

ChatGPT의 수학적 성능 분석: 국가수준 학업성취도 평가 및 대학수학능력시험 수학 문제 풀이를 중심으로

권 오 남 (서울대학교, 교수)

오 세 준 (이화여자대학교사범대학부속이화·금란고등학교, 교사)[†]

윤 정 은 (인천효성고등학교, 교사)

이 경 원 (단국대학교사범대학부속중학교, 교사)

신 병 철 (수원외국어고등학교, 교사)

정 원 (서울대학교 대학원, 학생)

이 연구는 수학교육에서의 ChatGPT의 활용 방안 도출을 위한 기초 연구로서 국가수준 학업성취도 평가 및 대학수학능력시험 문제에 대한 ChatGPT의 응답을 분석하였다. ChatGPT는 생성형 인공지능 모델로서 여러 분야에서 주목받고 있으며, 교육계에서도 ChatGPT 활용 방안에 대한 요구의 목소리가 높아지고 있다. 이에 이 연구에서는 3개년 국가수준 학업성취도 평가 및 대학수학능력시험 문제에 대한 ChatGPT 3.5의 응답에 대해서 정답률, 풀이 과정의 정확도, 오류 유형을 분류하여 분석하였다. ChatGPT의 국가수준 학업성취도 평가 문제 및 대학수학능력시험 문제의 정답률은 각각 37.1%, 15.97%로 나타났다. ChatGPT의 풀이 과정의 정확도는 5점 만점으로 산출하였을 때, 국가수준 학업성취도 평가는 3.44점, 대학수학능력시험은 2.49점으로 산출되었다. ChatGPT의 수학 문제를 풀이하는 데 나타나는 오류 유형은 절차적 오류와 기능적 오류로 나뉘었다. 절차적 오류는 다음 단계로의 식을 연결 짓는 과정이나 계산상의 오류를 가리키며, 기능적 오류는 ChatGPT가 텍스트를 인식, 판단, 출력하는 과정에서 발생하는 오류였다. 이러한 분석은 정답률만이 ChatGPT의 수학적 성능을 판단하는 기준이 되어서는 안 되며, 풀이 과정의 정확도나 오류 유형까지도 복합적으로 고려해야 함을 시사한다.

I. 서론

인공지능 기술의 발전과 함께 사전 훈련된 대규모 언어 모델(Large Language Model, LLM)은 대화형 상호작용 기반의 직관적인 사용법과 다양한 활용 가능성으로 주목받고 있다. 대규모 언어 모델은 BERT, LaMDA, Codex, ChatGPT, Bard 등 여러 가지가 있지만, ChatGPT는 OpenAI에서 개발한 인공지능 챗봇으로 전 세계 미디어와 인터넷 플랫폼에서 가장 많이 언급되고 있으며, 최근 과학기술 분야뿐만 아니라 사회, 경제, 의료, 마케팅 등 다양한 분야에서 주목하고 있는 기술이다. ChatGPT는 출시된 지 5일 만에 100만 명, 2주 만에 200만 명, 2개월 만에 1억 명의 사용자를 달성하였으며, 이는 페이스북이 100만 명의 사용자를 확보하는 데 10개월 이상이 걸린 것과 비교하면 전례 없는 기록이다. 자연어를 이해하고 생성하는 인공지능 모델의 한 종류인 ChatGPT는 다양한 목적에 따라 인간과 대화할 수 있는 기능을 제공한다. ChatGPT는 수학 문제 풀이를 제시하는 성능을 갖추고 있으며, LaTeX라는 전문적인 문서 조판 시스템을 이용하여 수학 기호와 공식을 표현할 수 있다.

* 접수일(2023년 5월 22일), 심사(수정)일(2023년 6월 10일), 게재확정일(2023년 6월 23일)

* MSC2000분류 : 97U99

* 주제어 : ChatGPT, 수학적 성능, 국가수준 학업성취도 평가, 대학수학능력시험

† 교신저자 : skyjune@sen.go.kr

ChatGPT의 핵심 기술은 자연어 처리와 생성이며, 대부분 딥러닝 기반의 모델로 구현된다. 사전 훈련된 언어 모델은 자연어 처리와 생성의 성능을 크게 향상하게 한다. 그러나 ChatGPT의 성능을 평가하고 개선하기 위해서는 단순히 대화의 자연스러움이나 유창성만을 고려하는 것이 아니라, ChatGPT가 생성하는 자연어가 논리적으로 정확하고 일관성 있는지도 검증해야 한다. 인간과 ChatGPT가 수학을 주제로 대화하는 상황이라면 수학적 논리와 절차의 적절성 및 정확성에 대해 점검해야 한다. 예를 들어, ChatGPT가 수학 문제를 풀거나 설명하는 경우 답이 올바른지, 과정이 타당한지, 표기법이 표준에 부합하는지 등을 확인할 필요가 있다. ChatGPT가 수학적 개념이나 용어를 사용하는 경우, 정의가 명확한지, 논리가 타당한지, 모순이 없는지 등을 검증할 필요가 있다.

교육계에서도 ChatGPT 기술을 주목하고 있다. ChatGPT가 여러 분야에서 활용되고 있는 만큼, 교육 현장에서 활용할 수 있는 방안을 마련해야 한다는 사회적 요구가 존재하며(정제영 외, 2023), ChatGPT 활용 지침 마련 연구 역시 필요하다(김근수, 2023 참조). 국어 및 영어 등의 언어 관련 교과목에서는 ChatGPT를 활용하여 학습 효과를 높이고 학습 동기를 제고하는 연구가 진행되고 있다(강동훈, 2023; 신동광 외, 2023; 장성민, 2023). 그러나 수학 교수·학습을 위한 인공지능 플랫폼 분석 연구들은 있지만(예, 김세영, 조미경, 2022; 박혜연 외, 2022), 새롭게 등장한 ChatGPT의 수학교육에서 수학적 성능 관련 연구는 부재한 실정이다.¹⁾

기술의 발전에 따라 사회적으로 ChatGPT가 큰 영향을 미치고 있는 만큼 수학교육 내부에서도 ChatGPT를 활용하는 방안에 대해서도 구체적인 논의가 필요하다. 미디어에서 ChatGPT에 대한 수학적 능력에 대해 보도한 바 있지만(정호진, 2023), 표본의 크기가 불충분하거나 그 결과를 도출하는 방법의 신뢰도와 타당도가 보장되어 있다고 보기는 어렵다. ChatGPT를 수학교육 현장에서 활용하기 위해서는 ChatGPT의 수학적 성능에 대한 점검과 연구가 충분히 뒷받침되어야 할 것이다. 수학교육의 맥락에서 충분히 고려되지 않은 채 ChatGPT가 교육 현장에 도입된다면, 도구의 역할에 한정될 뿐만 아니라 잘못된 정보를 제공하여 학생들에게 혼란을 유발할 수 있다. 수학 교실에서 ChatGPT를 활용할 수 있는 방안을 마련하기 위해, ChatGPT가 지닌 수학 문제 풀이 능력에 대한 체계적인 분석이 필요하다.²⁾

이에 이 연구에서는 인공지능 기술 중 언어 기반 생성 인공지능의 하나인 ChatGPT의 수학적 성능을 평가하고, 문제 풀이 과정에서 발생하는 오류와 그 원인을 분석하여 ChatGPT를 수학교육 분야에서 효과적으로 활용하는 방안을 모색하고자 한다. 이 연구는 인공지능 기반 언어 모델이 수학 문제를 풀이할 때 보이는 정답률과 오류 유형을 보고하는 초기의 연구가 될 것이다. 구체적인 연구 질문은 다음과 같다.

첫째, 수학 문제에 대한 ChatGPT의 답은 얼마나 정확한가?

둘째, 수학 문제에 대한 ChatGPT의 풀이 과정은 얼마나 정확한가?

셋째, 수학 문제에 대한 ChatGPT의 풀이 과정의 오류 유형은 무엇인가?

II. 연구의 배경

1. 생성형 인공지능 모델

생성형 인공지능 모델(generative AI model)은 이미지, 음성, 자연어 등의 입력된 데이터를 이용하여 새로운

1) '성능'의 사전적 정의는 '기계 따위가 지닌 성질이나 기능'이다. ChatGPT의 수학적 성능은 수학과 관련된 프롬프트에 대해서 ChatGPT가 맥락을 올바르게 파악하여 프롬프트에 알맞은 답을 제공하는 것으로서 조작적으로 정의할 수 있다. ChatGPT의 수학적 성능은 수학 개념을 묻는 프롬프트에 대해 올바르게 설명을 제시하는 것, 문제를 맥락에 맞게 이해하고 올바른 풀이를 제공하는 것뿐만 아니라 합당한 수학 문제를 만들어내는 것 등을 포함할 수 있다. 이 연구에서는 ChatGPT가 지닌 여러 수학적 성능 중에서 문제를 풀이하는 것에 초점을 둔다.

2) '문제'의 사전적 정의는 '해답을 요구하는 물음', '문항'의 사전적 정의는 '문제의 항목'이다. '문항'은 '문제'를 가리킬 때도 사용되지만 문항의 개수나 순서를 가리켜 이야기할 때의 용어가 있다. 이 연구에서는 '문제'와 '문항'을 구분하여 서술하였으며, '문제'는 수학과 관련한 해답을 요구하는 경우, '문항'은 '문제'의 개수나 순서를 가리켜 표현하는 경우에 사용하였다.

데이터를 생성하는 기술이다. 이미지 생성 인공지능 모델은 입력된 텍스트나 이미지로부터 새로운 이미지를 생성하는 기술로서 GANs(Generative Adversarial Networks)가 대표적이다. 음성 생성 인공지능 모델은 주어진 텍스트나 음성 입력을 바탕으로 새로운 음성을 생성하는 기술로서 WaveNet이나 Tacotron이 대표적이다. 자연어 생성 인공지능 모델은 사람이 사용하는 언어인 자연어의 입력에서 문장이나 단어 시퀀스를 기반으로 새로운 텍스트를 생성하는 기술로 GPT(Generative Pre-trained Transformer)가 대표적이다. 이 연구는 생성형 인공지능 모델 중에서 자연어를 처리하는 ChatGPT에 초점이 있으므로 자연어를 처리하는 생성형 인공지능 모델을 중심으로 살펴보면 다음과 같다.

생성형 자연어 처리 인공지능 모델은 입력된 데이터를 이용하여 새로운 텍스트를 생성하는 기술로 대화 시스템, 문장 생성, 번역 등의 다양한 분야에서 활용되고 있다. 이 인공지능 모델이 새로운 텍스트를 생성해내기 위해서는 자연어를 다루는 대규모 데이터에 기반한 인공지능 모델을 먼저 학습하는 기술인 사전 학습(pre-trained learning)이 필요하다. 구체적인 사례로 BERT(Bidirectional Encoder Representations from Transformers)가 있는데, 이는 대규모 텍스트 데이터인 위키피디아와 뉴스 기사 등을 이용하여 사전 학습을 진행하였다(Devlin et al., 2018). 이와 같은 인공지능 모델은 대규모의 자연어 데이터를 사전 학습하여 사람이 사용하는 자연어의 특징을 분석하여 텍스트를 생성해내는 것이다. 생성형 자연어 처리 인공지능 모델이 텍스트를 학습할 때, 텍스트를 토큰화(tokenization) 과정을 거쳐서 작은 단위로 분할되는 과정을 거치며, 이 단위는 주로 단어, 문자 등으로 구성된다. 각 토큰은 토큰의 의미와 특징을 표현한 벡터로 표현되는데, 이러한 변환의 과정을 임베딩(embedding) 기법이라고 한다. 이 임베딩 벡터는 트랜스포머(transformer)라는 구조에 입력된다.

