Journal of the Korean Association for Science Education
Journal homepage: www.koreascience.org

과학 교과의 학생 평가에서 ChatGPT의 활용 가능성 및 교사 인식 탐색

이동원, 심현표, 백종호*
한국교육과정평가원

Exploration on the Feasibility of Utilization and Teacher Perceptions of Using ChatGPT for Student Assessment in Science

Dongwon Lee, Hyeon-Pyo Shim, Jongho Baek*
Korea Institute for Curriculum and Evaluation

ARTICLE INFO

Article history:
Received 22 January 2024
Received in revised form
30 January 2024
Accepted 8 February 2024

Keywords:
student assessment, ChatGPT, teacher perception

ABSTRACT

This study explores the possibility of using a generative artificial intelligence, ChatGPT, for student assessment in science subjects. In order to achieve our goal, we developed assessment items, collected students' responses, and input them into ChatGPT to implement the assessment procedures. Subsequently, we shared the assessment results from ChatGPT with science teachers and compared them to the teachers' assessment process to investigate the use of ChatGPT in student assessment. Regarding the results, in terms of setting the scoring rubric, we found the rubric generated by ChatGPT to be generally appropriate. However, the consistency between the scoring results obtained from ChatGPT and those determined by the teachers was relatively low. This inconsistency was more pronounced in items with additional assessment components and a more intricate rubric. In regard to feedback on student responses, there were some instances where the feedback generated was scientifically incorrect or beyond the scope of the curriculum, but there were also some positives, such as the provision of exemplary answers to questions and additional examples that helped students learn further. From these results, the teachers perceived limitations in using ChatGPT to conduct assessment in terms of reliability, which is considered crucial in student assessment, but suggested that it could be used to support assessment. Finally, synthesizing these findings, implications for utilizing ChatGPT in student assessment were suggested.

Ⅰ. 서론

최근 인공지능(Artificial Intelligence: AI)은 우리의 일상에서 익숙하게 사용하고 있는 용어이자, 접할 수 있는 대상이 되었다. 2016년 3월에 있었던 알파고와 사람 사이의 바둑 대결은 인공지능에 대한 사람의 관심을 보다 촉발시키는 계기가 되었으며, 최근에는 국내외의 다양한 곳에서 다양한 인공지능 기술을 공개하고 사용할 수 있도록 제공하고 있다. 인공지능 용어는 1950년대에 처음 등장하여 사용되어 온 가운데, MIT, IBM, Google 등이 마련한 인공지능들이 공개되어 왔지만(Kim, 2021), 최근의 관심은 그간의 관심을 넘어 대중이 쉽게 사용할 수 있는 형태로 공개된 것에 기반하고 있음을 주목할 만한다.

인공지능은 사람이 사고하는 과정에서 사용하는 능력을 다양한 방법으로 구현하는 기술을 말하는데, 추론의 능력, 기능, 인식의 방식의 측면에서 다양하게 분류할 수 있다(Kim, 2021; Jho, 2023). 최근의 인공지능은 데이터 기반의 머신러닝의 방식을 택하고 있는데 학습을 위한 알고리즘에 따라서 여러 유형으로 구분하는 것이 가능하다. 다만, 중요한 점은 데이터의 양이 급속도로 누적, 증가되고, 학습의 방식이 고도화됨에 따라 인공지능의 발전도 비약적으로 이루어지고 있다는 점이다. 이와 같이 급속하게 인공지능이 발전하고 있는 가운데, 한계나 부작용에 대한 우려도 같이 제기되고 있으며, 인공지능의 도입과 활용에 대한 사회적 논의가 필요한 것이 사실이다. 이러한 맥락에서 교육 분야에서도 최근 인공지능에 대한 관심이 촉발되고 있는데, 이 기술이 교육 분야에서 어떻게, 왜 활용되어야 하는지에 대한 연구와 논의하는 것이 필요하다. 이와 동시에 여러 연구에서 보고하는 바와 같은 인공지능 활용으로부터 발생할 수 있는 교육 맥락에서의 문제들(Kang, 2023; Jang & So, 2023)의 해결 방안을 논의하고, 도출할 수 있는 활용 범위나 유의사항은 무엇인지 등에 대한 연구가 이루어질 필요도 있다.

인공지능과 관련하여 교육 분야에서의 활용은 AIED(Artificial Intelligence in Education)라고 불리는 데, 최근에는 두 범주로의 활용을 꾀하고 있다(Holmes, Bialik, & Fadel, 2019). 우선, 학습 대상으로의 관점을 반영하여 인공지능 자체에 대한 이해를 촉진하는 것이 첫 번째 관심 범주이며, 두 번째는 학습 지원 방법의 관점에서 인공지능을 활용하여 효과적인 교육을 수행하기 위한 방안을 탐색하고 개발하는 것이다(Hong *et al.*, 2020). 전자는 "Learning about AI"라고도 불리며(Holmes, Bialik, & Fadel, 2019), 이와 관련해서는 AI와 관련된 리터러시 함양을 위한 교육들이 강조되고 있다(Jang & So, 2003). 후자의 관점은 "Learning with AI"라고도 표현되는데(Holmes, Bialik,

* 교신저자 : 백종호 (hll14@kice.re.kr)
본 연구는 한국교육과정평가원의 '과학 교과에서 인공지능 챗봇의 학생 평가 결과 분석 및 활용 방안 탐색(ORM 2023-20-7)'의 내용을 기반으로 재구성 한 것임.
http://dx.doi.org/10.14697/jkase.2024.44.1.119

& Fadel, 2019), 다시 크게 둘로 구분하여 연구가 진행되는 것이 최근의 흐름이다. AI가 교육과 관련된 활동을 돕는 것은 교수・학습의 지원, 평가의 지원 측면으로 나누어 볼 수 있다. 교수・학습 지원 측면에서는 학생들의 학습을 관리하고 개별화된 학습을 촉진하는 도구로서의 활용을 의미하며, 최근의 학습 분석학의 발전과도 관련이 있다. 평가의 지원과 관련하여서는 평가 이후의 피드백, 성적 산정 및 채점과 관련된 자동화 등에 대한 연구들이 초점의 대상이다(González-Calatayud, Prendes-Espinosa, & Roig-Vila, 2021; Zhai *et al.*, 2021). 이 연구에서는 교육 활동의 지원의 측면 중, 특히 평가의 지원에 대한 인공지능 기술의 활용과 관련된 사항들을 탐색하는 것에 목적이 있다. 특히, 최근 높은 관심을 받는 ChatGPT를 중심으로 살펴보았는데, 오픈된 서비스가 높은 수준과 많은 데이터를 바탕으로 훈련한 것이면서도 챗봇의 형태로 개발되어 일반적으로 사용하기 용이하다는 점에 주목한 것이다.

ChatGPT는 OpenAI사에서 PFM(Pre-trained Foundation Model)을 이용하여 방대한 양의 데이터를 학습시킨 자연어 처리 인공지능 서비스로(OpenAI, 2023), 챗봇의 형태로 서비스가 이루어지고 있다. 챗봇은 특정 주제에 대해 인간 사용자와 대화 형식으로 상호작용 할 수 있는 소프트웨어를 의미하는데(Smutny & Schreiberova, 2020), 챗봇 형태로의 정보 입출력은 인공지능의 사용 편의성을 높인다고 볼 수 있다. 한편, ChatGPT가 동시대에 발표된 여러 인공지능 서비스 중에서 이목을 끈 이유는 트랜스포머 기반 모델들의 자연어 처리 성능에 더해 강화학습 방식을 적용하는 등의 이유가 있으며(Jho, 2023), 최근 가장 일반적으로 접하고 사용하는 서비스 중 하나임이 명확하다. 이러한 특성에 기반하여 기존의 챗봇이 가지고 있는 기술적인 한계를 크게 뛰어넘은 것으로 인식되어 다양한 산업 영역에서 활용 가능성에 대한 관심이 증대되고 있다(Zhou, 2023). 최근에는 GPT-3 모델에 이어 매개변수의 양을 대폭 증가시킨 GPT-4 모델이 공개되어 보다 복잡한 패턴에 대한 학습 능력이 증가할 것으로 기대됨과 동시에 한국어 기반의 처리와 응답도 보다 향상될 것으로 기대를 받고 있다. 아울러 여러 국내 업체에서도 한국어 기반의 인공지능 서비스를 시작하고 있어 GPT 기반의 언어 모델 활용은 더욱 활발해질 것으로 보인다.

이처럼 사회 전반에서 인공지능에 대한 관심이 높아짐에 따라 활용 방안에 대한 논의도 활발해지고 있는데, 교육 맥락에서의 활용에 대한 연구들도 최근 증가하는 추세에 있다. 특히 챗봇은 인공지능의 활용 관점에서 가장 관심을 많이 받는 기술 유형 중 하나로(Huh *et al.*, 2021), 국내에서는 주로 국어와 영어와 같은 언어의 교육과 관련한 연구들이 시작을 알리고 있었다(Chang, Park, & Park, 2021). 챗봇 기반의 교수・학습 활동의 효과를 보고하는 연구들은 인지적, 정의적 측면에서 학생들의 학습 지원에 긍정적 효과가 있음을 말하고 있다(Shin, 2019; Choi, & Nam, 2019; Hong *et al.*, 2020). 본 연구에서 사용하고자 한 ChatGPT를 활용한 연구들의 수도 짧은 기간이지만 증가하는 추세에 있다. ChatGPT가 공개된 이후, 2023년 5월까지 발표된 ChatGPT 활용 국내외 연구 문헌들을 조사한 Jang & So(2023)의 연구에 따르면, 개념연구와 실증연구가 비슷한 수로 이루어지고 있으며, 과학교육 분야의 연구들도 낮지 않은 비중을 차지하고 있는 것으로 나타났다.

