

Original Article

분수의 곱셈 오류에 대한 생성형 AI의 진단 및 피드백 역량 분석: ChatGPT, Gemini, DeepSeek를 중심으로

황성환¹, 이지은², 은혜 플라빈³, 강윤지^{4†}

¹춘천교육대학교 교수, ²Oakland University 교수, ³Georgia Institute of Technology 연구교수, ⁴서울소의초등학교 교사

Analysis of Generative AI's Diagnostic and Feedback Capabilities Regarding Fraction Multiplication Errors: Focusing on ChatGPT, Gemini, and DeepSeek

Sunghwan Hwang¹, Ji-Eun Lee², Eunhye Flavin³, Yunji Kang^{4†}

¹Professor, Chuncheon National University of Education, Chuncheon, South Korea, ²Professor, Oakland University, Rochester, MI, ³Senior Research Scientist, Center for Education Integrating Science, Mathematics, and Computing, Georgia Institute of Technology, Atlanta, GA, USA, ⁴Teacher, Seoul Soeui Elementary School, Seoul, South Korea

Abstract

본 연구는 생성형 AI의 수학교육적 활용 가능성을 탐색하고자, '분수의 곱셈' 영역에서 생성형 AI의 오류 진단 및 피드백 제공 능력을 분석하였다. 선행연구를 바탕으로 분수의 곱셈에서 발생하는 오류를 계산형 오류와 문장제 오류로 분류하고, ChatGPT-4, Gemini-2.0 Flash, DeepSeek-R1에 동일한 프롬프트와 오류 유형을 입력하여 각 모델의 응답을 수집하였다. 분석 결과, 세 종류의 생성형 AI는 전반적으로 복잡적이고 일관되지 않은 오류 진단 양상을 보였다. 계산형 오류의 경우, 세 모델 모두 낮은 진단 정확도를 보였으며, 모델 간 성능 차이는 크지 않았다. 반면 문장제 오류에서는 상대적으로 높은 진단 정확도를 보였고, AI 모델 간 역량 차이도 뚜렷하게 나타났다. 피드백 제공 측면에서도 문장제 오류에 대한 피드백이 계산형 오류보다 더 적절하고 구체적인 것으로 분석되었다. 이는 AI 모델의 유형별 강점과 약점이 교수적 맥락에 따라 구분되는 양상을 보이며, AI의 교육적 성공은 기술의 도입 자체가 아닌 그것을 선별하고 지도하는 교사의 전문적 역량에 달려있음을 시사한다.

핵심어: 생성형 AI, 분수의 곱셈, 수학적 오류, 피드백

This study aims to explore the potential of generative AI in mathematics education by analyzing its ability to diagnose student errors in solving fraction multiplication problems and provide feedback. Drawing on prior research, this study first categorized errors in fraction multiplication into computational errors and word problem errors. Prompts reflecting these error types were input into three generative AI models—ChatGPT-4, Gemini-2.0 Flash, DeepSeek-R1—and the responses from each model were collected for analysis. The results showed that the three generative AI models exhibited complex and inconsistent patterns in error diagnosis overall. In the case of computational errors, all three models demonstrated low diagnostic accuracy, with minimal differences in performance between models. In contrast, for word problem errors, the models showed relatively high diagnostic accuracy, and clear differences in capabilities were observed among the models. In terms of the feedback, the feedback for word problem errors was found to be more appropriate and specific than that for computational errors. This suggests that the strengths and weaknesses of AI models vary depending on the instructional context, and that the educational success of AI depends not on the adoption of the technology itself, but on the professional expertise of teachers who select and guide its use.

Keywords: generative AI, fraction multiplication, mathematical errors, feedback

Received: May 7, 2025 Revised: July 22, 2025 Accepted: July 23, 2025

†Correspondence: Yunji Kang, angie0718@sen.go.kr

ORCID: <https://orcid.org/0009-0006-4859-6255>

I. 서론

그동안 수학교육에서 테크놀로지 활용을 위한 다양한 연구들이 이루어져 왔다(Kang, 2023; National Council of Teachers of Mathematics [NCTM], 2024; Roschelle et al., 2017). 테크놀로지는 학생들이 수학 개념과 원리를 탐색하고 수학적 아이디어를 정당화하며 동료와의 실시간 협업을 가능하게 할 뿐만 아니라 수학 학습에 대한 흥미, 동기, 관심을 높이는 데에도 기여한다(Higgins et al., 2019; Roschelle et al., 2017). 이런 효과 외에도, 테크놀로지는 교사의 교수 활동을 지원하는 데 자동화된 채점과 피드백 기능 등을 통해 교사의 역할을 보완할 수 있다(Cullen et al., 2020; Drijvers, 2015). 이에 따라 연구자들은 수학교육에서 테크놀로지의 통합 필요성을 지속적으로 강조해 왔다(Kwon et al., 2023; NCTM, 2024).

최근에는 빅데이터 기반의 딥러닝 기술 발전에 힘입어 수학교육에서 인공지능(Artificial Intelligence, AI)을 활용하려는 시도가 더욱 활발히 이루어지고 있다(Bisconti et al., 2024; Oh et al., 2024). 지능형 교수 시스템(intelligent tutoring system), 적응형 학습 시스템(adaptive learning system), 자동 채점 및 피드백 시스템(Automated scoring and feedback systems), 생성형 AI(Generative AI)와 같은 다양한 AI 기반 플랫폼이 교육 분야에서 연구되고 있다(Pepin et al., 2025). 특히, ChatGPT, Gemini, DeepSeek와 같은 텍스트 기반 생성형 AI가 연구자들의 많은 관심을 받고 있는데(McGalliard & Otten, 2025), 생성형 AI는 거대 언어 모델(Large Language Model, LLM)을 기반으로 인간처럼 화자의 의도를 파악하고 문맥에 적합한 응답을 생성하는 기능을 갖추고 있다. 전통적인 테크놀로지(예: GeoGebra, GSP, 공학용 계산기)와 달리 생성형 AI는 어느 정도의 자율성과 적응성을 갖고 사람과 유사한 반응과 정보를 제공할 수 있으며 사용자와 실시간 상호작용이 가능하다(Bisconti et al., 2024). 이러한 특성으로 인해 생성형 AI는 학생들에게 개인화된 학습 기회와 즉각적인 피드백을 제공하며 학습자는 이를 통해 자신의 이해를 심화시킬 수 있는 교육적 기회를 갖게 된다. 이러한 가능성에 주목하여 학자들은 생성형 AI의 등장이 교수-학습 방식의 근본적인 전환을 이끌고 있다고 평가한다(Pepin et al., 2025). 그러나, 생성형 AI가 가진 기술적 한계, 부정확한 정보 제공, 환각(hallucination) 현상, 응답의 일관성 부족 등의 문제로 인해 교육 현장에서의 적극적인 활용에 대해 신중한 태도를 보이는 시각도 존재한다(Almarashidi et al., 2024; Lee et al., 2024; Pepin et al., 2025).

이러한 기대와 우려 속에서 생성형 AI의 역량을 분석하고자 다양한 연구들이 수행되어 왔다(Kang, 2023, 2024; Kwon et al., 2023; Lee et al., 2024; Son, 2023). 이들 연구는 여러 수준과 영역의 과제를 제시한 후, 생성형 AI가 이를 얼마나 정확하게 해결할 수 있는지 검토하는 데 중점을 두었다(Kang, 2024; Kwon et al., 2023; Oh et al., 2024). 그러나, Kwon et al.(2023)과 Oh et al.(2024)의 연구처럼 AI의 수학 문제 풀이 성능 검증에 집중된 기존의 접근을 넘어 수학교육의 본질적인 요소인 교수학적 관점에서 생성형 AI가 학생의 수학적 오류를 명확하게 진단하고 적절한 피드백을 제공할 수 있는지를 심층적으로 분석할 필요가 있다. 이에 본 연구는 분수의 곱셈 영역에서 나타나는 학생들의 오류를 중심으로 생성형 AI가 제시하는 진단 및 피드백의 역량을 분석하고자 한다.

II. 선행 연구

1. 생성형 AI의 종류와 특성

생성형 AI는 기존 데이터를 기반으로 학습하여 텍스트, 이미지, 오디오 등 다양한 형태의 콘텐츠를 생성하는 기술이다(Wardat et al., 2023). 생성형 AI는 특정 목적과 기능에 따라 설계되므로 모델의 종류와 특성이 다양하며 각 모델은 고유한 강점과 한계를 지닌다. ChatGPT는 OpenAI에서 개발한 텍스트 기반의 대화형 생성형 AI로, 질의응답, 창의적 글쓰기, 대화 처리 등 다양한 언어 기반 작업을 수행한다. 자연스러운 대화 흐름과 폭넓은 주제에 대한 정보 제공에서 강점을 보이지만 복잡한 수학 문제나 고도화된 기술적 문제 해결에는 한계를 보일 수 있다(Kang, 2024; Kwon et al., 2023). 2024년 5월에 공개된 GPT-4o는 텍스트, 이미지, 음성 등 다양한 형태의 데이터를 통합적으로 처리할 수 있는 멀티 모달(multi-modal) 생성형 AI이다. 이 모델은 기존의 ChatGPT-3.5에 비해 고차원적 추론 및 창의적 작업 수행 능력에서 우수한 성능을 나타내며 특히 다중 정보원을 활용한 복합 과제 처리에 효과적이다. 그러나 일부 기능은 특정 분야에서 충분한 검증이 이루어지지 않았으며 모든 기능을 무제한으로 활용하기 위해서는 유료 구독이 요구된다는 점에서 실용성에 제한이 존재한다.

Gemini는 Google DeepMind에서 개발한 멀티 모달 생성형 AI로 텍스트와 이미지 데이터의 통합 처리를 통해 고급 언어 이해 및 논리적 추론 능력을 제공한다. 다양한 산업 및 연구 분야에서 과제 분석과 구조화된 문제 해결에 강점을 보이며 높은 정확도와 효율성을 보이고 있다. 특히, Gemini는 구글 클래스룸(Google Classroom)과 연동되어 있어 학교 현장에서 교사가 학생들의 학습을 더욱 편리하게 관리하고 지도할 수 있다. 다만, 분석적·전문적 과제에 최적화된 설계로 인해 일상적인 대화 기능은 상대적으로 제한적일 수 있다(Ergene & Ergene, 2025; McGalliard & Otten, 2025).

DeepSeek는 중국의 한 스타트업체에서 개발한 생성형 AI로 수학 및 코딩과 같은 논리적 추론에 특화되어 있으며 데이터 기반의 문제 해결 상황에서 높은 정밀도를 보이는 것으로 알려져 있다. 이러한 특화된 역량으로 인해 학문적 연구뿐만 아니라 교육 영역에서도 그 활용 가능성이 주목 받고 있다. 그러나 일상적인 대화나 범용적 언어 생성에는 적합하지 않으며 최근에는 개인정보 보호와 관련된 우려로 인해 일부 국가에서는 규제 논의가 진행되고 있다(Lee, 2025).