트랜스포머는 딥러닝 기반의 인공지능망 구조³⁾로 자연어 처리에 사용되는 모델이다. 이 모델은 Vaswani 외(2017)에서 처음 소개되었는데, 기존의 순환 신경망(RNN, Recurrent Neural Network) 기반 모델보다 효과적인 학습과 생성을 위해 개발되었다. 트랜스포머는 인코더와 디코더라는 여러 개의 층을 포함하며, 이 층은 셀프 어텐션(self-attention)과 순전파 신경망(feed-forward neural)으로 구성된다. 인코더는 입력 문장을 임베딩 벡터로 변환한 후, 입력 문장 내의 단어 간의 상호 관계를 학습하기 위해 셀프 어텐션을 이용한다. 셀프 어텐션은 입력 문장에서 각 단어를 쿼리(query), 키(key), 값(value) 벡터로 표현한다. 쿼리 벡터는 주어진 입력에 대해 어떤 정보를 얻고자 하는지를 나타내는 벡터이며, 키 벡터는 입력 데이터의 각 요소의 식별자 역할을 하는 벡터이다. 쿼리와 키 벡터의 유사도를 내적이나 유클리드 거리 등을 이용해 계산하여 어텐션 스코어를 구한다, 소프트맥스(softmax) 함수를 통해 계산된 어텐션 스코어의 값을 확률로 변환한다. 각 키 벡터에 확률을 곱해 어텐션 가중치를 얻고, 이 가중치를 값 벡터에 곱하여 어텐션 벡터를 얻는다. 어텐션 벡터는 각 토큰이 다른 토큰들과 얼마나 관련이 있는지를 반영하므로, 이 과정을 통해 각 단어의 중요도를 계산하고 단어 간의 상호 관계를 파악할 수 있다. 디코더는 인코더에서 생성된 벡터를 기반으로 번역이나 생성 작업을 수행한다. 디코더는 인코더와 마찬가지로 셀프 어텐션 메커니즘을 사용하여 입력된 단어 사이의 상호 관계를 학습하며, 마지막 단계의 층에서 단어를 예측하여 생성하게 된다. 순전파 신경망은 인공지능망의 가장 기본적인 형태로 입력 데이터가 각 층의 뉴런을 통과하면서 가중치와 활성화함수를 통해서 처리되는데, 데이터를 처리하여 특징을 추출하고 분류 또는 회귀와 같은 예측 문제를 해결하는 데 사용된다. 트랜스포머 모델에서 셀프 어텐션과 함께 사용되어 입력 정보

3) 인공지능(Artificial Intelligence)은 컴퓨터를 사용하여 사람의 지적 능력을 모방하거나 구현하는 기술이다. 머신러닝(Machine Learning) 또는 기계학습은 사람이 정의한 모델과 특징 추출 방법을 이용하여 데이터를 기반으로 학습해서 추론할 수 있게 하는 기술이다. 딥러닝(Deep Learning)은 인공지능망을 사용한 머신러닝 기술 중의 하나로 대규모 데이터 학습에 효과적인 기술이다. 이들 사이의 관계는 딥러닝이 머신러닝에 포함되고, 머신러닝이 인공지능에 포함된다고 볼 수 있다. 인공지능망(Artificial Neural Network, ANN)은 복잡한 계산을 수행하고 패턴 학습이 가능한 알고리즘이며 여러 개의 뉴런의 층으로 구성되어 있다. 인공지능망이 딥러닝과 함께 발전하면서 딥러닝에서 수백 개 이상의 층을 가진 심층신경망(Deep Neural Network, DNN)의 개념으로 확장되었고 다양한 문제를 해결할 수 있게 되었다(Hinton et al., 2006).

를 효과적으로 처리할 수 있어서 텍스트 처리 및 생성 작업에서 성능을 발휘할 수 있다. 트랜스포머 모델에 특정 분야의 데이터를 추가하는 과정인 미세조정을 통해서 특정 분야에 관련된 지식을 학습하고 특정 작업에 우수한 성능을 발휘할 수 있다(Vaswani et al., 2017).

GPT는 자연어 처리 딥러닝 모델 중 하나인 트랜스포머를 활용하는 모델로서, 확률을 기반으로 하여 텍스트를 생성할 수 있도록 사전 학습된 것이다.⁴⁾ ChatGPT는 대규모의 비구조화된 텍스트 데이터를 이용하여 사전 학습을 수행하였으며, 사전 학습한 데이터는 뉴스 기사, 백과사전, 웹사이트, 책, 소셜, 논문, 블로그 등에서 수집된 것으로, 다양한 종류와 주제의 텍스트를 학습하였다. 사전 학습 과정에서 GPT는 마스크 언어 모델링(masked language modeling) 방식을 사용하여, 입력된 텍스트에서 임의로 일부 단어를 가리고(masked), 가려진 단어를 예측하는 작업을 수행한다. 이를 통해 GPT는 텍스트 데이터의 패턴과 구조를 학습하고, 다양한 문맥에서 단어의 순서에 대하여 확률을 파악할 수 있게 된다. 대규모 자연어 데이터를 사전 학습하여 인공지능 모델은 다음 단어가 나올 확률을 계산할 수 있으며, 새로운 문장이 입력되면 가장 높은 확률값을 갖는 단어를 생성할 수 있다.

대규모 언어 모델인 ChatGPT는 자연어 처리 기술로 사람의 언어인 텍스트를 학습하였고, 사람이 구사하는 수준의 텍스트를 생성한다. 하지만 텍스트를 학습한 대규모 언어 모델이 수와 식이 포함된 수학적 문장을 이해하고, 주어진 문제를 풀이할 수 있는지는 확인할 필요가 있다.

2. ChatGPT의 성능과 교육의 활용

정답이 있는 지도학습의 경우 인공지능 모델의 성능을 평가하기 위하여 활용되는 지표는 정확도(accuracy), 정밀도(precision), 재현율(recall) 등이 있지만, ChatGPT의 성능을 평가한 선행연구들은 대부분 정확도를 성능 평가의 지표로 활용해왔다. 이때 정답이 있는 공인된 시험의 경우 문제를 질문하고 그 답변의 정확도를 확인하였다. 예를 들어, Kung 외(2023)는 ChatGPT에게 미국 의사 면허 자격시험 문제를 질문하였으며, 답변을 채점하여 ChatGPT가 의사 면허 자격시험에서 합격할 수준임을 확인하였다. OpenAI(2023)는 GPT 3.5의 성능을 분석한 결과 미 변호사 시험 213점(상위 90%), 미 생물학 올림피아드 43점(상위 48%)을 획득하였다고 발표하였다. Azaria(2022)는 ChatGPT가 대부분의 대규모 언어 모델과 마찬가지로 큰 수 곱하기, 근 찾기, 거듭제곱(특히 분수) 계산, 무리수(예, π 또는 e)에서 더하기 또는 빼기와 같은 연산에서 오류가 있음을 확인하였다. 또한 Frieder 외(2023)는 대학원 수준의 문제, 증명 빈칸 추론, 수학 올림피아드 문제 등 6가지 유형의 수학 문제의 ChatGPT의 성능을 확인하였으며, 그 결과 ChatGPT의 수학적 능력이 평균적인 수학 대학원생보다 현저히 낮음을 확인하였다.

최근 ChatGPT를 활용한 교과 교육 연구도 시작되었다. 강동훈(2023)은 ChatGPT 등장으로 인한 국어교육의 변화 대응 방안으로 쓰기 교육 본질의 재정립, 비판적 읽기 지도 등을 제시하였다. 장성민(2023)은 ChatGPT 시대에 필요한 작문 능력의 요소들을 탐색하고 작문 교육의 방향에 대하여 논의하였으며, 질문 생성 능력, ChatGPT의 결과에 대한 메타적 읽기 능력, 출처 확인 및 보강 능력의 중요성을 강조하였다. 한편 신동광 외(2023)는 영어 교수 학습의 도구로 ChatGPT의 활용 가능성을 탐색하기 위하여 ‘사회적 대화 능력’, ‘다양한 분야의 정보 제공 능력’, ‘언어 구사 능력’, ‘창의적 능력’, ‘도덕적 분별력’의 5개 영역에서 총 17개의 과제를 제시하고 그 결과를 교육전문가들이 평가하였다. 연구 결과 ChatGPT는 정보 제공 능력과 학습 및 교육 자료를 만들어 내는 능력을 활용하여 문제 풀이 활동을 통해 콘텐츠 기반의 영어 학습 및 교육에 매우 효과적으로 사용될 수 있

⁴⁾ GPT는 기술을 업데이트하여 새로운 버전을 서비스한다. 2018년 6월 GPT-1이 발표되었고, 2019년 2월 GPT-2, 2020년 6월 GPT-3, 2022년 11월 GPT-3.5, 2023년 3월 GPT 4가 발표되었다. 2023년 3월 ChatGPT plugins도 공개되었는데 OpenAI 자체 플러그인뿐만 아니라 Wolfram Alpha와 같은 외부 플러그인도 적용될 수 있다.

음을 확인하였다. ChatGPT와 교육과 관련된 국내 선행연구들은 국어, 영어 교과교육에서 활용 가능성을 모색하였지만, ChatGPT의 성능에 대한 구체적인 확인은 없이 대답의 일부만을 확인한 것이었다. 이에 비해 수학 혹은 과학과 같은 자연과학 분야 교과교육과 관련된 ChatGPT 연구는 찾아보기 어렵다. 수학교육에서 ChatGPT 활용 가능성을 탐색할 연구가 필요하다.

3. 연구 방법 및 절차

이 연구는 ChatGPT의 수학 문제를 풀이하는 수학적 성능의 사례를 상세하게 이해하기 위해 수행된 사례연구(Stake, 1995)이다. ChatGPT의 프롬프트에 수학 문제를 입력하고, 이에 대한 응답을 분석하는 것이다. 이 연구에서 상정하는 수학 문제의 범위를 설정하고, 수학 문제의 풀이에 대해 분석의 기준을 세워 분석함으로써 ChatGPT의 수학 문제를 풀이하는 수학적 성능의 사례를 구체적으로 서술한다.

가. 연구 대상

ChatGPT의 수학적 성능을 분석하기 위하여 국가수준 학업성취도 평가와 대학수학능력시험의 문제에 초점을 두었다. 우리나라 교육과정 이해당사자들이 교육과정 실태 조사나 교육과정 개정 과정에서 제시하는 주요 의견 중 하나는 평가에 관한 내용이다. 따라서 이 연구에서 다루는 수학 문제를 상정하기 위해서는 우리나라의 평가 문제를 대표할 수 있는 것을 포함해야 한다. 국가수준 학업성취도 평가는 전국 중·고등학교 학생들의 성취도를 가늠할 수 있는 대표성을 갖춘 평가라고 볼 수 있으며, 대학수학능력시험은 학교 수학의 최종적인 단계의 평가에 해당하므로 수학 문제를 다루는 평가 중에서 가장 대표적인 평가이다. 이에 이 연구에서는 우리나라의 중·고등학교 수학과 평가를 대표할 수 있는 국가수준 학업성취도 평가 및 대학수학능력시험의 수학 문제에 대한 ChatGPT의 수학적 성능을 중심으로 분석한다.

국가수준 학업성취도 평가는 1998년 기본 계획이 수립된 후 학교 교육을 통해 학생들이 성취하기를 기대하는 목표라 할 수 있는 교육과정 성취기준의 도달 정도를 측정하여 교육과정의 개선을 위한 피드백을 제공해왔다. 이는 중학교 3학년, 고등학교 2학년을 대상으로 실시되며, 평가범위 및 내용 영역은 <표 II-1>과 같다.

<표 II-1> 국가수준 학업성취도 평가의 내용 영역

(구자욱 외, 2019; 이재봉 외, 2020a; 이재봉 외, 2020b; 이재봉 외, 2021a; 이재봉 외, 2021b 재구성)

년도	학년	평가범위	내용 영역
2018	중학교 3학년*	중학교 1, 2학년 전 범위 중학교 3학년 1학기 '제곱근과 실수', '근호를 포함한 식의 계산', '다항식의 인수분해'	수와 연산, 문자와 식, 함수, 기하, 확률과 통계
	고등학교 2학년*	고등학교 <수학 I>, <수학 II> 전 범위	다항식, 방정식과 부등식, 도형의 방정식, 집합과 명제, 함수, 수열, 지수와 로그
2019	중학교 3학년*	중학교 1, 2학년 전 범위, 중학교 3학년 1학기 범위	수와 연산, 문자와 식, 함수, 기하, 확률과 통계
	고등학교 2학년**	고등학교 <수학> 전 범위	
2020	중학교 3학년**	중학교 1, 2학년 전 범위, 중학교 3학년 1학기 범위	
	고등학교 2학년**	고등학교 <수학> 전 범위	

* 2009 개정 교육과정 적용
** 2015 개정 교육과정 적용

대학수학능력시험은 학생들이 수학의 기본 개념·원리·법칙을 이해하고, 이를 적용하여 계산하고 추론하며 문제를 해결하는 능력을 평가함으로써 대학 교육을 받는 데 필요한 수학적 사고력을 평가하는 시험이다(한국교육과정평가원, 2023). 수학 영역의 시험은 공통과목과 선택과목으로 구분되며 공통과목의 출제범위는 <수학Ⅰ>, <수학Ⅱ>이고, 선택과목의 출제범위는 <확률과 통계>, <미적분>, <기하>이다. 내용 영역은 <표Ⅱ-2>와 같이 2015 개정 수학과 교육과정의 과목을 기준으로 세분화된다.