국내에서 이루어진 연구들을 예를 들어 살피면, Byeon & Kwon(2023)은 과학적 개념 이해, 과학 탐구의 보조적 도구로서 ChatGPT 활용 방안을 협응의 관점에서 탐색하였고, Kim & Yu(2023)의 연구에서는 물리 수업의 진행 단계별 활용 방안을 조사하였다. 한편, Jho(2023)은 텍스트 기반의 생성성형 인공지능의 활용 방안과 한계 등을 논의하였다. 과학교육 분야 이외로 범위를 넓혀 살펴보면, Son(2023)은 수학교육 분야에서 ChatGPT를 이용해 학생들의 산출물과 예비교사-학생 사이의 담화를 분석하는 연구를 수행하였고, Shin, Jung, & Lee(2023)의 연구는 내용 중심 영어 교수・학습 도구로서의 활용 방안을 탐색하였다. 즉, 새로운 기술의 도입 방안에 대해 다양한 관점에서 접근을 시도하는 단계로 볼 수 있다. 또한 앞선 연구들은 텍스트 기반의 생성형 인공지능 기술을 교육 분야에 활용함에 있어 긍정적인 평가를 내리고 있으나, 후속 연구가 지속될 필요가 있음을 강조하고 있음이 특징이라고 할 수 있다. 이러한 기술이 교육 분야에 특화시켜 개발한 것이 아니며, 언어 처리에 초점을 둔 기술임을 감안하면, 교수・학습의 다양한 장면에서 산출한 데이터들을 축적하는 것뿐만 아니라, 현재의 상황에서 고려해야 할 사안들을 탐색해 보는 연구가 지속적으로 이루어질 필요가 있다. 즉, ChatGPT 활용의 가능성과 한계 등을 논의하기 위한 시작 단계임을 드러낸다고 볼 수 있다.

한편, 인공지능 기술의 활용을 평가의 영역으로 보다 국한시켜 살펴보면, 상대적으로 적은 수의 연구가 이루어졌다고 볼 수 있다. 2010년부터 2020년까지 수행된 인공지능을 활용한 연구들을 조사한 González-Calatayud, Prendes-Espinosa, & Roig-Vila(2021)에 따르면, 총 449개의 연구 중에 22개의 연구가 평가와 관련된 범주에 포함되며, 그 연구들의 주요 내용은 인공지능의 특성을 이해하거나, 자동채점과 관련된 것이었다. 여기에 속하는 연구의 다수가 형성평가의 관점에서 인공지능 도입을 꾀한 것은 주목할 만한 점이지만, 대다수의 경우 대학생을 대상으로 하는 연구였음을 고려한다면, 초・중등교육에서 어떻게 인공지능을 활용할 것인지 탐색할 필요성도 대두된다. 또한 우리나라의 교사들이 인공지능 기술을 활용할 때 가장 기대하는 점이 평가에서의 활용(Kim *et al.*, 2020)임을 고려하면 이에 대한 실천적인 전략을 마련하는 것도 필요하다. 평가 상황에의 활용평가의 관점에서 인공지능의 활용이라는 선상에서 Chang, Park, & Park(2021)은 탐구와 같은 교수・학습 활동의 개별화 학습을 위해 챗봇을 활용할 수 있다는 시사점을 제시하였는데, 이러한 시사점은 학생들의 반응을 즉각적으로 확인하여 그에 따른 적절한 처치가 필요함을 의미한다. 즉, 평가와 연동시켜 자동화된 채점 실시가 선결되었을 때 챗봇을 중심으로 한 개별화된 교수・학습은 그 효과를 높일 수 있으며, ChatGPT는 이러한 측면에서 관심을 받고 있다(Jho, 2023). 최근 해외 국가들에서 인공지능 기반 자동 채점 시스템과 연계한 적응형 학습 시스템을 준비하고 있는 흐름은(Hong *et al.*, 2020), 교수・학습을 위한 평가의 관점으로 인공지능 기술 도입이 초점을 맞추고 있음을 드러낸다. 아울러 ChatGPT가 텍스트 기반의 생성형 인공지능이라는 점은 이 기술이 평가 결과 확인과 피드백 등에 있어 즉각적인 산출물을 만들어 낼 수 있다는 기대를 낳는다.

ChatGPT를 활용한 평가 결과 확인, 채점과 관련된 연구들은 주로 인간 채점자의 채점 결과 비교에 초점을 두고 있다. 서답형 문항들의 자동채점이 주요 관심사인 가운데, 자동채점의 도입 가능성을 탐색하기 위해 사람의 채점 결과와 인공지능의 채점 결과 사이의 일치도를 비교하는 방식으로 연구들이 추진되고 있다. 우선 상대적으로 간단한 형태의 응답을 요구하는 단답형 문항을 활용한 연구들은 인공지능이

인간평가자에 비해 다소 관대한 경향을 보이는 것으로 나타났으며, 문항에서 묻는 내용이나 수준에 의해 일치도가 영향을 받는다고 보고하였다(Bhat *et al.*, 2022; Moore *et al.*, 2022). Park *et al.*(2023)의 연구는 GPT-4를 활용하여 초등학생들의 과학 탐구 보고서를 평가하고 가능성과 한계를 논하였는데, 서술형 답안을 평가하였다는 의미와 함께 단순히 지식 관점이 아닌 역량 관점에서 자동채점을 시도한 것이 특징적이다.

ChatGPT가 교육 분야를 위하여 특화시킨 인공지능이 아니기 때문에 교육분야로의 활용 가능성은 학습한 데이터에 전적으로 의존한다. 이에 인간 채점자와의 결과 비교가 관심을 받고 있지만, 이와 함께 인공지능이 생성한 결과물을 교육 분야에서 활용할 수 있는지의 여부는 이 결과물을 사용할 사람들인 교사의 인식 조사를 통해서도 탐색할 필요성이 있다. 또한 인공지능이 생성한 자료들의 특징을 유형, 특성에 따라 상세하게 비교하여 교사들의 인식이 무엇에 기인하는지 알아보는 것도 필요하다. 본 연구는 이러한 선상에서 과학의 평가 문항들을 다양한 유형과 내용으로 구성하여 교사의 채점 결과와 ChatGPT를 활용한 채점 결과를 비교하였다. 이와 함께 평가 문항의 응답별 피드백과 채점에 활용하기 위한 문항별 평가 기준을 ChatGPT를 이용해 생성하여 이에 대한 교사들의 활용 가능성과 한계에 대한 인식도 살펴보았다. 즉, 평가 상황에서 교사들이 생성해야 하는 여러 자료들을 인공지능을 활용해 생성하고, 활용 가능성에 대한 교사들의 인식을 조사하였다. 구체적으로는 다음의 두 질문을 중심으로 연구를 추진하였다. 첫째, 다양한 유형과 내용으로 구성한 과학과의 평가 문항에 대하여 인간채점과 ChatGPT를 활용한 채점은 어느 정도의 일치 수준을 보이며, 채점 결과의 특징은 무엇인가? 둘째, 과학과 평가 상황에서 생성하는 평가 기준, 평가 결과(채점), 피드백 자료에 대한 교사들의 인식은 어떠하며, 이유는 무엇인가? 이와 같은 연구 내용을 바탕으로 과학과의 평가 상황에서 ChatGPT 등의 인공지능 챗봇 활용에 대한 시사점을 탐색하고자 하였다.

II. 연구 방법

본 연구에서는 ChatGPT를 이용하여 학생 평가 과정의 일부를 수행해보고, ChatGPT 활용에 대한 교사들의 인식을 살펴보았다. 본 연구는 크게 평가 설계 및 문항 개발 단계, 현장 적용 단계, 채점 및 결과 분석 단계로 구분할 수 있으며, 각 단계별 구체적인 내용은 다음과 같다. 평가 설계 및 문항 개발 과정에서는 전체 평가 과정에 대한 구성 및 단계별 산출물 및 분석 대상을 설정하고 문항과 관련하여 평가 문항의 유형, 평가 요소 등을 구성하였다. 현장 적용 단계에서는 개발된 문항에 대해 학생들이 응답하고 그 결과를 수집하였다. 산출물 제작 및 분석 단계에서는 수집된 학생들의 응답에 대해 ChatGPT와 교사가 각각 채점을 실시하였으며, 이 과정에서 ChatGPT를 활용할 때 채점 결과뿐만 아니라 평가의 과정과 관련된 인공지능의 판단 과정을 확인할 수 있도록 프롬프트를 설계하여 입력하고 하고 평가 기준도 산출하도록 하였다. 이와 함께 각 학생들의 응답에 대한 피드백도 ChatGPT가 생성하도록 프롬프트를 구성하여 평가의 환류에 활용할 수 있는지 여부도 탐색하고자 하였다. 마지막으로 ChatGPT로부터 수집한 평가 결과를 채점을 수행한 교사들과 공유하고, ChatGPT의 활용에 대한 교사들의 인식을 탐색하였다.

1. 평가 설계 및 문항 개발

평가 설계 및 문항 개발 단계에서는 과학 교과에서 ChatGPT를 학생 평가에 활용하기 위해 평가 과정에 대한 단계와 세부 내용을 구성하고 문항에 대한 내용을 구체화 하였다. 본 연구에서 평가의 과정은 '평가 기준 수립', '평가 시행(채점)', '피드백 마련'의 세 단계로 구분하였고(교육부, 한국교육과정평가원, 2017), 평가 기준(채점 기준), 채점 결과(학생 응답 분석), 응답에 따른 피드백 등 평가의 각 단계에서 핵심이 되는 자료들을 ChatGPT를 사용하여 산출하고자 하였다. 이를 통해 과학 교과에서 ChatGPT를 학생 평가에 활용하고자 할 때 낮은 수준에서는 학생 응답을 채점하여 점수를 제시하는 정도에서 높게는 문항에 대한 채점 기준을 스스로 생성하고 학생들의 응답에 대해 적절한 수준으로 피드백까지 가능한지에 대한 활용 가능성을 탐색하고자 하였다.

본 연구에서 구성한 평가 요소, 문항의 형식, 평가 시행과정은 다음과 같다(Figure 1). 평가 요소와 글·그림의 유무로 구분하는 문항들에 대해 ChatGPT가 채점 기준을 수립하고, 채점을 실시하며, 피드백을 마련하는 것의 적절성과 한계를 확인하는 것이 본 연구의 주요 내용이며, 이를 위해 문항을 개발하는 교사가 작성한 채점 기준을 활용하여 학생 응답을 분석한 결과와 ChatGPT가 문항을 토대로 직접 마련한 채점 기준을 활용하여 학생 응답을 분석한 결과를 종합적으로 살펴보고자 하였다.

