.46397/JAIH.19.2>
- [5] A. Kerly, P. Hall, and S. Bull, "Bringing Chatbots into Education: Towards Natural Language Negotiation of Open Learner Models," *Knowledge-Based Systems*, Vol. 20, No. 2, pp. 177-185, 2007. <https://doi.org/10.1016/j.knsys.2006.11.014>
- [6] R. Winkler and M. Söllner, "Unleashing the Potential of Chatbots in Education: A State-of-the-Art Analysis," *Academy of Management Proceedings*, Vol. 2018, No. 1, 15903, 2018. <https://doi.org/10.5465/AMBPP.2018.15903aabstract>
- [7] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, ... and G. Kasneci, "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education," *Learning and Individual Differences*, Vol. 103, 102274, 2023. <https://doi.org/10.1016/j.lindif.2023.102274>
- [8] O. Zawacki-Richter, V. I. Marin, M. Bond, and F.

- Gouverneur, “Systematic Review of Research on Artificial Intelligence Applications in Higher Education,” *International Journal of Educational Technology in Higher Education*, Vol. 16, 39, 2019. <https://doi.org/10.1186/s41239-019-0171-0>
- [9] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, ... and V. Tseng, “Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models,” *PLOS Digital Health*, Vol. 2, No. 2, e0000198, 2023. <https://doi.org/10.1371/journal.pdig.0000198>
- [10] A. Gilson, C. W. Safranek, T. Huang, V. Socrates, L. Chi, R. A. Taylor, and D. Chartash, “How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment,” *JMIR Medical Education*, Vol. 9, e45312, 2023. <https://doi.org/10.2196/45312>
- [11] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, ... and P. Fung, “Survey of Hallucination in Natural Language Generation,” *ACM Computing Surveys*, Vol. 55, No. 12, 248, 2023. <https://doi.org/10.1145/3571730>
- [12] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, ... and D. Kiela, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada, pp. 9459-9474, 2020.
- [13] G. Lang and T. Gurplnar, “AI-Powered Learning Support: A Study of Retrieval-Augmented Generation (RAG) Chatbot Effectiveness in Online Course,” *Information Systems Education Journal*, Vol. 23, No. 2, pp. 4-13, 2025. <https://doi.org/10.62273/ZKLK5988>
- [14] H. Soliman, H. Kotte, M. Kravcik, N. Pengel, and N. Duong-Trung, “Retrieval-Augmented Chatbots for Scalable Educational Support in Higher Education,” in *Proceedings of the 2nd International Workshop on Generative AI for Learning Analytics (GenAI-LA)*, Dublin, Ireland, pp. 22-31, 2025.
- [15] J. Swacha and M. Gracel, “Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications,” *Applied Sciences*, Vol. 15, No. 8, 4234, 2025. <https://doi.org/10.3390/app15084234>
- [16] S. Yoon, S. Oh, S. Sohn, and M. Park, “Development of QA System with Education-specific RAG,” *The Journal of the Convergence on Culture Technology*, Vol. 11, No. 1, pp. 781-786, 2025. <https://doi.org/10.17703/JCCT.2025.11.1.781>
- [17] M. M. Cooper and M. W. Klymkowsky, “Let Us Not Squander the Affordances of LLMs for the Sake of Expedience: Using Retrieval Augmented Generative AI Chatbots to Support and Evaluate Student Reasoning,” *Journal of Chemical Education*, Vol. 101, No. 11, pp. 4847-4856, 2024. <https://doi.org/10.1021/acs.jchemed.4c00765>
- [18] S. Muller, A. Loison, B. Omrani, and G. Viaud, “GroUSE: A Benchmark to Evaluate Evaluators in Grounded Question Answering,” arXiv:2409.06595, 2024. <https://doi.org/10.48550/arXiv.2409.06595>
- [19] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating Text Generation With BERT,” in *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020. <https://doi.org/10.48550/arXiv.1904.09675>
- [20] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, “RAGAS: Automated Evaluation Of Retrieval Augmented Generation,” in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, St. Julians, Malta, pp. 150-158, 2024. <https://doi.org/10.18653/v1/2024.eacl-dem.o.16>
- [21] R. T. McCoy, P. Smolensky, T. Linzen, J. Gao, and A. Celikyilmaz, “How Much Do Language Models Copy from Their Training Data? Evaluating Linguistic Novelty in Text Generation Using RAVEN,” *Transactions of the Association for Computational Linguistics*, Vol. 11, pp. 652-670, 2023. [https://doi.org/10.1162/tacl\\_a\\_00567](https://doi.org/10.1162/tacl_a_00567)
- [22] M. Zhang, C. Li, M. Wan, X. Zhang, and Q. Zhao, “ROUGE-SEM: Better Evaluation of Summarization Using ROUGE Combined with Semantics,” *Expert Systems with Applications*, Vol. 237, 121364, 2024. <https://doi.org/10.1016/j.eswa.2023.121364>
- [23] W. Qiu, C. L. Su, N. B. Jamil, M. Thway, S. S. H. Ng, L. Zhang, ... and J. W. Lai, “A Systematic Approach to Evaluate the Use of Chatbots in Educational Contexts: Learning Gains, Engagements and Perceptions,” *Computers*, Vol. 14, No. 7, 270, 2025. <https://doi.org/10.3390/computers14070270>



**하오선(Oh-Sun Ha)**

2000년 : 동국대학교 대학원  
(가정학석사)

2005년 : 동국대학교 대학원  
(이학박사-의류학)

2021년 : 동국대학교 대학원  
(교육학박사-교육공학)

2008년~2012년: 동국대학교 교수학습개발센터 연구원

2012년~현 재: 동국대학교 교무처 교수법연구초빙교수

2015년~현 재: 동국대학교 학부 강사

2021년~현 재: 동국대학교 대학원 강사

※관심분야 : 교수법, 교수매체(생성형AI 활용 교육, 에듀테크 등), 교육과정(마이크로디그리 등)