<표Ⅱ-2> 2021~2023학년도 대학수학능력시험 시험 과목 및 내용 영역(한국교육과정평가원, 2023 재구성)

과목	내용 영역
<수학Ⅰ>	지수함수와 로그함수, 삼각함수, 수열
<수학Ⅱ>	함수의 극한과 연속, 미분, 적분
<확률과 통계>	경우의 수, 확률, 통계
<미적분>	수열의 극한, 미분법, 적분법
<기하>	이차곡선, 평면벡터, 공간도형과 공간좌표

※ 2021학년도 대학수학능력시험에는 <기하> 과목은 시험 과목에 포함되지 않았음.

국가수준 학업성취도 평가 및 대학수학능력시험은 공인된 정답이 있는 시험이다. 최근 3개년의 문제를 평가하고자 2018~2020학년도 국가수준 학업성취도 평가 문제(구자옥 외, 2019; 이재봉 외, 2020a; 이재봉 외, 2020b; 이재봉 외, 2021a; 이재봉 외, 2021b 참조)와 2021~2023학년도 대학수학능력시험 문제를 분석 대상으로 설정하였다.⁵⁾ 해당연도의 문제 중 그림이나 표가 필수 정보이거나 텍스트 기반이 아닌 문제는 제외하였으며, 그림 혹은 표가 보조적인 역할을 하여 인간이 그림, 표가 제시되지 않더라도 수학 문제를 풀이할 수 있는 경우에는 포함했다. 이때 표나 그림은 array 또는 table 코드로 텍스트화하여 프롬프트에 제시하였다.

국가수준 학업성취도 평가는 선다형 문제와 서답형 문제로 구성되며, 선다형 문제는 선지를 제외하고 서술형으로 질문을 변형하였으며, 선택지가 필요한 선다형 문제의 경우는 선택지의 참·거짓을 판단하고 판단의 근거를 설명하도록 질문을 수정하였다. 서답형 문제는 최대 세 문항의 보조 문제로 구성되는데, 각 보조 문제는 하나의 문항으로 간주하여 개수에 반영하였다. 국가수준 학업성취도 평가의 전체 문항 수와 표나 그림으로 인해 제외된 문항 및 선별된 문항의 수는 <표Ⅱ-3>과 같다.

<표Ⅱ-3> 국가수준 학업성취도 평가 분석 문항 수

학교급	연도	전체 문항 수(개)	제외 문항 수(개)	분석 문항 수(개)
중학교	2018	37	13	24
	2019	38	18	20
	2020	24	17	7
고등학교	2018	37	4	33
	2019	24	8	16
	2020	22	6	16
국가수준 학업성취도 평가 문항 수		182	66	116

⁵⁾ 국가수준 학업성취도 평가 문제 및 대학수학능력시험 문제에 대하여 파일럿 분석을 실시한 결과, ChatGPT의 풀이가 연도에 상관없이 같은 유형의 오류가 반복되고 있음을 확인하였다. 이에 한국교육과정평가원 홈페이지에서 수집할 수 있는 가장 최근의 3개년 시험을 분석 대상으로 선정하였다.

선별한 116개 문항은 한국교육과정평가원에서 제공한 문항정보표의 분류기준에 따라 문자와 식, 수와 연산, 함수, 기하, 확률과 통계, 다섯 개의 내용 영역으로 구분하였다. 2018학년도 국가수준 학업성취도 평가의 경우, 문항정보표에 제시된 내용 영역을 재분류하여, 집합과 명제 6문항, 지수와 로그 3문항은 수와 연산 영역으로, 다항식 3문항, 방정식과 부등식 8문항은 문자와 식 영역으로, 함수 3문항, 수열 4문항은 함수 영역으로, 도형의 방정식 6문항은 기하 영역으로 분류하였다. 이에 따라 최종 수집한 영역별 문항은 수와 연산 26문항, 문자와 식 45문항, 함수 18문항, 기하 17문항, 확률과 통계 10문항이다.

한편, 대학수학능력시험은 선다형 문제와 단답형 문제로 구성되며, 선다형 문제는 선지를 제외하고 모두 서술형으로 변형하였다. <표 II-4>와 같이 2021학년도 수학 가형, 수학 나형 각각 30문항, 2022, 2023학년도 공통과목(22문항), 선택과목인 <확률과 통계>, <미적분>, <기하>(각 8문항) 총 178문항을 대상으로 하였으며, 이 중 문제 이해에 표 혹은 그림이 필수적인 문제는 제외하였다.

<표 II-4> 대학수학능력시험 분석 문항 수

연도	응시 유형		전체 문항 수(개)	제외 문항 수(개)	분석 문항 수(개)
2021	가형		30	1	29
	나형		30	1	29
2022	공통		22	4	18
	선택 과목	<확률과 통계>	8	1	7
		<미적분>	8	0	8
		<기하>	8	0	8
2023	공통		22	0	22
	선택 과목	<확률과 통계>	8	1	7
		<미적분>	8	0	8
		<기하>	8	0	8
대학수학능력시험 문항 수			152	8	144

이에 분석 대상으로 <수학 I>(지수함수와 로그함수 16문항, 삼각함수 9문항, 수열 16문항), <수학 II>(함수의 극한과 연속 8문항, 미분 20문항, 적분 9문항), <확률과 통계>(경우의 수 11문항, 확률 12문항, 통계 6문항), <미분과 적분>(수열의 극한 6문항, 미분법 7문항, 적분법 8문항), <기하>(이차곡선 7문항, 평면벡터 4문항, 공간도형과 공간좌표 5문항) 문제를 수집하였다.

나. 자료 분석

수학 문제의 경우 수식이 포함된 경우가 많아서 수식을 LaTeX으로 변환한 뒤 OpenAI에서 제공하는 ChatGPT 3.5-turbo 모델에 입력하였다. ChatGPT의 기본적인 수학 문제 풀이 능력을 평가하기 위해 하나의 문제를 프롬프트에 입력하고 결과를 도출한 뒤 ChatGPT를 초기화한 후 새로운 문제를 입력하는 과정을 반복하였다. [그림 II-1]과 같이 OpenAI에서 제공하는 API를 활용하여 파이썬으로 함수를 만들어 2023년 4월 실험을 실시하였다. 이때 ChatGPT의 역할은 “You are a helpful assistant.”로 설정하고 답변을 한국어로 번역할 것을 요청하였다.

```

# 함수화 모델 - GPT 3.5 Turbo 선택
def chatmath(problem) :
    model = "gpt-3.5-turbo"
    # 질문 작성하기
    query = problem
    # 메시지 설정하기
    messages = [
        {"role": "system", "content": "You are a helpful assistant."},
        {"role": "user", "content": query+"이 질문에 답하고 그 결과를 한국어로 번역해주세요."}]
    # ChatGPT API 호출하기
    response = openai.ChatCompletion.create(
        model=model,
        messages=messages
    )
    answer = response['choices'][0]['message']['content']
    return answer

#최종코드 학업성취도
import csv
for i in range(len(data)): # 오류 이후 다시
    new_list[i].append(chatmath(data[i][6]))
with open("output_final.csv", "w", newline="", encoding='utf-8') as f:
    writer = csv.writer(f)
    writer.writerow(new_list)

```

[그림 II-1] 수학적 성능을 분석하기 위한 ChatGPT 활용 파이썬 코드

ChatGPT의 응답은 정답률과 풀이 과정의 정확도, 오류의 유형으로 나눠 분석하였으며, 구체적인 절차는 다음과 같다. 첫째, ChatGPT가 제시한 응답이 국가수준 학업성취도 평가와 대학수학능력시험의 정답과 일치하는지 확인하였다. 이때 풀이 과정에 일부 오류가 있더라도 정답과 일치하면 맞는 것으로 코딩하였는데, 오류가 있다 하더라도 실질적 정답을 출력할 수 있는가 여부를 알아보기 위함이었다. 또 국가수준 학업 성취도 평가는 인간 학습자의 정답률, 대학수학능력시험은 문항의 배점과 등급을 분석의 보조 자료로 활용하였다. 국가수준 학업 성취도 평가의 경우 정답률에 대한 분석을 연구보고서에서 공개하고 있으므로, 연구보고서를 분석의 참고 자료로 활용하였다(구자욱 외, 2019; 이재봉 외, 2020a; 이재봉 외, 2020b; 이재봉 외, 2021a; 이재봉 외, 2021b). 대학수학능력시험의 경우 모든 문제 배점은 2점~4점인데, 배점이 문제의 난도를 뚜렷한 상관관계가 있다고 하기는 어렵지만 2점에서 4점으로 갈수록 난도가 올라가는 경향이 있어, 문제 배점에 따른 정답률도 함께 제시하였다. 둘째, Frieder 외(2023)의 평가 기준을 참고로 5점 척도로 단계를 구분하여 부분 점수를 부여하였다(<표 II-5> 참조). 학생의 서술형 답안을 채점하는 것과 유사하게 ChatGPT의 풀이 과정과 답을 모두 검토하였다. 문제는 내용 영역을 기준으로 분류하여 각각의 정답률과 부분 점수를 산출하였다. ChatGPT가 제시한 답이 정답인지를 기준으로 정답률을 산출하였고, 국가수준 학업성취도 평가와 대학수학능력시험 문제 각각 2명씩 총 4명의 연구자가 부분 점수를 부여한 이후 모든 저자의 교차 검토 및 논의를 통해서 최종 부분 점수를 도출하였다.

<표 II-5> 부분 점수 및 평가 기준

부분 점수	평가 기준
0	ChatGPT가 프롬프트를 전혀 처리하지 못함
1	ChatGPT가 프롬프트의 맥락에 맞게 대답하는데 실패함
2	ChatGPT가 프롬프트의 맥락에 맞게 대답하였으나, 답이 완전히 틀림
3	ChatGPT가 프롬프트의 맥락에 맞게 대답하였으나, 답이 부분적으로만 맞음
4	ChatGPT가 프롬프트의 맥락에 맞게 대답하였으며, 답이 거의 맞음
5	ChatGPT가 프롬프트의 맥락에 맞게 대답하였으며, 답이 완전히 맞음

셋째, ChatGPT의 응답이 제시하는 오류 유형을 분석하는 작업은 ChatGPT가 국가수준 학업성취도 평가 문제 및 대학수학능력시험 문제들에 대한 응답을 개방 코딩의 방식으로 분류하고 나서(Strauss & Corbin, 1990), 공통된 특징을 보이는 오류들끼리 군집화하여 오류 유형을 명명하는 절차로 진행되었다. 4명의 연구자가 2명씩 코딩한 뒤 코딩 결과에 대한 교차 검토를 하였으며, 모든 저자의 논의를 통해서 분류의 범위를 통합하거나 조정하였다. 오류 유형은 ChatGPT의 절차적 오류와 기능적 오류로 분류할 수 있었으며, ChatGPT의 절차적 오류는 식 오류, 계산 오류 등의 오류를 가리키며, ChatGPT 기능적 오류는 인식 오류, 판단 오류 등과 같이 ChatGPT가 문제를 이해하지 못하거나 잘못된 출력을 보이는 오류를 가리킨다.

III. 연구 결과 및 논의

1. ChatGPT의 국가수준 학업성취도 평가 문제 처리 결과

가. 국가수준 학업성취도 평가 문제에 대한 ChatGPT의 정답률

2018~2020학년도 3개년의 중학교 3학년 및 고등학교 2학년 국가수준 학업성취도 평가에 대한 ChatGPT의 정답률은 <표 III-1>과 같다. 선별한 116개 문항 중 43문항에서 정답을 제시하였으며 37.1%의 정답률을 기록하였다. 이 중 중학교급 국가수준 학업성취도 평가는 총 51문항 중 16문항에서 정답을 제시하여 정답률 31.4%, 고등학교급 국가수준 학업성취도 평가에서는 총 65문항 중 27문항에서 정답을 제시하여 정답률 41.5%를 기록하였다. 고등학교급에서 정답률이 10.1% 더 높았다.