본 연구에서는 학생 평가에서 ChatGPT의 활용 가능성을 살펴보기 위해 문항을 개발하였다. 평가 문항은 현재 고등학교에서 통합과학을 지도하고 있는 교사 4인과의 협의회를 통해 제작하였다. 두 차례의 전문가협의회를 거쳐 문항과 문항별 채점 기준을 함께 마련하였다.

문항 개발을 위한 전문가 협의회를 통해 개발하는 문항의 기본적인 형식은 학교 현장에서 일반적으로 사용하는 평가의 양식을 준용하는 것으로 설정하였다. 이는 학교 현장에서 이루어지는 평가 과정과 관련하여 교사가 가지는 현실적인 요구와 관련이 있다. 학생 평가에서

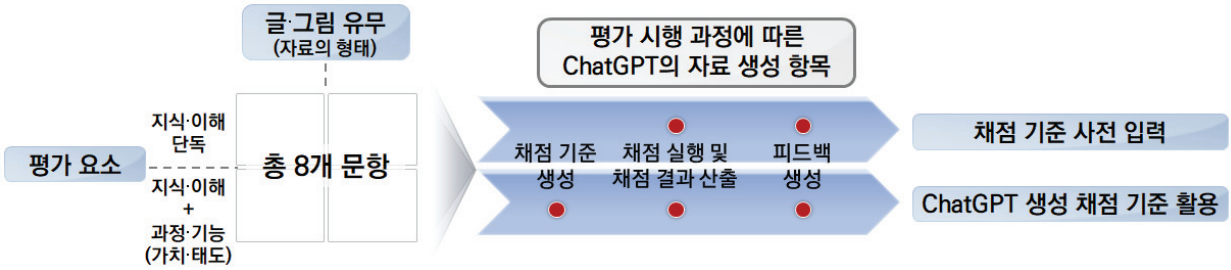

Figure 1. The use of ChatGPT and characteristic of the items in this study

Table 1. The number of question items developed in this study

문항 수(개)		평가 요소	
		지식・이해 중심	지식・이해 + 과정・기능
자료 형태	텍스트로만 구성	2	4
	텍스트, 그림으로 구성	2	-

Table 2. Information for each question item

문항 번호	영역	내용 요소	과정・기능 요소	그림 포함 여부	배점		
					지식・이해	과정・기능 (가치・태도)	계
1	시스템과 상호 작용	세포 내 정보의 흐름	-	O	2	-	2
2		효소의 작용	(가치・태도)과학의 유용성	X	2	1	3
3		물질대사	-	X	2	-	2
4		세포막을 통한 물질이동	탐구 설계와 수행	X	1	1	2
5	물질과 규칙성	화학 결합	-	X	2	-	2
6		알칼리 금속	결론 도출 및 평가	X	1	2	3
7		이온 결합	-	O	3	-	3
8		이온 결합, 공유 결합	자료의 수집・분석 및 해석	X	2	3	5

간단한 형태의 선택형 문항이나 서술형 중 단답형 문항은 ChatGPT를 활용하지 않고 다른 방법으로도 채점이 가능하거나 교사가 직접 채점하는 데 있어서 상대적으로 부담이 적다. 반면 서술형 문항의 경우 학생들의 응답을 평가하기 위해 장시간의 탐색을 요구하고 평가에 있어서 일관성을 유지하는 것이 어려운 면이 있기 때문에 ChatGPT의 활용 가능성을 살펴보기 위한 문항 개발에 있어서 서술형 형태의 문항 형식도 함께 활용하고자 하였으며 현장에서 이루어지는 평가의 맥락을 반영하여 단위 학교의 수행 평가에서 사용하는 것과 유사한 형태로 서술형 문항을 개발하였다.

서술형 문항에서 텍스트 형태로 작성된 학생들의 응답을 분석한 결과가 의미를 가지기 위해서는 기본적으로 문항별로 분석이 가능한 수준의 내용이 담겨있는 답안을 수집할 필요가 있다. 이에 따라 서술형 형태의 문항을 접한 경험이나 진술 능력을 고려할 때 고등학교 1학년을 대상으로 설정하여 문항을 개발하였고, 참여 교사와의 논의를 통해 연구가 진행된 시점 등을 고려하여 통합과학의 1학기에 진행되는 '시스템과 상호작용' 및 '물질과 규칙성' 단원에서 문항을 개발하였다.

최종적으로 개발한 문항은 크게 평가 요소와 문항의 자료 형태로 구분할 수 있다(Table 1). 먼저 평가 요소 측면에서 2022 개정 과학과 교육과정(교육부, 2022)의 내용 체계표의 틀을 준용하여 평가 요소를 도출하고, 각 요소의 내용은 2015 개정 과학과 교육과정(교육부, 2015)의 내용 체계표의 항목들을 활용하였다. 문항의 평가 요소의 측면에서 지식・이해 및 과정・기능[1]으로 구분한 평가틀을 활용하였고, 지식・이해만 단독으로 측정하는 문항과 지식・이해와 과정・기능을 같이 측정하는 문항을 고루 활용하였다. 이를 통해 ChatGPT를 단편적인 지식에 대한 평가가 아닌 다른 영역에 대한 평가에도 활용할 수 있는지 살펴보고자 하였다. 문항의 자료 형태 측면에서 문항의

다음은 화학 결합의 종류에 따른 물질의 성질을 비교하기 위한 실험이다.

[실험 과정]
(가) 홈판에 염화나트륨(NaCl) 2g과 설탕 2g을 각각 넣고, 간이 전기 전도성 측정기로 전류가 흐르는지 관찰한다.
(나) 2개의 비커에 염화나트륨(NaCl) 2g과 설탕 2g을 각각 넣고 증류수 50ml를 각각 넣어 녹인다.
(다) 과정 (나)에서 만든 수용액을 12홈판에 각각 넣고, 간이 전기 전도성 측정기로 전류가 흐르는지 관찰한다.

[실험 결과]

고체	전기 전도성	수용액	전기 전도성
염화나트륨	X	염화나트륨	○
설탕	X	설탕	X

위의 실험 결과를 바탕으로 **화학 결합의 종류**와 물질의 상태에 따라 **전기 전도성이 다른 이유**를 설명하시오.

조건 1. 위 물질들의 화학 결합 종류를 구분할 것
조건 2. 위 물질들의 상태를 구분하여 설명할 것

Figure 2. Example of an assessment item (8) that measures "knowledge and understanding" and "process and skills" elements

모든 요소가 텍스트 형태로만 제시된 경우와 그림이 포함된 경우로 나눌 수 있으며 실제 과학 교과에서 평가에 활용되는 문항들은 그림 등 다양한 형태의 자료들을 포함하기 때문에 텍스트 형태의 문항뿐만 아니라 그림이 포함된 문항의 평가에 대해서도 ChatGPT를 활용할 수 있는지 파악하고자 하였다. 단, 그림을 활용한 문항의 경우 ChatGPT에 그림을 입력하는 것에 제한이 있어 지식・이해만 측정하는 문항으로 개발하였고 ChatGPT에 그림이 포함된 문항에 대한 정보를 입력할 때 그림 대신 그림에 해당하는 설명을 텍스트 형태로 대체하여 입력하였다. Table 2는 개발한 문항의 세부 정보를 제시하였고, Figure 2는 물질과 규칙성 영역에서 개발된 문항의 예시이다.

채점 기준의 경우 문항 개발 이후 각 문항별로 개발되었으며 평가 요소 별로 수준을 구분하여 채점 기준과 배점을 제시하였다. 문항에 따라 평가 요소에서 지식・이해와 과정・기능을 동시에 측정하는 경우에는 교사가 제작한 채점 기준 역시 각 요소를 구분하여 채점 가능하도록 구성하였다. 예를 들어 Figure 3은 Figure 2에서 제시한 문항에 대해 교사가 제작한 채점 기준을 나타낸 것이다. 해당 문항은 화학

1) 본 연구에서는 주로 과정・기능을 하나의 평가 요소로 보았으나, 평가 범위의 측면에서 ChatGPT의 활용 가능성을 넓게 살피기 위해 1개 문항의 평가 요소로 가치・태도를 포함시켰다.

성취 기준	평가 요소	채점 기준	배점
[10통과01-05]	지식·이해	다음 두 가지 내용을 모두 올바르게 작성한 경우 • 염화나트륨 : 이온 결합 • 설탕 : 공유 결합	2
		위의 내용 중 한 가지 내용에 대해서만 올바르게 작성한 경우	1
		무응답, 모두 오답인 경우	0
	과정·기능	다음 세 가지 내용을 모두 올바르게 작성한 경우 • 물질의 종류에 따른 결과를 활용한 기술 • 물질의 상태에 따른 결과를 활용한 기술 1. 이온 결합 물질인 염화나트륨 고체 상태에서는 양이온과 음이온이 서로 강하게 결합하고 있어 이온이 이동할 수 없으므로 전기 전도성이 없다. 2. 이온 결합 물질인 염화나트륨 수용액 상태에서는 이온이 자유롭게 이동할 수 있기 때문에 전지 전도성이 있다. 3. 공유 결합 물질인 설탕 수용액은 설탕 분자가 이동할 뿐 전하를 띤 입자가 존재하지 않으므로 전기 전도성이 없다.	3
		위의 내용 중 두 가지 내용에 대해서 올바르게 작성한 경우	2
		위의 내용 중 한 가지 내용에 대해서만 올바르게 작성한 경우	1
		무응답, 모두 오답인 경우	0

Figure 3. Example of scoring criteria for assessment item (8) that measures “knowledge and understanding” and “process and skills” elements

결합의 종류와 물질의 상태에 따라 전기 전도성이 다른 이유에 대해 서술형으로 진술할 것을 요청하며, 평가 기준에서는 학생들이 작성한 글에 대해 지식・이해 측면에서 각 화합물이 어떤 화학결합을 하는지와 과정・기능 측면에서 화합물의 전기 전도성이 다른 이유를 타당하게 기술하였는지를 각각 모두 평가한다.