이 외에도 Claude, MathGPT, Khanmigo, Magic School AI 등 다양한 생성형 AI 모델이 교육 분야에서 활용되고 있다. 각 생성형 AI 모델은 고유의 기능과 적용 분야를 지니고 있으며 특정 목적에 최적화된 특성을 갖는다. 따라서 사용자(교사, 연구자 등)는 교육적 목적과 활용 맥락에 적합한 모델을 신중히 선택해야 하며 특히 새롭게 개발된 모델의 경우 기술적 안정성과 교육적 타당성, 윤리적 안전성을 확보하기 위한 충분한 검토와 검증 절차가 선행되어야 한다.

2. 수학 교육에서 생성형 AI에 대한 실증 연구

최근 들어 국내외 수학교육 분야에서 인공지능에 대한 관심이 급격히 증가하고 있다(예: Almarashidi et al., 2024; Kang, 2023, 2024). 특히 ChatGPT와 같은 대규모 언어 모델의 등장은 수학교육에 혁신적인 변화를 불러올 가능성을 제시하고 있으며 이러한 가능성과 잠재력에 대한 탐색은 생성형 AI의 수학교육적 활용을 중심으로

한 실증 연구로 이어지고 있다(Pepin et al., 2025).

현재까지 수행된 연구들은 대체로 생성형 AI의 문제 해결 능력에 초점을 맞추고 있다(Kwon et al., 2023; Lee & Park, 2024; Lee et al., 2024; Oh et al., 2024; Shin et al., 2024). 예를 들어, Kang(2024)는 ChatGPT 3.5에 초등학교 수학 교과서에 수록된 문장제를 해결하도록 한 후, 그 풀이 과정과 오류를 분석하였다. 그 결과, 정답률은 81.08%였으며, 문제 이해 오류, 식 수립 오류, 계산 오류 등의 한계가 발견되었다. 이와 유사하게 Kwon et al.(2023)는 ChatGPT-3.5의 국가수준 학업성취도 평가 및 대학수학능력시험 문항 해결 능력을 분석하였고 각각 37.1%와 15.9%의 정답률을 보였다. 이 과정에서 계산상의 실수, 단계 간 연결의 오류, 텍스트 인식 및 판단 오류가 나타나 정답률뿐만 아니라 풀이 과정의 질적 분석이 중요함을 시사하였다.

또한, AI 모델 간 성능 비교에 대한 관심도 증가하고 있다. Lee와 Park(2024)은 ChatGPT 3.5와 4.0의 수학 능력을 비교 분석한 결과 ChatGPT-4.0이 3.5보다 높은 연산 능력을 보였으며 특히 함수 영역과 기호 표현에서 우수한 성능을 나타냈다. 다만, 수와 연산 영역에서는 4.0이 3.5에 비해 상대적으로 낮은 성능을 보였으며 두 모델 모두 계산 과정에서 일부 오류가 나타났다. 또한, Oh et al.(2024)는 고등학교 1학년 수준의 수학 문제를 기반으로 ChatGPT-4, Claude 3 Opus, Gemini Advanced의 문제 해결 능력을 비교한 결과 ChatGPT 4가 가장 높은 정답률을 보인다고 보고하였다. 국외에서도 유사한 연구가 이루어졌는데 Ergene과 Ergene(2025)은 GPT-4o, GPT-4, GPT-3.5, MathGPT, Gemini를 대상으로 390개의 수학 문제 해결 능력을 비교하였다. 연구 결과, GPT-4o가 다소 높은 정답률을 보였으나 GPT-4와는 유사한 수준이었으며 MathGPT와 GPT-3.5의 정답률은 그보다 낮았다. 반면 Gemini는 가장 낮은 정답률을 보여주었다. 이를 종합하면, 생성형 AI 모델마다 정답률에 차이가 있으며 문제 해결 과정에서 오류를 범할 수 있음을 알 수 있다.

한편, 단순한 문제 해결 능력을 넘어서 생성형 AI의 오류 진단 및 교수학적 활용 가능성을 탐색하는 연구들도 진행되고 있다. Shin et al.(2024)는 교육 평가에 초점을 두고 GPT-4를 활용하여 고등학교 1학년 순열과 조합 단원의 서술형 문항 자동 채점 가능성을 검토한 결과 교사 채점과 GPT-4의 채점 간 높은 상관관계가 나타난다고 보고하였다. 또한, Lee et al.(2024)는 분수의 나눗셈 문장제에 나타나는 대표적인 오류 유형을 ChatGPT-4o가 진단하고 피드백할 수 있는지를 분석하였다. 연구 결과, 문제 이해 및 전략 진단 측면에서는 우수한 성능이 나타났으나 계획 실행 및 반성 과정, 탐색 역량과 관련된 오류에 대해서는 적절한 피드백을 제공하는 데 한계가 있었다. 아울러, 오류의 유형과 개수에 따라 진단 성능이 달라지는 경향도 확인되었다.

이러한 연구들은 생성형 AI가 수학교육에서 단순한 문제 해결 도구를 넘어 학습자의 이해를 심화시키고 교사의 교수 활동을 지원하는 도구로서의 역할이 더욱 확대될 수 있음을 보여준다. 또한, 생성형 AI가 특정 오류에 대해서는 효과적인 피드백을 제공할 수 있지만 모든 오류 유형에 대해 일관된 성능을 보이지 않는다는 점은 교육 현장에서 이를 활용하는 교사와 연구자들이 생성형 AI의 한계와 오류 가능성을 명확히 인식할 필요가 있음을 시사한다.

3. 분수의 곱셈 문제해결에서 나타나는 오류 유형

수학은 계통성이 강한 학문으로 학습 초기 단계에서 수학적 개념에 대한 명확한 이해 없이 학습이 진행될 경우 이후의 학습 과정에 부정적인 영향을 미칠 수 있다(Park & Park, 2017). 따라서 교사는 학생들이 문제 해결 과정에서 보이는 오류를 면밀히 분석하고 이를 효과적으로 지도할 수 있는 교육적 방안을 모색할 필요가

있다(An & Choi, 2016). 수학 교육자들은 오랫동안 학생들의 수학적 오류를 분석하기 위해 다양한 연구를 수행해 왔으며 특히 분수의 곱셈과 관련하여 학생들이 보이는 오류 유형을 계산형 오류와 문장제 오류로 구분하여 분석해왔다¹⁾(e.g., Park & Park, 2017; Simon et al., 2018; Van de Walle et al., 2023).

1) 계산형 오류

계산형 오류란 순수 계산 문제에서 부적절한 알고리즘이나 계산 절차를 적용함으로써 발생하는 오류를 의미한다. 다만, 문장제에서 올바른 수식을 세웠다 하더라도 계산 과정에서 잘못된 연산이 이루어진 경우 역시 계산형 오류로 분류된다(예: An & Choi, 2016; Park & Park, 2017; Park et al., 2023). 이러한 오류는 주로 분수 곱셈 개념에 대한 명확한 이해 부족이나, 자연수의 곱셈, 분수의 덧셈과 뺄셈 등 다른 연산에서 사용되는 계산 방식을 분수 곱셈에 부적절하게 적용할 때 발생한다. 본 연구에서는 7가지 유형의 계산형 오류를 살펴보았으며 자세한 설명은 Table 1에 제시되어있다.

Table 1 Error types in fraction multiplication computational problems

오류 유형(참고문헌)	예시
1. 대분수를 가분수로 바꾸지 않고 자연수, 분모, 분수를 각각 곱하는 오류 (An & Choi, 2016)	$1\frac{4}{7} \times 1\frac{1}{2} = 1\frac{4}{14}$
2. 대분수의 자연수, 분자, 또다른 자연수를 곱하는 오류 (Park & Park, 2017)	$6\frac{2}{5} \times 3\frac{36}{5} = 7\frac{1}{5}$
3. 약분 오류: 대분수를 가분수로 바꾸지 않은 상태에서 약분을 한 후, 곱셈을 수행하는 오류 (Park & Park, 2017)	$1\frac{4}{3} \times 2 = 1\frac{4}{3} \times 2 = \frac{7}{3} \times 2 = \frac{14}{3} = 4\frac{2}{3}$
4. 역수 변환 오류: 분수의 곱셈에서 역수를 취해 답을 구하는 오류 (An & Choi, 2016; Park & Park, 2017)	$3 \times 2\frac{1}{6} = \frac{3}{1} \times \frac{6}{13} = \frac{18}{13} = 1\frac{5}{13}$
5. 분수를 통분한 후 분모끼리는 곱하지 않고 분자끼리만 곱하는 오류 (Reys et al., 2015)	$\frac{2}{3} \times 4 = \frac{2}{3} \times \frac{12}{3} = \frac{2 \times 12}{3} = \frac{24}{3} = 8$
6. 분수의 덧셈과 적용해 통분을 활용해 해결하는 오류 (Park & Park, 2017)	$\frac{4}{9} \times \frac{5}{8} = \frac{32}{72} \times \frac{45}{72} = \frac{77}{72} = 1\frac{5}{72}$
7. 자연수와 분수의 곱셈에서 자연수를 분모에 곱하는 경우 (An & Choi, 2016; Park & Park, 2017; Reys et al., 2015)	$2 \times \frac{3}{5} = \frac{3}{10}$

2) 문장제 오류

문장제 오류란 문제의 맥락과 결과를 정확하게 이해하거나 해석하지 못해 발생하는 해석 오류로서 분수의 곱셈 알고리즘은 알고 있으나 이를 상황에 알맞게 적용할 줄 모르거나 분수의 곱셈에 대한 개념적 이해가 부족한 상황에서 발생한다. 분수의 곱셈 문제 해결에서 나타나는 문장제 오류 유형은 상황 인식 오류, 결과 해석의 오류 등이며 자세한 설명은 Table 2에 제시하였다(e.g., Haser & Ubuz, 2003; Mack, 2000; Prediger, 2008; Simon et al., 2018; Van de Walle et al., 2023).

1) 선행연구를 토대로 분수의 곱셈 오류 유형을 분류했지만, 오류 해석 및 분류는 연구자마다 관점이 다를 수 있어 일부 유형에서 중복되거나 경계가 명확하지 않은 경우가 존재한다.

