

수학과 ‘확률과 통계’ 영역에서 ChatGPT를 활용한 서답형 평가 피드백이 학생들의 피드백 리터러시에 미치는 영향*

The Influence of ChatGPT-Enhanced Evaluation Feedback on Students' Feedback Literacy in Descriptive Assessments within the 'Probability and Statistics' Domain of Mathematics Education

김육태[†] · 김영식^{†*}

Wooktae Kim[†] · Yungsik Kim^{†*}

요 약

본 논문은 수학과 ‘확률과 통계’ 영역의 서답형 평가에서 인간 교사가 제공하던 형성적 피드백을 ChatGPT가 보완 또는 대체할 수 있는지에 대한 탐구를 수행하였다. 연구 방법으로 학생들을 두 집단으로 나누어 서답형 평가를 시행하고, 이에 대한 피드백을 ChatGPT 및 인간 교사가 각각 제공하도록 실험을 진행하였다. 실험 데이터는 독립표본 t-검정을 이용하여 해석하였으며, 실험 결과 피드백 제공자가 인공지능이든 인간 교사든 피드백 리터러시의 하위 구성 요소인 피드업, 피드백, 피드포워드, 인지, 정서조절 및 피드백 구하기에서 통계적으로 유의미한 차이가 없음을 밝혀졌다. 이와 같은 연구 결과는 ChatGPT가 교육 분야에서 피드백을 제공하는 도구로 활용될 가능성을 시사한다.

주제어: 인공지능 자동채점 피드백, ChatGPT, 형성적 피드백, 피드백 리터러시, 수학과 서답형 평가

ABSTRACT

This thesis investigates how artificial intelligence, represented by ChatGPT, can replace the formative feedback traditionally provided by human educators in the field of mathematics, specifically in the realm of probability and statistics education. To conduct this study, students were divided into two groups and subjected to formative assessments, with feedback provided by both artificial intelligence, specifically ChatGPT, and human educators. The experimental results did not reveal statistically significant differences in students' feedback literacy, encompassing their ability to comprehend and utilize feedback, between the feedback providers, be it artificial intelligence or human educators. These research findings suggest the potential utility of artificial intelligence, ChatGPT in particular, as a tool for delivering feedback in the field of education.

Keywords: automated evaluation and scoring, ChatGPT, formative feedback, feedback literacy, descriptive essay problems in mathematics

1. 서론

교육과정에서 학생들의 학업성취도를 향상시킬 수 있는 중요한 교수학습 방법 중 하나는 피드백이다[1]. 최근의 교육 평가 패러다임에서는 학습과 성장을 위한 평가, 특히 학생들의 학습 향상과 성장을 목표로

한 형성적 피드백이 강조되고 있으며, 우리나라에서도 2015 개정 교육과정 이후 기존의 책무성 위주의 평가가 학습을 위한 평가로 변화 및 강조되면서 학생의 학습에 도움이 되는 피드백의 중요성이 더욱 주목받고 있다[2, 3].

그러나 이런 피드백의 중요성에도 불구하고 실제

[†]정 회 원: 한국교원대학교 교육대학원 컴퓨터교육과 석사과정

^{**}중신회원: 한국교원대학교 컴퓨터교육과 교수(교신저자)

논문투고: 2024년 01월 04일, 심사완료: 2024년 03월 13일, 게재확정: 2024년 03월 20일

* 본 논문은 제1저자의 한국교원대학교 교육대학원 석사학위논문 일부를 발췌하여 요약, 정리한 것임.

교육 현장에서는 이를 실현하는 데에 상당한 제약이 존재한다. 특히 교사들은 학생들에게 효과적이고 개인화된 피드백을 제공하기 위해 노력하지만, 이러한 노력은 교사의 역량과 자원, 시간 등 여러 요인에 의해 제한되곤 한다. 따라서 학생들에게 지속적이고 최적화된 피드백을 제공하는 것은 현실적으로 까다로운 일이며, 이는 교육의 질을 향상하는 데에 장애요인으로 작용할 수 있다. 이와 같은 문제점을 해결하기 위해 교육 기술 분야에서는 인간 교사를 대신할 수 있는, 인공지능을 기반으로 한 자동 평가 및 피드백 시스템에 관한 연구가 진행되었으며, 특히 최근에는 복잡한 자연어 처리를 필요로 하는 서답형 평가에서도 인공지능 기술을 적용하여 효과적인 평가와 피드백을 제공하는 방안을 모색하고 있는 것으로 보인다[4, 5].

이러한 상황에서 인공지능 기술의 혁신적인 진보는 기존 연구의 한계를 극복하고 효과적으로 피드백을 제공할 가능성을 제시한다. 따라서, 본 논문에서는 첨단 인공지능 기술 중 특히 자연어 처리 능력이 뛰어난 ChatGPT를 기반으로 인공지능이 실제 교육 현장에서 인간 교사의 피드백을 얼마나 효과적으로 대체할 수 있는지를 탐구하였다. 이를 위해 수학과 확률 및 통계 영역의 서답형 평가에서 학생들에게 제공된 ChatGPT의 피드백과 인간 교사의 피드백을 비교 분석하였으며, 피드백의 효과를 측정하기 위한 주요 도구로 피드백 리터러시라는 개념을 사용하였다[6, 7]. 피드백 리터러시는 학생들이 피드백을 이해하고 활용할 수 있는 능력을 의미하며, 학생들의 피드백 리터러시 수준을 정확히 파악하고 측정함으로써, 피드백 수업의 효과를 보다 명확하고 객관적으로 평가할 수 있을 것이다.

덧붙여 국내외의 기존 연구를 검토한 결과, 수학 교과에서 서답형 평가에 대한 인공지능 기반 자동 채점 및 피드백 연구는 상대적으로 부족함을 확인하였으며, 특히 국내 대다수의 관련 연구는 국어 교과나 영어 작문 채점 등에 집중되어 있었다[4]. 이러한 배경 속에서, 본 연구는 수학 교과에서 인공지능 기반의 서답형 평가 자동 채점 및 피드백 시스템 도입의 첫걸음으로서 의의를 지니며, 인공지능 기반의 피드백 시스템이 교육 현장에서 실질적으로 활용될 가능성을 탐색하고, 그 효과와 한계를 식별하는데 기여하고자 한다.

2. 이론적 배경 및 선행연구

2.1 서답형 평가 및 피드백에서 ChatGPT의 활용 가능성

서답형 평가는 학생들의 고차원적 사고 능력을 선다형 평가보다 정밀하게 측정할 수 있다는 점에서 이점을 가지고 있다. 그러나 이러한 서답형 평가는 실행 과정에서 상당한 자원을 필요로 한다. 특히, 각각의 서답형 응답을 철저히 검토하고 평가하기 위해선 전문가의 시간과 노력이 요구된다. 더불어, 이러한 평가를 수행하기 위한 물리적 자원, 장비 및 시설 등의 추가적인 비용이 발생하는 등 인적 및 물적 자원의 효율성 문제를 초래할 수 있다. 따라서, 서답형 평가의 교육적 가치는 충분하지만, 그 실행에 따른 부담이 크므로 이를 무분별하게 확대 및 적용하기는 어렵다는 점을 인정해야 한다. 이런 문제를 해결하기 위해 국외·국내에서 서답형 평가에 대한 인공지능 프로그램의 활용 방안이 연구되고 있다[5].