<표 III-1> 국가수준 학업성취도 평가 문제에 대한 ChatGPT 풀이의 정답률

학교급	연도	연도별 정답률 (정답 문항 수/선별 문항 수)	정답률 (정답 문항 수/선별 문항 수)
중학교	2018	20.8%(5/24)	31.4%(16/51)
	2019	40.0%(8/20)	
	2020	42.9%(3/7)	
고등학교	2018	36.4%(12/33)	41.5%(27/65)
	2019	50.0%(8/16)	
	2020	43.8%(7/16)	
국가수준 학업성취도 평가 문제			37.1%(43/116)

국가수준 학업성취도 평가 문제에 대한 ChatGPT의 내용 영역별 정답률은 <표 III-2>와 같다. 각 내용 영역에서의 정답률은 수와 연산(50.0%), 문자와 식(35.6%), 함수(33.3%), 기하(29.4%), 확률과 통계(30.0%)로 나타났다. 수와 연산 영역만 정답률이 평균(37.1%)보다 높게 나타나고 나머지 영역은 모두 평균 이하의 정답률을 보였다. 기하와 확률과 통계 영역은 다른 영역에 비해 낮은 정답률을 기록하였는데, 이는 기하 영역은 평면도형 및 입체도형을 다루는 경우, 확률과 통계 영역은 표나 그래프를 다루는 경우의 영역이 지닌 특수성과 관련이 있는 것으로 추측된다.

<표 III-2> 국가수준 학업성취도 평가 문제 ChatGPT 풀이 내용 영역별 정답률

내용 영역	정답률(정답 문항 수/전별 문항 수)
수와 연산	50.0%(13/26)
문자와 식	35.6%(16/45)
함수	33.3%(6/18)
기하	29.4%(5/17)
확률과 통계	30.0%(3/10)
전체 영역 정답률	37.1%(43/116)

국가수준 학업성취도 평가 문제 분석 결과 인간 학습자의 정답률이 80% 이상인 15문항 중에서 ChatGPT가 오답을 제시한 경우는 4문항(27%)이 있으며, 인간 학습자의 정답률이 40% 미만인 21문항 중에서 ChatGPT가 정답을 제시한 경우는 4문항(19%)이 있다. 전자의 4문항에 해당하는 영역은 문자와 식(4문항)이며, 후자의 4문항에 해당하는 영역은 수와 연산(2문항), 함수(1문항), 기하(1문항)이다.

국가수준 학업성취도 평가 문제에 대해 학습자의 정답률이 높다고 해서 ChatGPT가 낮은 정답률을 기록하거나 학습자의 정답률이 낮다고 해서 ChatGPT가 높은 정답을 제시한다고 보기는 어렵다. 또한, 수학 내용 영역에 초점을 두더라도 학습자와 ChatGPT의 정답률과 내용 영역 사이에 밀접한 관계는 없다고 볼 수 있다. 여기에서 논하는 ChatGPT의 정답률은 수학 문제에 대한 정답률이므로 문제의 난도나 학교급, 교육과정상의 위계 등을 연결 지어 어떤 관계를 단정 짓는 것에는 무리가 있다. 이 결과는 표본이 되는 수학 문제에 대한 현재 ChatGPT 버전의 정답률로서의 자료가 된다.

나. 국가수준 학업성취도 평가 문제에 대한 ChatGPT 풀이 과정의 정확도

ChatGPT가 국가수준 학업성취도 평가 문제에 대해 제시한 답안을 채점하고 부분 점수를 부여하여 계산한 평균값을 연도별, 학교급별로 정리하면 <표 III-3>과 같다. 국가수준 학업성취도 평가 문제 전체 부분 점수에 대한 평균은 5점 만점 중 3.44점이며, 중학교급은 3.41점 고등학교급은 3.47점으로 학교급에 따른 차이는 특징을 보이지 않는다.

<표 III-3> 국가수준 학업성취도 평가 연도 및 학교급별 부분 점수 평균

학교급	연도	연도별 문항 수(개)	연도별 부분 점수 평균(점)	문항 수(개)	부분 점수 평균(점)
중학교	2018	24	2.96	51	3.41
	2019	20	3.40		
	2020	7	3.86		
고등학교	2018	33	3.42	65	3.47
	2019	16	3.75		
	2020	16	3.26		
국가수준 학업성취도 평가 문제				116	3.44

국가수준 학업성취도 평가 문제의 부분 점수를 매기는 과정에서 ChatGPT가 응답을 유연하게 하지 못하거나 제시하지 못해서 부분 점수를 부여하지 않은 예는 없었다. 즉, 일반적인 차원에서 수학 개념에 관해 설명하거나 답을 할 수 없는 수학 문제라고 단언하는 것이 아니라 수학 개념에 기초한 문제 풀이 시도를 ChatGPT 답안에서 확인할 수 있었다. ChatGPT의 답안의 결과는 5점 만점 중의 3.44점을 기록했다. 또한 오답이지만 4점의 높은 부분 점수를 부여한 문항은 11개 있었으며, 0점을 부여한 문항은 없었다. 평균적으로 볼 때 ChatGPT가 국가수준 학업성취도 평가 문제에 대해서 문제 풀이 과정의 중간 이상을 풀이하는 것으로도 해석할 수 있다.

ChatGPT가 제시한 결과가 정답이지만 엄밀성이 떨어져 부분 점수에서 5점 만점으로 평가되지 않은 문항이 3개 있다. 이러한 문제는 문제 풀이 과정의 대부분을 완성하였으나 정답의 형태가 잘못된 경우와 중간 과정을 틀렸으나 답을 올바르게 낸 경우로 나뉜다. 구체적인 예를 살펴보면 <표 III-4>와 같다.

<표 III-4> 정답이지만 부분 점수에서 5점 만점을 부여하지 않은 문제 사례(구자옥 외, 2019 재구성)

문제 정보	2018년 국가수준 학업성취도 평가 문제 고등학교 2학년	
	5번	20번
문제	$\sqrt[3]{27} - \sqrt{-32}$ 의 값은?	전체집합 $U = \{x \mid x \text{는 } 5 \text{이하의 자연수}\}$ 의 두 부분집합 A, B 에 대하여 $A^C \cap B^C = \{1\}$ 이고 $A^C \cap B = \{4\}$ 일 때, 집합 A 의 모든 원소의 합은?
문제 프롬프트	$\sqrt[3]{27} - \sqrt{-32}$ 의 값을 구해 봅시다. 먼저 $\sqrt{-32}$ 를 양수와 음수로 나누면, $\sqrt{-32} = (-1) \cdot \sqrt{32}$ 이므로, $\sqrt[3]{27} - \sqrt{-32} = \sqrt[3]{27} - (-1) \cdot \sqrt{32} = \sqrt[3]{27} + \sqrt{32}$ 입니다. 따라서 $\sqrt[3]{27} = 3$ 이고 $\sqrt{32} = 4\sqrt{2}$ 이므로, $\sqrt[3]{27} - \sqrt{-32} = 3 + 4\sqrt{2}$ 입니다.	$U = \{1, 2, 3, 4, 5\}$ 의 두 부분집합 A, B 에 대하여 $A^C \cap B^C = \{1\}$ 이고 $A^C \cap B = \{4\}$ 일 때, 집합 A 의 모든 원소의 합은?
ChatGPT 풀이	$\sqrt[3]{27}$ 의 세제곱근은 3 이므로, $\sqrt[3]{27} = 3$ 입니다. $\sqrt{-32}$ 의 다섯제곱근을 구해 봅시다. 먼저 $\sqrt{-32}$ 를 양수와 음수로 나누면, $\sqrt{-32} = (-1) \cdot \sqrt{32}$ 이므로, $\sqrt[3]{27} - \sqrt{-32} = \sqrt[3]{27} - (-1) \cdot \sqrt{32} = \sqrt[3]{27} + \sqrt{32}$ 입니다. 따라서 $\sqrt[3]{27} = 3$ 이고 $\sqrt{32} = 4\sqrt{2}$ 이므로, $\sqrt[3]{27} - \sqrt{-32} = 3 + 4\sqrt{2}$ 입니다.	A^C 는 전체집합에서 A 에 속하지 않는 원소들의 집합을 의미합니다. 따라서, $A^C \cap B^C = \{1\}$ 은 A 에 속하지 않고 B 에도 속하지 않는 원소들의 집합이 됩니다. 문제에서 $A^C \cap B^C = \{1\}$ 이므로, 1 은 A 와 B 둘 다에 속하는 원소가 아니기 때문에 1 은 A^C 와 B^C 둘 중 하나에만 속합니다. 또한, $A^C \cap B = \{4\}$ 이므로, 4 는 A^C 에 속하고 B 에도 속합니다. 따라서 1 은 B 에, 4 는 A^C 에 속하게 되고, 이를 이용하여 A 를 알아낼 수 있습니다. 먼저 A^C 에는 $1, 4$ 를 제외한 $2, 3, 5$ 가 있습니다. 따라서 A 에는 $2, 3, 5$ 가 속합니다. 또한 4 는 A^C 와 B 둘 다 속하므로, 4 는 B 에 속합니다. 따라서 B 에는 $1, 2, 3, 5$ 가 속합니다. 따라서, A 의 모든 원소의 합은 $2+3+5=10$ 이 됩니다. 따라서, 답은 10 입니다.
부분 점수	4점	4점

2018학년도 국가수준 학업성취도 평가 고등학교 2학년 5번, 20번 문제에 대해 ChatGPT가 정답에는 도달하였 다. 5번 문항의 경우 정답의 형태가 잘못되었다. 이는 제곱근 계산을 묻는 문제로 $\sqrt{27} = 3$ 이고, $\sqrt{-32} = -2$ 이므로 $3 - (-2) = 5$ 의 결과를 얻어서 정답은 5가 되어야 한다. ChatGPT가 $\sqrt{27}$ 은 3으로 변환하였으나 $-\sqrt{-32}$ 를 $\sqrt{32}$ 로만 변환하였고, 이를 다시 $\sqrt{32} = 2$ 로 변환해내지 못하여 정답을 $3 + \sqrt{32}$ 라고 제시하였 다. 답을 5라고 내지 못하였을 뿐이며, $\sqrt{27}$ 과 $\sqrt{-32}$ 의 계산에 대한 수행은 이루어졌다. 20번 문항의 경우 중간 풀이 과정에 오류가 있다. 이는 집합의 연산을 묻는 문제로 문제의 조건에서 전체집합 $U = \{1, 2, 3, 4, 5\}$ 가 제시되었다. 문제의 조건 $A^C \cap B^C = \{1\}$ 에서 $A^C \cap B^C = (A \cup B)^C$ 이므로 $(A \cup B)^C = \{1\}$ 를 이용해 야 한다. 마찬가지로 문제의 조건 $A^C \cap B = \{4\}$ 에서 차집합의 정의에 따라 $B - A = \{4\}$ 이므로 집합 A 의 모든 원소의 합은 1, 2, 3, 4, 5에서 1과 4를 제외한 2, 3, 5의 합 $2+3+5=10$ 이 된다. $A^C \cap B^C = (A \cup B)^C$ 임을 이용하는 과정에서 $(A \cup B)^C$ 를 $A^C \cup B^C$ 로 인식하여 답을 내는 과정에 오류가 있 지만 마지막 단계에서 정답을 올바르게 냈다.

ChatGPT 풀이 과정의 정확도 분석은 ChatGPT의 정답률만이 ChatGPT의 수학적 성능을 판단할 수 있는 지 표가 아니라는 점을 시사한다. ChatGPT가 제시하는 응답이 정답에는 도달할 수 있어도, 그 과정에는 오류가 발 생할 수 있다. 뿐만 아니라 ChatGPT가 최종 단계에서 제시하는 응답이 오답이어도 전반적인 과정은 올바른 풀 이 절차를 제시할 수도 있다. 따라서 학습자가 ChatGPT에 수학 문제를 입력하였을 때 단순히 ChatGPT가 정답 을 도출했는가에 초점을 두는 것은 바람직하지 않으며, 수학 문제의 풀이 과정을 검증하는 과정이 필요할 것이 다. 이 과정에서 학습자는 문제를 해결하기 위한 핵심 개념은 무엇이며, 그것이 ChatGPT의 답에 반영되었는지 되돌아보아야 할 것이다.

다. 국가수준 학업성취도 평가 문제에 대한 ChatGPT의 오류 유형

국가수준 학업성취도 평가 문항 116개에 대해서 ChatGPT가 응답으로 제시한 문항의 수는 73개이다. 이 중에서 오류를 분류할 수 있는 것이 있으나 풀이가 과도하게 짧거나 빈칸을 채워서 선택지에 알맞은 것을 고르는 문제에 대해 ChatGPT가 단답형으로 답하여 응답의 근거를 판단할 수 없는 문항의 수가 7개였다. 정답이지만 풀이 과정에 오류가 발견된 문항은 3개였다. 73개 중에서 7개를 제외하고 3개를 포함한 69개의 문항에 대해 ChatGPT의 오류 유형을 정리하면 <표 III-5>와 같다.