Table 2 Error types in fraction multiplication word problems

오류 유형(참고문헌)	예시
8. 상황 인식 오류: 문제의 상황을 적절한 연산으로 표현하지 못하는 오류 (Simon et al., 2018; Van de Walle et al., 2023)	〈문제〉엄마가 피자의 절반을 지호에게 주었다. 지호는 엄마가 준 피자의 4분의 1을 먹었다. 지호는 전체 피자의 얼마만큼을 먹었는지 구하시오. 〈학생 답변〉 $\frac{1}{2} + \frac{1}{4} = \frac{3}{4}$
9. 결과 해석의 오류: 계산은 했으나, 상황을 설명하지 못하는 오류 (Simon et al., 2018; Van de Walle et al., 2023)	〈문제〉어떤 음식을 만들기 위해서는 설탕 3분의 2컵이 필요하다. 이 음식의 $\frac{3}{4}$ 만 만들고 싶다면, 설탕이 얼마나 필요한가? 〈학생의 답변〉 $\frac{6}{12}$ 만큼 필요해요. 설탕 $\frac{2}{3}$ 의 $\frac{6}{12}$ 만큼 필요해요.
10. 기준 단위 해석 오류: 곱셈상황에서 기준 단위를 잘못 해석하는 오류 (Haser & Ubuz, 2003; Simon et al., 2018)	〈문제〉호두 한 자루의 $\frac{3}{4}$ 이 69개라면, 그 자루의 $\frac{4}{23}$ 는 몇 개인가 〈학생 답변〉 $69 \times \frac{4}{23} = 12$
11. 단위 혼동 오류: 서로 다른 종류의 단위 간 연산 수행 오류 (Haser & Ubuz, 2003)	〈문제〉자동차의 속력이 시속 90km일 때, $3\frac{1}{5}$ 동안 얼마나 멀리 갈 수 있는가? 〈학생의 답변〉1시간의 5분의 1은 12이다. 3시간 동안 시속 90km로 달리면 270를 간다. 따라서 $12+270$ 인 282를 갈 수 있다. 답: 282km

4. 연구 문제

기존의 많은 연구들은 수학 문제에 대한 생성형 AI의 역량을 평가할 때 주로 정답률을 기준으로 분석하였다. 그러나 이러한 접근법만으로는 생성형 AI의 교수학적 활용 가능성을 충분히 조명하는 데 한계가 있다. 이를 보완하여 최근 Lee et al.(2024)은 교수학적 측면에서 분수의 나눗셈 오류를 중심으로 ChatGPT의 진단 및 피드백 제공 능력을 분석한 바 있다. 본 연구는 Lee et al.(2024)의 연구를 확장하여 분수의 곱셈 오류에 대해 세 가지 유형의 생성형 AI가 보이는 진단 및 피드백 역량을 분석하고자 한다.

분수의 곱셈은 여러 선행 연구(An & Choi, 2016; Park & Park, 2017; Simon et al., 2018; Van de Walle et al., 2023)를 통해 학생들이 자주 범하는 오류 유형이 밝혀져 있어, 생성형 AI가 학생의 오류 원인을 정확히 진단하고 교수학적으로 적절한 피드백을 제공할 수 있는지를 체계적으로 평가하기에 적합한 주제이다. 또한 분수의 곱셈은 분수의 나눗셈과 함께 초등학교 교사들이 지도하기 가장 어려운 영역으로 알려져 있어(Izsák, 2008) 교사들이 AI 활용을 통해 실질적인 도움을 받을 가능성이 높을 것으로 예상된다.

이에 본 연구는 NCTM(2024)이 강조한 바와 같이 교육 현장에서 AI 도구의 신뢰성과 활용 가능성을 검증해야 할 필요성을 고려해 분수 곱셈 영역에서 생성형 AI가 보이는 진단 및 피드백 제공 역량을 비교·분석하고자 한다. 이에 따라 현재 높은 점유율과 활용도를 보이는 AI 모델인 ChatGPT와 Gemini, 그리고 중국에서 개발되어 논리적 추론에 강점을 보이는 동아시아 기반의 DeepSeek를 분석에 포함하였다. 본 연구는 이렇게 특성과 개발 배경이 서로 다른 세 가지 생성형 AI 모델을 비교 분석함으로써 각 모델이 지닌 특성이 초등 수학 내용의 교수학적 진단과 피드백 생성에 어떠한 영향을 미치는지 다각적으로 탐구하고자 한다.

연구문제1. 분수의 곱셈 영역에서 나타나는 계산형 오류에 대해 생성형 AI(ChatGPT, Gemini, DeepSeek)는 어떠한 수준의 진단과 피드백 제공 역량을 보이는가?

연구문제2. 분수의 곱셈 영역에서 나타나는 문장제 오류에 대해 생성형 AI(ChatGPT, Gemini, DeepSeek)는 어떠한 수준의 진단과 피드백 제공 역량을 보이는가?

III. 연구 방법

1. 연구 설계 및 자료 수집

선행연구를 바탕으로 분수 곱셈 관련 오류 유형을 수집하였으며 연구자 간 다수의 논의 과정을 거쳐 계산형 오류 유형 7가지(Table 1)와 문장제 오류 유형 4가지(Table 2)를 선정하였다. 생성형 AI 모델은 시간의 경과에 따라 지속적으로 업데이트되므로 기능과 반응 방식에 변화가 발생할 수 있다. 따라서 자료 수집 시점은 연구 결과 해석에 있어 중요한 변수로 작용한다. 본 연구에서는 이러한 점을 고려하여 2025년 2월 약 1주일 동안 비슷한 시점에 ChatGPT-4, Gemini-2.0 Flash, DeepSeek-R1²⁾ 세 모델을 사용해 연구를 수행하였다.

구체적인 자료 수집 절차는 다음과 같다. 첫째, 선행연구를 바탕으로 각 오류 유형별 문제와 학생 응답을 선정하였다(Tables 1, 2 참고). 둘째, ChatGPT를 활용한 분수의 나눗셈 오류 진단 연구(Lee et al., 2024)를 참고하여 구조화된 프롬프트를 구안하였다(Table 3 참고). 이 프롬프트에는 생성형 AI의 역할(초등학교 교사), 상황(초등학생의 분수 곱셈 오류 분석 및 피드백 제공), 적절한 피드백(예: 학생의 사고를 확장하는 질문 제시)과 부적절한 피드백(예: 정답 도출 과정을 일방적으로 제공하는 방식), 출력 규칙(출력 시 순서와 내용)의 예시가 포함되었다.

셋째, 위 프롬프트를 ChatGPT에 입력한 후, 오류 유형별로 해당 문제와 학생의 응답을 함께 제시하였다. 계산형 오류는 생성형 AI가 수식을 정확하게 인식하고 처리하도록 LaTeX 형식으로 변환해 입력하였으며 문장제 오류는 다양한 언어 환경에서의 모델 반응을 살펴보기 위해 영어 원문을 그대로 사용하였다. 단, 분석의 일관성을 유지하고 국내 수학교육 맥락에서의 활용 가능성을 평가하고자 출력 언어는 한국어로 통일하였다. 생성형 AI는 반복 입력 시 응답에 변동성이 발생할 수 있기 때문에 동일한 질문을 여러 차례 제시하고 그 일관성을 분석하는 것이 타당하다. 이에 본 연구에서는 각 생성형 AI 모델에 동일한 내용의 프롬프트를 세 차례 제시하여 응답 간 일관성을 분석하였다. 그 결과, 동일한 날짜와 유사한 시간대에 입력된 응답들은 전반적으로 높은 유사성을 보이는 것으로 확인되었다. 이에 따라, 반복 입력 과정에서 발생할 수 있는 영향을 최소화하기 위해 최초로 입력된 1회차 응답 결과를 최종 분석 대상으로 활용하였다.

2) 이후 본 논문에서는 각 모델명을 ChatGPT, Gemini, DeepSeek로 통일하고, 버전 표기는 생략한다.

Table 3 Prompts presented to Generative AI (Adopted from Lee et al., 2024)

〈안내〉

당신은 초등학교 교사이다. 분수의 곱셈 문제를 해결하는 초등학교 학생의 풀이 과정을 보고 다음을 수행하라.

1. 학생의 문제해결 전략을 분석하여 어떤 오류가 발생하고 있는지 설명하라. 또한, 오류가 나타난 잠재적인 원인에 대해 상세히 설명하라.
2. 〈적절한 피드백과 부적절한 피드백의 특징〉을 참조하여, 학생이 자신의 문제해결 과정에서 발생한 오류를 명확히 이해하고, 스스로 문제를 해결할 수 있도록 적절한 피드백을 제공하라. 피드백의 의도를 자세히 설명하라.
3. 모든 답변은 한글로 출력하라.

〈적절한 피드백의 특징〉

1. 학생이 스스로 오류를 이해할 수 있도록 돕는 질문 혹은 정보를 제공하되, 구체적인 절차적 과정은 제공하지 않는 피드백
2. 학생의 사고를 확장하는 질문을 제공하는 피드백
3. 학생에게 자신의 생각에 대한 이유와 정당화를 요구하는 피드백

〈부적절한 피드백의 특징〉

1. 학생에게 문제해결의 모든 과정을 제공하는 피드백
2. 단순히 학생의 오류를 교정하는 피드백
3. 수학적 오류를 포함하는 피드백

〈출력 규칙〉

1. 학생의 문제해결 전략을 분석하여 어떤 오류가 발생하고 있는지 설명하라. 각 오류가 나타난 잠재적인 원인에 대해 상세히 설명하라.
오류:
오류가 나타난 잠재적인 원인:
2. 학생이 자신의 문제해결 과정에서 발생한 오류를 명확히 이해하고, 스스로 문제를 해결할 수 있도록 적절한 피드백 두 가지를 제공하라. 피드백의 의도를 자세히 설명하라.
피드백1:
피드백1의 의도:
피드백2:
피드백2의 의도:

2. 자료 분석

수집된 자료는 선행 연구(예: Ergene & Ergene, 2024; Kwon et al., 2023; Lee et al., 2024)를 바탕으로 수학적 오류 진단과 피드백 두 가지 범주로 나누어 분석하였다. 먼저, 오류 진단의 경우, 학생의 오류를 정확하게 식별하고 타당한 원인을 명확하게 설명한 응답 경우는 2수준으로, 오류는 식별하나 그 원인을 설명하지 못하는 경우는 1수준으로, 오류를 식별하지 못한 경우는 0수준으로 분류하였다. 한편, 생성형 AI에게 오류 진단을 요청하였을 때 동일한 프롬프트를 사용했음에도 불구하고 AI 모델에 따라 오류의 원인에 대한 설명의 수(가짓수)와 방식에서 차이를 보였다. 이에 따라 분석의 일관성을 위해 가장 먼저 제시된 응답만을 분석에 활용했다.

또한, 생성형AI가 제공하는 피드백 수준을 분석하기 위해 각 오류 유형별로 생성형 AI에게 두 개의 피드백을 요청하였으며, 각 피드백은 0, 1, 2 수준으로 구분하여 코딩하였다. 2수준은 학생의 사고를 확장하거나 오류를 이해할 수 있도록 돕는 질문을 제시하는 경우, 학생의 사고를 바탕으로 자신의 생각을 정당화하도록 유도하는 경우, 학생 스스로 오류를 인식할 수 있도록 탐색을 유도하는 피드백에 해당한다. 1수준은 문제 해결을 위한 절차적 과정만을 안내하거나, 오류의 수정을 직접 요구하는 경우(예: $\frac{2}{3} \times \frac{12}{3} = \frac{2 \times 12}{3}$ 의 오류에 대해 별다른 설명 없이 $\frac{2 \times 12}{3 \times 3}$ 로 곱하도록 요구하는 경우), 학생의 사고 과정은 고려하지 않고 계산 결과에만 초점을 둔 경우에 해당한다. 0수준은 오류를 적절히 진단하지 못하거나 피드백 내용이 비논리적이거나 부적절한 경우에 해당한다.