인공지능 기반 서답형 평가 및 피드백 프로그램에 대한 선행 연구들은 두 가지 주요한 한계점에 직면한 것으로 보인다[5]. 첫째, 이들 프로그램은 주로 특정 목적에 최적화되어 연구 및 개발됐다. 이는 교육과정, 학습자의 필요성, 교육환경 등 다양한 요인이 빠르게 변화하는 현대 교육 현장에서 유연성의 부족을 드러내며, 시간이 지남에 따라 투입된 자원에 대한 손실을 줄 가능성이 있다. 둘째, 현존하는 대부분의 인공지능 평가 프로그램은 정확도와 신뢰성을 높이기 위해 대량의 학습 데이터를 필요로 한다. 이는 데이터 확보에 대한 비용과 시간적 제약이 동반되며, 특히 특정 주제나 주제 영역에 대한 데이터가 제한적인 경우 그 한계가 더욱 명확해진다.

이런 문제에 대한 대안으로 Zero-Shot Learning 또는 Few-Shot Learning이 가능한 첨단 인공지능을 자동 채점 및 피드백 영역에 도입할 수 있다. Zero-Shot Learning은 인공지능이 학습한 적 없는 문제에 대해 인공지능이 예측할 수 있는 능력을 의미한다[8, 9]. 비슷한 개념으로 Few-Shot Learning이라는 용어가 있으며 이는 인공지능이 매우 적은 수의 훈련 예제를 사용하여 새로운 문제를 학습하는 능력을 의미한다[10].

Zero-Shot Learning 및 Few-Shot Learning 기술은 기존 인공지능 평가 프로그램의 제한 사항을 극복하는데 중요한 역할을 할 수 있다. 이러한 기술들은 특정 과목, 주제, 또는 학습자 그룹에 특화된 연구 및 개발 프로그램의 종속성을 줄이면서도 광범위한 교육적 맥락과 다양한 환경에서의 적용성을 확장할 수 있다.

따라서 교육자들은 주어진 상황과 환경에 맞게 인공지능 기반의 평가 도구를 맞춤화하거나 수정하는데 있어 더 큰 유연성을 경험할 수 있으며, 이는 교육 프로세스의 효율성 향상과 학습 결과의 질적 향상에 기여할 것으로 예상된다. 더군다나, 이러한 기술을 갖춘 인공지능 시스템은 대량의 학습 데이터에 크게 의존하지 않기 때문에, 데이터 수집 및 처리에 관련된 비용 및 시간을 절약할 수 있다. 특히, 데이터가 제한된 실세계의 많은 상황에서 이러한 시스템은 인공지능 평가 프로그램의 활용 범위를 크게 확장할 수 있을 것이다. 이는 교육 분야에서의 혁신적인 변화와 성과 향상을 위한 중요한 수단으로 간주될 수 있다.

ChatGPT는 많은 양의 텍스트 데이터로부터 전이학습을 통해 학습된 언어 모델인 GPT 엔진 기반의 채팅 서비스로서, 그 자체로는 명확한 Zero-Shot Learning 또는 Few-Shot Learning 기능을 목표로 만들어진 것은 아니다. 그러나, 다양한 언어 패턴을 학습하고 본적 없는 질문에 대한 합리적인 응답을 생성하는 능력을 가지고 있기 때문에, 이를 Zero-Shot Learning 또는 Few-Shot Learning 능력이라고 간주할 수 있다 [11]. 특히, ChatGPT의 Zero-Shot 학습 능력에 중점을 둔 선행연구[12]에서는 다양한 산술 및 자연어 추론 문제를 포함하는 데이터셋을 활용하여 실험하였고, 연구 결과 ChatGPT는 특정 작업에 특화된 인공지능과 비교하여 성능은 상대적으로 미흡하거나 비슷하지만 다양한 작업을 수행하는 능력을 지녔음이 확인되었다.

단순히 성능만을 기준으로 ChatGPT의 교육적 활용도를 판단한다면 그 가치는 제한적일 것이다. 그러나 ChatGPT는 특정 목적에 최적화된 기존의 인공지능 프로그램과 달리 다양한 교육적 맥락과 상황에서 적용 가능하며, 교사들이 상황과 환경에 따라 프로그램을 수정하거나 개인화하는데 유연하게 활용이 가능하다. 또한 대량의 학습 데이터에 크게 의존하지 않으며, 적은 양의 데이터로도 높은 성능을 발휘할 수 있다. 따라서 ChatGPT의 여러 장점을 종합적으로 고려했을 때, 일정 수준의 성능만 확보된다면 학교 현장에서 고부담 평가가 아닌 저부담 평가의 체점 및 피드백에 유용하게 사용될 수 있을 것이라 예상된다.

고부담 평가는 학생의 전반적인 성적과 입시에 큰 영향을 주는 평가로, 이러한 평가에서는 ChatGPT가 아직 신뢰성 있는 평가를 보장하는 데 한계가 있다. 이러한 고부담 평가에서는 미세 조정 및 최적화된 인공지능의 활용이 필요하다. 반면에, ChatGPT의 진정

한 가치는 일일 수행평가, 진단평가, 형성평가와 같은 저부담 평가에서 발휘될 것이다. 이러한 평가에서는 성능의 완벽성보다는 평가의 양과 다양성, 그리고 즉각적인 피드백이 중요하다. 따라서 ChatGPT와 같은 인공지능의 활용은 교사의 부담을 줄이고, 학생들에게 효과적인 피드백을 제공하는 데 큰 도움이 될 것이다. 본 논문의 주요 목적도 저부담 평가와 같은 상황에서 수학과 서답형 문제에 관한 ChatGPT의 평가 및 피드백이 인간 교사의 그것과 비교하여 어느 정도의 효과성을 발휘하는지 연구하는 것이다.

2.2 형성적 피드백

형성적 피드백은 학습 과정과 결과에 중점을 둔 피드백의 형태로, 학생의 학습 향상을 최우선으로 하는 목표를 가지고 있다. 이는 학습자의 현재 학습 수준과 필요한 개선사항에 초점을 맞추는 것으로, 학습자에게 앞으로 나아가야 할 방향을 제시하는데 효과적인 방법이다[13, 14]. 형성적 피드백은 단순한 정보 전달을 넘어, 학습자와 교사 사이의 상호작용을 통한 학습 진전에 초점을 둔다[15]. 본 연구는 실험 설계 단계에서 Figure 1과 같은 피드백 모형[16]을 사용하였다.

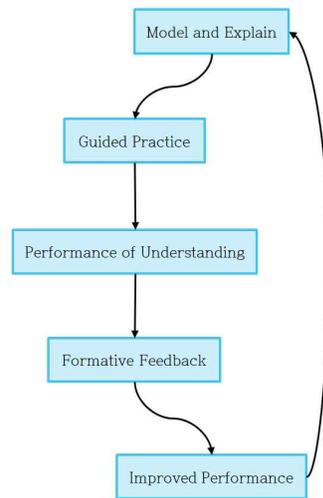


Figure 1. Circular process model of formative feedback

이 모형은 여러 단계를 통해 순환적인 구조를 형성하게 되는데, 초기 단계인 Model and Explain 단계에서는 설명과 모델링에 중점을 둔다. 이후 Guided Practice 단계에서 학습자는 교사의 지도하에 연습을

수행한다. Performance of Understanding 단계에서는 학습자의 이해도를 평가한 후, Formative Feedback 단계에서는 이러한 평가 결과를 바탕으로 학습자에게 피드백을 제공한다. Improved Performance 단계에서는 받은 피드백을 기반으로 학습자의 성과를 개선하는 활동을 진행하며, 이 단계의 완료 후에는 다시 Model and Explain 단계로 회귀하여 순환 구조를 계속 반복하게 된다.

