

Characteristics of GPT-4 automated scoring of scientific inquiry competency

Eun Hye Ham¹⁾ · So-Young Park²⁾ · ByungYoon Lee³⁾
Sunghye Lee⁴⁾ · You-kyung Lee⁵⁾ · Yujung Hong⁶⁾

¹⁾Associate Professor, Department of Education, Kongju National University

²⁾Professor, Division of Education, Sookmyung Women's University

³⁾Post-doctoral Research Fellow, Education Research Institute, Sookmyung Women's University

⁴⁾Research Professor, Global Institute for Talented Education, Korea Advanced Institute of Science and Technology

⁵⁾Associate Professor, Division of Education, Sookmyung Women's University

⁶⁾Research Professor, Soongsil University

This study aims to examine the effectiveness of GPT-4 in automated scoring system compared to human experts when assessing students' scientific inquiry reports. A total of 322 elementary students' science inquiry reports were evaluated using a GPT-4-based automated scoring system as well as by a group of human experts. The two sets of scoring data were compared to examine whether the internal structure of the GPT-scoring data resembled that of human scoring, and whether it aligned with theoretical expectations of scientific inquiry skills. The key findings are as follows: First, GPT-4's scoring was generally more lenient, particularly with more challenging criteria. Second, GPT-4 demonstrated higher consistency in scoring and internal consistency among items than human scoring. Third, the many-faceted Rasch model showed significant discrepancies in item difficulty when integrating GPT-4 scoring data with human scoring data, adversely affecting the internal and external fit of human raters, which suggests that the comparability between GPT-4 and human scoring outcomes is limited. Based on these findings, the limitations, possibilities, and challenges of using GPT-4 for automated scoring are discussed.

Keywords : GPT-4, Automated Scoring, Reliability, Many-faceted Rasch Model, Scientific Inquiry Competency

I. 서론

교육 장면에서 AI 활용에 대한 연구와 교육은 최근 빠르게 증가하고 있다. 2022년 정부가 발표한 디지털 인재양성 종합방안(관계부처 합동, 2022)에서는 교육현장에서의 AI 교육보조도구 도입, 디지털 교과서 등을 포함한 맞춤형 콘텐츠 개발, AI 기반 맞춤형 교육 플랫폼 구축 등의 방안 등을 강조하였으며, 이어 2023년 교육부는 디지털 기반 교육혁신 방안을 발표한 바 있다(교육부, 2023). 이 방안에서는 2028년까지 AI 디지털 교과서를 전면 도입하고 지능형 튜터링 시스템 등을 통해 학생들의 학습 유형에 따른 피드백 정보를 제공하는 교수-학습 모형을 제안하면서, 진단평가, 형성평가, 종합평가를 통한 개별 학생에 대한 이해와 개별화된 교수-학습을 강조하였다. 이는 에듀테크 및 AI 기술을 교수 및 학습 활동의 설계와 구현에 적용할 뿐만 아니라 개별 학생에 대한 진단과 평가에 적용할 가능성과 필요성을 시사한다.

한편, 학생평가는 최근 교수-학습 혁신 정책의 중요한 축이 되었다. 수업방식의 전환과 평가 혁신에 대한 논의는 과정중심평가와 서·논술형 평가의 확대, 교사의 평가 전문성 향상 정책을 통해 구체화되고 있다(교육부, 2024). 특히, 서·논술형 평가의 확대는 2022 개정 교육과정에서 학습의 과정을 중시하는 평가의 강조와 고등학교급에서 서술·논술 능력을 통한 창의적·비판적 사고 능력을 강조하는 과목 신설 등과도 흐름을 같이 한다(교육부, 2022). 그동안 학교에서의 평가는 평가의 형식이 선택형과 단답형 위주라는 이유로 학교에서 학생의 지식이나 정보의 수준만을 측정하고, 고등정신능력을 평가하지 못한다는 비판을 받아왔다. 박혜영과 동료들(2018)은 미래사회 대비 교육평가 비전 연구에서 미래사회에서 인간에게 필요한 역량이 무엇인지를 고찰하고, 교수-학습과 교육과정이 연계된 학습을 위한 평가, 과정을 지향하는 형성평가, 학습자 맞춤형 평가를 다시 한 번 강조하였다.

그러나 교수-학습 과정에서 도출되는 산출물 특히 수행형 평가자료는 개방적인 형태로, 텍스트, 도표·그림, 음성, 동영상 등 자료 유형이 다양하고, 평가해야 하는 영역이나 내용이 다차원적이고 복잡하기 때문에(백순근, 2000), 교사의 평가 부담이 매우 크다는 점에서 현장에서의 활용에 한계가 있다. 예를 들어, 가장 대표적인 수행형 평가인 서·논술형 평가의 경우, 채점자의 시간과 노력이 많이 들 뿐만 아니라 교사 개인 수준에서 채점기준을 정교화하기 어렵고, 채점자 내 판단의 비일관성과 채점자 간 판단의 편차를 통제하기 어려워(백유진, 2020; 지은림, 2008; 함은혜, 유예림, 2022), 학교 현장에서 확대되는 데 제한이 있다. 이처럼 채점에 대한 부담은 교수-학습과 연계된 과정중심평가 시행의 가장 큰 장애물 중에 하나로(박혜영 외, 2019), 이를 해소하기 위해 한국교육과정평가원을 시작으로 서·논술형 문항의 채점을 자동화하려는 시도가 꾸준히 있었다(진경애 외 2006; 노은희 외, 2012, 2014 등). 자동채점 연구는 한국어 자연어 처리기술 및 인공지능의 발전과 함께 전환기를 맞았으며, 최근 챗GPT를 포함한 대규모 언어모델의 상용화로 학생평가 방식은 새로운 국면을 맞을 것으로 기대되고 있다.

한편, 학교 학습에서 탐구역량과 이를 위한 활동과 평가가 지속적으로 강조되어 왔다. 탐구활동은 학생들의 고차원적 사고를 개발할 뿐만 아니라 학습참여도를 높이는 효과적인 학습 방법이자 평가방법이다. 특히, 2015 개정 교육과정에서는 학생 중심 수업이 과학적 탐구과정을 통해 이루어져야 함을 강조하고 있으며(교육부, 2015), 2022 개정 교육과정에서도 역시 학생의 탐구 및 추론, 통합적 사고, 문제해결력 등의 과학적 역량을 강조하고 있다(교육부, 2022). 탐구활동은 학생들이 구체적인 문제 상황에서 문제를 찾고, 문제를 해결하는 과정에 교과 내용 지식을 적용하도록 요구한다. 탐구보고서는 학생들이 자신의 문제해결 과정과 결과를 요약·기록할 뿐만 아니라 스스로 그 과정을 모니터링하는 동시에, 다른 사람이 그 과정을 이해할 수 있는 방식으로 자신의 사고 절차를 기술할 뿐만 아니라 구조화하는 과정을 포함한다는 점에서 고차원적인 학습활동의 집합체라고 할 수 있다. 학생들은 탐구보고서와 같은 과학적 글쓰기 과제를 통해 과학적 사고력과 과학적 탐구능력을 향상시킬 수 있다(박찬술, 손정우, 2020) 따라서 이와 같은 탐구보고서 평가의 과정을 일정 부분 자동화할 수 있다면 학교에서 탐구활동 평가를 확대하는 데 기여할 수 있을 것이다. 다만, 탐구보고서 평가는 형식상 서·논술형이기 때문에 글쓰기 평가를 포함하지만, 그에 더하여 학습목표를 반영하여 교과 지식이 적절하게 활용되고 있는지, 교과의 특성이 반영된 논증의 구조와 내용이 적절한지에 대한 평가가 통합될 필요가 있다는 점에서 매우 어려운 채점일 수 있다(김덕영, 박종원, 2015; 김현정, 김성기; 2021). 이런 맥락에서 박소영과 동료들(2023)의 연구에서 챗GPT를 활용하여 초등학생의 탐구보고서를 평가한 것은 새로운 시도로 이해할 수 있으며, 향후 탐구보고서 및 탐구역량에 대한 평가의 자동화 가능성과 과제를 지속적으로 탐색할 필요가 있다.

글쓰기 영역에서의 자동채점 연구는 상대적으로 활발하게 이루어졌다. 최근 이용상과 동료들(2022)의 연구에서는 랜덤포레스트 기법으로 자동채점 알고리즘을 구현하여 세종한국어평가 쓰기 도구에 대한 개발 및 연구를 진행하였다. 이 연구에서 인간-인간 채점과 인간-기계 채점 결과를 채점 모형의 정확도 측면에서 비교한 결과, ‘과제수행’, ‘언어사용’, ‘내용’의 세 가지 평가 영역 중 ‘언어사용’과 ‘내용’ 영역에서 양호한 수준이었다. 국립국어원에서도 인공지능을 활용하여 성인의 국어능력을 진단하는 자동채점 도구 개발을 진행하고 있다. 국립국어원은 2022년 인공지능 활용 국어능력 진단체계 개발 기초연구를 수행하고, 이를 기반으로 2023년에는 자동 채점을 지원하기 위한 인공지능 기반 모델을 개발하는 연구를 수행하였다. 이 연구에서는 사전학습 언어모델을 적용하여 신뢰도를 검증하였고(민병곤 외, 2023), 이 연구의 후속 연구에서는 인공지능 자동채점 도구를 개발하기 위한 작업을 진행할 것으로 예상된다. 이와 같은 흐름으로 이용상과 동료들(2023)은 ELECTRA 알고리즘을 적용하여 자동채점 도구 PASTA- I (Personalized Automated Scoring and Tutoring Assistant)를 개발하였다. 한국지능정보사회진흥원에서 제공한 에세이 글 평가 데이터를 적용하여 자동채점을 시행한 결과, 개발한 도구가 양호한 성능을 보이고 있음을 확인하였고, 향후 학교 현장에서 글쓰기 평가도구로 사용 가능할 것으로 전망하였다.

과학 교과에서 자동채점 활용 연구도 일부 수행되었다. 하민수와 동료들(2019)은 WA³I (Web-based Automated Assessment with Artificial Intelligence)에서 랜덤포레스트 기법을 활용한 과학 교과 서답형 문항 채점의 가능성을 보여주었다. 그러나 이 모형은 과학 교과의 내용 지식에 의존하는 짧은 답안을 평가한다는 점에서 논술형 평가의 활용에는 제한이 있다. 이만형과 유선아(2020; 2021)는 실제 교실에서의 학생의 발화에서 관찰되는 과학 논증 과정에 대한 자동채점을 시도하기도 하였다. 과학적 논증은 증거를 얻고 사용하는 과정에서 대화를 활용하여 설명과 예측을 개발하고 자신의 합리적 신념체계를 구축해가는 활동을 말한다. 이 연구에서는 학생들의 말하기 자료를 활용하여 기계학습 기반 자동채점 모델을 적용하였다. 이들은 후속 연구(이만형, 유선아, 2021)에서 자동 채점의 성능을 개선하기 위해, 논증 담화를 추가하여 논증 수준과 논증 패턴을 분석하였으며, 자료, 주장, 정당화의 논증 구성을 통해 학생들의 과학적 논증 수준을 높일 수 있는 교육적 시사점을 제시하였다.