먼저, ChatGPT의 응답에 대해 네 명의 연구자가 공동으로 코딩을 수행하고 충분한 논의를 거쳐 합의에 도달한 후, DeepSeek와 Gemini의 결과는 두 명의 연구자가 각각 나누어 코딩하였다. 연구자 간의 판단 일치도를 확인하기 위해 카파 계수(Kappa coefficient)를 산출한 결과 DeepSeek와 Gemini의 카파 계수는 각각 0.82와 0.85로 두 모델 모두 높은 수준의 일치도를 나타냈다(Landis & Koch, 1977). 먼저, 양적 코딩 결과를 기반으로 각 모델의 성능을 비교하였으며 이후 질적 분석을 통해 각 생성형 AI의 특징과 차이점을 보다 면밀히 검토하였다.

IV. 연구 결과

1. 계산형 오류에 대한 진단 및 피드백 수준

계산형 오류에 대한 생성형 AI의 진단 및 피드백 제공 수준을 분석한 결과는 Table 4, Figure 1과 같다. 오류 진단과 관련하여 세 모델 모두 7개의 문제 중 절반 가량만 오류 원인을 정확히 진단하였다. 총 14점 만점 기준으로 ChatGPT는 5점, DeepSeek는 6점, Gemini는 각각 7점을 획득하였다. 이는 세 생성형 AI 모델 모두 학생의 오류를 진단하는 데 일정한 한계를 보이며, 세 모델 간 성능 차이도 크지 않음을 시사한다. 주목할 점은 오류 유형에 따라 세 모델 모두 진단 정확도가 유사하게 높거나 낮게 나타나는 경향을 보였다는 것이다. 예를 들어, 분수의 곱셈이 아닌 다른 연산 알고리즘을 적용해서 생겨나는 오류 (오류 유형 4, 5)의 경우 세 모델 모두 정확하게 진단하지 못했다.

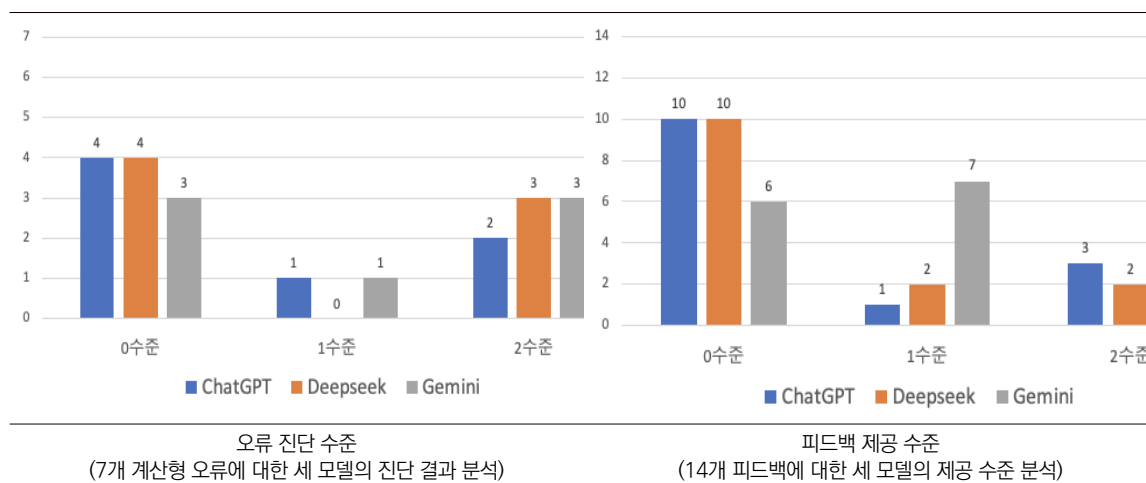
계산형 오류에 대해 생성형 AI가 제공한 피드백을 평가한 결과 Gemini가 9점으로 가장 높은 점수를 기록했고 ChatGPT는 7점, DeepSeek는 6점으로 뒤를 이었다. 피드백의 최고점이 28점임을 고려할 때, 세 모델 모두 절반 이하의 점수를 받은 것으로 확인된다. 높은 점수를 받기 위해서는, 생성형 AI가 학생의 사고를 확장시키고 오류를 이해할 수 있도록 유도하는 질문을 제시하거나, 학생이 자신의 생각을 정당화하도록 요구하는 피드백을 제공해야 한다. 그러나 피드백 수준에 대한 종합 평가 결과 본 연구에서 사용된 세 모델 모두 이러한 기준을 충족하지 못한 것으로 나타났다.

피드백 수준별 빈도를 살펴보면 ChatGPT는 0수준 10개, 1수준 1개, 2수준 3개, DeepSeek는 0수준 10개, 1수준 2개, 2수준 2개, Gemini는 0수준 6개, 1수준 7개, 2수준 1개로 분포되었다. ChatGPT와 Gemini를 비교해본 결과, ChatGPT는 부정확한 피드백(0수준)의 비율이 더 높은 반면, 2수준에 해당하는 높은 수준의 피드백도 많이 제시하는 등 피드백의 질에 편차가 나타나는 양상을 보였다. 반면 Gemini는 상대적으로 0수준과 1수준의 빈도가 균형을 이루며 보다 일관된 피드백을 제공한 것으로 분석되었다. 한편, DeepSeek는 상대적으로 낮은 수준의 피드백을 제공하는 경향을 보였다.

Table 4 Error diagnosis and feedback levels of three Generative AIs for computational errors

오류 유형	ChatGPT		DeepSeek		Gemini	
	진단	피드백	진단	피드백	진단	피드백
1. 대분수를 가분수로 바꾸지 않고 자연수, 분모, 분수를 각각 곱하는 오류	2	1) 1 2) 2	2	1) 1 2) 0	0	1) 1 2) 1
2. 대분수의 자연수, 분자, 또다른 자연수를 곱하는 오류	1	1) 0 2) 0	0	1) 0 2) 0	1	1) 0 2) 0
3. 약분 오류: 대분수를 가분수로 바꾸지 않은 상태에서 약분을 한 후, 곱셈을 수행하는 오류	0	1) 0 2) 0	0	1) 0 2) 0	2	1) 1 2) 1
4. 역수 변환 오류: 분수의 곱셈에서 역수를 취해 답을 구하는 오류	0	1) 0 2) 0	0	1) 0 2) 0	0	1) 0 2) 0
5. 분수를 통분한 후 분모끼리는 곱하지 않고 분자끼리만 곱하는 오류	0	1) 0 2) 0	0	1) 0 2) 0	0	1) 0 2) 0
6. 분수의 덧셈과 적용해 통분을 활용해 해결하는 오류	0	1) 0 2) 0	2	1) 2 2) 1	2	1) 2 2) 1
7. 자연수와 분수의 곱셈에서 자연수를 분모에 곱하는 경우	2	1) 2 2) 2	2	1) 2 2) 0	2	1) 1 2) 1
합계	5	7	6	6	7	9

Figure 1 Error diagnosis and feedback levels of three Generative AIs for computational errors



1) 계산형 오류 진단 사례

오류 유형1(대분수를 가분수로 바꾸지 않고 자연수, 분모, 분수를 각각 곱하는 오류)을 살펴보면 세 모델 모두 학생이 대분수를 곱하는 과정에서 자연수는 자연수끼리, 분자는 분자끼리, 분모는 분모끼리 곱하는 오류를 정확히 진단하였다. 더불어 대분수 곱셈 개념에 대한 학생의 이해 부족 가능성이라는 잠재적 오류 원인도 적절히 파악하였다. 반면, 세 모델 모두 오류 유형 4 (역수 변환 오류), 오류 유형 5 (분수를 통분한 후 분모끼리는 곱하지 않고 분자끼리만 곱하는 오류)에 대해서는 정확한 진단을 하지 못하였다.

오류 유형 $2(6\frac{2}{5} \times 3 = \frac{36}{5} = 7\frac{1}{5})$ 의 경우 학생은 첫 번째 대분수의 자연수와 분자(6×2)를 곱한 뒤 여기에 다시 두 번째 자연수인 3을 곱하는($6 \times 2 \times 3$) 오류를 범하여 36이라는 잘못된 결과를 도출하였다. 그러나 DeepSeek는 이러한 풀이 과정을 올바른 것("풀이는 정확합니다!")으로 판단하였다(Table 5 참조). 반면, ChatGPT와 Gemini는 학생의 오류를 식별하였지만 그 원인을 설명하는 데 있어 논리적 정확성이 부족했다. ChatGPT는 "정수 부분과 분수 부분을 각각 곱한 뒤"라고 표현하였지만 분자와 분모 중 어느 부분과 곱했는지에 대한 구체적 설명이 결여되어 있었다. 또한 Gemini는 "자연수 부분과 분자 부분을 더한 후 분모를 곱하는 방식($((6+2) \times 5 = 40)$)으로 계산했다"고 기술하여 실제 오류 발생 원인을 정확히 파악하지 못하였다. 오류 유형 4와 5의 경우, 세 모델 모두 오류의 원인을 식별하지 못했으며 이에 따라 오류 원인에 대한 설명도 정확하지 않았다.

Table 5 Comparison of error diagnosis by three Generative AIs for computational error type 2 ($6\frac{2}{5} \times 3 = 7\frac{1}{5}$)

모델 (오류 진단 수준)	오류 식별과 잠재적 원인 분석
ChatGPT (1수준)	<p>오류 1:</p> <p>학생은 대분수 $6\frac{2}{5}$에 3을 곱하는 문제에서 계산 순서를 일부 잘못 이해하고 적용했습니다.</p> <ul style="list-style-type: none"> 학생이 계산한 결과는 다음과 같습니다: $6\frac{2}{5} \times 3 = \frac{36}{5} = 7\frac{1}{5}$ <ul style="list-style-type: none"> 그러나 이는 틀린 결과입니다. <p>오류 1이 나타난 잠재적인 원인:</p> <ul style="list-style-type: none"> 학생이 $6\frac{2}{5}$을 가분수로 바꾸지 않고, 정수 부분과 분수 부분을 각각 곱한 뒤 이를 잘못 합친 것으로 보입니다.
DeepSeek (0수준)	<p>학생의 풀이는 정확합니다! $6\frac{2}{5} \times 3$을 계산한 결과 $\frac{36}{5}$로 정확히 계산했고, 이를 대분수로 변환한 $7\frac{1}{5}$도 올바르게 풀었습니다. 이 풀이 과정에는 오류가 없으며, 학생이 대분수를 가분수로 변환하고, 자연수와 분수의 곱셈을 정확히 수행한 후 결과를 대분수로 다시 변환하는 과정을 완벽히 이해하고 있음을 보여줍니다.</p>
Gemini (1수준)	<p>오류 1: 대분수를 가분수로 변환하는 과정에서 분자를 잘못 계산하는 오류</p> <ul style="list-style-type: none"> 오류 1이 나타난 잠재적인 원인: 학생은 대분수를 가분수로 변환하는 방법을 알고 있지만, 계산 과정에서 실수를 했을 가능성이 큼니다. $6\frac{2}{5}$를 $(6 \times 5 + 2)/5 = 32/5$로 계산해야 하는데, 분자를 36으로 잘못 계산했습니다. 아마도 자연수 부분과 분자 부분을 더한 후 분모를 곱하는 방식으로 계산했을 가능성이 있습니다.