2.3 ChatGPT를 활용한 형성적 피드백

ChatGPT를 서답형 평가 및 피드백에 적용하는 방법론은 다양하게 제시될 수 있지만, 본 연구에서는 기존의 형성적 피드백의 순환 모형 내에서 Performance of Understanding 단계와 Formative Feedback 단계 사이에 ChatGPT에 의한 서답형 평가 및 피드백 과정을 통합한 모형을 제안하며, 해당 모형은 Figure 2와 같다.

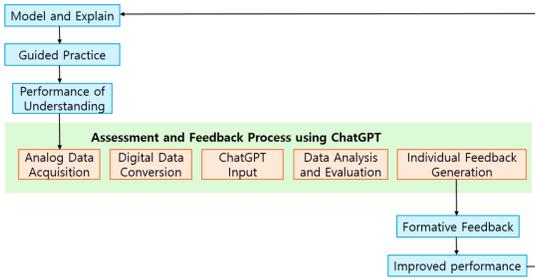


Figure 2. Feedback process model with ChatGPT

이 모형의 각 단계에서 일어나는 일은 구체적으로 다음과 같다. 먼저 Analog Data Acquisition 단계에서는 학생들의 서답형 평가에 대한 아날로그 형태의 데이터를 수집한다. 이후 Digital Data Conversion 단계에서는 다양한 기술, 예를 들면 스캐너, 카메라 및 OCR을 활용하여 아날로그 데이터를 디지털 형식으로 전환한다. ChatGPT Input 단계에서는 변환된 디지털 데이터를 ChatGPT에 적절하게 입력하기 위해, 인공지능 기술 중 하나인 프롬프트 엔지니어링을 적용한다. 이 기술은 학생의 개별적 특성 및 응답 스타일을 고려하여 맞춤형 피드백을 제공하는 데 중요하며, 출력 형식, 길이, 톤 등의 요구 사항을 정의한다. Data Analysis and Evaluation 단계에서는 ChatGPT를 활용하여 학생의 응답을 깊이 있게 분석하고 평가하며, GPT 엔진의 지속적인 발전은 이 단계의 효율성을 더

욱 향상시킬 것으로 예상된다. 마지막으로, Individual Feedback Generation 단계에서는 프롬프트 엔지니어링을 통해 설정된 기준에 따라 ChatGPT가 각 학생에게 개별적인 형성적 피드백을 생성한다.

2.4 피드백 리터러시

피드백의 효과를 측정하는 다양한 방법들이 존재하나, 본 연구에서는 피드백 리터러시라는 개념에 주목하고 이를 측정 도구로 사용하였다. 피드백 리터러시는 학생이 받은 피드백을 어떻게 해석하고, 그것을 자신의 학습에 어떻게 적용하는지에 대한 능력을 반영하는 지표이다. 이는 학생이 실질적으로 피드백을 통해 학습을 개선하는 과정에서 중요한 역할을 하는 요소로, 피드백의 효과를 정확하게 파악하는 데 있어 중요한 도구로 간주 된다. 특히, 본 연구에서는 선행 연구[7]에서 피드백 리터러시에 대한 정의와 그 하위영역을 바탕으로 개발한 피드백 리터러시 척도 문항(35 문항)을 사용하여 학생들의 피드백 리터러시를 측정하였다.

3. 연구 방법

3.1 연구 가설

본 연구는 수학과 ‘확률과 통계’ 영역의 서답형 평가에서 ChatGPT의 피드백 효과와 인간 교사에 의한 피드백 효과를 비교·분석함으로써 ChatGPT가 인간 교사의 역할을 보완 또는 대체할 수 있는지 검토하고자 한다. 따라서 다음과 같은 연구 가설을 설정하였다.

[연구문제] 수학과 ‘확률과 통계’ 영역의 서답형 평가에서 ChatGPT의 피드백을 받은 학생과 인간 교사의 피드백을 받은 학생들 간에 피드백 리터러시에 유의미한 차이가 있는가?

[연구가설] ChatGPT를 통해 피드백을 받은 학생과 인간 교사의 피드백은 학생들의 피드백 리터러시에 통계적으로 유의미한 차이가 없을 것이다.

3.2 연구 대상

본 연구의 연구 대상은 부산광역시 동래구에 소재하고 있는 S고 3학년 ‘확률과 통계’를 선택한 2개

학급을 대상으로 한다. 각 학급에서 임의로 절반의 학생을 선택하여 실험집단으로 지정하고, 나머지 절반을 통제집단으로 지정한 뒤, 실험집단에는 ChatGPT를 활용한 피드백을, 통제집단에는 인간 교사의 피드백을 제공한다. 이때, 학생들에게 수집된 데이터를 분석하는 과정에서 사전 또는 사후 검사 미응답, 불성실 응답, 불참자, 변수의 평균값을 기준으로 한 이상치(outlier) 등의 결측값을 제외하였다. 실제 통계 분석 대상자로 선정된 인원은 Table 1과 같다.

Table 1. Actual participants for statistical analysis

Categories	Feedback Provider	Male	Female	Total
Experimental Group	ChatGPT	5	11	16
Control Group	Human Teacher	7	10	17

3.3 연구 설계

본 연구에서는 연구 대상 학교인 S고등학교의 학사 일정을 고려하여 2023년 7월 초부터 여름 방학이 보충 수업이 진행되는 7월 말까지 실험 처치 기간으로 설정하였다. 이 시기에 설문지를 이용하여 피드백 리터러시에 관한 사전 검사를 한 뒤 서답형 평가를 2회 실시하고 각 집단에 대해 피드백 실시를 계획하였으며, 이후 사후 피드백 리터러시 검사를 실시하였다. 구체적인 연구 설계는 Table 2와 같다.

Table 2. Research design

G_1	O_1	X_1	O_3
G_2	O_2	X_2	O_4

G_1 : Experimental Group

G_2 : Control Group

X_1 : Evaluation and Feedback by ChatGPT

X_2 : Evaluation and Feedback by Human Teacher

O_1, O_2 : Pre-test (Feedback Literacy)

O_3, O_4 : Post-test (Feedback Literacy)

연구 대상을 선정하는 단계에서 ChatGPT를 통해 피드백을 제공받을 실험 집단과 인간 교사가 직접 피드백을 제공할 통제 집단을 임의로 나누어 사전 피드백 리터러시를 측정한다. 이때 연구 결과의 신뢰도를 높이기 위해 학생들 중 일부가 인공지능을 통해 피드백 받는다는 사실을 알려주고 학생들에게 동의를 받지만, 실제로 누가 인공지능을 통해 피드백 받는지는

알려주지 않는 블라인드 테스트를 실시한다.

이후 서답형 평가(4문제)를 실시하고 이에 관한 피드백을 한 뒤, 똑같은 유형의 다른 문제에 대해 서답형 평가를 실시하고 이에 관한 피드백을 반복한다. 이때 실험 집단은 ChatGPT에게 피드백을 제공받고 통제 집단은 인간 교사에게 피드백을 제공받은 뒤 두 그룹의 사후 피드백 리터러시를 측정한다.(학생별 피드백은 총 2회 실시)

마지막으로, 두 집단의 피드백 리터러시 측정 결과에 대해 독립표본 t-검정을 실시하여 인공지능을 통해 제공된 피드백의 효과와 인간 교사가 제공한 피드백의 효과에 차이가 있는지 분석한다.

위 실험의 결과 분석을 통해 학생들의 피드백 리터러시가 피드백 제공자에 따라 통계적으로 유의미한 차이가 없음을 확인된다면, 이는 인공지능 ChatGPT가 인간 교사의 피드백을 유효하게 대체할 수 있다는 것을 시사할 것이다. 반면, 통계적으로 유의미한 차이가 발견된다면, 이는 현재 단계의 인공지능 기술로는 아직 인간 교사의 피드백 능력을 대체하기에는 한계가 있음을 의미한다.