2. 현장 적용

본 연구에는 고등학교 과학교사 4명이 참여하였으며 이들은 모두 연구에 참여하는 시기에 고등학교에서 통합과학 수업을 담당하고 있었다. 이들은 문항 개발, 채점 기준 개발, 학생 응답 채점을 수행하였다. 또한 개발된 문항에 대해 연구 참여 교사가 소속되어 있는 세종특별자치시 소재 고등학교 1학년 중 2개 학급에 속하는 45명, 충청남도 소재 고등학교 1학년 중 1학급에 해당하는 26명이 학생들이 응답을 제출하여 총 71명의 학생들이 연구에 참여하였다.

연구에서 개발한 문항들은 웹에 탑재하여 평가 과정에서 학생들이 컴퓨터 등을 통해 웹에 접속하여 문항을 답변을 작성하여 제출하였다. 평가가 종료된 이후 학생들의 응답 결과에 대해 교사가 개발된 채점 기준을 바탕으로 문항별로 채점을 수행하고 이후 ChatGPT가 수행한 평가 결과와 비교하였다.

3. ChatGPT 산출물 생성 요청 및 분석

본 연구에서는 학생 응답별로 일관성 있는 형식의 결과를 얻기 위해 ChatGPT에 요청하는 프롬프트를 Table 3과 같이 구조화 하였다. ChatGPT를 활용한 산출물과 관련하여 채점 결과, 채점 결과에 대한 근거, 학생 응답에 대한 피드백을 학생별, 문항별로 개별화된 응답을 분석이 가능한 의미 있는 수준의 분량으로 얻고자 하였다. 문항 당 학생 응답 수는 71개이며, 모든 학생 응답을 한 번에 입력하는 경우 프롬프트에서 기술 분량에 대한 요청 내용이 포함되어 있더라도 ChatGPT가 개별 응답에 대한 응답이 아닌 요약된 형태로 답변을 제시하는 경우가 있어 요청하는 내용에 대한 충분한 답변을 얻기 위해 학생 응답별로 동일한 프롬프트에 학생 응답만 달리하여 반복하여 입력하는 형태로 자료를 수합하였다. 또한 이 과정에서 입력한 프롬프트에 따라 ChatGPT가 생성하는 응답을 확인하고 프롬프트에서 요청하는 내용과 비교하여 필요한 경우 프롬프트를 수정한 뒤 전체 학생 응답을 개별로 입력하는 과정을 반복하였다. 연구 진행 시점에서 ChatGPT 4의 경우 시간 당 응답 수 제약으로 분석이 필요한 다수의 응답을 개별적으로 반복하여 분석하는 데 한계가 있어 ChatGPT 3.5를 활용하였다.

프롬프트는 교사가 구성한 채점 기준을 활용한 경우와 ChatGPT가 생성한 채점 기준을 활용한 프롬프트로 나누어 구성하였다. 두 프롬프트는 모두 크게 문항, 채점 기준, 응답 제시 양식, 개별 학생 응답의 요소로 구성하였으며, 교사가 구성한 채점 기준을 활용한 경우와 대비하여 ChatGPT가 생성한 채점 기준을 활용한 프롬프트에서는 채점 기준 요소에서 ChatGPT가 생성한 채점 기준에 대한 기술을 요청하는 내용이 추가로 포함되어 있다.

Table 3. ChatGPT prompts used in scoring requests

요소	교사 생성 채점 기준 활용 프롬프트	ChatGPT 생성 채점 기준 활용 프롬프트
문항	・개발한 문항을 입력함 (그림이 포함된 경우 그림을 대체할 수 있는 정보 제공) [프롬프트 양식] Ⅰ. 다음은 문항이다. <문항> ...	
채점 기준	・교사가 작성한 문항별 점수별 평가 기준을 프롬프트에 입력함 [프롬프트 양식] 위의 문항을 다음의 평가 기준에 따라 2점, 1점, 0점으로 평가하고자 한다. 1) 2점인 경우 ...	・프롬프트에서 문항에 대한 평가 기준을 ChatGPT가 만들고 생성한 기준을 제시하도록 요청함 [프롬프트 양식] <평가 작성 양식> 1) 평가기준 : <문항>의 평가 기준을 2점, 1점, 0점 별로 각각 제시하시오. ...
응답 제시 양식	[프롬프트 양식] 위의 문항에 대한 <학생답안>을 아래 양식에 따라 평가하시오. <평가 작성 양식> a. 평가점수 : <학생답안>이 “2점, 1점, 0점”중 어디에 해당하는지 쓰시오. b. 평가 이유: ... 평가한 이유를 ... 기술하시오. c. 피드백: ...어떠한 점을 보완해야 하는지 ... 기술하시오. ...	

마지막으로 ChatGPT가 생성한 채점 기준, 학생 응답 분석 결과(채점 결과), 피드백에 대하여 현장 교사를 대상으로 질문지를 활용하여 각 산출물에 대한 적절성과 학교 현장에서의 활용 가능성 등에 대한 인식을 조사하였다.

Ⅲ. 연구 결과

본 연구는 학생 평가 상황에서 ChatGPT 등 인공지능 챗봇의 현장 활용 가능성을 살펴보기 위해 평가 문항을 제작하고 이에 대한 학생들의 응답을 교사와 ChatGPT가 평가하였다. 이 과정에서 ChatGPT가 수립한 평가 문항에 대한 채점 기준, 학생 답안에 대한 평가 결과, 피드백에 대한 특징과 이에 대한 교사의 인식을 살펴보았다.

1. 채점 기준 수립 단계에서의 ChatGPT의 활용

가. ChatGPT가 생성한 채점 기준의 주요 특징

ChatGPT가 생성한 채점 기준은 교사가 생성한 채점 기준과 비교하였을 때 수정・보완이 필요한 부분이 있기는 하지만 문항의 답안으로 요구하고 있는 평가 요소들을 어느 정도 갖추고 있는 것으로 보인다(Table 4). 형식적 측면에서 문항에서 답안으로 요구하는 사항이 분명하거나, 문항의 정답이 명확한 경우 배점에 따른 채점 기준을 생성하였으며, 부분 배점을 주어야 하는 경우 배점별 채점 기준을 수준에 따라 상이하게 제시하지만 '모두 정확하게 설명', '부분적으로 정확하게 설명', '완전히 잘못 설명' 등으로 구분하여 제시하는 경우가 많고 보다 구체적인 내용이 요구되는 경우가 있다. 예를 들어, Table 4의 5번 문항에 대해 GPT가 생성한 채점 기준을 살펴보면, 내용적 측면에서 교사가 생성한 채점 기준과 유사하게 나타나는 것을 살펴볼 수 있다. 또한 형식적 측면에서는 그 의미는 교사가 생성한 채점 기준과 크게 다르지 않지만, 평가 요소를 나눠서 분석적으로 평가 기준을 제시하기 보다는 '올바르게 설명', '자세한 설명이 필요', '설명을 제대로 이해하지 못함' 등으로 구분하여 나타내었다. 내용적 측면에서는 교사가 생성한 채점 기준과 비교하였을 때 과학적으로 옳은 내용을 전반적으로 진술하고 있었으나 자료에 내포되어 있는 과학적 개념을 분석하여 답안 작성을 요구하거나, 하나의 발문에 2가지의 평가 요소가 혼합되어 있는 경우 ChatGPT가 제시하는 배점에 따른 채점 기준의 타당성이 낮아지는 경향이 나타났다.

나. ChatGPT가 생성한 채점 기준에 대한 교사들의 인식 및 활용 방안

교사들은 채점 기준이 분명하게 제시되어 있는 문항의 경우 약간의 수정・보완을 통하여 실제 채점 기준을 수립하는 데 충분히 활용할 수 있다는 의견을 제시하였다. 또한, ChatGPT가 생성한 채점 기준을 검토하면서 개발한 문항에서 묻고자 하는 요소가 학생들에게 명확하게 전달될 수 있는지를 점검해 볼 수 있으며, 이를 통해 문항의 완성도를 높이는 작업에 도움이 될 것이라는 의견을 제시하였다.

"ChatGPT가 생성한 채점 기준들을 참고하여 교사가 채점 기준의 기본틀을 잡고 구체화할 수 있을 것이라고 생각했습니다. 또한 교사가 ChatGPT가 생성한 채점 기준들을 보면서 문항에서 묻고 있는 것이 모호하지는 않은지 등을 판단해 볼 수도 있을 것이라고 생각했습니다. ChatGPT가 생성한 채점 기준을 피드백 삼아 문항을 완성도 높게 수정할 수 있을 것이라고 생각했습니다."

"실제 현장에서 ChatGPT가 학생 응답을 위해 생성한 채점 기준을 바로 사용하기엔 무리가 있고 교사의 재검토가 필요함. 채점 기준을 수립하기 초기 단계에 (채점 기준 구성 수립단계) 활용이 적절해 보임."