오류 유형 3과 6에서는 모델 간 오류 진단 정확도에 차이가 있었다. 예를 들어, 오류 유형 3 (약분 오류)의 경우 Gemini는 오류를 정확하게 진단한 반면, ChatGPT와 DeepSeek는 오류를 정확히 식별하지 못하였다. 한편, 오류 유형 6(분수의 덧셈과 적용해 통분을 활용해 해결하는 오류)에서는 DeepSeek와 Gemini가 오류를 정확하게 진단한 반면, ChatGPT는 이에 실패하였다.

2) 계산형 오류에 대한 피드백 사례

오류 유형 1과 관련해 각 모델이 제공한 두 개의 피드백을 분석한 결과, ChatGPT와 Gemini는 총 3점, DeepSeek는

총 2점을 획득하였다(Table 6 참고). ChatGPT의 두 번째 피드백은 학생이 자신의 사고 과정을 되돌아보게 하고 유사한 원리를 적용할 수 있는 문제를 제시함으로써 학습을 촉진한다는 점에서 2수준으로 평가되었다(“예를 들어, $1\frac{1}{3} \times 1\frac{1}{3}$ 을 같은 방법으로 풀면 어떤 결과가 나오는지 한 번 해볼까요?”). Gemini 역시 대분수의 곱셈을 계산하기에 앞서 대분수를 어떤 형태로 변환해야 하는지를 학생에게 질문함으로써 2수준 피드백을 제공하였다(“대분수를 계산하기 전에 대분수를 어떠한 형태로 바꿔야 할까요?”). 반면, DeepSeek는 절차적 설명에 초점을 맞춘 피드백을 제공하였으며(“먼저 대분수를 가분수로 바꾸는 것이 중요해요”) 이는 1수준으로 분류되었다. 오류 유형 2, 4, 5의 경우 세 모델 모두 0점으로 분류되었으며, 오류 유형 3, 6, 7의 경우 모델별로 피드백 수준에 차이를 보였다. 이러한 결과는 특정 오류 유형에 따라 생성형 AI 모델의 오류 진단 및 피드백 제공 능력이 달라질 수 있음을 시사한다.

Table 6 Comparison of feedback by three Generative AIs for computational error type 1($1\frac{4}{7} \times 1\frac{1}{2} = 1\frac{4}{14}$)

모델 (피드백 제공 수준)	피드백 및 피드백 의도
ChatGPT (1, 2수준)	<p>피드백 1: "이 문제에서 대분수끼리 곱할 때 어떤 규칙을 사용했나요? 혹시 대분수를 가분수로 바꾸는 방법을 기억하고 있나요? $1\frac{4}{7}$을 가분수로 바꾸면 어떤 수가 될까요?"</p> <p>피드백의 의도: 이 질문은 학생이 대분수를 가분수로 변환하는 개념을 스스로 떠올릴 수 있도록 돕기 위한 것입니다. 기존에 사용한 잘못된 전략을 인지하게 하고, 옳은 전략을 찾아가도록 유도합니다.</p> <p>피드백 2: "대분수 $1\frac{4}{7}$과 $1\frac{1}{2}$를 곱했을 때, 정수 부분과 분수 부분을 각각 곱하는 방법이 항상 맞는 방법일까요? 예를 들어 $1\frac{1}{3} \times 1\frac{1}{3}$을 같은 방법으로 풀면 어떤 결과가 나오는지 한 번 해볼까요?"</p> <p>피드백의 의도: 이 질문은 학생이 자신이 사용한 전략의 한계를 깨닫고, 새로운 방법(가분수 변환)을 필요로 한다는 점을 인식하도록 돕기 위한 것입니다. 비슷한 상황에서의 실수를 통해 본인의 오류를 확실히 이해하도록 유도합니다.</p>
DeepSeek (1, 1수준)	<p>피드백1: "대분수를 곱할 때는 먼저 대분수를 가분수로 바꾸는 것이 중요해요. $1\frac{4}{7}$과 $1\frac{1}{2}$를 가분수로 바꿔보면 어떻게 될까요?"</p> <p>피드백의 의도: 학생이 대분수를 가분수로 변환하는 과정을 스스로 생각해볼도록 유도합니다. 이 질문은 학생이 대분수의 구조를 이해하고, 가분수로 변환하는 방법을 떠올리도록 돕는 데 목적이 있습니다.</p> <p>피드백2: "가분수로 바꾼 후에는 분수의 곱셈을 어떻게 해야 할까요? 분자와 분모를 각각 어떻게 처리하면 될지 생각해 보세요."</p> <p>피드백의 의도: 학생이 가분수로 변환한 후 분수의 곱셈을 수행하는 과정을 스스로 탐구하도록 유도합니다. 이 피드백은 학생이 분수의 곱셈 원리를 적용하고, 계산 과정을 명확히 이해하도록 돕는 데 목적이 있습니다.</p>

Table 6 Continued

모델 (피드백 제공 수준)	피드백 및 피드백 의도
Gemini (2, 1수준)	<p>피드백:</p> <ul style="list-style-type: none"> 대분수의 곱셈을 계산하기 전에, 대분수를 어떤 형태로 바꿔야 할까요? 분수의 곱셈은 어떻게 계산하는지 다시 한번 생각해 볼까요? 분자끼리, 분모끼리 어떻게 해야 할까요? <p>의도:</p> <ul style="list-style-type: none"> 학생 스스로 대분수를 가분수로 변환해야 한다는 것을 깨닫고, 올바른 곱셈 방법을 적용하도록 유도합니다. 분수의 곱셈 규칙을 다시 한번 상기시키고, 정확한 계산 과정을 이해하도록 돕습니다.

Note. 표기된 수준은 피드백 1과 2에 대한 평가 수준을 의미함

2. 문장제 오류에 대한 진단 및 피드백 수준

문장제 오류에 대한 생성형 AI의 진단 및 피드백 제공 수준을 분석한 결과는 Table 7, Figure 2와 같다. ChatGPT와 Gemini는 네 가지 문장제 유형에서 모든 오류를 정확히 진단하고 근거를 제시하였다. 반면, DeepSeek는 상황 인식 오류(오류 유형 8)에서 오류의 일부만을 진단해 1점을 받았고 기준 단위 해석 오류(오류 유형 10)에서 제시된 오답과 풀이를 정확하다고 진단하여 0점을 받았다. 따라서, 문장제 오류 진단에서 받을 수 있는 최고 점수인 8점 중 DeepSeek는 5점, ChatGPT와 Gemini는 각각 8점을 획득했다. 이는 DeepSeek의 오류 진단 성능은 다른 두 생성형 AI에 비해 정확성과 정교성 측면에서 다소 부족함을 의미한다.

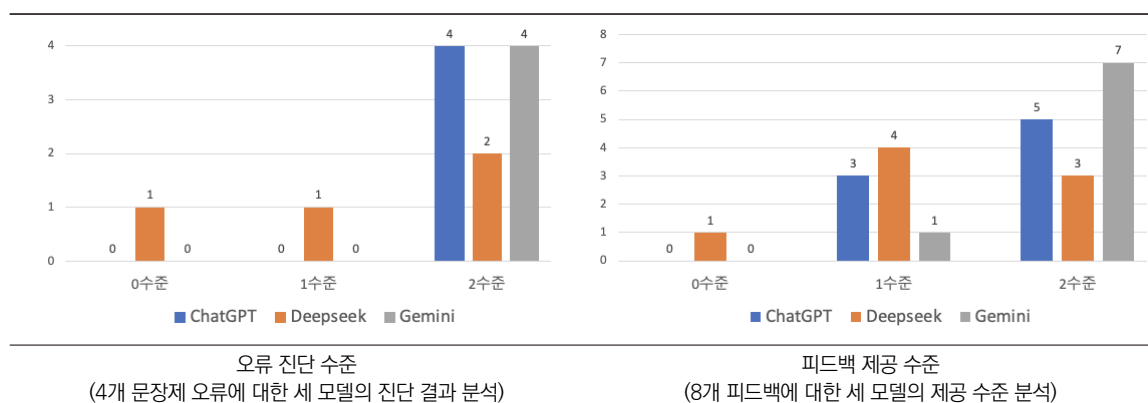
문장제 오류에 대한 생성형 AI의 피드백 수준을 분석한 결과 Gemini가 총 15점으로 가장 우수한 성과를 보였고, ChatGPT는 13점, DeepSeek는 10점으로 뒤를 이었다. 오류 유형별 수행 결과를 종합적으로 살펴보면 결과 해석 오류(오류 유형 9), 상황 인식 오류(오류 유형 8), 단위 혼동 오류(오류 유형 11), 기준 단위 해석 오류(오류 유형 10) 순으로 높은 피드백 수준을 보였다.

결과 해석 오류의 경우 DeepSeek의 피드백 중 하나가 1점으로 평가되었고 나머지 피드백은 모두 2점으로 채점되어 문장제 오류 유형에서 가장 높은 점수를 기록하였다. 상황 인식 오류에서는 ChatGPT와 DeepSeek의 피드백 중 각각 하나가 1점으로 평가되어 그 뒤를 이었고 단위 혼동 오류에서도 두 모델 모두 1점으로 평가된 피드백이 있었다. 반면, 기준 단위 해석 오류는 0점부터 2점까지 다양한 점수 분포를 보이며 전체 오류 유형 중 가장 낮은 수준의 피드백이 제공된 것으로 나타났다. 특히 DeepSeek가 제공한 피드백 중 하나는 유일하게 0점으로 평가되어 해당 오류 유형에 대한 이해와 대응에 한계가 있었음을 보여줬다. 이를 통해 동일한 생성형 AI 모델이라 하더라도 오류 유형에 따라 제공하는 피드백의 수준이 달라짐을 알 수 있다.