3.4 연구 수행

본 연구에서 서답형 평가는 총 2회 실시하며 한 회당 4문항을 출제하고 이에 대한 피드백은 1~2일 내에 제공한다. 이는 S 고등학교의 학사 일정 및 학생들의 문제를 검토하고 피드백을 제공하기 위한 연구자의 시간적 요소 등을 종합적으로 고려하여 결정하였다. 서답형 평가의 내용은 S고등학교 3학년 확률과 통계 과목의 1학기 2차 지필평가 시험 범위에서 출제하며, 구체적인 성취기준의 범위는 [확통02-05] ~ [확통03-07]이고, 단원명으로는 조건부 확률부터 통계적 추정까지이다. 이때 ChatGPT는 텍스트를 기반으로 만들어진 인공지능이기 때문에 문제 풀이에 함수의 그래프나 기하학적 요소가 필요한 연속확률분포, 정규 분포, 모평균의 추정 등은 제외하였다. 다만 표로 만들어진 데이터는 인식할 수 있기 때문에 통계 단원의 이산확률분포는 포함하였다. 최종적으로 결정된 서답형 평가 4개 문항은 각각 독립 사건, 독립 시행, 이산확률분포의 평균, 이항분포에 관련된 문제로 정하였으며 구체적으로는 Figure 3과 같다.

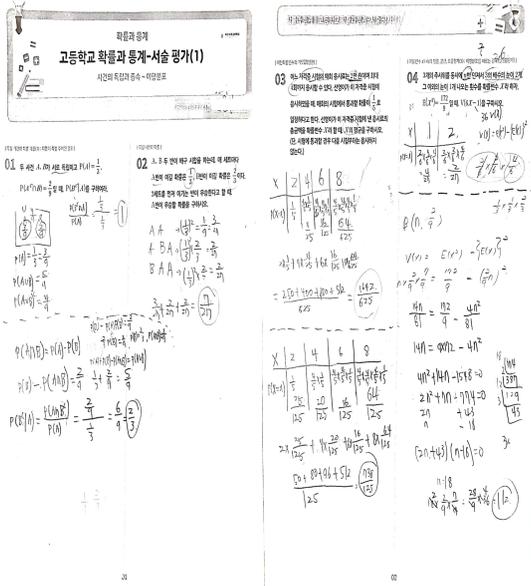


Figure 3. Descriptive essay problems for students

한편, 학생들의 서답형 평가 답안에 대한 ChatGPT의 피드백에 일관성을 부여하고 학생들의 수용성을 높이는 동시에 학생들을 학업 성취도 수준에 따라 맞춤형 피드백을 제공하기 위해, 적절한 프롬프트 엔지니어링을 통해 학생들의 서답형 답안을 ChatGPT에 입력하였으며, 내용은 Figure 4와 같다.

당신의 임무는 4개의 수학 서답형 문제에 대한 학생의 답이 올바른지 평가하고, 이에 대한 피드백을 제공하는 것입니다.

당신의 임무를 수행하려면 다음 지침을 따라주세요:

- 첫째, 주어진 문제의 해답과 학생의 해답을 비교해보고 학생의 해답이 올바른지 평가하세요.
- 둘째, 학생의 해답이 올바르지 않다면, 무엇이 문제인지 주어진 예시와 같은 느낌으로 피드백 해주세요.
- 셋째, 학생이 맞춘 문제 개수에 따라, 3개 이상을 맞췄으면 상 수준, 1개 이하를 맞췄으면 하 수준, 그 외는 중 수준으로 판단하고, 학생의 수준에 맞춰 학생을 격려하기 위한 종합 피드백을 주어진 예시의 느낌처럼 해주세요.

문제를 당신이 직접 풀이기 전에는 학생의 해답이 올바른지 판단하지 마세요.

문제는 교사가 학생에게 상냥하고 친절하게 가르쳐주는 어투로 써주세요.

종합 피드백은 200자 이내로 써주세요.

다음 예시와 같은 형식을 사용하세요:

문제:
...
이곳에 문제를 작성하세요.
...
문제의 해답:
...
이곳에 문제의 해답을 작성하세요.
...
학생의 해답:
...

Figure 4. Excerpts from ChatGPT input prompts

이때, 프롬프트 엔지니어링은 OpenAI사에서 공개한 ‘ChatGPT Prompt Engineering for Developers’ [17]의 가이드라인을 참조하였으며, 실제 ChatGPT가 수학적 문제를 해결하는데 프롬프트 엔지니어링의 존재 여부에 따른 효과는 미비하다는 연구 결과[18]에 따라, 문제 해결 과정에 프롬프트 엔지니어링을 집중하기보다는 ChatGPT가 피드백을 출력할 형식과 절차, 말투, 학생 수준에 따른 맞춤형 피드백에 집중했다. 추후 학생의 성취목표지향성이나 성별 등의 개인적 특성에 따라 어떤 피드백이 효과적인지에 대한 연구가 진행된다면, 이에 맞춰 개별적인 피드백을 출력하도록 추가할 수 있을 것이다.

또한, 인공지능과 인간 교사의 피드백 결과를 학생들에게 제공할 때, 그 형식은 Figure 5와 같이 일관되게 통일하였다.

학점과 통계-서울 평가(1)-개별 피드백		
		대상 학생 : 0000
	정답	학생 답
1번 문항	$\frac{2}{3}$	
2번 문항	$\frac{7}{27}$	
3번 문항	$\frac{738}{125}$ 만원 (=59040원)	
4번 문항	112	
학생 점수	/ 4 (맞은 문제 수 / 총 문제 수)	
1번 문제 피드백		
2번 문제 피드백		
3번 문제 피드백		
4번 문제 피드백		
종합 피드백		

Figure 5. Student Feedback Form

인간 교사의 피드백이든 인공지능의 피드백이든 모두 컴퓨터로 작성된 인쇄물(A4)의 형태로 배부하고 피드백 제공 주체는 밝히지 않는다. 이때 학생들의 서답형 답안을 프롬프트 엔지니어링을 통해 ChatGPT에 입력하고 처리하는 과정과 인간 교사가 피드백을 위해 필요한 시간 등으로 인해, 서답형 평가 후 1~2일 지연된 피드백이 학생들에게 제공되었으며, 인간 교사의 피드백 예시 및 ChatGPT의 피드백 예시는 Figure 6과 같다.

확률과 통계-서술 평가(1)-개별 피드백		
대상 학생 : 0000		
	정답	학생 답
1번 문항	$\frac{2}{3}$	1
2번 문항	$\frac{7}{27}$	7/27
3번 문항	$\frac{738}{125}$ 만원(=59040원)	1642/625
4번 문항	112	
학생 점수	1 / 4 (맞은 문제 수 / 총 문제 수)	
1번 문제 피드백	사건 A와 B가 독립이라는 명제는 사건 A'와 사건B가 독립이라는 명제, 사건 A과 사건 B'가 독립이라는 명제 등과 필요충분조건이라는 사실을 이용하여 문제를 쉽게 해결할 수 있습니다.	
2번 문제 피드백	매우 훌륭합니다. 주어진 문제를 몇 가지의 배반사건으로 나누어 논리적으로 문제를 해결하였습니다.	
3번 문제 피드백	이산확률분포표에서는 모든 확률의 합이 1이 되어야 합니다. 시간이 충분하다면 검산과정을 통해 혹시 자신이 실수한 부분이 있는지 점검해야 합니다.	
4번 문제 피드백	주어진 문제는 이항분포에 관한 문제입니다. 이항분포 문제는 그 문제가 이항분포를 따른다는 걸 인지하는 것부터가 매우 중요한 단계입니다.	
종합 피드백	2번 문제를 해결하는 과정을 보면 논리적으로 사고하고 문제를 해결하는 능력이 충분한 학생이라고 생각합니다. 중간 중간 개념에 대한 이해도가 부족한 경향이 있으니 이에 대해서 보완만 한다면 좋은 결과를 보여줄 거라 생각합니다.	