2. 평가 문항의 채점 단계에서 ChatGPT의 활용

가. ChatGPT가 생성한 문항 채점 결과의 특징

학생 답안을 ChatGPT를 활용하여 채점한 결과와 교사가 채점한

Table 4. Teacher-generated and ChatGPT-generated scoring criteria

문항 번호	교사가 생성한 채점 기준	ChatGPT가 생성한 채점 기준
1	2점: 과정 (가)는 번역 과정이며, RNA의 염기서열에 있는 코돈(유전암호)의 정보를 바탕으로 여러 종류의 아미노산들이 펩타이드 결합을 통해 단백질 분자로 합성됨을 옳게 설명함. 1점: 다음의 2가지 중 1가지만 옳게 설명함. - 과정 (가)의 명칭은 번역임. - RNA의 염기서열에 있는 코돈 정보를 바탕으로 여러 종류의 아미노산들이 펩타이드 결합을 이뤄 단백질 분자를 합성함. 0점: 설명을 제시하지 못함.	2점: 학생이 과정 (가)를 정확하게 '번역'이라고 기술하였고, <보기>에 제시된 용어를 모두 사용하여 설명한 경우. 1점: 학생이 과정 (가)를 '번역'이라고 기술하였으나, <보기>에 제시된 용어를 부분적으로 사용하거나 정확하지 않게 설명한 경우. 0점: 학생이 과정 (가)를 '번역'이라고 기술하지 않거나 <보기>에 제시된 용어를 사용하지 않거나 완전히 잘못된 설명을 제시한 경우.
5	2점: 다음 두 가지 내용을 모두 올바르게 작성한 경우 - 18족 이외의 원소들은 비활성 기체와 같이 안정한 전자 배치를 가지려는 경향이 있으며 이러한 경향을 가리켜 옥텟 규칙이라고 한다는 내용이 서술된 경우 - 18족 이외의 원소들은 다른 원자와의 결합(화학 결합)을 통해 비활성 기체와 같이 안정한 전자 배치를 가지려는 경향이 있다는 내용이 서술된 경우 1점: 위의 내용 중 한 가지 내용에 대해서만 올바르게 작성한 경우 0점: 무응답, 모두 오답인 경우	2점: 학생답안이 문항에 주어진 보기의 용어들(원자가 전자, 전자 배치, 옥텟 규칙)을 모두 포함하고 있으며, 18족 원소들이 안정한 상태를 유지하기 위해 화학 결합을 통해 전자 배치를 조정하는 이유와 옥텟 규칙에 따라 원자가 전자의 수를 8개로 맞추는 것이 옳은 이유를 올바르게 설명하고 있는 경우 1점: 학생답안이 보기의 용어들을 어느 정도 포함하고 있으며, 18족 원소들이 안정한 상태를 유지하기 위해 화학 결합을 통해 전자 배치를 조정하는 이유를 언급하고 있지만, 옥텟 규칙과 전자가 전자의 수를 8개로 맞추는 이유에 대해 더 자세한 설명이 필요한 경우 0점: 학생답안이 보기의 용어들을 전혀 포함하지 않거나, 옥텟 규칙과 18족 원소들의 안정한 상태에 대한 설명을 제대로 이해하지 못한 경우

결과를 비교하기 위해 ChatGPT 사이의 문항별 채점 결과의 일치도를 비교하였다. 세부적으로는 교사가 생성한 채점 기준을 활용한 경우에 대해 두 명의 교사와 ChatGPT 사이의 채점 결과를 각각 비교하였고 ChatGPT가 생성한 채점 기준을 활용한 경우에 대해서 동일한 방식으로 비교하였다. 마지막으로 교사가 생성한 채점 기준과 ChatGPT가 생성한 채점 기준을 활용하였을 때 ChatGPT의 채점 결과를 비교하였다. 일치도 비교 결과 Table 5와 같이 문항에 따라 편차가 크게 나타나며 ChatGPT와 교사와의 일치도 비교에서는 70%가 넘는 경우가 없었다. 특히, 부분 배점이 있는 문항의 경우 부분 배점이 없는 문항에 비해 일치도가 낮게 나타나는 것을 살펴볼 수 있었다.

한편 본 연구에서는 ChatGPT를 통해 학생 답안을 채점한 점수와 해당 점수를 부여하게 된 근거를 함께 출력하도록 프롬프트를 구성하였다. 이를 통해 챗봇이 생성하는 평가 근거를 채점 결과와 함께 고려하여 ChatGPT가 학생들의 답안을 평가하는 데 있어서 나타나는 특징들을 확인해 보았다.

ChatGPT의 경우 학생들의 응답이 '그럴듯하게' 작성되어 있는 경우, 실제로는 적절한 답안이 아님에도 불구하고 정답으로 채점하는 경우가 있었다. Table 6에 제시한 1번 문항의 사례와 같이 "펩타이드 결합으로 이루어진 RNA 3염기조합"이라는 문장이 잘못된 설명임에도 불구하고, 챗봇은 이러한 부분을 정확하게 파악하지 못하고 있다. 실제로 챗봇의 평가 근거 중 "RNA 코돈이 펩타이드 결합을 통해 아미노산으로 합성되고"라는 부분을 고려하였을 때, 학생이 잘못 설명한 부분과 동일한 오류를 평가 근거에서도 나타내고 있음을 알 수 있다.

다음으로, 문항의 맥락을 바탕으로 답안을 평가하기 보다는 답안에 제시되어 있는 "문장의 진위"를 중심으로 평가하는 사례도 발견할 수 있었다. Table 7의 2-(2)번 문항의 사례와 같이 이미 문항의 자료에 제시되어 있는 사례를 학생이 답안에서 동일하게 작성하였기 때문에 교사들은 0점을 부여한 반면, 챗봇의 경우 문항의 맥락을 고려하지 않은 채로 해당 내용이 틀린 예시가 아니기 때문에 답안의 정합성만을 바탕으로 채점하여 1점을 부여하였다.

교사가 생성한 채점 기준을 제시하여 평가를 진행하는 경우 제시된 채점 기준에 명시적으로 드러나 있지는 않지만, 학생의 응답에 문항의 채점 기준과 관계없이 과학적으로 잘못된 서술이 있는 경우 ChatGPT가 이를 정확하게 짚어내지 못하는 사례도 있었다. Table 8에 제시한 5번 문항의 사례에서 학생이 원자가 전자에 대한 정보를 잘못 기술하였기 때문에 교사는 0점을 부여했으나, ChatGPT는 2점을 부여하였다.

한편, 교사와 동일한 점수를 부여했다고 하더라도 교사와는 서로 다른 기준을 적용한 사례도 찾아볼 수 있었다. Table 9의 예시를 살펴

Table 5. Agreement between ChatGPT and teacher scoring of student assessments

문항 번호	영역	과정·기능 (가치·태도) 요소	그림 포함 여부	교사가 생성한 채점 기준 활용 방식			ChatGPT가 생성한 채점 기준 활용 방식		ChatGPT의 분석 결과 간 비교 (AI1-AI2)
				T1*-T2	T1-AI1**	T2-AI1	T1-AI2***	T2-AI2	
1	시스템과 상호 작용	-	O	80.3%	62.0%	45.1%	59.2%	42.3%	80.3%
2		(가치·태도) 과학의 유용성	X	93.0%	67.6%	66.2%	66.2%	67.6%	62.0%
3		-	X	80.3%	57.7%	53.5%	63.4%	60.6%	67.6%
4		탐구 설계와 수행	X	94.4%	38.0%	36.6%	39.4%	38.0%	71.8%
5	물질과 규칙성	-	X	76.1%	26.8%	43.7%	23.9%	38.0%	74.6%
6		결론 도출 및 평가	X	73.2%	52.1%	57.7%	54.9%	52.1%	67.6%
7		-	O	56.3%	33.8%	40.8%	26.8%	45.1%	60.6%
8		자료의 수집·분석 및 해석	X	53.5%	21.1%	28.2%	5.6%	18.3%	26.8%

* T1, T2: 교사1, 교사2 채점 결과
** AI1: 문항과 사전에 교사가 마련한 채점 기준을 ChatGPT에 적용하여 학생 응답을 채점한 결과
*** AI2: 문항만 입력한 후, ChatGPT가 생성한 채점 기준 적용하여 학생 응답을 채점한 결과

Table 6. Evaluation results for Item 1

채점 결과(교사/AI)	교사: 1점/ ChatGPT: 2점
학생 답안	(가)의 명칭은 번역이고 펩타이드 결합으로 이루어진 RNA 3염기조합이 하나의 코돈을 뜻하고 그 3염기 조합에 맞는 아미노산이 특정 단백질을 지정함
ChatGPT가 생성한 평가 근거	학생은 번역(과정 가)에 대해 정확하게 설명하고 있으며, RNA의 코돈이 펩타이드 결합을 통해 아미노산으로 합성되고, 이러한 아미노산들이 단백질을 구성하는 과정을 설명하고 있습니다.

Table 7. Evaluation results for Item 2-2

채점 결과(교사/AI)	교사: 0점/ ChatGPT: 1점
학생 답안	플라스틱을 효소로 분해시킨다.
ChatGPT가 생성한 평가 근거	학생은 플라스틱을 효소로 분해시키는 사례를 제시하였습니다. 플라스틱은 고분자 물질로서 자연적으로 분해되는 시간이 매우 길고, 분해되지 않은 플라스틱은 환경 문제를 일으키는데, 이러한 문제를 해결하기 위해 효소를 활용하여 플라스틱을 분해시키는 기술이 개발되고 있습니다.

Table 8. Evaluation results for Item 5

채점 결과(교사/AI)	교사: 0점/ ChatGPT: 2점
학생 답안	18족 이외에 원소들은 원자가 전자가 18개가 안되고 옥텟 규칙 때문에 원자가 전자가 18개가 되려는 성질을 가지고 있다 이때 원자가 전자가 18개가 안되면 전자 배치가 불안정 하여서 원자가 전자 18개를 만들기 위해서 화학 결합을 한다.
ChatGPT가 생성한 평가 근거	학생은 "18족 이외에 원소들은 원자가 전자가 18개가 안되고 옥텟 규칙 때문에 원자가 전자가 18개가 되려는 성질을 가지고 있다 이때 원자가 전자가 18개가 안되면 전자 배치가 불안정 하여서 원자가 전자 18개를 만들기 위해서 화학 결합을 한다"라는 내용을 제시하였습니다. 학생은 옥텟 규칙이 안정한 전자 배치를 가지려는 경향을 설명하고, 이러한 경향으로 인해 화학 결합을 하는 이유에 대해 언급하였기 때문에 2점을 주도록 합니다.

Table 9. Evaluation results for Item 2-1

채점 결과(교사/AI)	교사: 1점/ ChatGPT: 1점
학생 답안	(가), (나)에서 알 수 있는 효소의 특징은 효소가 물질의 분해를 돕는다는 것이다. 플라스틱으로 인환 환경 문제를 효소로 해결할 수 있는 이유는 효소가 플라스틱의 분해를 도울 수 있기 때문이다.
ChatGPT가 생성한 평가 근거	…(생략)… 답변이 좀 더 구체적이고 상세하게 설명되어야 합니다. 어떤 종류의 효소가 플라스틱의 어떤 성분을 분해하는지, 효소가 어떤 반응을 촉진시키는지 등에 대해서 더 자세하게 기술해야 합니다. 또한, 효소가 플라스틱 분해에 어떤 방식으로 기여하는지를 더 명확하게 설명해야 합니다.