그러나 이러한 자동채점 알고리즘 및 시스템 개발 연구들은 글쓰기 혹은 과학 교과지식 중 어느 한쪽을 평가하는 데 집중하고 있어서, 교과지식 및 교과기반 논증이 통합된 문제해결과정에서 산출된 탐구보고서를 평가하는 데 제한적이다. 게다가, 이러한 알고리즘 및 시스템 개발 연구들은 접근성의 측면에서 학교 현장에서 활용하기에는 한계가 있다. 챗GPT는 인간의 언어에 친숙하게 개발되고 누구나 시도해볼 수 있다는 점에서 접근성이 높아 교사와 학생 모두 현장에서의 활용가능성이 높다는 장점을 지닌다. 이러한 챗GPT의 장점을 활용하여, 최근 박소영과 동료들(2024)의 연구에서는 챗GPT의 자동채점화 가능성을 높이기 위해 평가플랫폼을 개발하기도 하였다. 이 연구에서는 기존의 챗GPT에서 학생들의 답안을 하나씩 채점하는 한계를 극복하기 위하여, OpenAI API 호출을 통해 여러 형식의 답안 파일을 자동으로 불러들여 채점하는 플랫폼을 개발하였다. 이 연구는 챗GPT의 기반이 되는 대규모 언어모델인 GPT를 활용한 자동채점 도구 활용의 가능성을 보여주었을 뿐만 아니라 현장에서의 활용가능성을 높였다는 점에서 의미가 있다.

이처럼 자동채점 알고리즘 및 시스템 개발이 다양하게 시도되고 있는 상황에서 중요한 과제 중 하나는 자동화된 채점의 신뢰도와 타당도를 평가하는 것이다. 기존의 연구들은 기계채점의 신뢰도를 분석하고 인간채점자와의 상관을 통해 평정의 정확성을 검토하여 타당도를 검증하였다(박소영 외, 2023; 이용상 외, 2022; 이용상 외, 2023 등). 그러나 이런 방식은 구체적으로 자동화 채점 도구의 개선 방향이나 가능성을 제시하는 데는 한계가 있다. 전체 총점에 대한 인간채점자와의 일관성을 중심으로 하는 채점의 정확성에 대한 분석만이 아니라 문항별 채점 결과의 내적 혹은 구조적 특성을 분석함으로써, 채점의 타당도 역시 검토할 필요가 있다. 이 연구에서는 GPT의 범용 모델¹⁾인 GPT-4를 활용한 자동채점 자료의 특성을 인간 채점자료의 특성과 비

1) 이 연구가 수행된 기간에는 GPT-4가 최신 모델이었으나, 2024년 5월 14일 GPT-4o가 오픈됨. GPT-4는 다양한 텍스트 생성 및 이해 작업에 사용되는 범용 언어 모델인 반면, GPT-4o는 멀티모달 기능을 강화하

교·분석하였다. 이를 통해 향후 GPT를 포함한 대규모 언어모델을 자동채점에 활용하고자 할 때 고려할 사항과 채점의 정확도와 타당성을 제고하기 위해 필요한 과제를 구체적으로 규명하고자 하였다. 구체적인 연구문제는 다음과 같다:

첫째, 과학탐구역량에 대한 GPT-4와 인간 채점 결과에서 채점항목의 난이도 분포는 어떻게 다른가?

둘째, 과학탐구역량에 대한 GPT-4와 인간 채점 결과의 신뢰도(채점항목 간 내적일치도, 채점자 간 일치도)는 어떻게 다른가?

셋째, 과학탐구역량에 대한 인간 채점자료에 GPT-4 채점자료를 추가하여 다국면채점자모형에 적합시킬 때, 모수 추정치와 적합도 지수의 안정성은 어떠한가?

II. 연구 방법

1. 분석 자료

본 연구에서는 한 대학에서 운영하는 온라인 교육 프로그램에 참여한 초등학교 5학년 학생 322명(남 213명, 여 109명)의 과학 탐구활동 보고서를 분석하였다. 이 보고서들은 PDF 파일 형식으로, 총 322편이 수집되었다. 학생들은 2015년 교육과정에 따라 초등학교 고학년 과학 및 기술 가정 교과와 내용을 바탕으로 ‘치즈는 왜 맛이 다를까?’라는 주제에 대해 학습하였다. 그 과정에서 학생들은 응고, 발효, 효소 같은 과학적 개념과 치즈의 다양한 종류, 영양소, 치즈를 이용한 요리법 등 기술가정의 개념을 배웠다. 이후, 치즈의 맛과 특성에 영향을 주는 요인을 실험적으로 조작해보고, 치즈를 실제로 만들어보며 그 결과를 관찰해 보고서로 작성하였다.

이 과제는 총 세 개의 미션으로 구성되었으며, 각 미션은 1-3개의 하위 문항을 포함하고 있다. 미션 1에서 학생들은 배운 내용을 바탕으로 치즈의 특징을 직접 관찰하고 기록했으며, 기존 레시피를 사용하여 치즈를 만든 후 그 특성을 보고서에 작성하였다. 미션 2에서는 치즈의 특성을 변화시킬 수 있는 두 가지 이상의 조건을 제시하고, 이러한 조건들이 치즈에 어떤 변화를 가져올지 예측하여 작성하였다. 미션 3에서는 제시된 조건들을 이용해 실제 실험을 진행하고, 그 결과를 분석하였다(구체적인 과제의 내용과 보고서 양식은 [부록 1] 참고).

고, 전문적인 분야의 데이터를 추가로 학습하여, 특정 작업이나 도메인에 최적화된 모델임.

2. 채점 시행

본 연구에서 사용한 채점항목은 함은혜와 동료들의 연구(2022)에서 개발된 기존의 25개 항목을 기반으로 하였다. 이 채점항목들은 과학적 탐구활동의 네 가지 절차 요소(가설 생성, 실험 설계 및 수행, 자료 분석 및 해석, 결론 도출 및 평가)와 네 가지 역량 요소(과학적 지식, 논리분석적 사고, 탐구적 태도, 의사소통)에 근거하여 개발된 것이다. 챗GPT를 활용하여 소규모로 예비채점을 반복 시행한 결과를 바탕으로, 기존에 개발된 채점항목 원안을 GPT 채점에 적합한 형식으로 재진술하였으며(부록 2) 참고), 이 과정에는 교육공학, 교육평가, 교육심리 교수 및 박사급 연구원 총 3인이 참여하였다. 이는 채점항목의 의미와 초점을 GPT가 보다 잘 이해하도록 하기 위한 것으로, 구체적인 개선 사항은 다음과 같다.

첫째, 주로 영어를 사용하는 AI인 GPT의 특성을 고려하여, 채점항목을 영어 문법에 맞게 재구성하였다. 예를 들어, 기존 항목들에는 주어(학생)가 빠진 경우가 많았지만, 본 연구에서는 주어를 명확히 하여 완성도 있는 문장들로 작성하였다. 둘째, 각 채점항목이 세 가지 미션의 내용 중 어떠한 내용을 바탕으로 평가되어야 하는지 명확히 밝히고자 하였다. 한 예로, 변화된 실험 조건의 명확성을 평가하는 항목(V9)에는 ‘미션 3에서’라는 구체적인 지침을 추가하여 어떤 부분을 평가해야 하는지 분명히 했다. 셋째, 기존 항목에서 불필요하게 사용된 괄호는 제거하고, 필요한 내용은 괄호를 사용하지 않고 문장 내에 풀어서 제시하였다(예: V2, V13, V22). 넷째, 과학적 지식이나 개념의 활용을 평가하는 문항에는 필요한 예시를 구체적으로 제시하여 평가의 정확도를 높였다. 기존 항목에서는 단순히 ‘과학적 개념’의 활용 여부를 묻는 수준이었다면, 본 연구에서는 과학적 개념을 구체적인 예시(예: 맛의 종류, 세기, 강도, 냄새, 생김새, 촉감 등)와 함께 설명하도록 개선하였다.

본 연구의 GPT 채점은 박소영과 동료들(2024)이 개발한 평가플랫폼을 활용하여 수행되었다. 이 플랫폼은 OpenAI API(<https://openai.com/index/openai-api>)를 기반으로 대규모 답안에 대한 평가를 자동화하기 위하여 개발되었다. 이 플랫폼에서는 사용자가 평가영역과 세부 평가항목, 각 평가항목별 점수 범위를 설정할 수 있으며, 이에 따라 GPT-4모델(GPT-4-0125-preview)이 반복 채점을 자동으로 시행하도록 할 수 있다. 최대 40,000자(A4 30쪽 분량)의 텍스트 입력을 지원하고, 다양한 파일 형식(hwp, docx, pdf)을 처리할 수 있으며, zip파일 형태로 한번에 여러 개 파일을 업로드할 수 있도록 설계되었다(부록 3) 참고).

이 연구를 위하여 해당 플랫폼에 25개의 채점항목을 입력하고, 점수를 0점 또는 1점으로 설정한 후, 학생들의 보고서를 업로드하였다. 이 때, 플랫폼의 텍스트 제한에 따라 한 번에 25개의 보고서를 업로드하여 자동채점을 하게 하였다. 시스템 프롬프트는 “너는 [과학적 탐구역량] 측정 분야의 전문가야. 주어진 학생의 리포트를 분석하고 이를 바탕으로 학생의 [과학적 탐구역량]을 다음 세부 평가기준별로 평가해줘”였고, Zero-shot 프롬프팅²⁾을 기반으로 채점을 시행하였다. 채

점 결과는 세부 평가 항목별 점수와 평가 기준별 점수 합산 결과, 그리고 간단한 서술형 피드백이 함께 제공된다(부록 4 참조).

한편, 인간채점 과정에는 총 11명의 전문가가 참여하였다. 이들은 교육학 분야에서 10년 이상의 경험을 가진 전문가 3명, 교육학, 아동학, 과학교육 분야의 박사과정생 및 박사수료생 8명으로 구성되었다. 모든 채점자들은 교사 자격증 소지자이거나, 초등학생 및 중등학생을 대상으로 한 교육 및 연구 경험이 있으며, 초등학생 대상의 과제 평가 경험을 가진 인원들로 선정되었다. 채점자들은 1시간 이내의 채점 훈련을 이수한 후, 채점자 1인당 50-60개 내외의 탐구보고서를 평가하였다. 탐구보고서 1개당 2인의 채점자를 배정하되, 채점자 간 엄격성 추정을 위해 체계적 연계 설계(Systematic links design) 원리에 따라(Guo & Wind, 2023), 채점자 간 25명에서 50명의 공통 채점대상이 배정되도록 하였다. 총점을 기준으로 채점 결과 간에 불일치가 크게 발생한 경우에는 추가적으로 2차 채점을 실시하여 평가의 일관성을 확보하였다.

3. 자료 분석

인간채점과 GPT-4 채점의 난이도 혹은 엄격성 분포를 비교하기 위하여, 답안별 인간채점자 2인의 평균 점수와 GPT-4 채점 2회의 평균 점수를 산출하여 총점 및 채점항목별 평균 점수 분포를 비교하였다. 채점항목별 평균 점수는 각 채점항목의 난이도 즉, 어렵고 쉬운 정도를 나타내는 동시에 해당 항목에 대한 채점자들의 평균적인 엄격성을 의미한다. 평균 점수가 높을수록 피험자들이 해당 채점항목에서 1점을 획득할 확률이 높다는 것을 의미하며, 동시에 채점자들이 해당 항목에 1점을 부여할 확률이 높은 것이다. 따라서 채점항목별 점수 분포를 비교함으로써, 동일한 채점항목의 난이도 차이가 어떠한지, 인간과 GPT-4 중 어느 쪽이 더 엄격하게 혹은 관대하게 점수를 부여하는지를 비교하였다. 동일 답안에 대한 인간채점자와 GPT-4 채점 간 총점 및 채점항목별 점수 차이의 통계적 유의성 검증을 위해 대응표본 t검증을 활용하였다.