확률과 통계-서술 평가(1)-개별 피드백		
대상 학생 : 0000		
	정답	학생 답
1번 문항	$\frac{2}{3}$	$\frac{2}{3}$
2번 문항	$\frac{7}{27}$	$\frac{7}{27}$
3번 문항	$\frac{738}{125}$ 만원(=59040원)	29264원
4번 문항	112	72
학생 점수	2 / 4 (맞은 문제 수 / 총 문제 수)	
1번 문제 피드백	두 사건 A, B가 서로 독립이면, 두 사건 A', B'도 서로 독립이라는 사실을 사용하여 두 사건 A, B'가 서로 독립이라는 것을 인지하셔야 합니다. 이를 통해 P(A'∩B')를 계산할 수 있습니다.	
2번 문제 피드백		
3번 문제 피드백	확률변수 X의 확률분포표를 만드는 과정에서 P(X=6000)을 계산하는 부분에서 실수가 있습니다. P(X=6000)은 이전 사항에 모두 불합격했을 확률 즉 (4/5)^3=64/125가 되어야 합니다. 이를 바탕으로 E(X)를 다시 계산해보시길 바랍니다.	
4번 문제 피드백	E(X^2)와 E(X)를 활용하여 식을 세우는 것은 맞습니다. 하지만 식을 정리하고 풀이하는 과정에서 오류가 발생하였습니다. E(X^2)의 값과 E(X)=n*(2/9)를 이용하여 n을 올바르게 계산하고 이를 바탕으로 E(X)를 다시 계산해야 합니다. 이후, 확률 변수의 선형변환의 공식에 대한 정의를 사용하여 V(aX-b)를 계산하시면 됩니다.	
종합 피드백	문제를 해결하는 과정에서 필요한 기본 원리를 적용하는 능력이 보입니다. 그러나 일부 문제에서는 확률의 개념에 대한 완전한 이해가 필요해 보입니다. 확률의 정의 및 확률변수에 대한 이해를 깊게 하여 이해도를 높이고, 다양한 유형의 문제에 대한 훈련을 통해 확률 문제를 더 잘 해결할 수 있도록 노력하세요. 기초 개념에 대한 충분한 이해가 후속 학습의 토대가 됩니다. 좋은 성과를 위해 계속 노력하세요!	

Figure 6. Example of Human Teacher Feedback (Top), Example of ChatGPT Feedback (Bottom)

한편, 학생들의 수준에 맞춰 ChatGPT가 생성한 종합 피드백의 예시는 Table 3과 같다. 이를 정성적으로 분석해보면, ‘상’ 수준 학생들에게는 이미 뛰어난 성과를 보인 부분에 대해 긍정적인 피드백을 부여하고, 동시에 높은 수준의 도전적인 과제를 수행할 것을 촉구한다. ‘중’ 수준 학생들에게는 이미 이해한 내용에 대한 긍정적인 피드백을 전달하면서 동시에 개선이 필요한 부분을 지적한다. ‘하’ 수준의 학생들에게는 기본적인면서 필수적인 개념을 강조하며, 해당 개념을 명확하게 이해할 수 있도록 격려한다.

Table 3. Examples of Comprehensive Feedback Generated by ChatGPT

Student Level	Feedback Generated by ChatGPT
High	“You really did a great job! You solved all the problems correctly and explained each step logically and wonderfully. This deep understanding and thorough preparation gives great results. Please continue to show us more like this. Your ability and passion for mathematics shines through. Wonderful!”
Medium	“You’re doing well! It seems like you already have a pretty good understanding of probability. You solved the problem perfectly, especially on problem number 2. Nevertheless, the problems that deal with more complex concepts are a little more difficult. In order to solve these problems, it’s important to make sure you understand and practice the basic concepts. If you want to study independence and the binomial distribution again and there are parts you don’t understand, Please feel free to ask any questions. If you keep working hard, you will get good results. I’m looking forward to it!”
Low	“I think you still lack understanding of the problem. But don’t worry, it may take some time to understand. Probability and statistics are complex concepts, so learn them slowly and one by one. is important. If you keep learning steadily, you will definitely grow. Good luck!”

3.5 검사 도구

피드백 리터러시를 측정하기 위한 검사 도구는 Figure 7과 같으며, 설문지에서 피드백 리터러시의 하위 요인은 피드백 활용능력과, 피드백 태도로 구성된다. 이때, 피드백 활용능력의 하위 요인은 피드업(1-3번 문항), 피드백(4-9번 문항), 피드포워드(10-14번 문항)으로 구성되고, 피드백 태도의 하위요인은 인지(15-24번 문항), 정서(25-27번 문항), 정서조절(28-31번 문항), 피드백 구하기(32-35번 문항)으로 구성되며,

각 문항은 리커트 5점 척도로 측정된다.

학번: _____ 이름: _____

아래는 인공지능을 이용한 피드백에 학생의 '피드백 리터러시'에 미치는 효과를 조사하기 위한 사전 설문지입니다. 각 문항에서 나와 가장 기쁘다고 생각하는 것에 체크하고 그 번호를 OMR 카드에 표시하세요.

* 피드백이란: 학습 결과 또는 학습 과정 등 학습과 관련하여 주어지는 다양한 정보 모두를 의미합니다.
예를 들면 시험을 치고 난 뒤에 받는 시험결과, 숙제나 학습지를 하고 난 뒤에 받는 의견, 수업시간 중 학습활동에 대한 조언 등 학습에 대한 다양한 정보 모두를 의미합니다.

순번	문항	(1) 매우 그렇지 않다	(2) 그렇지 않다	(3) 다	(4) 그렇다	(5) 매우 그렇다
1	나는 나의 학습이 평가되는 기준(상, 중, 하 등)이 무엇인지 파악할 수 있다.					
2	나는 나의 학습이 평가되는 방법(포트폴리오, 지필, 논술 등)이 무엇인지 파악할 수 있다.					
3	나는 수업에서 학습 목표가 무엇인지 파악할 수 있다.					
4	나는 피드백을 통해 내가 잘 몰랐던 부분이나 틀린 부분을 파악할 수 있다.					
5	나는 피드백을 통해 내가 더 노력해야 하는 부분을 파악할 수 있다.					
6	나는 피드백을 통해 내가 얼마나 잘하고 있는지 파악할 수 있다.					
7	나의 학습과정에서 더 노력해야 할 점을 스스로 찾을 수 있다.					
8	나는 친구의 학습과정에서 잘한 점을 찾을 수 있다.					
9	나는 평가 기준을 활용하여 나의 학습을 스스로 판단할 수 있다.					
10	나는 피드백을 활용하여 다음에 적용해야 할 학습전략을 생각할 수 있다.					
11	나는 피드백을 활용하여 학습 방법을 수정할 수 있다.					
12	나는 피드백을 활용하여 나에게 부족했던 학습 내용을 수정할 수 있다.					
13	나는 학습 능력 향상을 위하여 학습 계획을 스스로 판단하고 수정할 수 있다.					
14	나는 평가기준에 맞게 나의 학습내용과 학습방법을 스스로 수정할 수 있다.					
15	나는 피드백을 받으면 이를 정확히 이해하고 학습에 활용할 수 있을 것이라 기대한다.					
16	나는 친구들에 비해 다른 사람의 학습결과물을 보고 더 좋은 의견을 줄 수 있다고 생각한다.					
17	나는 다른 친구들에 비해 학습활동에 있어 더 좋은 피드백을 받을 것이라 기대한다.					

