SymbolicRAG: 추상 구문 트리 기반 구조 인지형 검색증강생성을 활용한 수학 단답형 자동채점

권재우[1], 김영재†
[1]서강대학교 AI・SW대학원
zeuskwon@nongshim.co.kr, youkim@sogang.ac.kr

SymbolicRAG: AST-Based Structure-Aware RAG for Automated Mathematical Short-Answer Grading

Jaewoo kwon[1], Youngjae Kim†
[1]Graduate School Of AI・SW, Sogang University

요 약

최근 Large Language Model(LLM)을 활용한 자동채점이 교육 현장에 도입되고 있으나, 수학 단답형 평가에서는 두 가지 한계를 보인다. 첫째, LLM이 (x-1)(x+1)과 x^2-1, √12와 2√3 같은 구조적으로 동치인 수식을 인식하지 못해 채점 오류가 발생한다. 둘째, 검색증강생성(RAG) 역시 임베딩 기반 검색으로 구조적 동치 풀이를 검색하지 못해 잘못된 참조를 제공한다. 본 연구는 이를 해결하기 위해 SymbolicRAG 프레임워크를 제안한다. SymbolicRAG는 (1) 추상 구문 트리(AST) 기반 재순위화로 수식의 구조적 유사성을 반영하여 관련성 높은 참조 풀이를 우선 검색하고, (2) 컴퓨터 대수 시스템(CAS) 심볼릭 검증으로 학생 답안과 정답의 수학적 동치성을 결정론적으로 판별하며, (3) 검증 결과와 참조 풀이를 통합하여 투명한 채점 근거를 제시한다. 중·고교 수학 150개 테스트 케이스 실험 결과, SymbolicRAG는 채점 정확도 98.7%, 구조 변형 정답 인식률(SER) 98.1%를 달성하여 Non-RAG 대비 정확도 8.3%p, SER 21.2%p 향상되었다. 본 연구는 공교육 AI 자동채점 시스템의 정확성과 설명 가능성을 동시에 확보하는 실용적 방법론을 제공한다.

1. 서론

교육부는 2025년부터 AI 디지털교과서 정책의 일환으로 자동채점 기술을 공교육에 도입하고 있으며, 채점의 설명 가능성과 투명성을 핵심 요구사항으로 제시하고 있다[1]. 교육 현장에서 자동채점 시스템이 실제로 활용되기 위해서는 단순한 정오답 판별을 넘어 채점 근거를 명확히 제시해야 하며, 이를 위해 검색증강생성(RAG)을 통한 유사 문제 풀이 참조가 필수적이다. 그러나 수학 단답형 평가에서 기존 RAG 시스템은 두 가지 한계를 보인다. 첫째, GPT-4o와 같은 최신 LLM도 "(x-1)(x+1)"과 "x^2-1"처럼 수학적으로 동치인 답안을 텍스트 형태 차이로 오답 처리한다[2]. 둘째, 채점 근거를 검색하기 위한 임베딩 기반 RAG 검색 역시 텍스트 유사도에 의존하여 구조적으로 동치인 수식을 유사 사례로 검색하지 못한다. 본 연구는 이러한 한계를 극복하기 위해 SymbolicRAG 프레임워크를 제안한다. 핵심은 두 가지이다. 첫째, 추상 구문 트리(AST) 기반 재순위화로 수식의 구조적 유사성을 반영한 검색을 수행한다. 둘째, 컴퓨터 대수 시스템(CAS)의 심볼릭 검증으로 수식의 수학적 동치성을 판별한다. 이를 통해 SymbolicRAG는 정확한 정답 판별과 구조적으로 관련성 높은 참조 풀이 검색을 동시에 달성하여, 교육 현장이 요구하는 정확하고 설명 가능한 자동채점을 실현한다.

본 연구의 기여는 세 가지로 요약된다. (1) 텍스트 기반 검색의 한계를 보완하기 위해 AST 기반 구조 인지형 재순위화를 제안한다. (2) SymPy 기반 CAS를 도입하여 수식의 대수적·수치적 동치성을 결정론적으로 판별한다. (3) 두 모듈을 결합한 SymbolicRAG가 기존 Non-RAG 및 임베딩 기반 RAG보다 구조 변형 정답에서 현저히 높은 성능을 보임을 실험적으로 입증한다.