신뢰도 분석을 위해 채점항목 간 내적일관성과 채점자 및 채점시점 간 평정의 일관성을 검토하였다. 채점항목 간 내적일관성 분석에는 채점항목별 인간채점자 2인의 평균 점수와 GPT-4 채점 2회의 평균 점수를 활용하여 Cronbach's alpha를 산출하였다. 한편, 채점자 및 채점시점 간 평정의 일관성 분석을 위해 Pearson 상관계수와 Spearman 등위상관계수를 산출하였다.

마지막으로, 피험자, 채점항목, 채점자의 3개 국면을 고려한 다국면 Rasch모형 분석을 활용하여 전반적인 채점자료의 특징을 분석하였다. Rasch모형은 개별 문항에 피험자가 정답 확률을 피험자의 능력과 문항의 난이도의 함수에 대한 수리적 모형으로, 피험자 능력과 문항 난이도를 직접 비교할 수 있도록 척도화하는 데 유용하다(Andrich & Marais, 2019). 다국면 Rasch모형은 Rasch 모형의 확장으로, 피험자의 능력과 문항난이도 이외에 문항별 점수에 영향을 줄 수 있는 채점자

2) 모델에게 특정 작업을 수행하도록 요청할 때 사전 학습된 예시 없이 지시하는 방식

의 엄격성이나 과제의 유형 등을 모형에 포함하여, 모형에 포함된 조건들과는 독립적인 피험자 능력 수준을 추정한다(Linacre, 2019). 다국면 Rasch모형 분석에 인간채점자의 채점자료만을 포함할 때와 GPT-4 채점자료를 추가할 때, 문항난이도 추정치와 채점자엄격성 추정치가 얼마나 혹은 어떻게 변화하는지, 문항난이도 추정 및 채점자엄격성 추정에 대한 적합도 지수가 얼마나 혹은 어떻게 변화하는지를 검토하였다. 이를 통해 GPT-4 채점자료가 인간 채점자료와의 통합하여 활용될 수 있는지 가능성과 한계를 탐색하고자 하였다. 다국면 Rasch모형 분석에는 R의 TAM package(Robitzsch, et al., 2023)가 사용되었다.

III. 연구 결과

1. 채점항목별 난이도 및 총점 분포 비교

인간과 GPT-4가 평가한 과학탐구역량의 채점항목별 점수 및 총점 분포(평균 및 표준편차)가 <표 1>에 제시되었다. [그림 1]-a)은 인간 채점과 GPT-4 채점에서의 항목별 난이도를 산포도로 나타낸 것이다. 인간 채점항목들의 난이도(X축)는 .02부터 .91까지 고르게 분포되어 있는 반면, GPT-4 채점에서 난이도(Y축)는 .5 미만이거나 .8 이상의 범위 안에 집중적으로 분포되어 있어 차이를 보였다. 전반적으로 인간 채점과 GPT-4 채점 간 항목별 난이도의 유사성이 낮았는데, 예를 들어, ‘제시한 조건은 사전자료나 개별 추가 자료를 통해 학습한 과학적 개념이나 원리와 관련이 있는가(V13)’에서 인간 채점에서는 평균 .73으로 쉬운 항목이었던 반면, GPT-4 채점에서는 평균 .29로 어려운 항목으로 나타나(<표 1> 참고), 큰 차이를 보였다.

구체적으로 난이도 차이에 차이가 없는 채점항목, 즉 [그림 1]-a)에서 대각선 근처에 분포하는 항목은 ‘치즈의 특징을 분석하기 위하여 해당 차시 학습에서 배운 것 외에, 자신이 가진 기타 사전 지식을 활용하여 응답하는가(V2)’, ‘자신이 분석한 결과를 이해(해석)하기 위해 자료를 추가로 탐색하는가(V18)’, ‘해당 차시에서 학습한 과학적 개념을 활용하여 치즈의 특징을 구체적으로 분석하는가(V1)’, ‘과학적 개념이나 원리를 활용하여, 제시한 조건 2의 변화가 왜 혹은 어떻게 치즈의 특징을 변화시키게 될지를 예측하는가(V6)’였다.

반면, 인간 채점 결과와 GPT-4 채점 결과 간에 난이도 차이가 유의한 채점항목이 더 많았고, 특히 GPT-4의 점수가 더 높은 항목들이 많아[그림 1]-a)에서 대각선 위쪽에 분포하는 항목들, 전반적으로 GPT-4가 더 관대하게 채점하는 경향이 있음을 알 수 있다. GPT가 인간 채점자와 비교하여 훨씬 관대하게 채점한 항목들은 ‘진술이 명료하고, 문단을 구조화하는 등 독자의 가독성을 고려하는가(V25)’, ‘자신이 제시한 조건에 따라 변화된 치즈의 특성을 기본 치즈와 비교하여 기술하는가(V14)’, ‘서로 다른 조건에 따라 변화된 치즈의 특성을 비교하여 기술하는가(V15)’, ‘치

즈의 특징에 대해 기술할 때 언어적 표현이 풍부한가(V22)', '분석 결과를 본인의 가설(예상했던 결과)과 비교하여 기술하는가(V16)', '참고한 자료의 출처를 명확하게 제시하는가(V23)' 등이었다. 반면, 인간보다 GPT-4가 더 엄격하게 채점한 항목은 '제시한 조건은 사전자료나 개별 추가 자료를 통해 학습한 과학적 개념이나 원리와 관련이 있는가(V13)', '조건을 변화시키는 과정에서 계량적 접근이 관찰되는가(V12)', '과학적 원리와 개념에 대한 사전지식을 활용하여 결과를 해석하는가(V17)' 등이었다.

<표 1> 채점항목별 점수 비교: 인간 채점과 GPT-4 채점 평균 차이 분석

	채점항목(0/1)	인간 채점		GPT 채점		대응표본 t-test	
		M	SD	M	SD	t	p
V1	해당 차시에서 학습한 과학적 개념을 활용하여 치즈의 특징을 분석하는가?	0.91	0.21	0.92	0.26	-1.26	0.209
V2	치즈의 특징을 분석하기 위하여 기타 사전 지식 (해당 차시 학습 이외)을 활용하는가?	0.05	0.17	0.04	0.19	0.85	0.393
V3	치즈의 특징을 변화시킬 수 있을 것으로 제시되는 조건을 2개 이상 명료하게 제시하는가?	0.79	0.37	0.96	0.18	-7.97	0.000
V4	각 조건의 변화가 치즈의 어떤 특징을 어떻게 변화시킬지를 예상(예측)하여 진술하는가?	0.58	0.46	0.86	0.33	-9.93	0.000
V5	(조건1) “각 조건의 변화가 왜 혹은 어떻게 치즈의 특징을 변화시키게 될지를” (사전에 학습한) 과학적 개념이나 원리를 활용하여 설명하는가?	0.32	0.42	0.23	0.02	3.47	0.000
V6	(조건2) “각 조건의 변화가 왜 혹은 어떻게 치즈의 특징을 변화시키게 될지를” (사전에 학습한) 과학적 개념이나 원리를 활용하여 설명하는가?	0.29	0.40	0.25	0.40	1.56	0.122
V7	(기타 자료를 참고하거나 스스로의 사고를 통해) 고려할만한 조건을 추가적으로 탐색하였는가?	0.07	0.22	0.01	0.11	5.29	0.000
V8	“각 조건의 변화가 왜 혹은 어떻게 치즈의 특징을 변화시키게 될지를” 기타 과학적 개념이나 원리를 활용하여 설명하는가?	0.11	0.25	0.05	0.20	4.01	0.000
V9	실험을 위해 변화시킨 조건이 명확한가?	0.90	0.25	0.94	0.22	-2.81	0.005
V10	선택한 조건을 변화시킬 수 있는 방법으로 치즈 제작 과정의 조건을 변화시켰는가?	0.83	0.30	0.90	0.29	-4.32	0.000
V11	자신이 선택한 조건 이외의 조건을 고려하여 통제하는가?	0.16	0.26	0.05	0.21	5.41	0.000

<표 1> 채점항목별 점수 비교: 인간 채점과 GPT-4 채점 평균 차이 분석 (계속)

	채점항목(0/1)	인간 채점		GPT 채점		대응표본 t-test	
		M	SD	M	SD	t	p
V12	조건을 변화시키는 과정에서 계량적 접근이 관찰되는가?	0.49	0.45	0.08	0.25	14.89	0.000
V13	제시된(선정된) 조건이 앞에서(사전 학습자료 혹은 개별 추가 학습자료) 학습한 과학적 개념이나 원리와 관련이 있는가?	0.73	0.35	0.10	0.29	27.63	0.000
V14	조건 변화에 따른 결과물의 특성을 기본치즈와 비교하여 기술하는가?	0.43	0.41	0.92	0.27	-20.58	0.000
V15	서로 다른 조건 변화에 따른 결과물의 특성을 비교하여 기술하는가?	0.43	0.39	0.90	0.29	-20.54	0.000
V16	가설을 바탕으로 결과를 분석하는가?: 분석결과를 본인의 가설(예상했던 결과)과 비교하여 기술하는가?	0.09	0.23	0.42	0.44	-12.69	0.000
V17	과학적 원리와 개념에 대한 사전지식을 활용하여 결과를 해석하는가?	0.36	0.39	0.06	0.22	13.24	0.000
V18	분석 결과를 이해(해석)하기 위해 추가 자료를 탐색하는가?	0.02	0.12	0.01	0.11	0.96	0.337
V19	분석 결과에 근거하여 특정 조건의 적절성이나 유용성을 평가하는가?	0.15	0.30	0.08	0.25	4.13	0.000
V20	실험을 통해 자신이 무엇을 배웠는지 기술하는가?	0.10	0.27	0.30	0.42	-8.95	0.000
V21	본인이 수행한 실험의 강점이나 보완할 점 등을 기술하거나 향후 탐구과제를 제시하는가?	0.05	0.18	0.08	0.25	-2.15	0.032
V22	치즈의 특징(관찰 결과)에 대한 언어적 기술이 풍부한가?	0.43	0.41	0.91	0.28	-19.49	0.000
V23	참고한 자료의 출처를 제시하는가?	0.02	0.11	0.26	0.40	-11.42	0.000
V24	(V12 외) 계량적 접근이 관찰되는가?	0.07	0.19	0.04	0.17	2.78	0.006
V25	진술이 명료하고, 문단을 구조화하는 등 독자의 가독성을 고려하는가?	0.36	0.40	0.88	0.32	-21.57	0.000
전체 총점		8.74	3.86	10.28	4.30	-7.82	.000

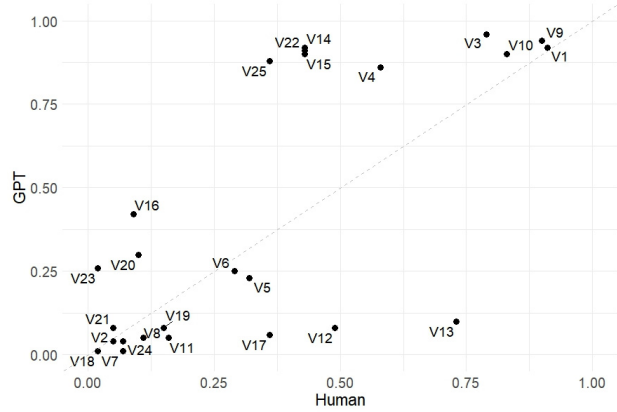
인간 채점 결과와 GPT-4 채점 결과에서 난이도가 유사한 항목들은 가설생성 단계에서의 과학 지식 활용 여부를 평가하는 항목(V1, V2, V9)과 평가 기준이 상대적으로 명확한 항목이었다(V18, V24). [그림 1]-a)를 살펴보면 이러한 항목들은 인간 채점에서 난이도가 매우 높거나 낮은 항목이라는 것을 확인할 수 있다. 반면, 난이도 차이가 크게 나타난 항목의 경우, GPT-4는 실험 설계 혹은 자료 해석 단계에서 과학 지식의 활용을 평가하는 항목(V12, V13, V17)을 더 엄격하게 채점하는 경향이 있었던 반면, 인간 채점자의 경우, 자료 해석 단계에서 결과 간 비교·분석을 평가하는 항목 (V14, V15, V16), 보고서의 전반적인 가독성과 언어적 표현(V22, V25) 등에 대해 더 엄격하게 채점하는 것으로 나타났다.