Table 7 Error diagnosis and feedback levels of three Generative AIs for word problem errors

오류 유형	의미	ChatGPT		DeepSeek		Gemini	
		진단	피드백	진단	피드백	진단	피드백
8. 상황 인식 오류	문제의 상황을 적절한 연산으로 표현하지 못하는 오류 (문제상황을 이해 못하거나 연산의 의미에 대한 이해가 부족한 경우)	2	1) 2 2) 1	1	1) 2 2) 1	2	1) 2 2) 2
9. 결과 해석 오류	계산은 했으나 상황을 설명하지 못하는 오류	2	1) 2 2) 2	2	1) 2 2) 1	2	1) 2 2) 2
10. 기준 단위 해석 오류	곱셈상황에서 기준 단위를 잘못 해석하는 오류	2	1) 2 2) 2	0	1) 1 2) 0	2	1) 2 2) 1
11. 단위 혼동 오류	서로 다른 종류의 단위 간 연산 수행 오류 (비와 관련된 상황에서)	2	1) 1 2) 1	2	1) 1 2) 2	2	1) 2 2) 2
합계		8	13	5	10	8	15

Figure 2 Error diagnosis and feedback levels of three Generative AIs for word problem errors



1) 문장제 오류 진단 사례

상황 인식 오류(오류 유형 8)의 경우, 세 모델 모두 공통적으로 곱셈이 적용되어야 할 문장제에 덧셈이 잘못 사용되었다는 점을 정확히 진단하였다(Table 8 참고). 그러나 분석의 구체성 수준에서는 차이를 보였다. DeepSeek는 “문맥을 잘못 이해함” 또는 “덧셈과 곱셈의 차이를 인지하지 못함” 등 일반적이고 포괄적인 진단을 제시한 반면, ChatGPT와 Gemini는 문제 상황에서 오류를 유발했을 가능성이 있는 구체적 문장 요소(예: “엄마가 $\frac{1}{2}$ 을 주었고, 그 중 $\frac{1}{4}$ 을 먹었다”)에 주목하여 보다 세부적이고 구조화된 분석을 제공하였다.

기준 단위 해석 오류(오류 유형 10)에 대해서는 ChatGPT와 Gemini가 ‘호두 한 자루’를 전체 양으로 간주해야 한다는 점을 정확히 인식하였고 문제 풀이 과정에서 전체의 $\frac{3}{4}$ 인 69개를 전체 단위로 오해한 오류를 명확히 지적하였다(Table 9 참고). 반면, DeepSeek는 이와 같은 오류를 식별하지 못하였으며, 학생의 풀이가 정확하다고 응답하였다(“학생의 풀이는 정확합니다”).

Table 8 Comparison of error diagnoses by three Generative AIs for word problem error type 8

모델 (오류 진단 수준)	오류 진단 및 오류의 잠재적 원인
ChatGPT (2수준)	<p>오류 1: 학생이 문제를 해결하는 방식에서, $\frac{1}{2}$와 $\frac{1}{4}$를 더하는 계산을 시도한 경우. 실제로 이 문제는 "몇 개의 부분을 먹었는지" 묻고 있기 때문에, 더하기가 아니라 "$\frac{1}{2}$의 피자에서 $\frac{1}{4}$을 먹었다"는 사실을 반영해야 한다.</p> <p>오류 1이 나타난 잠재적인 원인: 이 오류는 학생이 문제에서 두 부분을 더하는 방식($\frac{1}{2} + \frac{1}{4}$)으로 접근했기 때문입니다. 문제는 "엄마가 $\frac{1}{2}$을 주었고, 그 중 $\frac{1}{4}$을 먹었다"라는 내용이므로, 학생이 그 계산을 잘못 이해하여 더하는 방식으로 해결하려고 했을 수 있다. 문제의 본질은 $\frac{1}{2}$에서 $\frac{1}{4}$만큼을 먹은 것이므로, 덧셈이 아닌 곱셈이나 비율 계산을 사용해야 한다는 점을 놓쳤을 가능성이 있다.</p>
DeepSeek (1수준)	<p>오류 1-1: 학생이 문제를 잘못 이해하고, 분수의 덧셈을 사용하여 문제를 풀었습니다. 문제는 분수의 곱셈을 사용해야 하는 상황이지만, 학생은 $\frac{1}{2} + \frac{1}{4} = \frac{3}{4}$로 계산했습니다.</p> <p>오류 1-1이 나타난 잠재적인 원인: 학생이 문제의 문맥을 잘못 이해했거나, "먹은 양"을 계산할 때 덧셈이 아니라 곱셈을 사용해야 한다는 개념을 이해하지 못한 것으로 보입니다. 또한, 분수의 덧셈과 곱셈의 차이를 명확히 구분하지 못한 것도 원인일 수 있습니다.</p>
Gemini (2수준)	<p>오류 1: 문제 상황을 곱셈이 아닌 덧셈으로 해석하는 오류</p> <ul style="list-style-type: none"> 오류 1이 나타난 잠재적인 원인: 학생은 "% of what she gave me"라는 문구를 단순히 두 분수를 더하는 것으로 이해했을 가능성이 큼. 문제에서 요구하는 것은 "$\frac{1}{2}$의 $\frac{1}{4}$" 즉, 두 분수의 곱셈인데, 이를 덧셈으로 잘못 해석한 것입니다. 문제의 핵심을 정확히 파악하지 못하고, 단편적인 정보에만 집중했을 가능성이 있습니다.

Note. 오류 유형 8 문제: 엄마가 주문한 전체 피자의 $\frac{1}{2}$ 을 내게 주셨다. 나는 엄마가 주신 것의 $\frac{1}{4}$ 을 먹었다. 나는 전체 피자의 얼마를 먹었는가?

Table 9 Comparison of error diagnoses by three Generative AIs for word problem error type 10

모델 (오류 진단 수준)	오류 진단 및 오류의 잠재적 원인
ChatGPT (2수준)	<p>오류 1: 학생이 문제를 해결하는 과정에서 69에 $\frac{4}{23}$을 곱하는 방법을 사용했으나, 문제에서 요구하는 것은 "$\frac{3}{4}$의 69가 무엇인지"를 기반으로 $\frac{4}{23}$을 계산하는 것이므로, 단순히 69에 $\frac{4}{23}$을 곱하는 방식은 적절하지 않습니다.</p> <p>오류 1이 나타난 잠재적인 원인: 이 오류는 학생이 문제의 조건을 제대로 해석하지 않고, 주어진 숫자 69에 곱셈을 적용한 것에서 발생합니다. 실제로 문제는 "$\frac{3}{4}$가 69일 때, 전체 양을 구하고 그 후 $\frac{4}{23}$을 구하라"는 의도를 담고 있습니다. 학생이 문제를 풀 때, 69는 $\frac{3}{4}$에 해당하는 양이라는 사실을 간과했을 수 있습니다.</p>
DeepSeek (0수준)	<p>학생의 풀이는 정확합니다! $\frac{4}{23}$을 계산하기 위해 $69 \times \frac{4}{23}$을 수행한 결과 12로 정확히 계산했습니다. 이 풀이 과정에는 오류가 없으며, 학생이 분수의 곱셈을 정확히 이해하고 적용한 것을 보여줍니다.</p>
Gemini (2수준)	<p>오류 1: 문제에서 69가 호두 전체의 양이 아닌 $\frac{3}{4}$에 해당하는 양임을 간과하는 오류</p> <ul style="list-style-type: none"> 오류 1이 나타난 잠재적인 원인: 학생은 문제에서 주어진 69가 호두 전체의 양이라고 잘못 생각했을 가능성이 큼. 문제에서 69는 호두 한 봉지의 $\frac{3}{4}$에 해당하는 양이므로, 먼저 호두 한 봉지의 전체 양을 구한 후 $\frac{4}{23}$을 계산해야 합니다.

Note. 오류 유형 10문제: 만약 호두 한 자루의 $\frac{3}{4}$ 이 69개라면, 한 자루의 $\frac{4}{23}$ 은 호두 몇 개인가?

2) 문장제 오류에 대한 피드백 사례

Table 10은 기준 단위 해석 오류 및 단위 혼동 오류 유형에 대해 세 가지 생성형 AI 모델이 제공한 피드백의 예시이다. DeepSeek의 경우 잘못된 오류 진단에 기반한 피드백을 제시하고 있다. 즉, 69개의 호두를 ‘한 자루 전체의 양’으로 잘못 판단하여 이를 바탕으로 ‘호두 한 자루의 $\frac{4}{23}$ 이 12개’라는 학생의 오류를 그대로 수용한 채 전체 양을 계산하라는 절차적 지시를 제시하였다. 해당 피드백은 수학적 오류를 포함하고 있으며 학생의 개념 이해를 돕지 못한다는 점에서 0수준으로 분류되었다. 반면, ChatGPT는 ‘69’라는 수가 호두 한 자루 전체가 아닌 $\frac{3}{4}$ 에 해당하는 양임을 학생이 스스로 판단할 수 있도록 유도하는 피드백을 제공하였다(“이를 구하려면 69에 어떤 계산을 해야 할까요?”). Gemini 또한 ‘69’라는 수가 전체인지 일부인지를 학생 스스로 판단할 수 있도록 유도하는 방식으로 피드백을 구성하였다(“69는 호두 전체의 양일까요? 아니면 호두 한 봉지의 일부일까요?”).

전반적으로 ChatGPT와 Gemini는 문장제 오류 피드백에서 학생의 오류를 바탕으로 Polya(1945)의 문제 해결 4단계 중 첫 번째 단계인 ‘문제 이해’에 집중하도록 유도하는 공통적인 경향이 나타났다. 그러나, ChatGPT는 학생이 문제 해결 계획을 수립할 수 있도록 구체적인 힌트 중심의 피드백을 제시하였으며 Gemini는 개방형 피드백을 통해 ‘학생이 스스로 오류를 이해하도록 돕는다’는 프롬프트에 충실한 응답을 생성하는 경향을 보였다. DeepSeek는 모든 피드백이 일관되게 낮은 수준은 아니었으나 학생의 실제 풀이 과정에 기반하지 않은 유도 질문이나 구체성이 결여된 피상적인 피드백을 제공하는 경향이 상대적으로 더 많이 나타났다.

나머지 세 오류 유형의 경우(오류 유형 8, 9, 11) 모든 피드백이 최소 1점 이상으로 분류되었으며, 1점 보다 2점으로 분류된 경우가 더 많이 나타났다. 이러한 결과는 계산형 오류 분석 결과와 동일하게 동일한 생성형 AI 모델이라 하더라도 오류 유형에 따라 제공하는 피드백의 수준이 달라짐을 의미한다.

Table 10 Comparison of feedback by three Generative AIs for word problem error type 10

모델 (피드백 제공 수준)	피드백 및 피드백 의도
ChatGPT (2, 2수준)	피드백1: "먼저, 3/4가 69라고 주어졌을 때, 전체 양은 얼마일까요? 이를 구하려면 69에 어떤 계산을 해야 할까요?"
	피드백의 의도: 이 피드백은 학생이 먼저 69가 3/4에 해당하는 양임을 인식하고, 전체 양을 구하는 데 필요한 계산을 하도록 유도합니다. 문제를 푸는 데 필요한 첫 번째 단계(전체 양 구하기)를 강조합니다.
	피드백2: "전체 양을 구한 후, 4/23을 계산하는 방법은 무엇일까요? 그 계산을 통해 최종적으로 무엇을 구할 수 있을까요?"
	피드백의 의도: 이 피드백은 학생이 전체 양을 구한 후, 그 양에서 4/23을 어떻게 구할 수 있는지 이해할 수 있도록 돕습니다. 학생이 문제를 단계적으로 해결하도록 유도하여, 올바른 순서로 계산을 진행하게 합니다.