Figure 7. Excerpts from Feedback Literacy Survey Form

4. 연구 결과

4.1 사전 검사 결과

먼저, 실험 집단(ChatGPT의 피드백을 받은 집단)과 통제 집단(인간 교사의 피드백을 받은 집단) 간의 사전 피드백 리터러시에 대해 독립표본 t-검정을 실시하여 두 집단의 동질성 여부를 확인하였으며, 결과는 Table 4와 같다.

Table 4. Pre-Assessment Results for Each Element of Feedback Literacy

Feedback Literacy	G	N	Mean	SD	Norm	HOV	t	p	MWU
Feedup	CG	17	12.24	2.047	O	O	-0.018	0.986	0.736
	EG	16	12.25	2.595	X				
FeedBack	CG	17	24.88	3.638	O	O	0.135	0.894	.
	EG	16	24.69	4.644	O				

FeedForward	CG	17	19.71	3.478	O	O	0.905	0.372	.
	EG	16	18.56	3.777	O				
Cognition	CG	17	42.00	4.924	X	X	0.966	0.342	0.533
	EG	16	39.94	7.197	O				
Emotion	CG	17	12.35	2.714	X	O	1.036	0.308	0.204
	EG	16	11.38	2.705	O				
Emotion Regulation	CG	17	16.12	3.498	O	O	-0.244	0.809	0.929
	EG	16	16.38	2.419	X				
Feedback Acquisition	CG	17	16.88	3.806	X	O	1.800	0.082	0.058
	EG	16	14.75	2.910	O				

독립표본 t-검정의 신뢰성 확보를 위한 전제 조건인 분산의 동질성을 검정하기 위해 Levene의 등분산성 검정을 실시하였으며, 피드백 리터러시의 하위 요소 중 피드업의 유의확률 0.660, 피드백의 유의확률 0.140, 피드 포워드의 유의확률 0.681, 정서의 유의확률 0.870, 정서조절의 유의확률 0.195, 피드백 구하기의 유의확률이 0.242로 유의 수준 0.05에서 각각 통계 집단과 실험 집단 사이의 등분산성 가정을 충족시켰다. 다만 인지 요소에서는 유의확률이 0.048로 등분산성 가정(약어 HOV로 표기)를 충족시키지 않았다.

한편 실험집단과 통제집단의 표본의 크기가 30이하로 작았기 때문에 정규성을 검정하기 위해 Shapiro-Wilk의 정규성 검정을 실시하였으며, 피드백 리터러시의 하위 요소 중 실험집단 피드업의 유의확률 0.042, 통제집단 인지의 유의확률 0.002, 통제집단 정서의 유의확률 0.006, 실험집단 정서조절 유의확률 0.028, 통제집단 피드백 구하기의 유의확률 0.002로 유의 수준 0.05에서 각각 정규성을 충족시키지 않았고, 나머지 요소들은 정규성 가정(약어 Norm으로 표기)를 충족시켰다.

독립표본 t-검정을 실시한 결과 피드백 리터러시 하위 요소 중 피드업의 유의확률 0.986, 피드백의 유의확률 0.894, 피드 포워드의 유의확률 0.372, 인지의 유의확률 0.342, 정서의 유의확률 0.308, 정서조절의 유의확률 0.809, 피드백 구하기의 유의확률 0.082로 실험 집단과 통제집단은 유의 수준 .05에서 통계적으로 유의미한 차이가 없는 동질 집단임을 확인하였다. 이때, 측정값이 등분산성 또는 정규성을 만족시키지 않는 경우, 비모수 검정 방법 중 하나인 Mann-Whitney U 검정을 추가하여 실험의 신뢰도를 높이고자 하였으며, Mann-Whitney U 검정 결과(약어 MWU으로 표기)에서도 피드백 리터러시의 각 요소에서 실험집단과 통제집단은 통계적으로 유의미한 차이가 없는 동질 집단임을 확인하였다.

4.2 사후 검사 결과

사전 검사 이후 2회에 걸쳐 수학과 ‘확률과 통계’ 영역의 서답형 평가에서 형성적 피드백을 실시하였으며, 실험집단에는 ChatGPT가 피드백 하였고 통제집단에는 인간 교사가 피드백을 하였다. 이후 피드백 리터러시의 차이를 분석하기 위해 사후 검사를 실시하고 독립표본 t-검정을 통해 비교한 결과는 Table 5와 같다.

Table 5. Post-Assessment Results for Each Element of Feedback Literacy

Feedback Literacy	G	N	Mean	SD	NormHOV	t	p	MWU
Feedup	CG	17	13.24	1.954	X	0.702	0.488	0.423
	EG	16	12.75	2.017	X			
FeedBack	CG	17	26.59	3.890	O	0.675	0.504	.
	EG	16	25.63	4.303	O			
FeedForward	CG	17	21.53	4.110	X	0.670	0.508	0.382
	EG	16	20.63	3.612	O			
Cognition	CG	17	44.82	6.682	X	0.842	0.406	0.363
	EG	16	42.88	6.602	O			
Emotion	CG	17	13.29	2.339	X	1.251	0.220	0.345
	EG	16	12.13	3.008	X			
Emotion Regulation	CG	17	17.76	3.437	X	0.526	0.603	0.309
	EG	16	17.19	2.810	X			
Feedback Acquisition	CG	17	17.65	3.334	X	1.666	0.106	0.074
	EG	16	15.69	3.420	X			

Levene의 등분산 검정을 실시한 결과 피드업의 유의확률 0.635, 피드백의 유의확률 0.433, 피드 포워드의 유의확률 0.730, 인지의 유의확률 0.423, 정서의 유의확률 0.087, 정서조절의 유의확률 0.623, 피드백 구하기의 유의확률 0.466으로 유의수준 0.05에서 피드백 리터러시의 모든 하위 요소가 등분산성을 충족시켰다.

Shapiro-Wilk의 정규성 검정을 실시한 결과 통제집단과 실험집단의 피드업의 유의확률이 각각 0.019, 0.013, 피드백의 유의확률이 각각 0.103, 0.134, 피드 포워드의 유의확률이 각각 0.040, 0.088, 인지의 유의확률이 각각 0.029, 0.162, 정서의 유의확률이 각각 0.012, 0.015, 정서조절의 유의확률이 각각 0.000, 0.187, 피드백 구하기의 유의확률이 각각 0.008, 0.061으로 유의수준 0.05에서 피드백 요소 등 일부를 제외한 대부분의 피드백 리터러시의 하위 요소가 정규성 가정을 충족시키지 못했다.

피드백 리터러시의 하위 요소별로 두 집단의 차이

를 살펴보기 위해 독립표본 t-검정을 실시한 결과 피드업의 유의확률 0.488, 피드백의 유의확률 0.504, 피드 포워드의 유의확률 0.508, 인지의 유의확률 0.406, 정서의 유의확률 0.220, 정서조절의 유의확률 0.603, 피드백 구하기의 유의확률 0.106으로 두 집단은 유의수준 0.05에서 피드백 리터러시의 모든 요소에 통계적으로 유의미한 차이가 없었다. 이때, 측정값이 정규성을 만족시키지 않는 경우 실시한 Mann-Whitney U 검정 결과에서도 피드백 리터러시의 모든 요소에서 두 집단은 통계적으로 유의미한 차이가 없었다. 따라서 연구 가설인 “ChatGPT를 통한 피드백을 받은 학생과 인간 교사의 피드백은 학생들의 피드백 리터러시에 통계적으로 유의미한 차이가 없을 것이다.”가 채택되었다.