표준편차를 살펴본 결과, 인간 채점에서 편차가 큰 항목은 순서대로 V4, V12, V5, V14, V22, V25, V6 등 이었으며, GPT-4 채점에서 편차가 큰 항목은 V16, V20, V23, V6 등이었다. 편차가 큰 항목 중 인간 채점과 GPT-4 채점 간에 공통적으로 편차가 큰 항목은 ‘과학적 개념이나 원리를 활용하여, 제시한 조건 2의 변화가 왜 혹은 어떻게 치즈의 특징을 변화시키게 될지를 예측하는가(V6)’, ‘각 조건의 변화가 치즈의 어떤 특징을 어떻게 변화시킬지 예상(예측)하여 진술하는가(V4)’, ‘진술이 명료하고, 문단을 구조화하는 등 독자의 가독성을 고려하는가(V25)’ 등이었다. GPT-4 채점의 경우 표준편차 0.3 이상이 6개 항목, 인간 채점의 경우 13개 항목으로 인간 채점자 간의 편차가 크게 나타나는 항목이 더 많았다.

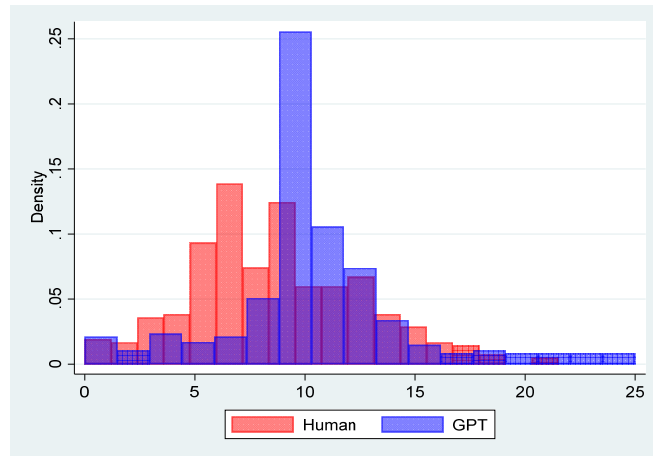
인간 채점과 GPT-4 채점 간 차이, 각 채점 간 편차를 종합적으로 고려할 때, 인간 채점과 GPT-4 채점 간 편차($t=-21.57$) 뿐만 아니라 각 채점 내 편차(인간 채점 $SD=.4$, GPT-4 채점 $SD=.32$)가 큰 항목은 ‘진술이 명료하고, 문단을 구조화하는 등 독자의 가독성을 고려하는가(V25)’ 항목이었다. 반면, 인간 채점과 GPT-4 채점 간 편차, 각 채점 간 편차가 전반적으로 작은 항목은 ‘해당 차시에서 학습한 과학적 개념(맛의 종류/세기/강도, 냄새, 생김새, 촉감 등)을 활용하여 치즈의 특징을 구체적으로 분석하는가(V1)’, ‘치즈의 특징을 분석하기 위하여 해당 차시 학습에서 배운 것 외에, 자신이 가진 기타 사전 지식을 활용하여 응답하는가(V2)’, ‘실험을 위해 변화시킨 조건을 명확하게 기술하였는가(V9)’, ‘자신이 분석한 결과를 이해(해석)하기 위해 자료를 추가로 탐색하는가(V18)’, ‘미션 3 이외의 과정에서 계량적 접근이 관찰되는가(V24)’ 등이었다.

인간 채점과 GPT-4 채점의 총점 분포는 [그림 1]-b)와 같다. 인간 채점은 왜도 0.396, 첨도 3.015로 나타났으며, GPT-4 채점은 왜도 0.724, 첨도 5.403으로 나타나, 인간채점 결과가 정규분포에 더 가까운 형태를 보였다. GPT-4 채점이 인간 채점보다 오른쪽으로 꼬리가 긴 분포를 보여 GPT-4가 더 관대하게 점수를 부여하는 경향이 관찰되었다. 한편, 인간 채점에서는 20점 이상의 최상위 점수를 받은 비율이 희소했지만, 20점 미만에서는 전체 점수대에서 종모양으로 상대적으로 고른 분포를 보였다. 반면, GPT-4 채점에서는 10점을 중심으로 점수가 크게 집중되어 있는 경향을 보였다. 즉, 전체 채점대상의 35% 이상이 10점 전후에 몰려있었다.

(a) 인간-GPT-4채점에서
채점항목별 난이도 비교



(b) 인간-GPT-4채점에서
총점 분포



[그림 1] 인간-GPT-4채점에서 채점항목별 점수와 총점 분포 비교

2. 채점항목 간 내적일관성 및 채점자 간 신뢰도 지수 비교

항목별 점수에 대한 내적일관성 정도를 알아보기 위해 평가영역별 Cronbach's alpha 계수를 검토하였다(<표 2>). 이는 인간채점자 혹은 GPT-4가 평가영역별 혹은 전체 25개 채점항목들을 피험자 평가에 얼마나 유사하게 혹은 변별되게 사용하였는지를 보여준다. 수과학지식, 논리분석적 사고, 탐구적 태도, 의사소통을 포함한 4개 항목에 대한 내적일치도 범위는 인간 채점자의 경우 0.533~0.777, GPT-4의 경우 0.695~0.864로 나타나 GPT-4가 인간 채점자보다 높은 내적일관성을 보였다. 전체 평균 역시 인간 채점자는 0.868, GPT-4는 0.929로 나타나 둘 다 채점의 내적일관성이 양호하였지만, GPT-4 채점에서 채점항목 간 일관성이 더 높았다. 이는 인간채점과 비교하여 GPT-4 채점에서 채점항목들 간의 상관이 높다는 것을 의미한다.

평가영역별로 좀 더 자세히 살펴보면, 인간 채점자와 GPT-4 채점에서 모두 논리분석적 사고 채점에 대한 내적일관성이 가장 높게 나타났다. 그 다음으로 내적일관성이 높았던 평가영역 역시 인간채점자와 GPT-4에서 모두 탐구적 태도로 동일하였다. 의사소통 항목에 대해서는 인간 채점자와 GPT-4 채점자가 Cronbach's alpha 0.6 정도 수준으로 유사한 수준을 보였다. 인간 채점자와 GPT-4 사이에 가장 차이가 크게 나타났던 평가영역은 수과학지식이었다. GPT-4 채점자는 0.723으로 네 개의 항목 중 세 번째로 높은 수준을, 인간 채점자는 0.533으로 네 개의 항목 중 가장 낮은 수준의 내적일관성을 보였다. 인간 채점자와 GPT-4 모두 논리분석적 사고, 탐구적 태도 순으로 가장 높은 내적일치도를 보였는데 두 항목 모두 과학적 영역에서의 체계적이고 확장적인 생각의 기술, 즉 상대적으로 고차원적인 사고과정을 다룬다는 점에서 공통점이 있다. 인간 채점자와 GPT-4 모두 고차원적인 사고과정의 산출물을 평가하는 것에 대해 채점 일관성이 높다고 볼 여지가 있는데, 이는 나머지 두 영역이 정보의 심층적인 처리과정을 다루기보다는 비교적 표면적인 수준의 지식 기재 여부(수과학지식)와 보고서 작성 기술 수준(의사소통)을 평가한다는 점에서 그러하다. 하지만 논리분석적 사고와 탐구적 태도의 채점항목 수가 모두 다른 두 영역보다 많다는 점도 유의해서 보아야 한다. 특히 논리분석적 사고는 10개의 항목으로 다른 항목보다 월등히 많은 항목을 가지고 있어서 이렇게 많은 항목의 수가 채점항목 간 내적일치도를 높이는 데 기여했을 가능성도 있다.

정리하면, 인간 채점과 GPT-4 채점에서 채점항목들 간 내적일치도는 전반적으로 양호하게 나타나, 인간과 GPT-4 채점 모두 주어진 채점항목을 과학탐구역량이라는 능력 차원을 평가하는데 일관되게 사용하였다고 볼 수 있다. 그러나 수과학지식 채점에 대한 인간 채점의 내적일관성 지수는 GPT-4와 달리 0.6 미만으로 다소 낮은 수준을 보였다.

<표 2> 평가영역별 채점항목 간 내적일치도(Cronbach's alpha)

영역	채점항목 수	인간	GPT-4
수과학지식	4	0.533	0.723
논리분석적 사고	10	0.777	0.864
탐구적 태도	6	0.708	0.808
의사소통	5	0.611	0.695
전체	25	0.868	0.929

다음으로 25개 모든 문항에 대하여 채점자 간, GPT-4 채점 시점 간 상관계수를 살펴본 결과가 <표 3>에 제시되었다. Spearman 등위상관계수를 기준으로, 인간 채점자 간 평균 상관계수는 .785~.802, GPT-4 채점 시점 간 상관계수는 .834~.906으로, GPT-4 채점 시점 간 상관($p=.906$)이

더 높은 경향을 보였다. 특히 인간 채점자 간 상관의 경우, 채점대상에 배정된 채점자 쌍(pair)에 따라 상관이 다르게 나타나 상관계수의 범위가 .628부터 .955까지 매우 넓게 나타났다. 즉, 채점자 A와 B의 점수 간 상관은 .628로 상대적으로 낮은 반면, 채점자 C와 D의 점수 간 상관은 .955로 매우 높게 관찰되어, 채점자 쌍에 따라 채점자 간 일치도에 편차가 있었다.

다음으로, 인간 채점자 11명 각각의 점수와 GPT-4로 채점한 두 번의 점수 간 상관을 각각 도출한 결과, Pearson 상관계수를 기준으로 평균 .618, Spearman 등위상관계수를 기준으로 평균 .588로 인간 채점자 간 일치도나 GPT-4 채점 간 일치도와 비교하여 낮게 나타났다. 한 개인 채점자와 한 GPT-4 시점의 채점 간 상관은 이처럼 평균적으로 낮을 뿐만 아니라, 편차도 크게 나타났다. 상관이 높을 때에는 .802~.852였지만, 낮은 경우에는 .375~.391로 나타나 GPT-4 채점과의 일치도에서도 개인 간 편차가 큰 것으로 보인다. 마지막으로, 인간 채점자 11명의 평균과 GPT-4 두 번의 채점 평균 간의 상관은 .630~.639로 양호한 수준을 보였다.