보면, 2-(1)번 문항의 학생 응답에 대해 교사와 ChatGPT가 모두 2점 중 1점을 부여하였다. 그러나 교사는 효소가 물질을 "빠르게" 분해시킨다는 특성이 누락되어 1점을 부여한 것이지만, 챗봇은 효소의 구체적인 명칭과 반응 과정에 대한 내용이 누락되어 1점을 부여하였다고 평가 근거를 출력하였다. 결과적으로, 교사와 챗봇의 채점 점수가 일치하더라도, 동일한 채점 근거에 기반한 채점이 이루어지지 않았을 가능성도 있다는 점은 주목할 필요가 있다.

학생에게 요구하는 응답이 특정 개념 등의 용어로 구체화 되어 있는 문항이 아니라, 학생 답안의 다양성이 높게 나타나는 문항의 경우, 각 사례에 채점 기준 및 근거를 유연하게 적용해야 하지만, 이러한 능력이 ChatGPT에서는 잘 나타나지 않았다. 6번 문항 세 번째 채점 기준과 관련된 사례에서 교사가 생성한 채점기준에서는 결론이 무엇이고 결과가 어떻게 다른지 제시되어 있는데 해당 기준으로 교사와 ChatGPT가 사례에 제시된 학생 답안을 평가하였을 때 교사는 결론에서의 일반화를 기술했다는 관점에서 해당 응답을 정답으로 평가하였으나 ChatGPT는 가설과 일치여부가 중요하다는 점을 언급하며 0점을 부여하였다.

이 외에도 답안에서 용어를 잘못 사용하였거나 오타가 있는 경우, 개념을 명확하게 제시하지 않았지만 설명이 충분한 경우 교사의 채점과 챗봇의 채점이 서로 일치하지 않는 현상이 나타나기도 하였다.

나. ChatGPT가 수행한 문항 채점 결과에 대한 교사들의 인식

교사들은 ChatGPT가 채점한 결과가 교사들이 채점한 결과에 비추었을 때 일치도가 높지 않고, 채점 근거의 일관성이 다소 부족하다는 점을 이유로 ChatGPT를 채점에 직접적으로 활용하는 것이 어려울 것이라고 인식하고 있었다. 구체적으로 한 교사는 핵심 용어나 주요 내용에 대해 초점을 맞추는 경향이 있는 반면 ChatGPT의 경우 문항의 의도보다는 학생 답안 내용 자체의 정합성에 초점을 맞추는 경향을 보인다는 의견이 있었다.

"저는 핵심 용어나 동의어 또는 동의 표현이 포함되어 설명하면 가점을 하였지만, GPT는 핵심 용어 포함 유무는 기준에 두지 않고, 맞는 말이라면 모두 가점을 한 것으로 해석됩니다."

다만, 교사의 채점도 ChatGPT의 채점과 마찬가지로 항상 정확하지 않고, 문항에 대한 채점 기준이 일관되지 않거나 모호한 경우도 있을 수 있기 때문에 이러한 부분을 점검하여 채점의 신뢰성을 높이기 위한 보조 도구로서 ChatGPT의 활용 가능성을 제시하기도 하였다.

"교사가 서술형 답안지를 채점할 때 재검 단계에서 ChatGPT의 채점 결과와 교사의 초검 결과를 비교해 보고, 불일치하는 사례 위주로 그 점수 차이가 나타나는 이유를 고민해본다면 초검이 제대로 이루어졌는지 검증하기에 유용할 것 같습니다. … ChatGPT의 분석 근거들이 구체적으로 제시되어 있으므로 교사가 채점할 때 참고하기에 유용할 것이라고 생각했습니다. 예를 들어 교사가 채점 단계에서 점수 부여를 어떻게 하는 것이 옳을지 모호한 부분이 있을 때 ChatGPT의 분석 근거를 읽어본다면 도움이 될 것이라고 생각했습니다."

또한 채점 결과 분석에서 나타나는 ChatGPT의 부정확한 반응의 원인을 파악하고, 문항의 출제 및 프롬프트의 입력 측면에서 이러한 부분들을 줄일 수 있는 방안을 마련한다면 효율적인 활용이 가능할 것이라는 응답도 있었다.

"사전 채점이 일부 일치하는 ChatGPT의 분석 결과 다수가 교사 간의 학생 답안 해석에 대한 협의 과정을 거쳐 의견을 조율하는 과정을 거친다면, 교사와 ChatGPT의 채점 결과 일치도가 높아질 것으로 사료됨."

3. 평가 결과에 대한 피드백 단계에서의 ChatGPT의 활용

가. ChatGPT가 생성한 피드백 사례의 특징

ChatGPT가 생성한 문항의 피드백 사례를 살펴보면 타당도와 신뢰도를 완전히 갖추었다고 보기에는 어려움은 있으나, 활용 측면에서 긍정적인 부분도 있었다.

구체적으로, 답안의 내용 중 일부가 잘못되어 있는 경우 정확한 피드백을 제공하지 못하는 사례가 있었고, 과학적으로 타당한 진술을

타당하지 않다고 인식한 응답도 있었다. 또한 교육과정 수준에서 학생들이 답안을 올바르게 썼다고 하더라도, ChatGPT의 피드백이 교육과정을 넘어서는 수준에서의 학습을 권장하는 경우도 있었다(Table 10).

1번 문항의 예시와 같이 학생의 답안에서 '번역'이 '복사'로 잘못 서술되었으나, 이에 대한 지적을 하지 못한 반면, '물이 빠져나오게 된다.'와 같은 진술은 타당함에도 불구하고 과학적이지 못한 진술이라는 피드백을 제공하기도 하였다. 2-1번 문항의 예시와 같이 교육과정을 넘어서는 수준에서의 학습을 요구하는 피드백을 제공하기도 하였다. 이러한 부분이 심화/보충 학습이라는 측면에서 잘못된 것이 아닐 수도 있지만, 과도한 학습을 요구하는 것이 적절하지 않을 수도 있다. 또한 문항에서 요구하는 답안의 요소 및 조건들이 ChatGPT가 인식하기에 어려움이 있는 경우, ChatGPT의 채점 결과 및 피드백의 내용이 각 답안의 요소들을 분명하게 구분하여 이루어지기 보다는 총괄적으로 생성되는 경향도 발견할 수 있었다. 이러한 경우 학생이 실제로 '잘못 응답'한 부분을 분명하게 짚어주지 못하게 된다.

반면, Table 11에 제시한 사례들과 같이 ChatGPT가 생성한 피드백이 적절하다고 판단되는 사례들도 확인할 수 있었다. 1번 문항의 사례와 같이 학생들의 답안이 비어있거나 내용이 간결하게만 제시된 경우 '모범 답안'을 제공하는 등의 적절한 피드백을 제공하기도 하였으며, 2-2번 문항과 같이 특정 사례를 조사하여 응답하도록 하는 문항의 경우, 학생의 응답과 관련되어 있는 추가적인 사례들을 제공해 주기도 하였다. 또한 6번 문항의 경우 답변 기술시 명확한 표현을 위해 수정해야 하는 점을 피드백으로 제공하여 단순히 학생들이 제시한 내용적인 측면에서의 피드백뿐만 아니라 형식적인 측면에서도 개선점을 제공하는 사례가 있었다. 이러한 부분은 학생 응답에 대한 개별화된 피드백을 제공하는 데 있어서 인공지능 챗봇의 활용 방안으로서 긍정적인 부분이라고 할 수 있다.

나. ChatGPT가 생성한 피드백에 대한 교사들의 인식 및 활용 방안

교사들은 ChatGPT가 생성한 피드백의 불완전성에 주목하여 학생들에게 직접적으로 제공하는 데 한계가 있음을 인식하고는 있었다. 그러나 학생들 각각에게 피드백을 제공하는 것 자체가 어려움이 있기 때문에 ChatGPT를 통해서 생성된 피드백을 수정・보완하여 학생들에게 제공하는 방안이 평가 업무를 경감하는 데 도움이 될 것이라는 의견도 있었다.

Table 10. Feedback from ChatGPT about student responses that needs teacher review before use

문항 번호	학생 응답	ChatGPT가 생성한 피드백
1	복사이다. 코돈이 아미노산의 종류를 결정한다. 아미노산은 펩타이드 결합을 통해 단백질을 형성한다. 이때 물이 빠져나오게 된다.	학생의 답안은 과정 (가)를 정확하게 설명하였지만, "물이 빠져나오게 된다"라는 부분은 과학적 개념 측면에서 정확하지 않은 정보입니다. …
2-1	플라스틱으로 인한 환경문제를 효소로 해결할 수 있는 이유는 바로 효소의 물질분해 능력 때문입니다. 가와 같이 밥의 포도당을 녹말로 목말에서 엿당으로 바꾸어 단맛을 내는 것과 나와 같이 인공적인 고열 고압의 환경에서 염산이라는 위험한 물질 없이도 일반 자연환경에서 효소를 이용하여 단백질을 분해 가능하기 때문에 이 능력을 플라스틱 환경문제에서 유용하게 활용할 수 있을 것 같습니다.	…(생략)… 그러나, 효소의 특징과 플라스틱 분해에 대한 더 명확하고 자세한 설명이 필요합니다. 예를 들어, 효소가 물질을 분해하는 원리에 대해 좀 더 구체적으로 설명하고, 플라스틱 분해 효소가 플라스틱의 고분자 구조를 어떻게 분해하는지에 대해 더 자세히 설명하는 것이 도움이 될 것입니다. …(생략)…
4	(라)의 식물 상태 결과표를 봤을 때 (가)에서 예측한 '비료를 권장량보다 많이 주면 식물이 더 잘 자랄 수 있을 것'이라는 예측은 옳지 못하며, A에서 식물이 잘 자라지 못한 이유는 과도해진 비료 성분에 막혀 정작 식물의 성장에 가장 중요한 요소들은 세포막을 통과하지 못했기 때문이다.	…(생략)… (예측의 타당성 대한 피드백) 학생은 (라)의 결과와 (가)의 예측 사이의 관계를 과학적인 이유와 논리로 설명해야 합니다. 비료를 권장량보다 많이 주는 것이 식물의 성장에 불리하다는 점을 잘 이해하고 있지만, 더 정확한 이유를 설명해야 합니다. 식물에 과도한 비료가 주어지면 뿌리와 줄기 등의 세포막을 통과하는 물질의 농도 차이가 식물의 수분 흡수를 방해할 수 있으며, 이로 인해 식물이 물과 영양분을 올바르게 흡수하지 못하여 성장에 방해가 됩니다. …(전생략)…