5. 결론 및 제언

본 연구에서는 수학과 확률과 통계 영역에서 ChatGPT의 서답형 평가 피드백이 학생들의 피드백 리터러시에 미치는 영향을 분석하였다. 연구의 결론으로 피드백 제공자가 인간 교사든 ChatGPT이든 피드백 리터러시의 하위 구성 요소인 피드업, 피드백, 피드포워드, 인지, 정서, 정서조절 및 피드백 구하기에서 통계적으로 유의미한 차이가 없음을 밝혀졌다. 이는 ChatGPT가 인간 교사와 동등한 수준의 피드백을 제공할 수 있음을 의미하며, 향후 교육 현장에서 인공지능 기반의 피드백 도구를 효과적으로 활용하기 위한 교수법 및 교육전략의 개발에 유용한 지침을 제공할 수 있을 것으로 기대된다.

본 연구는 국내 수학 교육 분야에서 학생들의 서답형 평가 답안을 인공지능 기술로 평가하고 피드백하는 과정을 체계적으로 분석한 최초의 시도로 그 중요성을 띠고 있다. 본 연구에서 활용된 ChatGPT는 수학과 ‘확률과 통계’ 영역에서 학생들의 서답형 답안. 여기에 포함된 다양한 수식, 확률과 통계 용어, 확률 분포표 등의 복잡한 요소들을 정밀하게 평가하고 피드백하는 능력을 갖추고 있으며, 인공지능을 통한 평가와 피드백은 인간 교사의 평가와 피드백과 비교하여 학생들의 피드백 리터러시에 유의미한 차이를 보이지 않았다. 이러한 연구 결과는 수학 교과의 서답형 평가 및 피드백에서 인공지능의 활용에 대한 가능성을 열어놓았음을 시사한다.

더욱이, 본 연구에서 활용된 ChatGPT는 GPT-4 엔

진을 기반으로 하며, 이 기술의 지속적인 발전과 함께 그 성능도 향상될 것으로 예상된다. 또한 인공지능을 활용한다면 피드백의 즉시성, 일관성을 충족시킬 수 있으며, 학생의 개별적인 학습 상황과 조건에 맞춰 맞춤형 피드백을 제공할 수 있다. 특히 한 번의 정밀한 설정과 구축을 통해 교육 과정에서의 자원 손실을 크게 줄이며, 교육자의 역할을 효과적으로 보완하고 교육의 효율성을 높일 수 있는 잠재력을 보유하고 있다.

본 연구 과정에서 몇 가지 한계와 제한 사항을 인식하였으며, 이러한 제한들은 ChatGPT의 기술적 한계와 연구 실행에서의 한계로 구분된다.

ChatGPT의 기술적 한계로는 첫째, 연구의 진행 시점인 2023년 7월 OpenAI에서 발표한 가장 최신버전인 GPT-4 모델을 사용하는 ChatGPT를 기반으로 실험을 진행하였으나, 현재의 ChatGPT는 텍스트 기반의 데이터만 처리할 수 있다는 한계를 가지고 있다. 이러한 제약은 수학 교과에서 그래프, 기하학적 도형 등의 비텍스트 요소가 핵심적인 역할을 하는 경우 ChatGPT의 활용을 제약한다. 또한 학생들이 문제 해결 과정에서 도식화나 그림을 활용하는 경우에도 이러한 정보를 ChatGPT에 입력하거나 분석하는 것은 현재로서는 불가능하다. 그러나 ChatGPT와 관련된 기술은 지속적으로 발전하고 있으며, 향후 이미지나 그래픽을 처리할 수 있는 능력이 향상될 가능성이 있다.

둘째, ChatGPT의 정확성에 관한 한계이다. 본 연구 과정에서 ChatGPT의 출력 결과를 세밀하게 검토한 결과, 일부 경우에 ChatGPT의 정확성에 문제가 있음을 발견하였다. 예를 들어, 학생이 4개의 문제 중 3개를 올바르게 해결했음에도 불구하고, ChatGPT는 2개의 문제만 올바르게 답했다고 평가하는 경우와 학생의 답과 풀이 과정이 정확하다고 피드백을 제공하면서 실제로는 오답으로 평가하는 상황이 있었다. 학생들에게는 이러한 부정확한 평가가 포함된 상태로 피드백이 제공되었기 때문에 연구의 결과는 여전히 유효하다. 그러나 교육 평가의 신뢰도와 정확성이 중요한 상황, 특히 고부담 시험과 같은 중요한 평가 상황에서는 ChatGPT의 활용에 신중을 기해야 하며, 수학 교과에 적용되는 ChatGPT의 정확성을 평가하고 최적화하는 추가 연구가 필요하다.

한편, 연구 실행에서의 한계로는 첫째, 표본 규모가 33명으로 상대적으로 작아 데이터의 정규성 확보가 어려웠다. 이 때문에 다양한 변수에 따른 세부적인 분석을 실시하지 못했으며, 통계적 효력이 작은 비모수

검정 방법을 도입해야 했다.

둘째, 학생들이 수기로 작성한 아날로그 형태의 데이터를 디지털 형태로 변환하는 과정에서 어려움이 있었다. 현재로서는 데이터를 변환하는 과정에서 상당한 시간과 노력이 소요되었으며, 이러한 한계가 ChatGPT의 교육 현장에서의 활용을 제약하는 요인 중 하나로 작용할 수 있다. 추후 스캐너, 카메라, OCR 기술 등이 더욱 발전하고 관련 기술이 개발된다면, 이러한 데이터 변환 과정을 자동화하고 효율적으로 처리할 수 있을 것이다.

셋째, 본 연구는 수학 교과의 일부 영역, 특히 확률과 통계에 집중하여 실시되었다. 연구의 범위가 이러한 특정 주제에 국한되었기 때문에, 그 결과를 수학 교과의 다른 영역에 일반화하는 것에는 주의가 필요하다. 대수, 기하, 함수와 같은 복잡한 수학적 개념과 문제에 대해 인공지능 기반 피드백이 어떠한 효과를 나타내는지 평가하는 추가적인 연구가 필요하다.

참고문헌

- [1] Mcmillan, J. H., Venable, J. C., & Varier, D. (2013). Studies of the effect of formative assessment on student achievement: so much more is needed. *Practical Assessment, Research, and Evaluation, 18*, Article 2. DOI : 10.7275/tmwm-7792
- [2] Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153-189. DOI : 10.3102/0034654307313795
- [3] The Ministry of Education. (2015). *Comprehensive Overview of Elementary and Middle School Curriculum (No. 2015-74)*. <https://www.moe.go.kr/boardCnts/viewRenew.do?boardID=141&boardSeq=60747&lev=0&searchType=null&statusYN=C&page=1&s=moe&m=040401&opType=N>
- [4] Lee, G., & Ha, M. (2020). The Present and Future of AI-based Automated Evaluation: A Literature Review on Descriptive Assessment and Other Side. *Journal of Educational Technology, 36*(2), 353-382. DOI : 10.17232/KSET.36.2.353
- [5] Kim, J., & Park, I. (2023). *Comparison on Educational Effectiveness between Artificial Intelligence's Automatic Scoring Feedback and Human Instructor's Feedback*. Master's Thesis. Korea University.
- [6] Carless, D., & Boud, D. (2018). The development of student feedback literacy: enabling uptake of

feedback. *Assessment & Evaluation in Higher Education* Vol.43, Issue 8. DOI : 10.1080/02602938.2018.1463354

[7] Park, M., & Sohn, W. (2019). *Development and Validation of Feedback Literacy Scale for Students*. Ph.D. dissertation. Kyungpook National University.