<표 3> 채점자 간, 채점 시점 간 일치도

구분	Pearson 상관계수	Spearman 등위상관계수
인간 채점자 간	.802 (.635~.938)	.785 (.628~.955)
GPT-4 채점 시점 간	.906	.834
인간(11명, 개인)-GPT-4 간	.618 (.391~.802)	.588 (.375~.852)
인간 평균-GPT-4 평균 간	.630	.639

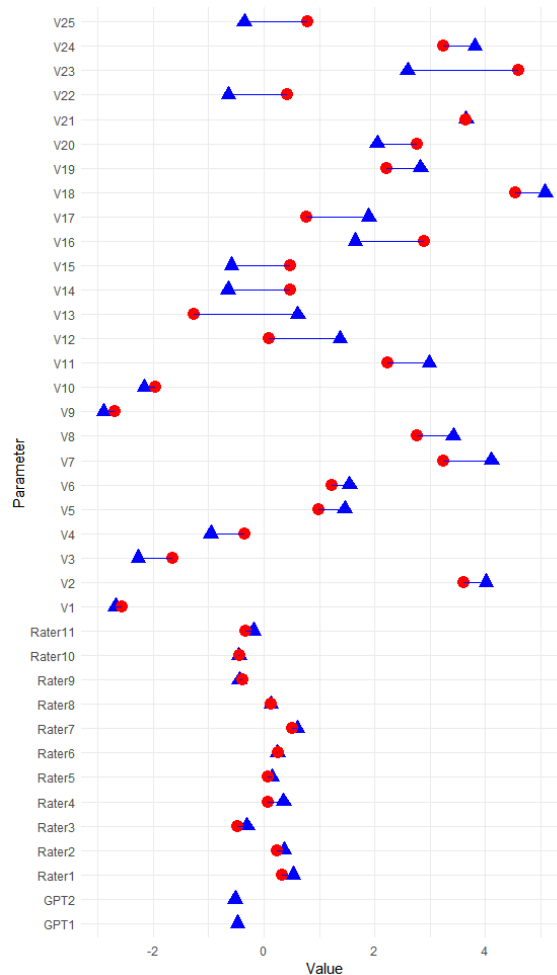
3. 다국면 Rasch 모형 분석

인간 채점 자료와 GPT-4 채점 자료의 비교가능성을 탐색하기 위하여 다국면 Rasch 모형 분석을 활용하였다. <표 4>는 다국면 Rasch 모형의 적합도 지수로, M1은 인간 채점만 포함한 모형이고, M2는 인간 채점과 GPT채점을 통합한 모형이다. 인간과 GPT-4 채점자료를 분석한 모형(M2)은 인간 채점만 분석한 모형(M1)과 비교하여 전반적인 모형적합도 지수가 크게 손상되는 것을 확인할 수 있다. 또한, GPT-4 채점을 포함하는 경우, 인간 채점자료만 포함하는 모형과 비교하여 모형적합도가 통계적으로 유의하게 감소하였다($\Delta X^2=10899.23(2)$, $p<.001$).

<표 4> 다국면 라쉬모형 모형 적합도: 인간 채점자만 포함한 경우와 GPT 채점을 포함한 경우

	Log likelihood	Deviance	#par	AIC	BIC
M1. 인간 채점만 포함	-6710.39	13438.78	36	13510.78	13649.67
M2. GPT-4 채점 포함	-12169.01	24338.01	38	24414.01	24560.61

다음으로 M1과 M2에서 채점자 엄격성 및 문항 난이도 추정치가 어떻게 변화하는지를 검토하였으며, 그 결과를 [그림 2]에 제시하였다. 채점자 엄격성 추정치([그림 2] 하단 Rater1 부터 Rater11)를 먼저 살펴보면, M1에서 인간 채점자 11인의 엄격성 추정치는 최소 -0.47에서 최대 0.51까지로 나타났으며, M2에서 GPT-4 2회 채점의 엄격성 추정치는 각각 -0.52, -0.48로 인간 채점자와 비교하여 상당히 낮게 나타나, GPT-4가 인간채점자와 비교하여 학생들의 수행을 관대하게 평가한 것을 알 수 있으며, 이는 앞서 채점항목별 점수 분포(<표 1>)에서도 관찰된 바 있다. M1과 M2 간 채점자 엄격성 추정치의 차이를 살펴보면, GPT-4 채점 결과를 포함할 때 인간 채점자 엄격성 추정치가 최소 -0.45에서 최대 0.60으로 약간 변화하였지만, 개별 채점자의 엄격성



※ 빨간동그라미: 인간채점자만 포함한 모형(M1), 파랑세모: GPT-4채점을 포함한 모형(M2)

[그림 2] 채점항목 난이도 및 채점자 엄격성 모수 변화

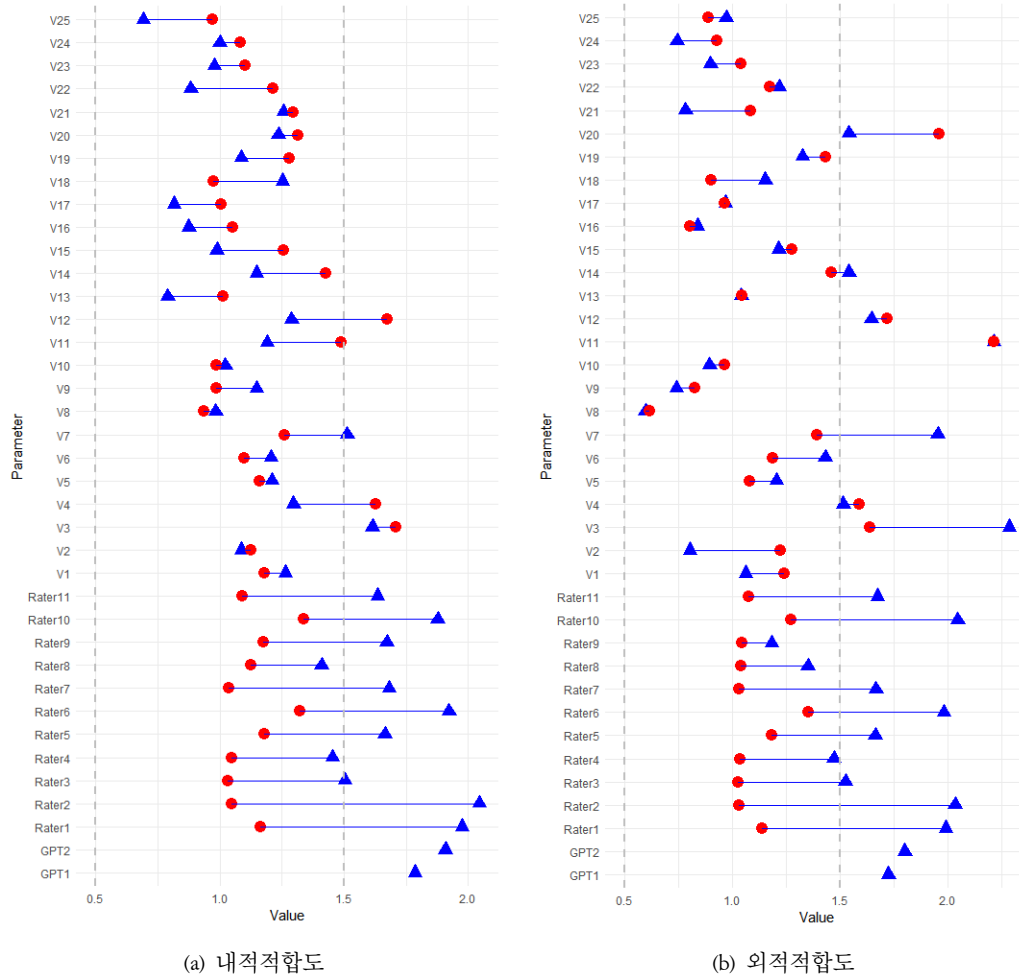
추정치의 변동이 눈에 띄게 관찰되지는 않았다.

이와 달리, 채점항목별 난이도 추정치는 M1과 M2 간 변동이 크게 나타났다(그림 2] 하단 V1부터 V25). 즉, GPT-4채점을 추가하는 경우, 채점항목별 난이도 추정치가 크게 변화하였다. 변동의 크기와 방향은 채점항목의 내용적 특성에 따라 차이를 보였다. 전반적으로 하위역량 중 ‘과학적 지식’과 ‘의사소통’을 측정하는 채점항목에서의 변동이 크게 나타난 반면, ‘탐구적 태도’를 측정하는 채점항목은 GPT-4 채점 추가에 따른 난이도 모수 변화가 적은 경향을 보였다. 한편, ‘논리분석적 사고’역량을 측정하는 채점항목은 과학탐구 절차에 따라 다른 양상을 보였는데, 자료분석 및 해석 절차와 관련된 채점항목(V14, V15, V16)은 모두 추정치의 변화가 크게 나타난 반면, 가설 생성(V3, V4, V5, V6), 실험설계 및 해석(V9, V10) 절차에 해당하는 채점항목은 상대적으로 변동이 적었다.

M1과 M2 간의 난이도 추정치의 변화 방향을 살펴보면, 인간 채점자와 GPT-4 채점의 점수 분포 차이를 그대로 반영하는 경향을 보였다. 구체적으로 M1 (인간채점 모형)과 비교하여 M2 (인간채점+GPT-4채점 모형)에서 난이도 추정치가 높아지는 항목들은, 공통적으로 인간 채점의 평균이 높았으며, 난이도 추정치가 낮아지는 항목들은 GPT-4 채점 평균이 높게 나타난 항목들이었다. 예를 들어, V25의 경우 인간 평균은 0.36점, GPT 평균은 0.88로, 인간에 비해 GPT-4가 관대하게 채점하였으며, 이에 따라 M2에서 난이도 추정치가 낮아졌다. 반면, V17은 인간 평균 0.36점, GPT-4 평균 0.06점으로 GPT가 인간에 비해 엄격하게 채점하였고, 이에 따라 GPT 채점을 포함하는 경우 난이도 추정치는 높아졌다.

한편, 채점자 엄격성 및 채점항목 난이도 추정치의 적합도를 살펴본 결과는 [그림 3]과 같다. M1에서 인간 채점자 11인의 엄격성 추정치(Rater1부터 Rater11)의 내적적합도는 최소 0.99부터 최대 1.34, 외적적합도는 최소 0.99부터 최대 1.36까지로 양호하게 나타났다. 그러나, M2에서 GPT-4 채점을 포함하면, GPT-4 2회 채점 추정치의 적합도는 내적적합도와 외적적합도 모두에서 1.7이상으로 양호도 판단 기준(0.5이상 1.5미만)을 벗어나 있었다. 인간 채점자의 엄격성 추정에서도 내적적합도가 최소 1.42에서 최대 2.07, 외적적합도가 최소 1.70에서 최대 2.06으로 그 값이 크게 증가하여, 적합도가 대부분 크게 손상되었다.

상대적으로 채점항목별 난이도 추정치의 적합도는 GPT-4 채점 포함 여부에 따라 변화가 크지 않았다. 다만, 두 적합도 지표에서 변화 정도가 다소 다르게 관찰되었다. 내적적합도의 경우, GPT-4를 추가한 후에도 내적적합도가 여전히 적절한 수준으로 유지되는 경향을 보였다. 내적적합도가 1.5 바깥에 있던 V4, V12 문항은 GPT-4 채점 추가 후 오히려 내적적합도가 적절한 범위(0.5이상 1.5 미만)로 변화하였으며, V3 문항은 여전히 1.5 바깥의 값이긴 하나 적절한 범위에 가깝게 이동하는 모습이 나타났다. 반면, 외적적합도의 경우, 대부분의 문항에서 외적적합도 변화가 작은 편이었으나, 일부 문항에서는 적합도 변동이 크게 관찰되었다. 구체적으로 V3, V7 문항은 GPT-4 채점 추가 후 외적적합도가 상당히 손상되는 결과가 나타나, GPT-4 채점이 Rasch모형



※ 빨강동그라미: 인간채점자만 포함한 모형(M1), 파랑세모: GPT-4채점을 포함한 모형(M2)

[그림 3] 채점항목 난이도 및 채점자 엄격성 적합도 지수 변화

에서 예측되는 결과와 매우 다른 채점 결과를 보인 것으로 해석되는 반면, V20 문항은 오히려 GPT-4 채점 추가 후에 외적적합도가 적절한 수준에 가까워지기도 하였다.