2. 연구배경

2-1. 관련연구

자동 단답형 채점(Automatic Short Answer Grading, ASAG) 시스템은 1960년대 통계 기반 텍스트 매칭[6]에서 시작하여 기계학습[7], 딥러닝[8]을 거쳐 최근 대규모 언어모델(LLM) 기반 접근법[9,10]으로 진화해왔다.

그리고 검색증강생성(RAG)은 교육 분야에서 채점 근거 생성과 자동채점[5,17]에 적용되고 있다. Chu et al.(2025)의 GradeRAG[5]는 과학 교육 단답형 문제에서 RAG의 효과를 실증했다. GradeRAG는 (1) 도메인 특화 지식 베이스 구축, (2) 전문가 채점 근거(scoring rationale) 검색을 통한 채점 논리 학습, (3) 단계별 프롬프트 구조를 통해 채점 정확도를 향상시키고 투명한 채점 설명을 제공했다. 이는 RAG가 LLM의 채점 성능뿐 아니라 설명 가능성도 개선할 수 있음을 보여주었다. 그러나 GradeRAG는 임베딩 기반 의미 검색에 의존하여 수학 영역에는 적용이 어렵다. 수학 답안은 수식으로 구성되어 텍스트 형

*본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단(RS-2025-00564249)의 지원을 받아 수행되었다.
† Corresponding Author

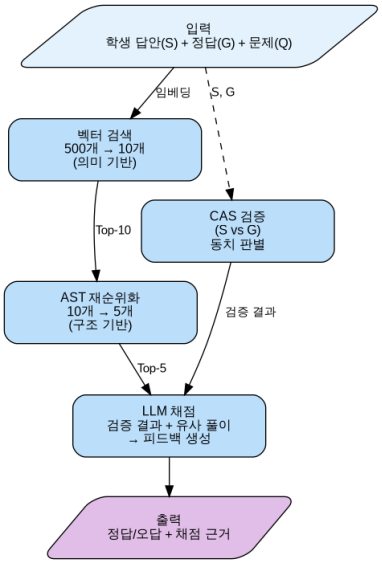
태만으로는 구조적 동치성을 포착할 수 없다.

2-2. 본 연구의 차별점

본 연구는 수식 기반 검색을 잘 할 수 있도록 다음과 같은 방법을 제안한다. 첫째, **AST 기반 구조 인지형 검색**을 도입해서 표현은 다르지만 구조적으로 동치인 풀이를 우선 검색한다. 둘째, SymPy를 활용하여 **CAS 기반 심볼릭 검증**을 통해 학생 답안과 정답의 수학적 동치성을 확인한다. 셋째, **심볼릭 검증 우선 프롬프트**를 설계해서 LLM모델에 입력한다. 이러한 접근은 LLM의 언어 이해 능력과 CAS의 형식적 정확성을 결합하여, 수학 자동채점의 정확성과 투명성을 동시에 확보한다.

3. 제안방법

본 연구에서 제안하는 SymbolicRAG 프레임워크는 수학 단답형 자동채점의 정확성과 투명성을 확보하기 위해 추상 구문 트리(Abstract Syntax Tree, AST) 기반 구조 인지형 검색과 컴퓨터 대수 시스템(Computer Algebra System, CAS) 심볼릭 검증을 통합한 새로운 접근법이다.



3.1. 컴퓨터 대수 시스템(Computer Algebra System, CAS)

컴퓨터 대수 시스템(CAS)은 수학적 표현식을 기호(symbolic) 형태로 처리하여 정확한 대수적 연산을 수행하는 소프트웨어 시스템이다. 일반적인 수치 연산과 달리, CAS는 변수를 포함한 표현식을 심볼릭하게 조작하여 수학적 동치성(equivalence)을 결정론적으로 판별할 수 있다. 본 연구는 Python의 SymPy 라이브러리[20]를 활용하여 다음 두 가지 유형의 검증을 수행한다. SymPy는 순수 Python으로 구현된 오픈소스 CAS로, AST 기반의 표현식 파싱과 다양한 대수적 변환 규칙을 제공한다.