<표 III-5> 국가수준 학업성취도 평가 문제에 대한 ChatGPT의 오류 유형

대분류	오류 유형		문항 수(개)		비율
	소분류		소계	합계	
ChatGPT 절차적 오류	식 오류	식 풀기 오류	13	32	46.38%
		식 세우기 오류	7		
		식 전개 오류	8		
		식 해석 오류	4		
	계산 오류	사칙계산 오류	7	17	24.63%
		경우의 수 및 확률 계산 오류	5		
		기타 계산 오류	5		
ChatGPT 기능적 오류	인식 오류	기하적 대상 인식 오류	6	10	14.49%
		수 인식 오류	4		
	판단 오류		5	5	7.25%
	수 출력 오류		5	5	7.25%
	합계			69	100%

ChatGPT가 보인 오류 유형은 크게 두 가지로 나뉘어 볼 수 있다. 첫째, ChatGPT의 절차적 오류에 관한 것으로 식과 계산의 오류에 관한 것이다(총 49문항 - 71.01%). 절차적 오류 중에서 식의 오류가 절반 가까이 차지한다. 식의 오류는 방정식과 부등식을 풀이하는 것(13문항), 문제 풀이 과정에서 식을 세우는 것(7문항), 식을 전개하는 것(8문항), 주어진 식을 함수식으로 해석하는 것(4문항)의 오류로 나뉘어 볼 수 있다(총 32문항 - 46.38%). 식의 오류 중에서는 방정식이나 부등식을 푸는 과정에서 등식의 성질이나 부등식의 성질을 잘못 사용하여 오류가 생긴 경우가 가장 많았다. 계산의 오류는 문제를 풀이하는 과정에서 사칙계산의 오류(7문항), 경우의 수 세기 및 확률 계산의 오류(5문항), 기타 계산 오류(5문항)로 나뉘어 볼 수 있다(총 17문항 - 24.63%). 기타 계산 오류는 제곱근 계산의 오류, 수열의 합 계산의 오류, 집합의 연산 오류로 각각 2개, 1개, 2개 문항이다. 계산의 오류 중에서 덧셈, 뺄셈, 곱셈, 나눗셈의 사칙계산의 오류가 가장 많았으며, 경우의 수를 구하는 과정에서 일부의 경우를 고려하지 않는 경우가 그 뒤를 이었다.

둘째, ChatGPT의 기능적 오류에 관한 것으로 인식 오류, 판단 오류, 수 출력 오류에 관한 것이다(총 20문항 - 28.99%). 인식 오류로는 그림이나 표, 그래프를 제외하고 풀이할 수 있는 문제 중에서 기하적 대상에 대한 인식의 오류(6문항), 수 인식의 오류(4문항)가 있다(총 10문항 - 14.49%). 이는 좌표평면에서 좌표를 잘못 인식하거나, 직각삼각형으로 인식하여 피타고라스 정리를 이용해야 할 것을 직각삼각형으로 인식하지 못하는 경우이다. 소인수분해 결과에서 지수를 지수로 인식하지 못하는 것과 같은 수에 대한 인식 오류도 있었다. 판단 오류로는 삼각형의 결정 조건에 대한 판단, 수의 대소 관계, 국문으로 제시된 함수 상황 설명에 대한 판단을 잘못된 오류가 있다(5문항 - 7.25%). 수 출력 오류는 ChatGPT의 풀이 과정에서 단순한 오타를 가리키는 것으로 답안을 출력하는 과정에서 수의 오류에 관한 것이다(5문항 - 7.25%).

대분류	절차적 오류	기능적 오류
소분류	식 오류 식 풀기 오류	인식 오류 기하적 대상 인식 오류
오류 설명	방정식과 부등식을 풀이하는 과정에서 등식의 성질이나 부등식의 성질을 사용하는 과정에서의 오류	기하적 대상(그림이나 그래프 등)가 없더라도 풀이할 수 있는 문제이지만, 문제의 기하적 대상을 인식하는 과정에서의 오류
문제 정보	2018년 국가수준 학업성취도 평가 문제 중학교 3학년 5번	2018년 국가수준 학업성취도 평가 문제 중학교 3학년 26번
문제	일차방정식 $3x - 7 = -2x + a + 4$ 의 해가 $x = 2$ 일 때, 상수 a 의 값은?	그림은 좌표평면 위에 두 점 $A(0,6)$, $B(3,0)$ 을 지나는 직선을 나타낸 것이다. 직선 AB 와 x 축, y 축으로 둘러싸인 삼각형 AOB 를 직선 AB 를 축으로 하여 1회전시킬 때 생기는 입체도형에 대하여 <보기>에서 옳은 것을 모두 고른 것은? <div style="border: 1px solid black; padding: 5px;"> <보기> ㄱ. 원뿔이다. ㄴ. 회전축에 수직인 평면으로 자를 때 생기는 단면은 원이다. ㄷ. 회전축을 포함하는 평면으로 자를 때 생기는 단면은 직사각형이다. </div>
문제 프롬프트	<code>\text{5. 일차방정식} 3x-7=-2x+a+4 \text{의 해가 } x=2 \text{일 때, 상수 } a a \text{의 값은?}</code>	<code>\text{26. 그림은 좌표평면 위에 두 점 } \mathrm{A}(0,6), \mathrm{B}(3,0) \text{을 지나는 직선을 나타낸 것이다. 물음에 답하시오. } \text{26. 직선 } \mathrm{AB} \text{와 } x \text{축, } y \text{축으로 둘러싸인 삼각형 } \mathrm{AOB} \text{를 직선 } \mathrm{AB} \text{를 축으로 하여 1회전시킬 때 생기는 입체도형에 대하여 <보기>에서 옳은 것을 모두 고른 것은?}</code> $\begin{array}{l} \text{ㄱ. 원뿔이다.} \\ \text{ㄴ. 회전축에 수직인 평면으로 자를 때 생기는 단면은 원이다.} \\ \text{ㄷ. 회전축을 포함하는 평면으로 자를 때 생기는 단면은 직사각형이다.} \end{array}$
ChatGPT 풀이	주어진 일차방정식을 정리하면 다음과 같습니다. $\begin{aligned} 3x - 7 &= -2x + a + 4 \\ 5x - 7 &= a + 4 \end{aligned}$ 또한, $x=2$ 일 때, 위의 방정식의 해가 되므로 $x=2$ 를 대입해서 a 의 값을 구합니다. $\begin{aligned} 5(2) - 7 &= a + 4 \\ 10 - 7 &= a + 4 \\ 3 &= a \end{aligned}$ 따라서, 상수 a 의 값은 3입니다.	우선 지문에서 주어진 직선 AB 는 기울기가 -2 인 직선이므로, x 축과 y 축과 만나는 점을 구해보면 $A(0,6)$, $B(3,0)$, $O(0,0)$ 이 됩니다. 삼각형 AOB 를 직선 AB 를 축으로 1회전시켰을 때 생기는 입체도형은 원뿔이 아니기 때문에 ㄱ은 옳지 않습니다. 회전축에 수직인 평면으로 자를 때 생기는 단면은 직선 AB 와 수직이므로, 단면은 원이 아니라 직사각형입니다. 따라서 ㄴ은 옳지 않고, ㄷ은 옳습니다. 따라서 정답은 ㄷ입니다.
오류의 근거	등식의 성질을 이용해서 올바른 값을 구하지 못함.	직선을 1회전시켰을 때 생기는 입체도형을 기하적 대상(원뿔)으로 인식하지 못함.

[그림 III-1] 국가수준 학업성취도 평가 문제에 대한 ChatGPT의 오류 사례(구자욱 외, 2019 재구성)

국가수준 학업성취도 평가 문제에 대해서는 ChatGPT의 절차적 오류 중에 식 오류의 비율이 가장 높았고 그 중에서 식 풀기 오류가 차지하는 비율이 가장 높았다. ChatGPT의 기능적 오류 중에 인식 오류의 비율이 가장 높았고, 기하적 대상 인식 오류가 차지하는 비율이 가장 높았다. 구체적인 사례는 [그림 III-1]과 같다. ChatGPT의 절차적 오류가 나타나는 비율은 기능적 오류가 나타나는 비율에 비해 상대적으로 높다. 오류를 나타낸 69개의 문항 중 49개 문항이 절차적 오류에 관한 것이었다(71.01%). 기능적 오류는 ChatGPT의 수나 식에 대한 인식이나 수를 출력하는 과정의 오류에 관한 것이므로 ChatGPT의 기능이 발전되거나 수학과 관련된 응용 프로그램을 결합할 수 있는 모델을 활용한다면 개선될 수 있는 오류로 보인다.

2. ChatGPT의 대학수학능력시험 문제 처리 결과

가. 대학수학능력시험 문제에 대한 ChatGPT의 정답률

2021~2023학년도 대학수학능력시험 수학 문제에 대한 ChatGPT의 정답률은 <표 III-6>과 같다. ChatGPT는 선별한 144개 문항 중 23문항에 대하여 정답을 제시하였으며 15.97%의 정답률을 기록하였다. 이 중 2021학년도 대학수학능력시험 가형에서는 총 29문항 중 3문항에서 정답을 제시하여 정답률은 10.34%, 2021학년도 대학수학능력시험 나형에서는 총 29문항 중 8문항에서 정답을 제시하여 정답률이 27.59%이다. 2022학년도 대학수학능력시험 공통 18문항의 정답률은 11.11%, <확률과 통계> 7문항의 정답률은 14.29%, <미적분> 8문항의 정답률은 12.50%, <기하> 8문항의 정답률은 0%이다. 2023학년도 대학수학능력시험 공통 22문항의 정답률은 27.27%, <확률과 통계> 7문항의 정답률은 14.29%, <미적분> 8문항의 정답률은 12.50%, <기하> 8문항의 정답률은 0%이다.

<표 III-6> 2021~2023학년도 대학수학능력시험 수학 문제의 연도별 정답률

연도	응시 유형		문항 수(개)	정답 문항 수(개)	정답률
2021	가형		29	3	10.34%
	나형		29	8	27.59%
2022	공통		18	2	11.11%
	선택과목	<확률과 통계>	7	1	14.29%
		<미적분>	8	1	12.50%
		<기하>	8	0	0.00%
2023	공통		22	6	27.27%
	선택과목	<확률과 통계>	7	1	14.29%
		<미적분>	8	1	12.50%
		<기하>	8	0	0.00%
2021 ~ 2023학년도 대학수학능력시험 문제			144	23	15.97%

<표 III-7>은 2021~2023학년도 대학수학능력시험에 대한 ChatGPT의 답안을 채점하여 등급으로 환산한 것이다. 2021학년도 대학수학능력시험에서는 가형의 점수는 9등급인 것에 비해 나형에 대한 풀이는 6등급에 해당한다. 또한 공통 문제와 선택과목 문제로 분리된 2022~2023학년도 대학수학능력시험에서는 2022학년도에 <확률과 통계>, <미적분>, <기하> 모두 9등급인 것에 비해 2023학년도에는 각각 6등급, 6등급, 7등급에 속한다.

<표 III-7> 2021~2023학년도 대학수학능력시험 수학 문제의 연도별, 응시 유형별 점수 및 등급

연도	응시 유형	ChatGPT 대학수학능력시험 점수(점)	ChatGPT 대학수학능력시험 등급
2021	가형	9	9등급
	나형	24	6등급
2022	공통 및 <확률과 통계>	9	9등급
	공통 및 <미적분>	9	9등급
	공통 및 <기하>	6	9등급
2023	공통 및 <확률과 통계>	21	6등급
	공통 및 <미적분>	21	6등급
	공통 및 <기하>	18	7등급

ChatGPT의 2021~2023학년도 대학수학능력시험 수학 문제의 배점별 정답률은 <표 III-8>과 같다. ChatGPT는 배점 2점짜리 24문항 중 9문항에 정답을 제시했고(37.50%), 3점짜리 68문항 중 13문항에 정답을 제시했으며(19.11%), 4점짜리 52문항 중 1문항만 정답을 제시하였다(1.92%).