한편, 학생 능력 추정의 안정성(precision)은 인간 채점만 활용하였을 때는 .897이었으나, GPT-4 채점을 추가한 후에는 .944로 향상되어, GPT-4 채점 결과를 추가함으로써 추정값의 안정성이 상대적으로 증가하였다. GPT-4가 점수를 부여하는 양상이 인간과 다름에도 불구하고, 학생 능력 추정의 안정성이 향상된 것은, 개별 학생에 대한 평가에 배정된 채점자가 2인(인간)에서 4인(인간 2인과 GPT-4 2회)로 증가한 효과이다.

IV. 논의 및 결론

이 연구는 OpenAI API를 활용한 GPT-4 자동채점시스템을 활용하여 학생들의 과학탐구보고서를 평가한 채점자료와 인간 채점자료를 비교함으로써, 교실기반 수행평가에 GPT-4 자동채점을 활용하는 것의 가능성과 한계를 구체적으로 탐색하기 위하여 수행되었다. 이를 통해 향후 GPT-4를 포함한 대규모 언어모델을 자동채점에 활용하고자 할 때 고려할 사항과 채점의 정확도와 타당성을 제고하기 위해 필요한 과제를 구체적으로 규명하고자 하였다. 주요 연구결과와 논의사항을 정리하면 다음과 같다.

첫째, GPT-4가 인간채점자보다 학생들의 과학탐구역량을 전반적으로 관대하게 평가했으며, 특히, 인간이 엄격하게 채점했던 항목에 대해서 GPT-4가 관대하게 채점하는 경향이 뚜렷하게 관찰되었다. 구체적으로, 전반적인 보고서의 가독성과 언어적 표현(V22, V23, V25), 가설 생성 및 자료 분석 단계에서의 결과에 대한 예측·비교·종합(V3, V4, V14, V15, V16), 자신의 탐구과정에 대한 성찰과 평가(V20, V21)에서 GPT-4 채점이 인간 채점보다 관대하게 나타났다. 반면, 인간 채점자보다 GPT-4가 엄격하게 채점한 항목은 실험 수행 및 자료 분석 단계에서의 수과학적 지식 활용을 평가하는 항목들(V12, V13, V17)이었다. 인간과 GPT-4 채점의 난이도가 유사했던 항목은 가설 생성 단계에서 과학적 지식의 활용 여부에 대한 평가(V1, V2, V6, V9 등)였으며, 해당 채점항목에서 인간과 GPT-4 간의 편차, 인간 채점자 간 편차, GPT-4 채점시점 간 편차가 가장 작은 경향을 보였다.

둘째, 과학탐구역량에 대한 인간 채점과 비교하여, GPT-4 채점은 중간 수준의 수행을 적절하게 변별하지 못하는 것으로 나타났다. 인간 채점에서는 25개 채점항목의 난이도가 상당히 고르게 분포되어 있는 반면, GPT-4 채점에서는 중간 수준의 난이도를 가진 채점항목이 존재하지 않았다. 총점 분포에서도, 인간 채점에서는 최상위 점수를 제외하고 전체 점수대에 걸쳐 종모양의 비교적 고른 분포를 보인 반면, GPT-4 채점에서는 10점 전후로 30% 이상이 몰려있었다. 이러한 경향은 GPT-4가 주어진 채점항목을 활용하여 중위권 학생들의 수행을 변별하는 데 매우 제한적이었음을 시사한다. 또한, GPT-4의 채점 결과와 비교하여 인간 채점결과가 정규분포에 더 가까운 형태를 보였다. 인간 채점자는 채점기준에 비추어 여러 학생들의 수행을 비교하면서 채점을 진행한 반면, GPT-4는 채점기준에 비추어 학생들의 수행을 독립적으로 채점했기 때문에, 인간 채점자가 학생들의 수행을 상대적으로 평가하는 데 유리했을 가능성을 시사한다.

셋째, 채점항목의 내적일관성을 살펴본 결과, 수과학지식, 논리분석적 사고, 탐구적 태도, 의사소통 모든 평가영역에서 인간 채점자료에서 보다 GPT-4 채점자료에서 내적일관성이 높게 나타났다. 이는 인간 채점자가 GPT-4와 비교하여 채점항목들을 다소 변별되게 활용하는 반면, GPT-4는 채점항목들을 유사하게 활용하는 경향이 있음을 의미한다. 이론적으로 채점항목의 내적일관성은 채점항목들이 평가하고자 하는 잠재특성의 일차원성(unidimensionality)을 전제하기 때문에

(Raykov & Marcoulides, 2019), 내적일관성 지수가 높을수록 여러 채점항목들이 측정하는 능력이 단일한 차원의 능력으로 수렴된다고 본다. 이러한 관점에서 GPT-4는 전체 혹은 평가영역별 채점항목들이 단일 차원의 능력을 대표한다는 가정에 충실하게 학생들의 수행을 평가한 것이다. 게다가 내적일관성 지수는 평가도구의 신뢰도를 대표하는 것이 아니라, 신뢰도 추정의 한 가지 방법일 뿐이라는 점도 유념할 필요가 있다(Sijtsma, 2009; Raykov & Marcoulides, 2019)

그러나 다수의 채점항목들이 평가하는 능력이 다차원적일 가능성이 제기될 때 혹은 능력의 다차원성(multidimensionality)을 탐색하고자 할 때, GPT-4 채점이 평가의 타당화 과정을 왜곡할 수 있다. 인간 채점에서 나타나는 채점항목들 간의 낮은 일관성은 그 기저에 기능하는 또 다른 능력의 차원 혹은 채점항목들 간의 변별되는 속성을 시사하며 추가 탐색의 단서를 제공한다. 반면, GPT-4 채점에서의 높은 일관성은 채점항목들 간의 일관성을 확인해준다. 따라서, 평가하고자 하는 능력의 단일차원성이 이론적 혹은 경험적으로 확실히 전제될 때, GPT-4 채점의 일관성은 매우 유용한 특성이 된다. 반면, 평가하고자 하는 능력의 단일차원성에 대한 이론적 혹은 경험적 증거가 명확하지 않다면, GPT-4의 내적일관성은 오히려 평가의 타당화를 위협하는 요인이 될 수 있다.

넷째, 인간 채점자 간, GPT-4 채점 시점 간 총점의 일치도를 살펴본 결과, GPT-4 채점 시점 간 일치도가 인간 채점자 간 일치도보다 평균적으로 더 높았다. 채점항목별 점수 분포에서도, 인간 채점자 간의 편차보다 GPT-4 채점 간의 편차가 적은 경향을 보였다. 특히, 언어적 표현의 풍부성, 가독성을 포함한 의사소통 항목들의 경우, 인간 채점자 간의 편차는 크고, GPT-4 채점 시점 간 편차는 적었다. 다만, GPT-4는 동일한 채점자가 2회 반복 채점한 것인 반면, 인간 채점자는 두 명의 다른 채점자가 독립적으로 채점하였다는 점을 고려할 필요가 있다. 그럼에도 불구하고, 인간 채점자는 300여 개 답안을 한번에 채점하기 어렵기 때문에, 채점자-피험자 설계를 통해 답안 혹은 보고서를 여러 명의 채점자에게 적절한 분량으로 배정하여 채점을 시행하는 것이 불가피하다는 점을 고려할 때, 피로감 없이 여러 학생의 보고서를 빠르고 일관되게 평가하는 GPT-4 채점은 매우 유용할 수 있다. 한편, 향후 인간 채점자와 GPT-4 채점의 안정성을 좀 더 정밀하게 비교하기 위해서는 동일한 인간 채점자가 서로 다른 시점에서 반복 채점을 한 결과를 GPT의 다른 시점에서의 채점과 비교해 볼 필요가 있다.

주목할 점은 인간 채점자의 경우, 짝(pair)이 되는 채점자 구성에 따라서 일치도의 편차가 크게 나타났는데, 일부 채점자 짝의 경우, 채점자 간 점수 상관이 .9 이상으로 GPT-4 채점 시점 간 상관보다 높았다. 이는 채점자 간 채점항목에 대한 이해가 잘 공유되고, 채점대상 보고서의 질이 고르게 분포될 경우, 인간 채점에서도 상당히 높은 수준의 채점자 간 신뢰도를 확보할 수 있음을 시사한다. 반대로, 채점자 간 채점항목에 대한 이해가 잘 공유되지 못한 상태로, 채점대상 보고서 또한 특정한 수준에 몰려있는 경우, 인간 채점의 안정성이 GPT-4보다 낮을 수 있다. 따라서 자동채점의 성능을 평가하고 개선하고자 할 때, 인간 채점 결과를 절대적인 정확도의 기준

으로 삼기보다, 인간 채점자 간 일치도가 확보된 평가기준이나 영역 혹은 개선이 필요한 평가기준이나 영역을 규명하여, 자동채점의 학습 과정을 세부적으로 조정할 필요가 있다. 이와 관련하여 국립국어원에서 수행한 민병곤과 동료들(2023)의 연구에서도 채점자 교육을 통한 인간채점자 간 신뢰도 향상이 중요함을 언급한 바 있다.

다섯째, 평가영역별로 살펴보면, 인간 채점자는 수과학지식보다는 의사소통 항목에서 조금 더 높은 내적일치도를, GPT-4는 의사소통보다는 수과학지식 항목에서 조금 더 높은 내적일치도를 보였다. 언어적 표현의 풍부성, 표와 그림을 활용한 가독성, 출처 표기 등을 통해 보고서 전반에서 보이는 형식이 인간 채점자들에게는 조금 더 일관성 있게 판단된 것으로 보인다. 의사소통 항목이 보고서 형식에서 기술적인 부분들을 많이 다루기 때문에 GPT-4의 내적일치도가 더 높은 것으로 예상한 결과와 다소 다른 부분이기도 하다. 특히, 출처 표기 관련 항목에서, 인간 채점자보다 GPT-4 채점 간 불일치가 크게 나타난 점은 주목할 만하다. 향후 형식, 표현 등 평가에 대한 GPT-4의 신뢰도는 후속 연구에서 더 살펴볼 필요가 있다.

여섯째, 다국면채점자모형 분석 결과, 개별 채점자의 채점 기준과 GPT-4의 채점 기준은 다소 다르게 기능하고 있을 가능성을 보여 GPT-4 채점이 개별 채점자의 채점 결과를 완전히 대체하기에는 제한이 있다. 인간 채점에 GPT-4 채점 결과를 포함(M2)하여 다국면 Rasch 모형을 분석하고 이를 인간 채점만 활용한 경우(M1)와 비교한 결과를 살펴보았을 때, GPT-4 결과를 포함하였을 때 전반적인 모형적합도가 손상되는 것이 확인되었다. 또한 항목의 난이도 추정치 역시 변동이 크게 나타나고 있어 GPT-4가 인간과는 유사한 기준으로 채점을 한다고 판단하기에는 다소 어려웠다. 하지만 채점문항의 적합도에서는 GPT-4 결과를 추가하였을 때 오히려 적합도가 양호한 값을 보여주거나 큰 변동이 없는 문항들도 존재하고 있다는 점에서, GPT 채점이 잘 기능하는 채점 대상, 채점 문항 및 진술의 특징에 대한 연구가 뒷받침된다면 일부 문항에서는 GPT-4 채점을 보조적으로 적용해볼 수 있는 방안을 모색할 수 있을 것이다.