Table 11. Examples of feedback from ChatGPT on student responses that you might want to utilize

문항 번호	학생 답안	ChatGPT가 생성한 피드백
1	번역, 펩타이드결합을 했다.	…(생략)… 하지만 문제에서 요구한대로 <보기>에 제시된 용어들(코돈, 아미노산, 펩타이드 결합)을 모두 사용하여 설명해야 합니다. 코돈과 아미노산에 대한 설명이 없기 때문에 1점으로 평가되었습니다. 학생은 '번역'과 관련하여 코돈은 RNA 염기 3개의 조합이고, 이 코돈은 아미노산을 지정하며, 이 아미노산들이 펩타이드 결합을 통해 단백질을 형성한다는 점을 추가적으로 설명해야 합니다.
2-2	동식물의 신체를 모방하여 실생활에 사용하는 사례들이 있다. 도꼬마리에서 영감을 받아 제작한 벨크로 테이프가 그 대표적인 예이다.	…(생략)… 또한, 다양한 동식물의 신체를 모방하여 개발된 제품들을 추가적으로 소개할 수 있습니다. 예를 들어, 참새의 날개를 모방하여 개발된 덕트 시스템, 상어의 비늘 구조를 모방하여 개발된 무인 잠수함 등 일상생활에서 활용되는 동식물 모방 기술에 대해 추가로 언급할 수 있습니다. …(생략)…
6	알칼리 금속의 원자 번호가 클수록 알칼리 금속의 반응성은 더 크다.	…(생략)… 하지만, 더 정확한 표현을 위해 '클수록'이라는 표현 대신 '증가함에 따라' 또는 '증가하는 경향을 보였다'와 같이 구체적인 수량적인 변화를 반영할 수 있는 표현이 더 적절합니다. 또한, 실험에서 어떤 원자번호를 사용했는지와 실험에서 관찰한 결과를 구체적으로 언급하는 것도 과학적인 내용을 보다 명확히 표현하는 데 도움이 됩니다. 이러한 점을 고려하여 답안을 보완하는 것이 좋습니다.

"ChatGPT를 활용한 채점에 대한 피드백 부분은 대단히 유의미한 부분이라고 생각됩니다. 실제 학교 현장에서 평가 후 수업 시간 부족과 함께 교사의 업무 과중으로 개인별 피드백의 제공이 잘 이루어지지 않고 있는 상황입니다. 학생들에게 모범 답안만 제시하고 학생들 자신의 답변이 틀린 이유만 제공하는 경우가 많은데 피드백에서 학생들의 틀린 답안에 대한 이유 그리고 올바른 정답에 대한 예시까지 학생들에게 제공하는 것은 유의미한 의미가 있는 피드백인 것 같습니다. 다만 피드백에 교육과정을 넘어서는 부분에서는 수정이 있어 보이는데 이 부분은 교사가 피드백을 받고 다시 수정하는 것으로 보완할 수 있는 부분입니다. GPT에 의한 피드백은 교사의 업무의 경감이 있을 것으로 기대되어 많은 도움을 받을 수 있는 부분이라고 생각됩니다."

Ⅳ. 결론 및 제언

최근 상용화된 ChatGPT 등의 인공지능은 특정 분야에 특화시킨 모델이라기보다 범용성을 갖고 있다는 측면에서 평가 등의 교육 분야에로의 활용을 위해서는 가능성과 한계에 대해 면밀하게 살펴보는 것이 필요하다. ChatGPT를 활용한 본 연구의 결과에서는 인공지능의 생성물을 평가 상황에 활용하는 것은 생성물의 종류, 문항의 형태 등에 따라 그 가능성과 한계가 다소 달랐다. 채점과 관련하여 ChatGPT의 채점 결과는 사람의 채점 결과에 비해 상대적으로 관대하였다는 다른 연구들(Bhat *et al.*, 2022; Moore *et al.*, 2022; Park *et al.*, 2023)과 유사하면서도 특징적인 결과가 나타났다. 응답을 위해 특정 과학 용어를 사용해야 하는 경우, 문장이나 문단의 연결이 구조상 적절하게 작성된 경우에 사용한 용어가 적절하지 않아도 정답으로 채점하는 사례들이 나타났다. 반면, 과학 용어가 구체적이지 않다면, 학생 응답의 맥락을 해석하여 채점해야 하나, ChatGPT의 결과에서는 다소 경직된 방식의 채점이 나타나기도 하였다.

한편, 중요한 사용자인 교사들은 피드백, 평가 기준에 대해서는 사용 가능성을 높게 보았으나, 응답에 대한 채점 결과에 대한 신뢰도는 낮게 평가하였다. 즉 현재 학교에서 진행되는 일련의 평가 과정에 ChatGPT 등의 생성형 인공지능이 생성한 결과를 수정하지 않고 사용하는 것은 평가 결과의 타당성, 신뢰성 측면에서 적절하지 않음이 자명하다. 다만, 교사들은 채점 결과를 포함해 모든 생성물을 주요한 자료로 사용하는 것보다 보조 도구로, 즉 교수학습의 업무 지원 측면에서 활용할 수 있겠다고 생성물을 바라보았다. 이는 자료 자체의 신뢰도가 낮더라도 사전 또는 사후에 교사 자신의 생성물을 검토하기 위한 목적에서 사용할 수 있다고 판단하였음을 의미한다. 구체적으로 평가 기준은 문항 제작 단계에서 평가 요소가 명확한지 검토하는 용도로, 채점 결과는 교사 자신의 채점 결과를 검토하기 위한 용도로, 피드백은 ChatGPT의 초안을 수정, 보완하여 학생들에게 제공하는 방식으로 활용할 수 있다고 보았다. 특히, 학생들의 응답 별로 피드백을 준비할 수 있다는 점은 교사의 업무를 크게 경감할 수 있어, 큰 이점을 가져올 수 있기 때문에 고무적이다.

본 연구에서 살펴본 바와 연구진의 연구 진행 과정에 따른 경험을 바탕으로 몇 가지 사항들을 논의하고자 한다. 현재 공개되어 익숙한 생성형 인공지능 서비스들은 교육 분야에 특화되어 있지 않다. 또한 그 특성상 일관적인 반응을 이끌어 내는 것도 쉽지 않다. 일반적인 학교의 맥락을 고려하면 한 개 학급, 또는 학년에 동일한 문항을 제시하고 답안을 채점하게 되는데, 앞서 말한 특성에 의해 일정한 채점 결과를 얻는 것은 분명한 한계가 있을 것으로 예상된다. 따라서 개별적인 응답의 신뢰도와 타당성을 확보하고 이후 일관적인 응답을 이끌어내는 방법이 필요하다. 연구진들이 ChatGPT를 활용해 학생 응답을 채점하는 과정에서 경험한 것은 프롬프트가 구체화되고 정교화 될수록 더 일정하게 기준에 근거하고 형식이 동일한 결과를 얻을 수 있었다는 점이다. 즉, 상용화된 인공지능 시스템 활용을 위해서는 사용자들의 반복적인 경험을 토대로 한 최적화된 프롬프트 구축이 이루어져야 한다. 이를 보다 효과적으로 달성하기 위한 방안 중 하나로는 전문적 학습 공동체와 같은 집단에서 이에 대한 경험 공유의 장을 마련하는 것이 있을 수 있으며, 이외에는 Jho(2023)의 연구에서 언급한 파인튜닝된 모델의 제공을 고려할 필요도 있을 것이다. 이러한 선상에서 Park *et al.*(2022)의 연구가 추진하고 있는 한국어 기반의 서·논술형 평가 자동채점 시스템과 같은 모델들을 지속적으로 개발하고 활성화할 필요가 있다.

다른 측면에서 살펴보면, 문항에서 요구하는 과학 개념의 수가 많을수록 ChatGPT가 생성한 채점 기준과 교사의 채점 기준 사이의 간극이 커지는 것을 확인할 수 있었다. 아울러, 이 때 교사의 채점 기준과 다르게 표현의 정합성 등에 초점이 맞추어지는 경향도 나타났다. 따라서 ChatGPT를 평가에 활용하고자 한다면, 모든 문항에 적용하기보다 주어진 특정 조건 내에서 설명을 요구하는 문항과 같이 제한 범위가 명확한 문항에서 활용하는 것도 활용 가능성을 높이는 하나의 방안이 될 수 있다. 이외에도 문항 개발 단계에서 인공지능과의 상호작용을 통해 개발한 문항이 인공지능을 활용한 채점에 활용될 수 있는지 탐색하는 것도 도움이 될 것이다.

평가를 실시하기 위한 문항 개발 단계에서 ChatGPT는 다른 방식으로도 활용 가능하다. ChatGPT와 같은 텍스트 기반의 생성형 인공지능은 정보의 입력에 대응하는 글을 계속 생성하기 때문에 평가 문항 개발 단계에서 문항의 타당성을 높이기 위해 활용 가능하다. 예를 들어, 서술형 문항 개발 과정에서 교사가 예상하기 어려운 학생들의 응답을 사전에 파악하는 목적에서 답안 작성을 요청하는 용도로 인공지능을 사용할 수 있다. 이와 유사한 관점에서 교사가 채점 기준을 작성할 때, 교차 검토의 목적에서 인공지능에게 채점 기준 생성을 요청하는 등의 방안으로 활용 가능성을 살펴볼 수 있다. 실제 본 연구에 참여한 교사들 역시 이러한 방식의 활용을 언급하곤 하였다.