<표 III-8> 2021~2023학년도 대학수학능력시험 수학 문제의 배점별 정답률

배점(점)	문항 수(개)	ChatGPT 정답 문항 수(개)	정답률
2	24	9	37.50%
3	68	13	19.11%
4	52	1	1.92%
계	144	23	15.97%

ChatGPT는 대학수학능력시험 응시 유형의 난도가 높을수록 ChatGPT의 정답률이 낮아지는 경향이 있다. 2021학년도 대학수학능력시험에서는 나형의 정답률(27.59%)이 가형의 정답률(10.34%)보다 2배 이상 높다. 또한 ChatGPT가 2021학년도 대학수학능력시험 가형에서 9등급인 것에 비해 2021학년도 대학수학능력시험 나형에서 6등급으로 3등급의 차이가 있다. 또한, 배점별 정답률의 차이를 고려하면 상대적으로 배점이 높은 문제들을 풀이하기에는 ChatGPT의 수학 문제에 대한 학습이 상대적으로 부족하였으며 그 결과 생성된 답안에 오류가 많은 것으로 분석할 수 있었다. 한편 간단한 문제에 대해서는 문제를 이해하고 올바른 과정으로 풀이를 제시하였지만, 부정확한 계산이 자주 보였다. 다음으로 ChatGPT는 다른 영역에 비해 기하 영역 문제 풀이의 정확성이 떨어지는 것으로 보인다. 2022~2023학년도 대학수학능력시험 기하 영역 문제들은 모두 표나 그림이 없이도 풀이할 수 있는 문제였음에도 불구하고 정답률이 0%였다.

나. 대학수학능력시험 문제에 대한 ChatGPT 풀이 과정의 정확도

2021~2023학년도 대학수학능력시험 수학 문제에 대한 ChatGPT 풀이 과정의 정확도를 나타낸 결과는 <표 III-9>와 같다. 2021학년도 대학수학능력시험 가형 29문항의 부분 점수 평균은 2.89점이고, 나형 29문항의 부분 점수 평균은 2.72점이다. 2022학년도 대학수학능력시험 공통 18문항의 부분 점수 평균은 1.77점, <확률과 통계> 7문항의 부분 점수 평균은 1.71점, <미적분> 8문항의 부분 점수 평균은 2.37점, <기하> 8문항의 부분 점수 평균은 1.25점이다. 2023학년도 대학수학능력시험 공통 22문항의 부분 점수 평균은 3.18점, <확률과 통계> 7문항의 부분 점수 평균은 2.57점, <미적분> 8문항의 부분 점수 평균은 2.12점, <기하> 8문항의 부분 점수 평균은 2.37점이다. 2021~2023학년도 대학수학능력시험 수학 문제 전체의 부분 점수 평균은 2.49점이다.

<표 III-9> 2021~2023학년도 대학수학능력시험 수학 문제의 부분 점수

연도	응시 유형	문항 수(개)	부분 점수 평균(점)
2021	가형	29	2.89
	나형	29	2.72
2022	공통	18	1.77
	<확률과 통계>	7	1.71
	<미적분>	8	2.37
	<기하>	8	1.25
2023	공통	22	3.18
	<확률과 통계>	7	2.57
	<미적분>	8	2.12
	<기하>	8	2.37
전체		144	2.49

2021~2023학년도 대학수학능력시험 수학 문제의 과목별, 내용 영역별 부분 점수 평균은 <표 III-10>과 같다. <수학 I>의 41문항의 부분 점수 평균은 2.57점, <수학 II>의 37문항의 부분 점수 평균은 2.61점, <확률과 통계>의 29문항의 부분 점수 평균은 2.55점, <미적분>의 21문항의 부분 점수 평균은 2.42점, <기하>의 16문항의 부분 점수 평균은 1.81점이다. <수학 II> 과목의 함수의 극한과 연속 내용 영역 문제의 부분 점수 평균이 3.37점으로 가장 높았고, <기하> 과목의 평면벡터 내용 영역 문제의 부분 점수 평균이 1.25점으로 가장 낮았다. <기하> 과목 문제 중 평면벡터, 공간도형과 공간좌표 내용 영역을 제외한 나머지 내용 영역 문제들의 평균은 2점대 중반이다.

<표 III-10> 2021~2023학년도 대학수학능력시험 수학 문제의 과목별, 내용 영역별 부분 점수 평균

과목	내용 영역	문항 수(개)	내용 영역별 부분 점수 평균(점)	과목별 부분 점수 평균(점)
<수학 I>	지수함수와 로그함수	16	2.43	2.57
	삼각함수	9	2.5	
	수열	16	2.93	
<수학 II>	함수의 극한과 연속	8	3.37	2.61
	미분	20	2.5	
	적분	9	2.2	
<확률과 통계>	경우의 수	11	3	2.55
	확률	12	2	
	통계	6	2.83	
<미적분>	수열의 극한	6	2	2.42
	미분법	7	3.14	
	적분법	8	2.12	
<기하>	이차곡선	7	2.14	1.81
	평면벡터	4	1.25	
	공간도형과 공간좌표	5	1.8	

한편, 대학수학능력시험 수학 문제에 대한 ChatGPT 풀이에서 정답이지만 부분 점수 5점 만점을 부여하지 않은 문제는 144개의 문항 중 2개의 문항이다. 2021학년도 대학수학능력시험 가형 1번 문제와 2023학년도 대학수학능력시험 6번 문제는 정답이지만 풀이 과정에 논리적인 비약과 계산 실수가 있어 부분 점수 4점을 부여하였다(<표 III-11> 참조). 2021학년도 대학수학능력시험 가형 1번 문제에서는 무리수의 근호를 사용하여 계산하는

대신 무리수를 무한소수로 바꾸어 계산하였으나 이는 근삿값을 이용한 추측일 뿐 정확한 풀이로 볼 수 없다. 2023학년도 대학수학능력시험 6번 문제에서는 b 의 값을 구하는 대신 b 의 범위로부터 b 의 값을 추론하였다.

<표 III-11> 정답이지만 부분 점수에서 5점 만점을 부여하지 않은 문제 사례

문제 정보	2021학년도 대학수학능력시험 가형 1번	2023학년도 대학수학능력시험 6번
문제	$\sqrt[3]{9} \times 3^{\frac{1}{3}}$ 의 값은? [2점]	함수 $f(x) = 2x^3 - 9x + ax + 5$ 는 $x = 1$ 에서 극대이고, $x = b$ 에서 극소이다. $a + b$ 의 값은? (단, a, b 는 상수이다.) [3점]
문제 프롬프트	root 3 of 9 times 3^(1 over 3)의 값은?	함수 f LEFT (x RIGHT) =2x^3-9x+ax+5는 x=1에서 극대이고, x=b에서 극소이다. a+b의 값은? (단, a, b는 상수이다.)
ChatGPT풀이	이 문제에서 우리는 먼저 $3^{\frac{1}{3}}$ 를 계산해야 합니다. 이 값은 1.4422495703074083과 같습니다. 그런 다음 $\sqrt[3]{9}$ 를 계산하면 2.080083823051904와 같습니다. 마지막으로 두 값을 곱합니다. 따라서, $3^{\frac{1}{3}} \times \sqrt[3]{9} = 1.4422495703074083 \times 2.080083823051904 \approx 3$ 이므로 답은 3입니다.	함수 $f(x)$ 의 극댓값은 $f'(x)=0$ 일 때이고, 극솟값은 $f''(x)>0$ 일 때입니다. $f(x)$ 의 도함수는 $f'(x)=6x^2-18x+a$ 입니다. $x=1$ 에서 극댓값이기 때문에 $f'(1)=0$ 이 되어야 합니다. 따라서 $6(1)^2-18(1)+a=0$ 이므로 $a=12$ 입니다. $f(x)$ 의 이계도함수는 $f''(x) = 12x - 18$ 입니다. $x=b$ 에서 극솟값이기 때문에 $f''(b) > 0$ 이어야 하므로, $12b - 18 > 0$ 입니다. 따라서 $b > 3/2$ 이므로 $b=2$ 가 된다. 따라서 $a+b=12+2=14$ 입니다. 따라서 $a+b$ 의 값은 14입니다.
부분 점수	4점	4점

2021학년도 대학수학능력시험 결과에서 가형과 나형 문제의 부분 점수의 평균은 각각 2.89점, 2.72점으로 차이가 크지 않았다. 이는 가형과 나형 문제 모두 평균적으로 ChatGPT가 대부분의 문제에 대한 주요 아이디어는 잘 파악했지만, 계산상의 오류가 있거나 풀이 과정이 길어져 문제의 맥락을 파악하지 못해 정답을 도출하지 못한 것으로 보인다.

다. 대학수학능력시험 문제에 대한 ChatGPT의 오류 유형

대학수학능력시험 문제 144문항에 대해서 ChatGPT가 오답으로 제시한 문항의 수는 121개이다. 나머지 23문항은 정답이었고, 그중 2문항은 정답이지만 추론 과정에서 오류가 있었다. 144문항 중 123문항에 대한 ChatGPT의 오류 유형을 정리하면 <표 III-12>와 같다.

<표 III-12> 2021~2023학년도 대학수학능력시험 수학 문제에 대한 ChatGPT의 오류 유형

오류 유형			문항 수(개)		비율
대분류	소분류		소계	합계	
ChatGPT 절차적 오류	식 오류		17	17	13.82%
	계산 오류	사칙계산 오류	28	41	33.33%
		개수 중복 및 누락 오류	13		
ChatGPT 기능적 오류	인식 오류	조건 인식 오류	13	19	15.44%
		발문 인식 오류	3		
		기호 인식 오류	3		
	판단 오류	개념 사용의 판단 오류	27	42	34.14%
		접근 방식의 판단 오류	15		
	출력 오류		4	4	3.27%
합계				123	100%

ChatGPT의 오류 유형은 국가수준 학업성취도 평가와 마찬가지로 절차적 오류와 기능적 오류로 나뉘어 볼 수 있다. 먼저, ChatGPT의 절차적 오류에는 식 오류와 계산 오류가 있다(총 58문항 - 47.15%). 식 오류는 식 세우기, 풀이 맥락에 맞는 적절한 식 선택하기, 식에 수를 대입하기 등의 과정에서 오류를 말한다(총 17문항 - 13.82%). 식 오류로 판단된 문제들은 그 풀이 과정이 불필요하게 길다는 특징이 있고, 긴 문장을 출력하면서 조건식을 누락하거나 착각한 것이 대부분이다. 계산 오류는 사칙계산 오류, 개수 중복 및 누락 오류로 구분할 수 있다(총 41문항 - 33.33%). 이 중 사칙계산 오류는 덧셈, 뺄셈, 곱셈, 나눗셈에서의 단순 실수를 말하고, 개수 중복 및 누락 오류는 경우의 수를 구하는 문제에서 모든 경우를 다 세지 못하는 경우 또는 개수를 중복하여 세는 경우 등이 해당된다.

다음으로, ChatGPT의 기능적 오류는 인식 오류, 판단 오류, 출력 오류로 분류할 수 있었다(총 65문항 - 52.85%). 인식 오류에는 조건 인식 오류(13문항), 발문 인식 오류(3문항), 기호 인식 오류(3문항)가 있다(총 19문항 - 15.44%). 조건 인식 오류는 문제에서 주어진 조건을 활용하지 못하거나 잘못된 조건을 활용하는 오류이고, 발문 인식 오류는 구해야 하는 답의 값에 맞춰 계산하지 않는 오류이며, 기호 인식 오류는 $a_{\{11\}}$ 를 a_1 으로, $\log_3(3)$ 8을 $\log 38$ 로 착각하는 등 기호를 잘못 인식하여 발생하는 오류이다. 판단 오류는 개념 사용의 판단 오류와 접근 방식의 판단 오류로 세분화된다(총 42문항 - 34.14%). 출제 의도와 다른 수학적 개념을 사용하였거나 불필요한 개념을 사용하여 오류를 낸 경우로 주로 수학 개념의 다른 정의, 혹은 교과과정 외 개념을 사용하여 풀이한 사례이다(27문항). 접근 방식의 판단 오류는 문제 풀이 방향 자체가 잘못 설정된 경우로 대부분 그림 또는 그래프를 이용하여 기하적 특징을 파악하여 풀이해야 하는 문제를 대수식에서 출발하여 해석적으로만 풀이를 시도한 오류이다(15문항). 마지막으로 출력 오류는 풀이 과정에서 일부분의 풀이를 제시하지 않거나 출력 과정에서 오류이다(총 4문항 - 3.27%). 대부분 합답형 문제에서 \neg , \cup , \cap 중 일부를 풀이하지 않았거나 설명을 출력하는 과정상의 오류이다.