종합하면, GPT-4 채점에서 채점항목 활용의 일관성 및 채점 시점 간 판단의 일치도가 양호하고, 일부 채점기준에서 인간 채점자와의 판단의 일관성이 확보되었으며, 빠르고 효율적이라는 점에서, 향후 인간 채점의 보조자로서 GPT-4의 역할을 기대할 만하다. 또한, GPT-4 채점자료가 인간 채점자의 판단을 검토·평가하는 기회를 제공한다는 점을 고려하여, 인간 채점자료와의 연계 및 활용 가능성을 구체적으로 탐색할 필요가 있다. 그러나, GPT-4는 학생 수행 중 특히 중간 수준의 다양한 질적 차이를 적절하게 변별하지 못했고, 다수의 평가영역과 채점항목에서 학생 수행 수준에 대한 판단준거(standards)가 인간 채점자들과 다르게 나타나, 이를 개선하기 위한 연구와 실험이 요청된다.

연구의 제한점과 향후 연구 과제는 다음과 같다.

첫째, 이 연구에서는 GPT-4 채점자료의 구조적 특성에 초점을 맞추어 인간 채점자와의 유사성을 분석함으로써, GPT-4가 개별 채점항목을 어떻게 활용했는지를 구체적으로 검토하였다. 향

후 연구에서는 인간-GPT 간의 평가 결과의 불일치가 큰 사례 혹은 불일치가 적은 사례를 살펴봄으로써 GPT 채점을 개선하거나 활용도를 높일 수 있는 방안을 탐색할 필요가 있다.

둘째, 이 연구에서는 GPT-4 채점에 Zero-shot 프롬프팅을 활용하였다. GPT 성능에 대한 선행연구들에서 프롬프트의 방식과 내용이 GPT 응답의 질에 영향을 준다고 알려져 있기 때문에(한진영, 이민정, 2024; Li et al., 2024), 향후 프롬프트에 채점 사례를 추가하는 Few-shot 프롬프팅³⁾ 혹은 Chain-of-thought 프롬프팅⁴⁾ 등 다양한 프롬프팅 전략의 활용이 평가 결과를 얼마나 개선시키는지 조사할 필요가 있다.

3) 여러 개의 예시를 제공하여 모델이 작업을 이해하고 수행할 수 있도록 돕는 방법

4) 모델이 복잡한 문제를 해결할 때, 단계별로 사고 과정을 설명하도록 유도하는 방법

참고문헌

- 관계부처 합동 (2022.08.22.). 디지털 인재양성 종합방안.
(Translated in English) Consortium of Ministries(2022.08.22.). *A comprehensive plan for digital human resources*.
- 교육부 (2015). 2015 개정 교육과정.
(Translated in English) Ministry of Education (2015). *2015 revised national curriculum*.
- 교육부 (2022.12.21.). 2022 개정 초·중등학교 및 특수교육 교육과정 확정·발표. 교육부
(Translated in English) Ministry of Education(2022.12.21.). *Announcement of 2022 revised national curriculum for elementary, secondary, and special education*.
- 교육부 (2023.02.23). 디지털 기반 교육혁신 방안
(Translated in English) Ministry of Education(2023.02.23.). *Digital based educational reform*.
- 교육부 (2024.01.24.). 2024년 주요정책 추진 계획.
(Translated in English) Ministry of Education(2024.01.24.). *Implementation plan for 2024 educational policy*.
- 김덕영, 박종원 (2015). 학생의 열린 과학 탐구 보고서 작성을 돕기 위한 점검표 개발. 한국과학교육학 회지, 36(6), 1075-1083.
(Translated in English) Kim, D., Park, J. (2015). Development of a checklist for helping students' open scientific inquiry report writing. *Journal of the Korean Association for Science Education*, 35(6), 1075-1083.
- 김현정, 김성기 (2021). 탐구보고서에 기반한 화학교사의 과학 역량 평가 실태 분석. 대한화학회지, 65(3), 209-218.
(Translated in English) Kim, H., & Kim, S. (2021). Analysis on actual condition of chemistry teachers' scientific competency assessment based on inquiry report. *Journal of the Korean Chemical Society*, 65(3), 209-218.
- 노은희, 심재호, 김명화, 김재훈 (2012). 대규모 평가를 위한 서답형 문항 자동채점 방안 연구. 한국교육과정평가원.
(Translated in English) Noh, E., Shim, J., Kim, M., & Kim, J. (2012). *Research on automatic scoring methods for short-answer questions in large-scale assessments*. KICE.
- 노은희, 이상하, 임은영, 성경희, 박소영 (2014). 한국어 서답형 문항 자동채점 프로그램 개발 및 실용성 검증. 한국교육과정평가원.
(Translated in English) Noh, E., Lee, S., Lim, E., Sung, K., & Park, S. (2014). *Development and validation of a Korean short-answer question automatic scoring program*. KICE.
- 민병곤, 남가영, 김선희, 장성민, 이성준, 권은선 (2023). 2023년 국민의 글쓰기 능력 진단 체계 개발. 국립국어원.
(Translated in English) Min, B., Nam, K., Kim, S., Jang, S., Lee, S., & Kwon, E. (2023). *Development of Korean writing assessment system 2023*. National Korean Agency.
- 박소영, 이병윤, 홍유정 (2024). ChatGPT를 활용한 AI 글쓰기 의사소통 역량 평가 도구 개발에 대한 연구: 기술전문가와의 상호소통을 중심으로. 실천공학교육논문지, 16(1), 1-11.
(Translated in English) Park, S., Lee, B., Hong, Y. (2023). An exploratory study on developing the AI essay test tool based on ChatGPT: Focusing on the interaction with the engineer. *Journal of Practical Engineering Education*, 16(1), 1-11.

- 박소영, 이병윤, 함은혜, 이유경, 이성혜 (2023). ChatGPT-4의 과학적 탐구 역량 평가 가능성 탐색: 인간평가자와의 비교를 중심으로. *교육학연구*, 61(4), 299-332.
- (Translated in English) Park, S., Lee, B., Ham, E. H., Lee, Y., & 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.
- 박찬술, 손정우 (2020). 탐구적 과학 글쓰기를 통한 데이터 기반 과학 탐구학습이 초등학교 학생의 과학과 핵심역량에 미치는 영향. *교사교육연구*, 59(2), 245-258.
- (Translated in English) Park, C., & Son, J. (2020). The effects of data-based scientific inquiry linked with science writing heuristic(SWH) on elementary school students' science core competencies. *Teacher Education Research*, 59(2), 245-258.
- 박혜영, 김성숙, 김경희, 이명진, 김광규, 김지영 (2019). 수업-평가 연계 강화를 통한 서·논술형 평가 내실화 방안. 한국교육과정평가원.
- (Translated in English) Park, H., Kim, S., Kim, K., Lee, M., Kim, K., & Kim, J. (2019). *Substantializing methods of restricted and extended response essay assessment through enforcing the instruction-assessment alignment*. KICE.
- 박혜영, 이명애, 이명진, 김부연, 임해미, 이현숙, 이동엽 (2018). 미래사회 대비 교육과정, 교수학습, 교육평가 비전 연구(III): 초·중등학교의 교육평가 방향을 중심으로. 한국교육과정평가원.
- (Translated in English) Park, H., Lee, M., Lee, M., Kim, B., Yim, H., Lee, H., & Lee, D. (2018). *Education vision for the future curriculum, instruction, and evaluation in South Korea (III): New directions for Korean elementary and secondary school assessment*. KICE.
- 백순근 (2000). 수행평가의 원리. 교육과학사.
- (Translated in English) Baek, S. (2000). *Principles of performance assessment*. Kyoyookbook.
- 백유진 (2020). 논술문 채점에 나타난 국어 교사의 채점 편향의 특성 분석: 텍스트 특징에 따른 채점 편향 분석을 중심으로. *청람어문교육*, 76, 67-101.
- (Translated in English) Baek, Y. (2020). An analysis on the characteristics of rating bias of Korean language teachers in rating persuasive writing. *Journal of CheongRam Korean Language Education*, 76, 67-101.
- 이만형, 유선아 (2020). 전문가의 형태소 분류를 활용한 과학 자동 채점. *한국과학교육학회지*, 40(3), 321-336.
- (Translated in English) Lee, M., & Ryu, S. (2020). Automated scoring of scientific argumentation using expert morpheme classification approaches. *Journal of the Korean Association for Science Education*, 40(3), 321-336.
- 이만형, 유선아 (2021). 기계 학습을 활용한 논증 수준 자동 채점 및 논증 패턴 분석. *한국과학교육학회지*, 41(3), 203-220.
- (Translated in English) Lee, M., & Ryu, S. (2021). Automated scoring of argumentation levels and analysis of argumentation patterns using machine learning. *Journal of the Korean Association for Science Education*, 41(3), 203-220.
- 이용상, 신동광, 김현정 (2022). 한국어 쓰기 평가를 위한 자동채점의 가능성 탐색. *이중언어학*, 86, 171-191.
- (Translated in English) Lee, Y., Shin, D., & Kim, H. (2022). Applying an automated essay scoring to a writing test of Korean language. *Bilingual Research*, 86, 171-191.
- 이용상, 최윤석, 이승현 (2023). 한국어 논술답안 자동채점 프로그램 PASTA- I 개발. *교육평가연구*,

- 36(4), 711-730.
- (Translated in English) Lee, Y., Choi, Y., & Lee, S. (2023). The automated scoring program for Korean essays, PASTA-I. *Journal of Educational Evaluation*, 36(4), 711-730.
- 지은림 (2008). 논술고사의 신뢰성에 영향을 미치는 채점자 특성 분석. *교육평가연구*, 21(2), 97-113.
- (Translated in English) Chi, E. (2008). Analyzing rater characteristics affecting the reliability of essay examinations. *Journal of Educational Evaluation*, 21(2), 97-113.
- 진경애, 남명호, 김명화, 오상철, 김민정, 주형미 (2006). 서답형 문항 자동채점 프로그램 도입 방안 연구(I). 한국교육과정평가원.
- (Translated in English) Jin, K., Nam, M., Kim, M., Oh, S., Kim, M. J., Joo, H. (2006). *Study on the introduction of short answer question automatic scoring program*. KICE.
- 하민수, 이경진, 신세인, 이준기, 최성철, 주재걸, 박지선 (2019). 학습지원 도구로서의 서술형 평가 그리고 인공지능의 활용: WA3I 프로젝트 사례. *현장과학교육*, 13(3), 271-282.
- (Translated in English) Ha, M., Lee, K., Shin, S., Lee, J., Choi, S., Choo, J., Park, J. (2019). Assessment as a learning tool and utilization of artificial intelligence: WA3I project case. *Practical Science Education*, 13(3), 271-282.
- 한진영, 이민정 (2024). 문제중심학습과 챗 GPT: 프롬프트와 문제해결력에 대한 탐색. *교양학연구*, 26, 111-145.
- (Translated in English) Han, J., & Lee, M. (2024). Problem-based learning and ChatGPT: Explorative analysis of the relationship between chatgpt prompts and problem-solving skills. *The Journal of General Education*, 26, 111-145.
- 함은혜, 유예림 (2022). 텍스트 마이닝 기법을 활용한 대학생 세계이해 논술형 평가 답안의 수행 수준별 특성 분석. *교육평가연구*, 35(4), 687-717.
- (Translated in English) Ham, E. H., & Yu, Y. (2022). Text-mining analyses of undergraduates' essays on global perspectives by performance levels in Korea. *Journal of Educational Evaluation*, 35(4), 687-717.
- 함은혜, 이유경, 박소영, 박혜진, 이성혜 (2022). 초등학교 과학 탐구과제 수행 특성 분석 및 채점기준 개발. *한국과학교육학회지*, 42(2), 239-252.
- (Translated in English) Ham, E. H., Lee, Y., Park, S., Park, H., Lee, S. (2022). Analysis on the characteristics and criteria development in performing science inquiry tasks for elementary school students. *Journal of the Korean Association for Science Education*, 42(2), 239-252.
- Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: measuring in the educational, social, and health sciences*. Springer.
- Guo, W., & Wind, S. A. (2023). The effects of rating designs on rater classification accuracy and rater measurement precision in large-scale mixed-format assessments. *Applied Psychological Measurement*, 47(2), 91-105.
- Li, Z., Xie, B., Hilsabeck, R., Aguirre, A., Zou, N., Luo, Z., & He, D. (2024). Effects of different prompts on the quality of GPT-4 responses to dementia care questions. *arXiv preprint arXiv:2404.08674*.
- Linacre, J. M. (2019). *Facets computer program for many-facet Rasch measurement, version 3.81.2*. Winsteps.com.
- Raykov, T., & Marcoulides, G. A. (2019). Thanks coefficient alpha, we still need you! *Educational and Psychological Measurement*, 79(1), 200-210.