3.2. 벡터 기반 검색 및 AST 재순위화

AST(Abstract Syntax Tree)는 수식의 연산 구조를 계층적 트리로 표현하는 자료구조이다[18]. 학생 답안이 입력되면 임베딩 모델을 사용하여 학생 답안과 문제를 임베딩하고, 벡터 공간에서 코사인 유사도가 높은 상위 10개의 후보 문서를 검색한다. 검색된 10개의 후보 문서에 대해 AST 기반 재순위화를 수행한다. 재순위화 점수는 구조 유사도(70%)와 메타데이터 정합도(30%)의 가중 합으로 계산된다.

$$Score = 0.70 \times f_{구조} + 0.30 \times f_{메타}$$

구조 유사도는 SymPy를 사용하여 두 수식의 동치성을 확인하며, 동치이면 1, 동치가 아닌 경우에는 수식의 구성 요소(변수, 상수, 연산자)를 추출하여 Jaccard 유사도를 계산한다. 메타데이터 정합도는 학생의 학년 및 단원 정보와 문서 메타데이터의 일치도를 평가한다. 재순위화 후 상위 5개의 문서를 선정하여 LLM 프롬프트에 포함한다.

3.3 심볼릭 검증(Symbolic Verification)

학생 답안이 입력되면 답안과 정답의 수학적 동치성을 검증하기 위해 2단계 검증을 순차적으로 수행한다. 첫 번째로 대수적 동치성 검증을 시도하는데, 두 수식의 차이를 계산하고 대수적으로 간소화한다. 간소화 결과가 정확히 0이면 수학적으로 동치라고 판단하며 신뢰도 1.0을 부여한다. 대수적 검증이 실패하면 수치적 검증을 수행하여, 수식의 모든 변수에 무작위 값을 10회 대입하고 결과값을 비교한다. 각 테스트에서 오차가 허용 범위(10^{-10}) 이내이면 통과로 간주하며, 90% 이상 통과 시 동치로 판정한다. 검증 결과는 동치 여부(is_equivalent), 검증 방법(method), 신뢰도(confidence)를 포함하는 구조화된 형태로 생성되며, 재순위화된 유사 풀이와 함께 LLM에 제공되어 최종 채점 및 피드백 생성에 활용된다. 특히 is_equivalent=true인 경우 LLM에게 반드시 정답으로 판정하도록 명시하여, CAS의 결정론적 검증 결과가 최우선 판정기준으로 작동하도록 설계하였다.

3.4 LLM 채점 및 근거 생성

심볼릭 검증 결과와 재순위화된 유사 풀이 5개를 통합한 프롬프트를 GPT-4o(Temperature=0.0)에 입력한다. 프롬프트는 CAS 검증 결과를 최우선 판정 기준으로 명시하며, "is_equivalent=true인 경우 반드시 정답으로 판정"하도록 지시한다. LLM은 정오답 판정, 채점 근거(CAS 검증+유사 풀이 참조)를 JSON 형식으로 반환한다. 이를 통해 정확한 판정과 투명한 설

명을 동시에 제공한다.

4. 평가

본 연구는 OpenAI GPT-4o 모델을 사용하여 수학 단답형 문제 자동 채점 실험을 수행하였다. 실험 데이터셋은 중·고등학교 수학 교육과정을 기반으로 구성된 150개 테스트 케이스를 포함하며, 각 문제는 표준 정답, 구조 변형 정답, 오답의 세 가지 변형을 포함한다. RAG 데이터베이스는 AI Hub 수학 문제 데이터셋과 자체 구축한 전문가 풀이를 포함하여 총 500개 문서(1000개 변형)로 구성하였으며, OpenAI text-embedding-ada-002 모델을 사용하여 임베딩을 생성하였다

4.1 평가지표

실험은 채점 정확도는 Accuracy, Precision, Recall, F1-score로 평가하였으며, 구조 변형 인식률(SER)은 구조적으로 변형된 정답을 올바르게 인식한 비율로 정의하였다.

$$SER = \frac{\textit{올바르게 인식한 구조 변형 정답 수}}{\textit{전체 구조 변형 정답 수}}$$

답안 유형별 인식 성능은 표준 정답, 구조 변형 정답, 오답의 세 범주로 분류하여 각각의 정확도를 측정하였다.