대학수학능력시험에 대한 ChatGPT 오류의 유형의 비율은 국가수준 학업성취도 평가의 오류 세부 유형의 비율과 차이가 있다. 대학수학능력시험의 경우 ChatGPT의 기능적 오류 중 판단 오류(34.14%)가 가장 높은 비율을 차지하였고, 절차적 오류 중 계산 오류(33.33%)가 두 번째로 높은 비율을 차지하였다. 구체적인 사례는 [그림 III-2]와 같다. 이에 비해 국가수준 학업성취도 평가에서는 절차적 오류 중에 식 오류(46.38%)와 사칙계산 오류(24.63%)가 각각 첫 번째, 두 번째로 높은 비율을 차지하였다. 국가수준 학업성취도 평가에 비해 대학수학능력시험에서 기능적 오류의 비율이 훨씬 더 높은 것은 국가수준 학업성취도 평가 문제에 비해 대학수학능력시험 문제의 난도가 높아 단순 계산상의 오류보다는 수학 개념 사용이나 풀이 접근 방식상 오류가 나타난 결과로 보인다.

대분류	절차적 오류	기능적 오류
소분류	계산 오류 개수 중복 및 누락 오류	판단 오류 접근 방식의 판단 오류
오류 설명	조건을 고려하지 않아 개수를 중복·누락하여 발생한 오류	출제 의도와 다른 접근 방식으로 풀이하여 발생한 오류
문제 정보	2021학년도 대학수학능력시험 나형 15번	2023학년도 대학수학능력시험 <기하> 28번
문제	세 학생 A, B, C를 포함한 6명의 학생이 있다. 이 6명의 학생이 일정한 간격을 두고 원 모양의 탁자에 다음 조건을 만족시키도록 모두 둘러앉은 경우의 수는? (단, 회전하여 일치하는 것은 같은 것으로 본다.) [4점] (가) A와 B는 이웃한다. (나) B와 C는 이웃하지 않는다.	두 초점이 $F(c, 0)$, $F'(-c, 0)$ ($c > 0$)인 쌍곡선 C 와 y 축 위의 점 A 가 있다. 쌍곡선 C 가 선분 AF 와 만나는 점을 P , 선분 AF' 과 만나는 점을 P' 이라 하자. 직선 AP 는 쌍곡선 C 의 한 점근선과 평행하고 $AP : PP' = 5 : 6$, $PF = 1$ 일 때, 쌍곡선 C 의 주축의 길이는? [4점]
문제 프롬프트	세 학생 m A,B,C를 포함한 6명의 학생이 있다. 이 6명의 학생이 일정한 간격을 두고 원 모양의 탁자에 다음 조건을 만족시키도록 모두 둘러앉은 경우의 수는? (단, 회전하여 일치하는 것은 같은 것으로 본다.) 조건 (가) mA와 mB는 이웃한다. (나) mB와 mC는 이웃하지 않는다.	두 초점이 mF LEFT (itc,0 RIGHT), mF' LEFT (it-c,0 RIGHT)인 쌍곡선 C와 y축 위의 점 mA가 있다. 쌍곡선 C가 선분 mAF와 만나는 점을 mP, 선분 mAF'과 만나는 점을 mP'이라 하자. 직선 mAP는 쌍곡선 C의 한 점근선과 평행하고 $\overline{mAP}:\overline{mPP'}=5:6$, $\overline{mPF}=1$ 일 때, 쌍곡선 C의 주축의 길이는?
ChatGPT 풀이	처음에 아무 학생이나 원 탁자에 마음대로 앉을 수 있다고 생각해봅시다. 그러면 첫 번째 자리에 6가지 학생 중 아무나 앉을 수 있습니다. 두 번째 자리에는 첫 번째 자리에서부터 시계 방향으로 인접한 자리에만 앉을 수 있고, 조건 (가)에 따라 첫 번째 자리에 앉은 학생의 이웃인 mB가 들어올 수 있으므로 1가지 경우가 있습니다. 세 번째 자리에는 첫 번째 자리와 두 번째 자리에서부터 시계 방향으로 떨어진 4개의 자리 중 하나에만 앉을 수 있습니다. 조건 (나)에 따라 첫 번째 자리와 두 번째 자리에 앉은 학생의 이웃인 mC는 앉을 수 없으므로 3가지 경우가 있습니다. 네 번째 자리부터는 첫 번째 자리와 두 번째 자리에서부터 시계 방향으로 떨어진 4개의 자리 중 하나에만 앉을 수 있습니다. 따라서 2가지 경우가 있습니다. 마지막 다섯 번째 자리에서는 마지막 첫 번째, 두 번째, 세 번째 자리에서만 선택할 수 있습니다. 그러므로 1가지 경우가 있습니다. 마지막 여섯 번째 자리에서는 그 자리에도 첫 번째 자리에서부터 시계 방향으로 떨어진 4개 자리 중 하나에만 앉을 수 있습니다. 따라서 2가지 경우가 있습니다. 따라서 총 경우의 수는 $6 \times 1 \times 3 \times 2 \times 1 \times 2 = 72$ 입니다.	먼저, 쌍곡선 C의 정의에 따라 두 개의 초점을 연결한 선분의 중간점 O를 중심으로 하는 쌍곡선입니다. 따라서, 직선 mAP는 점 O를 지나고, 쌍곡선 C와 평행합니다. 또한, 쌍곡선 C의 정의에 따라 y축과 만나는 점이 있으므로, 쌍곡선 C는 아래로 열린 모양을 가지는 것을 알 수 있습니다. ... (중략) ... 이제, 쌍곡선 C의 방정식에서 $y = 0$ 인 경우를 생각해봅시다. $\sqrt{(x - itc)^2 + 0^2}/c - \sqrt{(x + itc)^2 + 0^2}/c' = 1$ 위 식에서 좌변은 초점 mF와 mF'의 거리차와 같습니다. 이 거리차는 쌍곡선 C의 주축의 길이와 같으므로, 위 식을 변형하면 다음과 같습니다. $\sqrt{(x - itc)^2}/c - \sqrt{(x + itc)^2}/c' = 1$ 위 식을 제곱하면, 다음과 같은 이차방정식을 얻을 수 있습니다. $(c'/c)x^2 + 2itc(c'-c)x - (c'^2 - itc^2) = 0$ 이차방정식의 해는 다음과 같습니다. $x = \frac{(-2itc(c'-c) \pm \sqrt{4itc^2(c'-c)^2 + 4(c'/c)(c'^2 - itc^2)})}{(2c'/c)}$ 이제, x의 값을 구하여 쌍곡선 C의 주축의 길이를 구하면 됩니다.
오류의 근거	회전시켜 일치하면 같은 배열로 보는 원순열의 전제를 고려하지 않아 개수를 중복하여 셈.	기하적(그림, 그래프)으로 접근해야 하는 문제를 대수적(식)으로 풀이함.

[그림 III-2] 대학수학능력시험 문제에 대한 ChatGPT의 오류 사례

IV. 결론 및 제언

이 연구는 국가수준 학업성취도 평가 문제 및 대학수학능력시험 문제에 대한 ChatGPT 풀이의 정답률, 정확도, 오류 유형을 분석하여 생성형 인공지능 기술로 주목받고 있는 ChatGPT의 수학적 성능을 보고하고, ChatGPT를 수학교육의 맥락에서 활용 방안에 대한 아이디어를 제공하는 것을 목표로 한다. 연구 결과에 대한 요약 및 결론을 제시하면 다음과 같다.

첫째, ChatGPT의 국가수준 학업성취도 평가 문제 및 대학수학능력시험 문제의 정답률은 각각 37.1%, 15.97%로 나타났으나, 학교급이나 내용 영역, 문제 배점을 기준으로 ChatGPT의 수학적 성능을 판가름하기에는 무리가 있다. 국가수준 학업성취도 평가 문제의 경우 중학교급보다 고등학교급의 문제에 대한 정답률이 높고, 영역별로는 수와 연산 영역의 정답률이 가장 높았다. 대학수학능력시험 문제의 경우 6~9등급을 기록하였고 배점이 높을수록 정답률이 낮은 경향을 보였다. 하지만 학년군이나 내용 영역, 교육과정상의 과목 위계는 ChatGPT의 문제 풀이 능력을 판단하는 기준이 될 수는 없다. 또한, 이 연구에서는 다수의 선다형 문제를 서술형 문제로 변형하여 ChatGPT가 풀게 하였기 때문에 선다형 문제에서 요구하는 답을 만족하는 경우 정답으로 인정하였기 때문에 정답률이 ChatGPT의 성능을 판단할 수 있는 유일한 지표가 될 수는 없다.

둘째, ChatGPT의 국가수준 학업성취도 평가 문제 및 대학수학능력시험 문제에 대한 풀이 과정의 정확도는 5점 만점 중 각각 3.44점, 2.49점이므로 정답 여부와 관계없이 국가수준 학업성취도 평가 문제 및 대학수학능력시험 문제 풀이 과정의 정확도는 중간 정도를 기록한다. 국가수준 학업성취도 평가 문제의 경우 중학교급 51개 문항의 부분 점수 평균은 3.41점이며, 고등학교급 65개 문항의 부분 점수 평균은 3.47점이다. ChatGPT가 제시한 결과가 정답이지만 부분 점수가 5점이 아닌 문제에는 정답의 형태가 잘못되었거나 정답을 내는 과정에 오류가 있었다. 대학수학능력시험 문제의 경우 과목에 따라 1.81 ~ 2.61점 사이의 부분 점수 평균을 나타냈다. 국가수준 학업성취도 평가 문제와 마찬가지로 오답이지만 부분 점수를 높게 부여하거나 정답이지만 풀이 과정에 오류가 발생한 경우가 있었고, 이 문제들도 답의 형태와 정답을 내는 과정에서 오류가 있었다. 이에 따라 풀이 과정의 정확도와 관계없이 ChatGPT가 생성하는 수학 답안의 오류 유형 분류가 필요했다.

셋째, ChatGPT의 국가수준 학업성취도 평가 문제 및 대학수학능력시험 문제에 대한 오류 유형은 절차적 오류와 기능적 오류로 나눌 수 있는데, 기능적 오류는 ChatGPT의 기능 발전에 따라 개선될 수 있는 오류이고, 절차적 오류는 ChatGPT의 수학 학습에 의해 개선될 수 있는 오류이다. ChatGPT의 절차적 오류는 국가수준 학업성취도 평가 문제, 대학수학능력시험 문제에서 각각 71.01%, 47.15%를 나타냈다. 국가수준 학업성취도 평가 문제와 대학수학능력시험 문제 모두에서 절차적 오류 중 식 오류와 계산 오류가 나타났는데, 식 오류는 식 풀기, 세우기, 전개, 해석 오류에 관한 것이며 계산 오류는 사칙계산, 경우의 수 및 확률 등의 계산 오류이다. ChatGPT의 기능적 오류는 국가수준 학업성취도 평가 문제와 대학수학능력시험 문제에서 각각 28.99%, 52.85%를 나타냈다. 두 문제 모두에서 인식 오류, 판단 오류, 출력 오류가 나타났는데, 인식 오류는 질문, 발문, 기호 인식의 오류, 판단 오류는 개념 사용이나 문제 접근 방식의 판단 오류, 출력 오류로 구분할 수 있었다. 국가수준 학업성취도 평가 문제보다 대학수학능력시험 문제에서 판단 오류의 비율이 높았는데, 이는 문제가 다루는 수학 개념의 범위에 기초한 것으로서 고등학교 문제를 대학 수준의 개념을 끌어와 접근하되 풀이하지 못하는 경우나 수학 개념의 적용이 잘못된 경우의 오류였다.

연구 결과에 대한 요약 및 결론에 대한 제언을 제시하면 다음과 같다. 첫째, ChatGPT 기능 발전에 따른 정답률을 비교하여 보고하는 후속 연구를 지속해서 수행할 필요가 있다. ChatGPT 언어 모델은 시간이 지남에 따라 기능과 프롬프트 입출력 속도가 개선되고 있다. 이 연구에서 ChatGPT가 읽어낼 수 없어서 배제하였던 표나 그림이 포함된 일부 문제들도 이후 개선된 ChatGPT에서는 수학적 성능을 점검할 수 있을 것이다. 이러한 변화

에 따른 ChatGPT의 수학적 성능의 발전 양상을 보고하는 연구는 해당 ChatGPT의 수학적 성능에 대한 기록을 나타내는 것뿐만 아니라 독립된 인공지능 모델로서 수학 문제 풀이를 수행할 수 있는지를 제안한다. 또한 Wolfram Alpha와 같은 수학 기반 프로그램이 탑재된 ChatGPT의 수학 문제 풀이 수행 능력을 확인해 보는 것 역시 의미가 있을 것이다.