Table 10 Continued

모델 (피드백 제공 수준)	피드백 및 피드백 의도
DeepSeek (1, 0수준)	피드백1: "정말 잘 계산했어요! $69 \times \frac{4}{23}$ 을 정확히 계산하여 12라는 답을 도출했군요. 그렇다면 $\frac{3}{4}$ 가 69일 때, 전체 양은 어떻게 구할 수 있을까요?"
	피드백의 의도: 학생의 정확한 계산을 칭찬하고, 동일한 문제를 확장하여 전체 양을 구하는 방법을 생각해보도록 유도합니다. 이를 통해 학생의 이해를 더욱 확장시키는 데 목적이 있습니다.
Gemini (2, 1수준)	피드백2: "만약 $\frac{3}{4}$ 가 69라면, $\frac{1}{4}$ 은 얼마일까요? 그리고 전체 양은 어떻게 될까요?"
	피드백의 의도: 학생이 분수의 개념을 활용하여 단계별로 문제를 해결하도록 유도합니다. 이 피드백은 학생이 문제 해결 과정을 체계적으로 이해하고, 수학적 사고를 확장하는 데 목적이 있습니다.
	피드백: <ul style="list-style-type: none">문제에서 69는 호두 전체의 양일까요? 아니면 호두 한 봉지의 일부일까요?69가 호두 한 봉지의 $\frac{3}{4}$에 해당하는 양이라면, 호두 한 봉지의 전체 양은 어떻게 구할 수 있을까요?
	의도: <ul style="list-style-type: none">학생 스스로 문제에서 주어진 정보를 정확하게 이해하고, 69가 호두 전체의 양이 아님을 깨닫도록 유도합니다.호두 한 봉지의 전체 양을 구하는 방법을 스스로 생각하고, 문제 해결 과정에 적용하도록 격려합니다.

Note. 지면의 제약으로 인해, 두 개의 피드백 중 하나만 제시함.

V. 결론 및 논의

본 연구는 분수 곱셈 문제에서 발생하는 계산형 오류와 문장제 오류를 중심으로 세 가지 생성형 AI 모델(ChatGPT-4, Gemini-2.0 Flash, DeepSeek-R1)의 오류 진단 능력과 피드백 제공 역량을 비교·분석하였다. 본 연구는 다음과 같은 두 가지 제한점을 가진다. 첫째, 분수 곱셈과 관련된 모든 오류 유형을 포괄하지 못하였으며 실제 학생 데이터를 사용하지 않고 기존 문헌을 기반으로 가상의 학생 응답을 구성하여 분석을 수행하였다. 따라서 문제의 난이도나 오류 유형이 달라지거나, 실제 학생 데이터를 활용할 경우 연구 결과에 차이가 발생할 수 있다. 둘째, 각 오류 유형별로 하나의 사례만을 제시하였기 때문에 유형별로 여러 사례를 제시했을 경우 분석 결과가 달라질 수 있는 가능성이 존재한다.

그럼에도 불구하고 본 연구는 생성형 AI의 문제 해결 능력에 집중한 기존 연구들(예: Kwon et al., 2023; Oh et al., 2024)과 달리 교수학적 관점에서 생성형 AI가 학생의 수학적 오류를 어떻게 진단하고 피드백하는지를 살펴보았다는 점에서 의의가 있다. 특히 기존 연구들이 대부분 ChatGPT 단일 모델의 분석에 국한된 것(예: Shin et al., 2024)과 달리 본 연구는 Gemini와 Deepseek의 교수학적 역량을 비교 분석하였다는 점에서 학문적으로 기여한다.

1. 오류 진단 역량: 혼재된 결과로 인한 비교의 제한성

생성형 AI의 오류 진단 결과는 전반적으로 복합적이고 일관되지 않은 양상을 보였다. 계산형 오류의 경우 세 모델 모두 진단 정확도가 낮았으며 AI 모델 간 성능 차이도 크지 않았다. 반면, 문장제 오류에서는 전반적으로 진단 성능이 우수하게 나타났고 AI 모델 간 역량 차이도 상대적으로 뚜렷하게 드러났다. 특히 ChatGPT와 Gemini는 모든 문장제 오류에서 오류 원인을 명확하게 지정한 반면, DeepSeek는 피상적이고 일반적인 수준의 언어로 진단하는 경향을 보였다. 이러한 결과는 문제 유형(계산형 vs. 문장제)에 따라 AI의 오류 진단 능력에 차이가 나타날 수 있음을 시사한다.

오류 유형(유형1~유형11)에 따른 진단 성능을 분석한 결과 계산형 오류의 경우 세 모델 모두 오류 유형에 따라 유사한 정확도를 보였다. 반면, 문장제 오류에서는 DeepSeek만 일부 문항에서 낮은 성능을 보였고 ChatGPT와 Gemini는 모든 문항에서 성공적인 오류 진단을 수행하였다. 이는 오류 유형 자체가 AI의 진단 성능에 영향을 미칠 수 있으며 단순히 모델 간 비교를 넘어 오류 유형별 분석이 병행되어야 함을 보여준다. 이러한 복합적인 양상은 AI 모델 간의 오류 진단 성능을 단정적으로 비교·판단하기 어렵다는 점을 시사하며 다양한 변수들을 통제한 추가적인 실증 연구의 필요성을 제기한다. 또한, AI 모델이 오류 유형을 보다 정확하게 진단하기 위해서는 각 오류 유형별로 보다 다양한 사례를 제시할 필요가 있음을 알 수 있다.

2. 피드백 제공 역량: 모델 간 뚜렷한 성능 격차

피드백 제공의 측면에서는 계산형과 문장제 오류 모두 세 AI 모델 간 성능 차이가 비교적 명확하게 나타났으며 전반적으로 Gemini, ChatGPT, DeepSeek 순으로 높은 피드백 수준을 보였다. 계산형 오류의 경우 전반적으로 세 모델 모두 높은 수준의 피드백을 제공하지 못했으며 피드백의 일관성 측면에서는 Gemini가 가장 우수한 성능을 보였다. 문장제 오류의 경우 세 모델 모두 계산형 오류보다 높은 수준의 피드백을 제공하였다. 이처럼 본 연구에서는 생성형 AI 모델들이 전반적으로 계산형 오류보다 문장제 오류 상황에서 더욱 효과적이고 적절한 피드백을 제공하는 경향을 확인할 수 있었다. 이러한 차이점은 문장제의 경우 문제에 대한 보다 풍부한 언어적 정보가 제시되기 때문에 정보가 간소화되어 제시되는 계산형에 비해 생성형 AI가 학생의 오류를 보다 정확하게 분석할 수 있었던 것으로 판단된다. 다만, 연구에 사용된 문제들의 난이도가 동일하지 않다는 점에서 단순 비교에는 한계가 있음을 유의해야 한다.

문장제 오류에 대한 피드백에서 나타난 특성을 살펴보면 ChatGPT와 Gemini는 학생의 오류를 기반으로 피드백을 구성하며 문제 상황에 대한 이해를 중심으로 접근하고 있었다. 이는 기존 연구에서 지적된 바와 같이 학생들이 문장제를 분수 곱셈 상황으로 인식하지 못하고 다른 연산을 적용하는 오류(Mack, 2000; Prediger, 2008; Simon et al., 2018)를 보일 경우, 이 두 모델이 의미 있는 교육적 지원을 제공할 수 있음을 시사한다. 한편, DeepSeek는 제공된 피드백의 일관성이 낮았으며 피드백이 항상 학생의 사고 과정을 바탕으로 구성된 것은 아니었다. 일부 피드백은 학생이 자신의 오류를 자각하고 수정하는 데 실질적인 도움을 줄 수 있었지만 다른 경우에는 문제 해결과 직접적 관련성이 부족한 피상적인 피드백이 주를 이루었다. 이처럼 AI모델 간의 피드백 제공 역량의 차이점은 각 모델의 알고리즘 구조와 피드백 설계 방식의 차이에서 비롯된 것으로 판단된다. 후속 연구에서는 각 생성형 AI모델별 작동 알고리즘과 피드백 생성 방식에 어떠한 차이점이 있는지를 분석하고 다양한 교육적

맥락에서 어떤 모델을 사용하는 것이 보다 효과적일지에 대해 규명할 필요가 있다.

3. 시사점

본 연구는 AI 기술이 발전하고 있음에도 불구하고 생성형 AI 모델의 활용만으로 모든 학생에게 자동화된 양질의 수학적 피드백을 제공하는 데 한계가 있음을 보여준다. 이는 과거 연구에서 GeoGebra, 계산기, 다양한 물리적 교구(manipulatives) 등의 도입이 학습 효과를 자동적으로 보장하지 않으며, 해당 도구를 어떻게 활용하고 지도하느냐에 따라 학습 효과가 달라진다는 점이 강조된 맥락과 유사하다(Clements & McMillen, 1996). AI의 도입 자체가 수학 학습의 질을 자동적으로 향상시키는 것은 아니며 AI를 교육적으로 어떻게 활용할 것인지에 대한 교사의 판단과 개입이 여전히 핵심 변수로 작용함을 보여준다. 따라서 생성형 AI를 효과적으로 활용하기 위해서는 교사의 전략적 개입과 교육적 역량이 필수적이다.

본 연구의 시사점은 다음과 같다. 첫째, AI 모델이 모든 문제 유형과 오류 유형에 대해 일관되고 정확한 진단과 피드백을 제공하지 않는다는 점에 주목해야 한다. 동일한 프롬프트를 제공했음에도 불구하고 각 모델의 기술적 특성으로 인하여 세 AI 모델의 오류 진단과 피드백 수준에 차이가 나타났다. 따라서 교사는 AI가 때로 오판을 하거나 실질적인 도움이 되지 않는 피드백을 제공할 수 있음을 염두에 두어야 하며, 생성형 AI 모델의 응답을 무조건적으로 수용하기보다는 해당 수학 분야에 대한 전문성을 바탕으로 AI의 답변을 비판적으로 분석하고 선별하여 학생의 오류 진단과 피드백 제공에 활용할 수 있는 역량을 갖추어야 한다.

둘째, 각 AI 모델의 장점은 강화하고 단점은 보완하는 방향의 교수 전략이 필요하다. 생성형 AI는 모델에 따라 문제 해결 접근 방식이나 사고 촉진 방식이 상이하게 나타났으며 이는 각 모델의 알고리즘과 응답 방식의 차이에서 비롯된 것으로 보인다. 예를 들어, 일부 모델은 구조적 안내와 명확한 절차 제공을 통해 문제 해결을 촉진하고 다른 모델은 자기주도적 사고를 유도하는 개방형 피드백을 제공하는 특징을 보였다. 이러한 차이는 학생들의 사고 과정, 학습 성향, 수학적 표현 양상에 영향을 줄 수 있으므로 모든 오류 진단과 피드백을 AI에 전적으로 의존하는 것은 바람직하지 않다. 따라서 교사는 각 모델의 장단점을 파악하고 AI 도구를 교수학적 도구로 효과적으로 활용할 수 있어야 한다.