5. 실험 결과

5.1 RAG 조건별 채점성능

조건	Accuracy	Precision	Recall	F1	SER
Non-RAG	0.904	0.989	0.865	0.923	0.769
General-RAG	0.833	0.988	0.760	0.859	0.558
SymbolicRAG	**0.987**	**1.000**	**0.981**	**0.990**	**0.981**

<표1 : RAG 조건별 채점 성능>

실험은 표 1은 세 가지 채점 조건의 성능을 비교한 결과이다. SymbolicRAG는 Accuracy 0.987, SER 0.981로 가장 높은 성능을 달성하였으며, Non-RAG 대비 정확도가 8.3%p, SER이 21.2%p 향상되었다. General RAG는 오히려 Non-RAG보다 낮은 성능(Accuracy 0.833)을 보였는데, 이는 임베딩 기반 검색만으로는 수식의 구조적 동치성을 충분히 포착하지 못하기 때문으로 분석된다.

5.2 답안 유형별 인식 성능

조건	표준정답	구조변형정답	오답
Non-RAG	0.96	0.77	0.98
General-RAG	0.96	0.56	0.98
SymbolicRAG	**0.98**	**0.98**	**1.00**

<표2 : 답안 유형별 인식 성능>

표 2는 답안 유형에 따른 인식 성능을 나타낸다. Non-RAG는 표준 정답(0.96)과 오답(0.98)에서 높은 정확도를 보였으나, 구조 변형 정답 인식률은 0.77에 그쳤다. General RAG는 구조 변형 정답 인식률이 0.56으로 더욱 저하되었다. 반면 SymbolicRAG는 모든 유형에서 0.98 이상의 정확도를 달성하였으며, 특히 구조 변형 정답 인식률(SER)이 0.98로 Non-RAG(0.77) 대비 21%p 향상되었다. 이는 심볼릭 검증이 텍스트 형태와 무관하게 수학적 동치성을 정확히 판별할 수 있음을 보여준다.

6. 결론

본 실험 결과 SymbolicRAG는 모든 평가 지표에서 기존 방법을 상회하는 성능을 달성하였다. 특히 구조 변형 정답 인식률이 0.98로 Non-RAG(0.77) 대비 21.2% 향상되어, "(x-1)(x+1)"과 "x²-1" 같은 대수적으로 동치인 표현을 정확히 인식할 수 있음을 입증하였다. 이는 AST 기반 재순위화와 CAS 기반 심볼릭 검증의 결합이 수식의 구조적 동치성 판별에 효과적임을 보여준다.

참고문헌

[1] 교육부, "AI 디지털교과서 추진방안", 2023.

[2] Li, Y., et al., "Evaluating Mathematical Reasoning in Large Language Models", ACL, 2024.

[3] Gao, Y., et al., "Retrieval-Augmented Generation for AI-Generated Content: A Survey", arXiv, 2024.

[4] Chu, X., et al., "GradeRAG: Enhancing Automatic Short Answer Grading with Retrieval-Augmented Generation", 2025.

[5] Page, E. B., "The Imminence of Grading Essays by Computer", Phi Delta Kappan, 1966.

[6] Mohler, M., et al., "Semantic Textual Similarity for Short Answer Grading", NAACL, 2011.

[7] Sultan, M. A., et al., "Fast and Easy Short Answer Grading with High Accuracy", NAACL, 2016.

[8] Brown, T., et al., "Language Models are Few-Shot Learners", NeurIPS, 2020.

[9] OpenAI, "GPT-4 Technical Report", arXiv, 2023.

[10] Meurer, A., et al., "SymPy: Symbolic Computing in Python", PeerJ Computer Science, 2