한편, 입력 정보에 대응하는 글을 생성한다는 점은 평가와 교수·학습의 유기적 연계를 보다 원활하게 이루는 데 활용 가능하다. 많은 연구들이 언급한 것처럼 맞춤형, 개별화 교육의 지원은 ChatGPT와 같은 유형의 인공지능에 기대하는 것 중에 하나이다(Chang, Park, & Park, 2021; Kim & Yu, 2023; Jho, 2023). 본 연구에 참여한 교사들도 ChatGPT가 학생들의 각 응답별로 피드백을 생성할 수 있음에 관심이 높았고, 사용자의 검토를 통해 일부 내용만 수정하면 실제 수업에서 활용 가능할 것으로 이야기하였다. 이러한 언급은 피드백이 교수·학습의 과정과 중첩되는 활동이기 때문에 정확성의 측면에서 보다 유연한 기준을 가지고 있음에 기인한다고 볼 수 있지만, 채점이 정확하고 이루어질 때 피드백도 더욱 효과적일 수 있음을 고려하면 응답에 대한 해석과 무관하게 사용하는 것은 다소 유의할 필요도 있다.

구체적인 실천 방안을 예로 들어 제안하면 다음과 같다. 교사가 사전에 준비한 문항과 함께 답안에 대한 평가 및 피드백을 받을 수 있도록 구성한 프롬프트를 준비하여 학생들에게 제공한다. 학생들은

문항에 대한 정답과 프롬프트를 ChatGPT에 입력하여 자신이 입력한 응답에 대한 채점 결과와 피드백을 받는다. 이러한 과정을 교사의 수업 계획에 따라, 또는 학생 스스로 반복하여 수행하는 과정을 통해 개인별로 응답에 부족한 부분을 파악하거나, 관련된 과학 내용이나 정보들을 탐색하도록 안내하는 방식으로 개별화된 지도를 계획할 수 있다.

본 연구는 최근 가장 널리 알려진 인공지능 활용 기술 중 하나인 ChatGPT를 평가 상황에 적용하고 활용 방안에 대한 교사들의 인식을 확인하였다는 측면에서 선행 연구들과는 차별성을 갖는다. 다만, 최신의 모델인 GPT-4가 아닌 GPT-3.5를 이용하였다는 점, ChatGPT 자체가 교육 영역이나 평가 자체에 특화된 모델이 아니라는 점은 본 연구의 한계라고 볼 수 있다. 즉, 본 연구는 교육 맥락에서 인공지능 챗봇의 활용 가능성을 단정적으로 판단하기 보다는, 탐색적으로 살펴보는 데 초점을 두고 있다. 이러한 측면에서 향후 인공지능 챗봇 기술의 발달과 함께 지속적인 연구가 수행되어야 하며, 향후 생성형 인공지능 기술을 교육의 여러 영역에서 활용하기 위해서는 교육 분야에 특화시킬 수 있는 데이터 수집 및 처리를 거친 모델의 구축과 활용 방안 모색이 필요하다. 이를 위해서는 인공지능 기술의 전문가, 교과 교육 전문가, 평가 분야의 전문가 등의 지속적인 협업이 필수적이다. 이와 함께 앞으로 학급 단위나 더 소규모 단위에서 학습을 지원하기 위한 방안을 모색하고 효과를 살펴보는 연구들이 추진될 수 있을 것이다. 소규모 집단을 대상으로 한 연구를 통해 인공지능을 활용한 교수·학습과 평가의 연계 방안을 마련하고 그 효과를 검증하는 것이 추후 이루어져야 할 것이다. 이와 같은 현장 적용과 연구를 통해 발전하고 있는 여러 기술들의 활용 방안을 탐색하고 기대하는 개별화, 맞춤형 교육을 위한 방안을 지속적으로 탐색하는 것이 요구된다.

국문요약

본 연구는 과학 교과의 학생 평가에서 생성형 인공지능인 ChatGPT의 활용 가능성을 탐색하는 데 목적이 있다. 이를 위해 평가 문항을 개발하여 학생들로부터 문항에 대한 응답 자료를 수집하였고, 응답 결과를 ChatGPT에 입력하여 평가의 과정을 수행하도록 하였다. 또한 교사들에게 ChatGPT로부터 얻은 평가 결과를 공유하고, 교사들의 평가 과정과 비교하여 학생 평가에서의 ChatGPT의 활용 가능성을 탐색해 보았다. 연구 결과, 채점 기준의 설정 측면에서 ChatGPT가 생성해 내는 채점 기준이 전반적으로 타당성을 갖추고 있는 것으로 볼 수 있었다. 그러나 ChatGPT가 수행한 채점 결과는 교사들 채점한 결과와 비교하였을 때 일관성이 다소 낮았고, 특히 문항에 포함된 평가 요소가 많고, 평가 요소별 배점 기준이 복잡할수록 채점 결과의 일치도가 더 낮게 나타났다. 학생 응답에 대한 피드백 측면에서는 학생 응답의 과학적 타당성을 평가하는 과정에서 일부 잘못된 내용을 제공하거나 교육과정을 넘어서는 수준의 피드백이 생성되는 경우도 있었으나, 문항에 대한 올바른 답안을 알려주고, 추가적인 사례들을 제공하는 등의 긍정적인 부분도 확인할 수 있었다. 이러한 결과들을 바탕으로 교사들로부터 ChatGPT의 활용 가능성에 대한 인식을 살펴보았을 때, 학생 평가에서 중요하게 인식되는 '신뢰성' 측면에서 부족한 면이 있기는 하지만, 교사들의 평가를 지원하는 측면에서 활용 가능성을 발견할 수도 있었다. 마지막으로, 이러한 연구 결과를 종합하여 학생 평가에서의 ChatGPT의 활용에 대한 시사점을 제언하였다.

주제어 : 학생평가, ChatGPT, 교사 인식

References

Bhat, S., Nguyen, H., Moore, S., Stamper, J., Sakr, M., & Nyberg, E. (2022). Towards automated generation and evaluation of questions in educational domains. Paper presented at the 15th International Conference on Educational Data Mining, Durham.

Byeon, J.-H., & Kwon, Y.-J. (2023). An investigation of generative AI in educational application: Focusing on the usage of ChatGPT for learning biology. Brain, Digital, & Learning, 13(1), 1-17.

Chang, J., Park, J., & Park, J. (2021). An analysis on the trends of education research related to 'artificial intelligence chatbot' in Korea: Focusing on implications for use in science education. Journal of Learner-Centered Curriculum and Instruction, 21(13), 729-743.

Choi, S.-W., & Nam, J.-H. (2019). The use of AI chatbot as an assistant tool for SW education. Journal of the Korea Institute of Information and Communication Engineering, 23(12), 1693-1699.

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial intelligence for student assessment: A systematic review. Applied Sciences, 11(12), 5467-5482.

Holmes, W., Bialik, M., & Fadel, C. (2019). Artificial intelligence in education: Promise and implications for teaching and learning. Boston: Center for Curriculum Redesign.

Hong, S., Cho, B., Choi, I., Park, K., Kim, H., Park, Y., & Park, J. (2020). Artificial Intelligence and EduTech in School Education. RRI 2020-2, Korea Institute for Curriculum and Evaluation.

Huh, M., Bae, Y., Seok, H., & Lee, J. (2021). Domestic research trends of learning with AI. Journal of The Korean Association of Information Education, 25(6), 973-985.

Jang, H., & So, H.-J. (2023). The analysis of research trends and topics about the educational use of ChatGPT. Journal of Research in Curriculum & Instruction, 27(4), 387-401.

Jho, H. (2023). Understanding of generative artificial intelligence based on textual data and discussion for its application in science education. Journal of the Korean Association for Science Education, 43(3), 307-319.

Kang, D. (2023). The advent of ChatGPT and the response of Korean language education. Korean Language and Literature, 82, 469-496.

Kim, H. (2021). The artificial intelligence era and science education - With a focus on the autonomy and relatedness of artificial intelligence -. The Journal of Yeolin Education, 29(6), 1-23.

Kim, H., Park, J., Hong, S., Park, Y., Kim, E., Choi, J, & Kim, Y. (2020). Teachers' perceptions of AI in school education. Journal of Educational Technology, 26(3), 905-930.

Kim, J. I., & Yu, H. (2023). Exploring applications of ChatGPT for physics education: Focusing on high school and general physics class. School Science Journal, 17(3), 216-239.

Moore, S., Nguyen, H. A., Bier, N., Domadia, T., & Stamper, J. (2022). Assessing the quality of student-generated short answer questions using GPT-3. Paper presented at the 17th European Conference on Technology Enhanced Learning, Toulouse.

OpenAI. (2023). ChatGPT: Optimizing language models for dialogue. Retrieved December 05, 2023 from https://openai.com/blog/chatgpt/

Park, J.-I., Lee, S., Song, M. H., Lee, M., Lee, M., & Choi, S. (2022). A Study on the Development of Automated Scoring Method for Computer-based Essay and Short Answer Question Type Assessment (Ⅰ). RRE 2022-6, Korea Institute for Curriculum and Evaluation.

Park, S.-Y., Lee, B., Lee, Y. Ham, E. H., & Lee, S. (2023). Exploring the possibility of science-inquiry competence assessment by ChatGPT-4: Comparisons with human evaluators. Korean Journal of Educational Research, 61(4), 299-332.

Shin, D. (2019). Feasibility and constraints in applying an AI chatbot to english education. Brain, Digital, & Learning, 9(2), 29-40.

Shin, D., Jung, H., & Lee, Y. (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.

Smutny, P., & Schreiberova, P. (2020). Chatbots for learning: A review of educational chatbots for the Facebook Messenger. Computers & Education, 151, 1-11.

Son, T. (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.
Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Li, J., Yuan, J., & Li, Y. (2021). A review of Artificial Intelligence (AI) in education from 2010 to 2020. Complexity. 2021, 1-18.
Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., Peng, H., Li, J., Wu, J., Liu, Z., Xie, P., Xiong, C., Pei, J., Yu, P. S., & Sun, L. (2023). A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. arXiv preprint arXiv:2302.09419. Retrieved May 23, 2023 from https://arxiv.org/abs/2302.09419

저자정보

이동원(한국교육과정평가원 부연구위원)
심현표(한국교육과정평가원 부연구위원)
백종호(한국교육과정평가원 부연구위원)