[8] Palatucci, M., Pomerleau, D., Hinton, G., & Mitchell, T. M. (2009). Zero-shot learning with semantic output codes. *Advances in Neural Information Processing Systems*, 22, 1410-1418.

[9] Xian, Y., Christoph, H. L., Schiele, B., & Akata, Z. (2019). Zero-shot learning - a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41, Issue 9. DOI : 10.48550/arXiv.1707.00600

[10] Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. *31st Neural Information Processing Systems*. DOI : 10.48550/arXiv.1703.05175

[11] Brown, T.B., Mann, B., & Ryder, N. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33. DOI : 10.48550/arXiv.2005.14165

[12] Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., & Yang, D. (2023). Is chatgpt a general-purpose natural language processing task solver?. *arXiv.org*. DOI : 10.48550/arXiv.2302.06476

[13] Brookhart, S. M. (2017). *How to Give Effective Feedback to Your Students*. ASCD.

[14] Fluckiger, J., Tixier y Vigil, Y., Pasco, R. & Danielson, K. (2010). Formative feedback: involving students as partners in assessment to enhance learning. *College Teaching*, 58, Issue 4. DOI : 10.1080/87567555.2010.484031

[15] Carless, D. (2015). Exploring learning oriented assessment processes. *Higher Education*, 69, 963-976. DOI : 10.1007/s10734-014-9816-z

[16] Moss, C. M., & Brookhart, S. M. (2012). *Learning Targets: Helping Students Aim for Understanding in Today's Lesson*. ASCD. p.22.

[17] Fulford, I., & Ng, A. (2023, July 28). *ChatGPT Prompt Engineering for Developers*. <https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/>

[18] Frieder, S., Pinchetti, L., Griffiths, R. R., Salvatori, T., Lukaszewicz, T., Petersen, P. C., Chevalier, A., & Berner, J. (2023). Mathematical capabilities of chatgpt. *arXiv.org*. DOI : 10.48550/arXiv.2301.13867



김 옥 태

2011년 한국교원대학교
수학교육과(교육학사)
2024년 한국교원대학교 교육대학원
컴퓨터교육전공(교육학석사)

2024년 ~ 현재 부산 사직고등학교 교사
관심분야: 컴퓨터 교육, 수학 교육, 융합 교육
E-Mail: kwt052@naver.com



김 영 식

1982년 서울대학교 전기공학과(공학사)
1987년 노스캐롤라이나주립대학교 전기
및 컴퓨터공학(공학석사)
1993년~ 노스캐롤라이나주립대학교
전기 및 컴퓨터공학(공학박사)

1994년 ~ 현재 한국교원대학교 컴퓨터교육과 교수
관심분야: 컴퓨터 교육, 프로그래밍 교육, 피지컬 컴퓨팅, AI 교육
E-Mail: kimys@knue.ac.kr

부 록



[그림 2] 인공지능 ChatGPT를 활용한 피드백 과정 모형

<표 1> 연구 대상 중 통계 분석 대상자

구분	피드백 주체	남학생	여학생	합
실험집단	ChatGPT	5	11	16
통제집단	인간 교사	7	10	17

〈표 2〉 연구 설계

G_1	O_1	X_1	O_3
G_2	O_2	X_2	O_4

G_1 : 실험집단

G_2 : 통제집단

X_1 : ChatGPT에 의한 서답형 평가 및 피드백(2회 반복)

X_2 : 인간 교사에 의한 서답형 평가 및 피드백(2회 반복)

O_1, O_2 : 사전 검사(피드백 리터러시)

O_3, O_4 : 사후 검사(피드백 리터러시)

〈표 3〉 ChatGPT에 의해 생성된 종합 피드백 예시

학생 수준	ChatGPT에 의해 생성된 피드백
상	“정말 훌륭한 작업을 했어요! 모든 문제를 올바르게 풀었고, 각 단계별로 논리적으로 훌륭하게 설명했습니다. 이런 깊은 이해와 철저한 준비는 훌륭한 결과를 가져옵니다. 계속해서 이런 모습을 보여주세요. 수학에 대한 당신의 능력과 열정이 빛을 발하고 있습니다. 멋져요!”
중	“너무 잘하고 있어요! 이미 확률에 대한 이해도가 상당한 것 같습니다. 특히 2번 문제에서는 문제를 완벽하게 해결해주셨습니다. 그럼에도 불구하고 조금 더 복잡한 개념을 다루는 문제에서는 약간 힘들어하는 모습을 보았습니다. 이런 문제들을 해결하기 위해 기본 개념을 확실하게 이해하고 연습하는 것이 중요합니다. 독립성과 이항 분포에 대해 다시 한번 공부해 보시고, 이해가 안 가는 부분이 있다면 꼭 질문해주세요. 계속해서 노력하면 좋은 결과를 얻을 수 있을 것입니다. 기대하겠습니다!”
하	“아직 문제에 대한 이해가 부족한 것 같아요. 하지만 걱정하지 마세요, 이해하는 데 시간이 조금 걸릴 수 있습니다. 확률과 통계는 복잡한 개념이니, 천천히 하나씩 배워나가는 것이 중요해요. 꾸준히 학습을 계속하시면 반드시 성장하실 수 있을 거예요. 힘내세요!”

〈표 4〉 피드백 리터러시 각 요소별 사전 검사 결과

피드백 리터러시	집단 구분	N	평균	표준 편차	정규성 여부	등분산성 여부	t	p	Mann-Whitney 검정
피드업	통제	17	12.24	2.047	O	O	-0.018	0.986	0.736
	실험	16	12.25	2.595	X				
피드백	통제	17	24.88	3.638	O	O	0.135	0.894	·
	실험	16	24.69	4.644	O				
피드포워드	통제	17	19.71	3.478	O	O	0.905	0.372	·
	실험	16	18.56	3.777	O				
인지	통제	17	42.00	4.924	X	X	0.966	0.342	0.533
	실험	16	39.94	7.197	O				
정서	통제	17	12.35	2.714	X	O	1.036	0.308	0.204
	실험	16	11.38	2.705	O				
정서조절	통제	17	16.12	3.498	O	O	-0.244	0.809	0.929
	실험	16	16.38	2.419	X				
피드백 구하기	통제	17	16.88	3.806	X	O	1.800	0.082	0.058
	실험	16	14.75	2.910	O				

〈표 5〉 피드백 리터러시 각 요소별 사후 검사 결과

피드백 리터러시	집단 구분	N	평균	표준 편차	정규성 여부	등분산성 여부	t	p	Mann-Whitney 검정
피드업	통제	17	13.24	1.954	X	O	0.702	0.488	0.423
	실험	16	12.75	2.017	X				
피드백	통제	17	26.59	3.890	O	O	0.675	0.504	·
	실험	16	25.63	4.303	O				
피드포워드	통제	17	21.53	4.110	X	O	0.670	0.508	0.382
	실험	16	20.63	3.612	O				
인지	통제	17	44.82	6.682	X	O	0.842	0.406	0.363
	실험	16	42.88	6.602	O				
정서	통제	17	13.29	2.339	X	O	1.251	0.220	0.345
	실험	16	12.13	3.008	X				
정서조절	통제	17	17.76	3.437	X	O	0.526	0.603	0.309
	실험	16	17.19	2.810	X				
피드백 구하기	통제	17	17.65	3.334	X	O	1.666	0.106	0.074
	실험	16	15.69	3.420	X				