셋째, 향후 생성형 AI의 활용 효과를 명확히 검증하기 위해 사용 빈도, 문제 유형, 난이도, 오류 유형, 피드백의 내용 및 형식 등 다양한 요인들을 통제된 실증적 연구가 필요하다. 본 연구에서는 문장제에서의 AI 성능이 계산형보다 상대적으로 우수하게 나타났으나 문제 수와 난이도가 동일하지 않았기 때문에 명확한 결론을 도출하는 데에는 제한점이 있었다. 후속 연구에서는 급변하는 AI 기술의 가능성과 한계, 교사의 지식과 전문성에 기반한 교육적 판단과 개입이 학생들의 수학 학습에 미치는 영향, 교사가 AI의 응답을 교육적으로 해석하고 활용하는 방안 등에 대한 연구가 중요한 과제가 될 것이다.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

- Almarashidi, H. S., Jarrah, A. M., Khuma, O. A., & Gningue, S. M. (2024). Unveiling the potential: A systemic review of ChatGPT in transforming mathematics teaching and learning. *EURASIA Journal of Mathematics, Science and Technology Education*, 20(12), em2555.
- An, S. H., & Choi, C. W. (2016). A study of diagnosis and prescription of errors of fractional multiplication and division. *Journal of Elementary Mathematics Education in Korea*, 20(3), 457-477.
- 안소현, 최창우(2016). 분수의 곱셈과 나눗셈 오류 유형 진단 및 지도방안 연구. **한국초등수학교육학회지**, 20(3), 457-477.
- Bisconti, P., McIntyre, A., & Russo, F. (2024). Synthetic socio-technical systems: Poïèsis as meaning making. *Philosophy & Technology*, 37(3), 94.
- Clements, D. H., & McMillen, S. (1996). Rethinking concrete manipulatives. *Teaching Children Mathematics*, 2(5), 270-279.
- Cullen, C. J., Hertel, J. T., & Nickels, M. (2020). The roles of technology in mathematics education. *The Educational Forum*, 84(2), 166-178.
- Drijvers, P. (2015). Digital technology in mathematics education: Why it works (or doesn't). In S. Cho (Ed.), *Selected regular lectures from the 12th international congress on mathematical education* (pp. 135-151). Springer.
- Ergene, O., & Ergene, B. C. (2025). AI ChatBots' solutions to mathematical problems in interactive e-textbooks: Affordances and constraints from the eyes of students and teachers. *Education and Information Technologies*, 30(1), 509-545.
- Han, Y. J., & Jeon, I. H. (2023). An analysis of error types in fractional division word-problems solving process. *Journal of Elementary Mathematics Education in Korea*, 27(4), 297-324.
- 한유진, 전인호(2023). 분수의 나눗셈 문장제 해결 과정에서 나타나는 오류 유형 분석. **한국초등수학교육학회지**, 27(4), 297-324.
- Haser, Ç., & Ubuz, B. (2003). Student's conception of fractions: A study of 5th grade students. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 24, 64-69.
- Higgins, K., Huscroft-D'Angelo, J., & Crawford, L. (2019). Effects of technology in mathematics on achievement, motivation, and attitude: A meta-analysis. *Journal of Educational Computing Research*, 57(2), 283-319.
- Izsák, A. (2008). Mathematical knowledge for teaching fraction multiplication. *Cognition and Instruction*, 26(1), 95-143.
- Kang, Y. J. (2023). Exploring the use of AI language models in terms of pedagogical practice and research in mathematics education. *Journal of Educational Research in Mathematics*, 33(4), 1023-1040.
- 강윤지(2023). 수학교육의 교수학적 실천과 연구 측면에서 AI 언어 모델 활용 방안 탐색. **수학교육학연구**, 33(4), 1023-1040.
- Kang, Y. J. (2024). A study on the didactical application of ChatGPT for mathematical word problem solving. *Communications of Mathematical Education*, 38(1), 49-67.
- 강윤지(2024). 수학 문장제 해결과 관련한 ChatGPT의 교수학적 활용 방안 모색. **수학교육 논문집**, 38(1), 49-67.
- Kwon, O. N., Oh, S. J., Yoon, J., Lee, K., Shin, B. C., & Jung, W. (2023). Analyzing mathematical performances of ChatGPT: Focusing on the solution of national assessment of educational achievement and the college scholastic ability test. *Communications of Mathematical Education*, 37(2), 233-256.
- 권오남, 오세준, 윤정은, 이경원, 신병철, 정원(2023). ChatGPT의 수학적 성능 분석: 국가수준 학업성취도 평가 및 대학수학능력시험 수학 문제 풀이를 중심으로. **수학교육 논문집**, 37(2), 233-256.
- Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 33(2), 363-374.
- Lee, H. S., & Park, Y. Y. (2024). A comparative analysis of the mathematical capabilities of ChatGPT 3.5 and 4.0: Focused on the 8th grade mathematics textbooks. *Journal of the Korea Contents Association*, 24(7), 436-449.
- 이현수, 박영용(2024). ChatGPT 3.5와 4.0의 수학적 능력 비교 분석: 중학교 2학년 수학 교과서를 중심으로. **한국콘텐츠학회논문지**, 24(7), 436-449.
- Lee, K. H., Shin, D. J., Park, J. H., Kim, Y. J., & Lee, S. H. (2024). Exploring the use of ChatGPT in mathematical

- problem solving competency education: An analysis of error diagnosis and feedback performance in solving word problems on division of fractions. *Journal of the Korean School Mathematics Society*, 27(4), 639-666.
- 이경화, 신동조, 박진형, 김영준, 이송희(2024). 수학적 문제해결 역량 교육을 위한 ChatGPT 활용 방안 탐색: 분수의 나눗셈 문장제 해결 과정에서 나타나는 오류 진단 및 피드백 성능 분석. *한국학교수학회논문집*, 27(4), 639-666.
- Lee, S. J. (2025, February 6). Ministry blocks access to 'DeepSeek' across industries, diplomacy, and defense. *Donga Ilbo*. <https://www.donga.com/news/Society/article/all/20250206/130977078/2>
- 이소정(2025.02.06.). 산업-외교-국방부 '딥시크' 접속 차단. **동아일보**. <https://www.donga.com/news/Society/article/all/20250206/130977078/2>
- Mack, N. K. (2000). Long-term effects of building on informal knowledge in a complex content domain: The case of multiplication of fractions. *Journal of Mathematical Behavior*, 19(3), 307-332.
- McGalliard, W., & Otten, S. (2025). AI responses to challenging problems and educator responses to AI availability. *Digital Experiences in Mathematics Education*, 11(2), 319-332.
- National Council of Teachers of Mathematics (2024). *NCTM position statement on artificial intelligence and mathematics teaching*. <https://www.nctm.org/standards-and-positions/Position-Statements/Artificial-Intelligence-and-Mathematics-Teaching/>
- Oh, S., Yoon, J., Chung, Y., Cho, Y., Shim, H., & Kwon, O. N. (2024). Analysis of generative AI's mathematical problem-solving performance: Focusing on ChatGPT 4, Claude 3 Opus, and Gemini Advanced. *The Mathematical Education*, 63(3), 549-571.
- 오세준, 윤정은, 정유진, 조운주, 심효섭, 권오남. (2024). 생성형 인공지능의 수학 문제 풀이에 대한 성능 분석: ChatGPT 4, Claude 3 Opus, Gemini Advanced를 중심으로. **수학교육**, 63(3), 549-571.
- Park, M. Y., & Park, Y. H. (2017). An analysis on the error according to academic achievement level in the fractional computation error of elementary sixth graders. *Journal of Elementary Mathematics Education in Korea*, 21(1), 23-47.
- 박미연, 박영희(2017). 초등학교 6학년 학생이 분수 계산문제에서 보이는 오류의 학업성취수준별 분석. **한국초등수학교육학회지**, 21(1), 23-47.
- Park, S. S., Lew, S. R., Kim, S. M., Kwon, S. Y., Kim, N. K., Kang, H. J., Kim, K. T., Kim, B. K., Kim, Y. J., Kim, Y. S., Kim, E. K., Sung, C. K., Oh, H. J., Lee, M. H., Jeong, Y. S., Choi, B. H., & Choi, J. Y.(2023). *Elementary school mathematics 5-1 guidebook*. YBM.
- 박성선, 류성림, 김상미, 권성룡, 김남균, 강호진, 김경탁, 김보경, 김영진, 김용성, 김응관, 성창근, 오혜진, 이명희, 정인수, 최병훈, 최주영(2023). **초등학교 수학 5-1 지도서**. YBM.
- Pepin, B., Buchholtz, N., & Salinas-Hernández, U. (2025). A scoping survey of ChatGPT in mathematics education. *Digital Experiences in Mathematics Education*, 11, 9-41.
- Polya, G. (1945). *How to solve it: A new aspect of mathematical method*. Princeton University Press.
- Prediger, S. (2008). The relevance of didactic categories for analyzing obstacles in conceptual change: Revisiting the case of multiplication of fractions. *Learning and Instruction*, 18(1), 3-17.
- Reys, R. E., Lindquist, M. M., Lambdin, D. V., & Smith, N. L. (2015). *Helping children learn mathematics* (11th ed.). John Wiley & Sons.
- 박성선, 김민경, 방정숙, 권점례 공역(2017). **초등교사를 위한 수학과 교수법**. 경문사.
- Roschelle, J., Noss, R. L., Blikstein, P., & Jackiw, N. (2017). Technology for learning mathematics. In J. Cai (Ed.), *Compendium for research in mathematics education* (pp. 853-878). NCTM.
- Shin, B. C., Lee, J. S., & Yoo, Y. J. (2024). Exploring automatic scoring of mathematical descriptive assessment using prompt engineering with the GPT-4 model: Focused on permutations and combinations. *The Mathematical Education*, 63(2), 187-207.
- 신병철, 이준수, 유연주(2024). 프롬프트 엔지니어링을 통한 GPT-4 모델의 수학 서술형 평가 자동 채점 탐색: 순열과 조합을 중심으로. **수학교육**, 63(2), 187-207.
- Simon, M. A., Kara, M., Norton, A., & Placa, N. (2018). Fostering construction of a meaning for multiplication that

subsumes whole-number and fraction multiplication: A study of the learning through activity research program. *The Journal of Mathematical Behavior*, 52, 151-173.

Son, T. K. (2023). Exploring the possibility of using ChatGPT in mathematics education: Focusing on student product and pre-service teachers' discourse related to fraction problems. *Education of Primary School Mathematics*, 26(2), 99-113.

손태권(2023). ChatGPT의 수학교육 활용 가능성 탐색: 분수 문제에 관한 학생의 산출물과 예비교사의 담화 사례를 중심으로. **초등수학교육**, 26(2), 99-113.

Van de Walle, J. A., Karp, K. S., & Bay-Williams, J. M. (2023). *Elementary and middle school mathematics: Teaching developmentally* (11th ed.). Pearson.

Wardat, Y., Tashtoush, M. A., AlAli, R., & Jarrah, A. M. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, 19(7), em2286.