둘째, ChatGPT의 영역이나 과목, 문제의 유형에 따른 수학 문제 풀이 과정의 정확도에 대한 연구가 마련될 필요가 있다. 이 연구에서는 영역이나 과목에 따라서 풀이 과정의 정확도에 차이가 있음을 제시하였다. 수학 문제 풀이 과정의 정확도를 높이기 위해서는 ChatGPT의 수학 문제풀이의 추론 능력에 대한 학습이 기반이 되어야 할 것이다. ChatGPT의 추론 능력의 개선을 탐구하기 위해 같은 문제에 대해서 프롬프트의 변환을 통한 재실행을 거친 풀이 과정의 변화를 탐구하는 연구가 마련될 필요가 있다. 또한, 수학 문제 유형에 따른 ChatGPT의 풀이 과정을 탐구하는 연구가 수행되어야 할 것이다. 이 연구에서는 활용한 선다형 문제 중에는 표나 그림이 포함된 문제뿐만 아니라 함답형, 빈칸 추론 문제가 포함되어 있었다. 이러한 문제 유형은 프롬프트 구성뿐만 아니라 정확도에도 영향을 줄 수 있는 요인이 된다. 문제 유형에 따른 ChatGPT의 수학 문제 풀이 과정에 대한 연구는 ChatGPT의 수학 영역이나 과목별 풀이 능력의 편차를 완화하는 데 도움을 줄 수 있고, 프롬프트 유형과 답안 사이의 관계를 이해하는 데 도움을 줄 것이다.

셋째, ChatGPT의 절차적 오류 및 기능적 오류에 대하여 수학에서의 영역과 연결성이 있는지에 대한 점검이 필요하며, 인간 학습자가 보이는 오류와 유사한지에 관한 후속 연구가 마련될 필요가 있다. 수학교육 연구 분야에서는 수학 개념에 따라서 학생들이 보이는 오류에 관한 연구가 수행되었다. 이 연구에는 국가수준 학업성취도 평가 및 대학수학능력시험의 여러 영역의 문제에 대해 다루었지만, 특정 내용 영역에 초점을 두지 않았기 때문에 인간 학습자의 오류와의 비교를 수행하지는 않았다. 따라서 내용 영역에서의 초점을 두어 절차적 오류 및 기능적 오류의 유형을 구분하고, 인간 학습자가 보이는 오류 유형과의 비교 분석하는 연구가 수행될 필요가 있다. 이는 ChatGPT가 보이는 오류로부터 인간 학습자의 오류를 더 깊이 이해할 수 있는 연구가 될 것이다.

수학교육에서 ChatGPT의 활용은 위기인가? 또는 기회인가? 교수자와 학습자는 프롬프트를 구성하여 입력하고, ChatGPT의 반응에 대해 검토하면서 표현의 방식이나 답안의 논리성에 대한 점검을 수행하게 된다. 새로운 문제를 만들어 내는 것, 그 문제와 풀이의 정확성을 평가하는 과정은 수학 문제 풀이나 추론 능력뿐만 아니라 수학 용어 및 기호를 바탕으로 한 의사소통 능력 향상에 도움을 줄 것이다. 기존에 시간 제약이 있어 도전하지 못했던 수학 교실에서 여러 가지 정보를 ChatGPT로 생산해 처리하는 과정은 교수자와 학습자에게 새로운 길을 열어줄 것으로 보인다. 절차적 오류 및 기능적 오류가 개선된 ChatGPT 모델에 문제 풀이 과정을 정교화하는 프롬프트를 입력한다면 이는 단순히 생성형 인공지능 모델이 아닌 수학교육의 맥락에서 수학 교수·학습 및 평가를 지원할 수 있는 보조교사의 역할로 활용하는 방안을 마련하는 데 도움이 될 것이다. 따라서 수학교육에서 ChatGPT의 활용은 교수·학습 및 평가에서 새로운 기회로 작용할 수 있을 것이다.

참고문헌

- 강동훈 (2023). 챗지피티(ChatGPT)의 등장과 국어교육의 대응. 국어문학, **82**, 469-496.
- Kang, D. H. (2023). The advent of ChatGPT and the response of Korean language education. *Korean Language and Literature*, **82**, 469-496.
- 구자옥 · 박지현 · 이광상 · 박수민. (2019). 2018년 국가수준 학업성취도 평가 결과 분석-수학-. 한국교육과정평가원. 연구자료 ORM 2019-45-3.
- Ku, J., Park, J. H., Lee, K. S., & Park, S. (2019). *Analysis of the 2018 National Assessment of Educational Achievement Results: Mathematics*. Korea Institute for Curriculum and Evaluation. Research Report ORM 2019-45-3.
- 김근수 (2023, March 17). 교육부, '챗GPT 활용 원칙' 만든다...이대에 정책연구 의뢰. 뉴스시스. https://www.newsis.com/view/?id=NISX20230316_0002230010 2023. 5. 21. 검색.
- Kim, K. S. (2023, March 17). Ministry of Education is going to construct 'Guidelines for Utilizing ChatGPT' for policy research commissioned by Ewha Womans University. *NEWSIS*. Retrieved from https://www.newsis.com/view/?id=NISX20230316_0002230010.
- 김세영 · 조미경 (2022). 개별화 맞춤형 수학 학습을 지원하는 AI 기반 플랫폼 분석. 수학교육논문집, **36(3)**, 417-438.
- Kim, S., & Choi, M. K. (2022). AI-Based educational platform analysis supporting personalized mathematics learning. *Communication of Mathematics Education*, **36(3)**, 417-438.
- 박혜연 · 손복은 · 고호경 (2022). 수학 교수·학습을 위한 인공지능 플랫폼 분석 연구. 수학교육논문집, **36(1)**, 1-21.
- Park, H. Y., Son, B. E., & Ko, H. K. (2022). Study on the mathematics teaching and learning artificial intelligence platform analysis. *Communication of Mathematics Education*, **36(1)**, 1-21.
- 신동광 · 정혜경 · 이용상 (2023). 내용중심 영어 교수 학습의 도구로서 ChatGPT의 활용 가능성 탐색. 영어교과교육, **22(1)**, 171-192.
- Shin, D. K., Jung, H. K., & Lee, Y. S. (2023) Exploring the potential of using ChatGPT as a content-based English learning and teaching tool. *Journal of the Korea English Education Society*, **22(1)**, 171-192.
- 이재봉 · 박지현 · 손예림 (2020a). 2019년 국가수준 학업성취도 평가 결과 분석: 중학교 수학. 한국교육과정평가원. 연구자료 ORM 2020-23-3.
- Lee, J. B., Park, J. H., & Son, Y. R. (2020a). *Analysis of the 2019 National Assessment of Educational Achievement Results: Middle school mathematics*. Korea Institute for Curriculum and Evaluation. Research Report ORM 2020-23-3.
- 이재봉 · 이광상 · 손예림 (2020b). 2019년 국가수준 학업성취도 평가 결과 분석: 고등학교 수학. 한국교육과정평가원. 연구자료 ORM 2020-23-7.
- Lee, J. B., Lee, K. S., & Son, Y. R. (2020b). *Analysis of 2019 National Assessment of Educational Achievement Results: High school mathematics*. Korea Institute for Curriculum and Evaluation. Research Report ORM 2020-23-7.
- 이재봉 · 박지현 · 유혜원 (2021a). 2020년 국가수준 학업성취도 평가 결과 분석: 중학교 수학. 한국교육과정평가원. 연구자료 ORM 2021-51.
- Lee, J. B., Park, J. H., & Yoo, H. W. (2021a). *Analysis of 2020 National Assessment of Educational Achievement Results: Middle school mathematics*. Korea Institute for Curriculum and Evaluation. Research Report ORM 2021-51.
- 이재봉 · 정혜윤 · 유혜원 (2021b). 2020년 국가수준 학업성취도 평가 결과 분석: 고등학교 수학. 한국교육과정평가원. 연구자료 ORM 2021-55.
- Lee, J. B., Jung, H. Y., & Yoo, H. W. (2021b). *Analysis of 2020 National Assessment of Educational Achievement Results:*

- High school mathematics*. Korea Institute for Curriculum and Evaluation. Research Report ORM 2021-55.
- 장성민 (2023). 챗GPT가 바꾸어 놓은 작문교육의 미래 - 인공지능 시대, 작문교육의 대응을 중심으로 -. 작문연구, **56**, 7-34.
- Chang, S. M. (2023) ChatGPT has changed the future of writing education- Focusing on the response of writing education in the era of artificial intelligence -. *Writing Research* **56**, 7-34
- 정제영 · 조현명 · 황재운 · 문명현 · 김인재 (2023). 챗GPT 교육 혁명. 포르체.
- Jung, J. Y., Cho, H. M., Hwang, J. W., Moon, M. H., & Kim, I. J. (2023). *ChatGPT educational revolution*. Porche.
- 정호진 (2023, February 6). 美 의사시험 합격한 챗GPT, 수능 수학 성적은 '9등급'. 한국경제 TV.
<https://www.wowtv.co.kr/NewsCenter/News/Read?articleId=A202302060139>. 2023. 5. 21. 검색
- Jung, H. J. (2023, February 6). ChatGPT passed the US medical exam, and the mathematics score on the Korean university scholastic ability test was '9th grade'. *Korea Economy TV*. Retrieved from
<https://www.wowtv.co.kr/NewsCenter/News/Read?articleId=A202302060139>
- 한국교육과정평가원 (2023). 2024학년도 대학수학능력시험 학습 방법 안내. 수능 CAT 2023-2-1.
- Korea Institute for Curriculum and Evaluation (2023). *Guide to studying for the 2024 College Scholastic Ability Test*. Research Report CAT 2023-2-1.
- Azaria, A. (2022). ChatGPT usage and limitations. *HAL Open Science*. hal-03913837.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Frieder, S., Pinchetti, L., Griffiths, R. R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., ... & Berner, J. (2023). Mathematical capabilities of ChatGPT. *arXiv preprint arXiv:2301.13867*.
- Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, **18(7)**, 1527-1554.
- Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digital Health*, **2(2)**, e0000198.
- OpenAI. (2023). GPT-4 technical report. *arXiv:2303.08774* [cs.CL]
- Stake, R. E. (1995). *The art of case study research*. Sage publications.
- Strauss, A., & Corbin, J. (1990). *Basics of qualitative research*. Sage publications.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, **30**.
- Weaver, W. (1955). *Machine translation of languages*. In W. N. Locke & A. D. Booth (Eds.), MIT Press. (Reprint, Original work published 1949)

Analyzing Mathematical Performances of ChatGPT: Focusing on the Solution of National Assessment of Educational Achievement and the College Scholastic Ability Test

Kwon, Oh Nam

Seoul National University
E-mail: onkwon@snu.ac.kr

Oh, Se Jun[†]

Ewha Womans University High School
E-mail: skyjune@sen.go.kr

Yoon, Jungeun

Incheon Hyoseong High School
E-mail: yoonhoho1004@snu.ac.kr

Lee, Kyungwon

Dankook University Middle School
E-mail: kyungwon.lee.snu@gmail.com

Shin, Byoung Chul

Suwon Academy of World Languages
E-mail: sinbc2003@snu.ac.kr

Jung, Won

Seoul National University
E-mail: garden@snu.ac.kr

This study conducted foundational research to derive ways to use ChatGPT in mathematics education by analyzing ChatGPT's responses to questions from the National Assessment of Educational Achievement (NAEA) and the College Scholastic Ability Test (CSAT). ChatGPT, a generative artificial intelligence model, has gained attention in various fields, and there is a growing demand for its use in education as the number of users rapidly increases. To the best of our knowledge, there are very few reported cases of educational studies utilizing ChatGPT. In this study, we analyzed ChatGPT 3.5 responses to questions from the three-year National Assessment of Educational Achievement and the College Scholastic Ability Test, categorizing them based on the percentage of correct answers, the accuracy of the solution process, and types of errors. The correct answer rates for ChatGPT in the National Assessment of Educational Achievement and the College Scholastic Ability Test questions were 37.1% and 15.97%, respectively. The accuracy of ChatGPT's solution process was calculated as 3.44 for the National Assessment of Educational Achievement and 2.49 for the College Scholastic Ability Test. Errors in solving math problems with ChatGPT were classified into procedural and functional errors. Procedural errors referred to mistakes in connecting expressions to the next step or in calculations, while functional errors were related to how ChatGPT recognized, judged, and outputted text. This analysis suggests that relying solely on the percentage of correct answers should not be the criterion for assessing ChatGPT's mathematical performance, but rather a combination of the accuracy of the solution process and types of errors should be considered.

* 2000 Mathematics Subject Classification : 97U99

* Key words : ChatGPT, mathematical performances, national assessment of educational achievement, college scholastic ability test

[†] corresponding author