- Robitzsch, A., Keifer, T., & Wu, M. (2023). *TAM: Test Analysis Modules {R package}*.
<https://github.com/alexanderrobitzsch/TAM>.
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of cronbach's alpha. *Psychometrika*, 74(1), 107-120.

투고일 : 2024.05.09. / 심사일 : 2024.05.16. / 심사완료일 : 2024.06.11.

<요 약>

GPT-4를 활용한 과학탐구역량 자동채점의 특성 분석*

함은혜¹⁾ · 박소영^{2)†} · 이병윤³⁾ · 이성혜⁴⁾ · 이유경⁵⁾ · 홍유정⁶⁾

¹⁾국립공주대학교 교육학과 부교수 · ²⁾숙명여자대학교 교육학부 교수

³⁾숙명여자대학교 교육연구소 전임연구원 · ⁴⁾한국과학기술원 과학영재교육연구원 연구교수

⁵⁾숙명여자대학교 교육학부 부교수 · ⁶⁾숭실대학교 대학혁신원 연구교수


이 연구는 GPT-4기반 자동채점시스템을 활용한 과학탐구역량 채점자료가 인간전문가의 채점자료와 어떻게 다른지를 비교·분석한 것이다. 이를 위해 연구진이 개발한 GPT-4기반 자동채점시스템을 활용하여 초등학생 과학탐구활동보고서 322개를 평가하였으며, 산출된 채점자료의 내적 구조가 인간 채점자료의 내적 구조와 유사한지, 과학탐구역량에 대한 이론적 가정을 지지하는지를 검토하였다. 주요 연구결과는 다음과 같다. 첫째, GPT-4를 활용한 채점은 인간 채점과 비교하여 관대하였으며, 특히 난이도가 높은 채점항목에 대해서 더 관대한 경향을 보였다. 둘째, 채점의 일관성과 채점항목 간 내적일치도는 인간 채점보다 높은 경향을 보였다. 셋째, 다국면 채점자 모형 분석 결과, GPT-4 채점 자료를 인간 채점 자료와 통합하는 경우, 채점항목의 난이도에서 변동이 크게 나타났으며, 인간채점자의 내적적합도와 외적적합도를 상당히 손상시키는 것으로 나타나, GPT-4 채점 결과와 인간채점 결과의 비교가능성이 지지되지 않았다. 연구 결과를 바탕으로, GPT-4를 활용한 자동채점의 한계, 가능성과 과제를 논의하였다.

주제어 : GPT-4, 자동채점, 신뢰도, 다국면채점자모형, 과학탐구역량

* 이 논문은 2020년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2020S1A3A2A02095447).

† 교신저자: 박소영, 숙명여자대학교, 교수 (E-mail: syngprk@sookmyung.ac.kr)


부록 1. 과학탐구보고서 작성 양식

● 미션1 

앞서 나온 레시피를 따라 '기본 리코타 치즈'를 만들어 보고 특징을 적어 봅시다.


사진	
특징	

여러분은 기본 리코타 치즈 레시피를 충실히 따랐나요? 이 기본 레시피는 어떻게 만들어진 것일까요? 끓이는 시간, 불의 세기, 우유/생크림의 비율, 레몬즙의 양, 물기를 짜주는 정도 등의 조건들은 치즈의 특성에 어떠한 영향을 미칠까요?

● 미션2 

리코타 치즈 레시피에서 치즈의 특성을 다르게 변화시킬 수 있는 조건을 적어도 두 개 이상 생각해 봅시다. 변화시킨 조건에 따라 치즈에 어떤 결과가 기대되는지 역시 설명해 주세요. (조건 ex. 끓이는 시간)

사진	
특징	

● 미션3 

[미션 2]에서 제시했던 조건 중 하나를 선정하여, 그 조건에 변화를 주며 (ex. 물을 덜 넣는다 / 더 넣는다.) 치즈를 직접 만들어 보고 결과를 분석해 봅시다.)

	A	B
변화 조건		
사진		
조건을 변화시킴으로써 치즈에 나타나는 현상		
치즈의 특징		
분석 결과	I	

부록 2. 과학탐구과제의 역량요소별 채점항목과 GPT-4채점을 위한 수정안

과학탐구 절차 요소	과학탐구 역량 요소	번호	채점항목 원안	GPT-4 채점을 위한 수정안
가설 생성	과학적 지식	V1	해당 차시에서 학습한 과학적 개념을 활용하여 치즈의 특징을 분석하는가?	미션 1에서 학생은 해당 차시에서 학습한 과학적 개념(맛의 종류/세기/강도, 냄새, 생김새, 촉감 등)을 활용하여 치즈의 특징을 구체적으로 분석하는가?
	과학적 지식	V2	치즈의 특징을 분석하기 위하여 기타 사전 지식(해당 차시 학습 이외)을 활용하는가?	미션 1에서 학생은 치즈의 특징을 분석하기 위하여 해당 차시 학습에서 배운 것 외에, 자신이 가진 기타 사전 지식을 활용하여 응답하는가?
	논리분석적 사고	V3	치즈의 특징을 변화시킬 수 있을 것으로 제시되는 조건을 2개 이상 명료하게 제시하는가?	미션 2에서 학생은 치즈의 특징을 변화시킬 수 있는 조건을 2개 이상 명료하게 제시하는가?
	논리분석적 사고	V4	각 조건의 변화가 치즈의 어떤 특징을 어떻게 변화시킬지를 예상(예측)하여 진술하는가?	미션 2에서 학생이 제시한 각 조건의 변화가 치즈의 어떤 특징을 어떻게 변화시킬지 예상(예측)하여 진술하는가?
	논리분석적 사고	V5	(조건1) “각 조건의 변화가 왜 혹은 어떻게 치즈의 특징을 변화시키게 될지를” (사전에 학습한) 과학적 개념이나 원리를 활용하여 설명하는가?	미션2에서 학생은 과학적 개념이나 원리를 활용하여, 제시한 조건 1의 변화가 왜 혹은 어떻게 치즈의 특징을 변화시키게 될지를 예측하는가?
	논리분석적 사고	V6	(조건2) “각 조건의 변화가 왜 혹은 어떻게 치즈의 특징을 변화시키게 될지를” (사전에 학습한) 과학적 개념이나 원리를 활용하여 설명하는가?	미션 2에서 학생은 과학적 개념이나 원리를 활용하여, 제시한 조건 2의 변화가 왜 혹은 어떻게 치즈의 특징을 변화시키게 될지를 예측하는가?
	탐구적 태도	V7	(기타 자료를 참고하거나 스스로의 사고를 통해) 고려할만한 조건을 추가적으로 탐색하였는가?	미션 2와 미션 3에서 학생은 다른 자료를 참고하거나 스스로의 생각을 통해 치즈의 특징을 변화시키는 조건을 추가적으로 탐색하는가?
	탐구적 태도	V8	“각 조건의 변화가 왜 혹은 어떻게 치즈의 특징을 변화시키게 될지를” 기타 과학적 개념이나 원리를 활용하여 설명하는가?	미션 2와 미션 3에서 학생은 각 조건의 영향을 예측할 때, 참고자료를 활용하여 과학적 개념이나 원리를 설명하는가?
실험 설계 및 시행	논리분석적 사고	V9	실험을 위해 변화시킨 조건이 명확한가?	미션3에서 학생이 실험을 위해 변화시킨 조건을 명확하게 기술하였는가?
	논리분석적 사고	V10	선택한 조건을 변화시킬 수 있는 방법으로 치즈 제작 과정의 조건을 변화시켰는가?	미션3에서 학생은 자신이 선택한 조건을 변화시킬 수 있는 방법으로 치즈 제작 과정의 조건을 변화시켰는가?
	논리분석적 사고	V11	자신이 선택한 조건 이외의 조건을 고려하여 통제하는가?	미션3에서 학생은 자신이 선택한 조건 이외의 조건을 고려하여 통제하는가?
	의사소통	V12	조건을 변화시키는 과정에서 계량적 접근이 관찰되는가?	미션3에서 조건을 변화시키는 과정에서 계량적 접근이 관찰되는가?
	과학적 지식	V13	제시된(선정된) 조건이 앞에서(사전 학습자료 혹은 개별 추가 학습자료) 학습한 과학적 개념이나 원리와 관련이 있는가?	미션3에서 학생이 제시한 조건은 사전자료나 개별 추가 자료를 통해 학습한 과학적 개념이나 원리와 관련이 있는가?

부록 2. 과학탐구과제의 역량요소별 채점항목과 GPT-4채점을 위한 수정안 (계속)

과학탐구 절차 요소	과학탐구 역량 요소	번호	채점항목 원안	GPT-4 채점을 위한 수정안
자료분석 및 해석	(논리분석적 사고)	V14	조건 변화에 따른 결과물의 특성을 기본치즈와 비교하여 기술하는가?	미션3에서 학생은 자신이 제시한 조건에 따라 변화된 치즈의 특성을 기본 치즈와 비교하여 기술하는가?
	논리분석적 사고	V15	서로 다른 조건 변화에 따른 결과물의 특성을 비교하여 기술하는가?	미션3에서 학생은 서로 다른 조건에 따라 변화된 치즈의 특성을 비교하여 기술하는가?
	논리분석적 사고	V16	가설을 바탕으로 결과를 분석하는가?: 분석결과를 본인의 가설(예상했던 결과)과 비교하여 기술하는가?	미션3에서 학생은 분석결과를 본인의 가설(예상했던 결과)과 비교하여 기술하는가?
	과학적 지식	V17	과학적 원리와 개념에 대한 사전지식을 활용하여 결과를 해석하는가?	미션3에서 학생은 과학적 원리와 개념에 대한 사전지식(지방구, 단백질, 응고, 산성, 결합반응, 대류 현상 등)을 활용하여 결과를 해석하는가?
	탐구적 태도	V18	분석 결과를 이해(해석)하기 위해 추가 자료를 탐색하는가?	미션3에서 학생은 자신이 분석한 결과를 이해(해석)하기 위해 자료를 추가로 탐색하는가?
결론 도출 및 평가	탐구적 태도	V19	분석 결과에 근거하여 특정 조건의 적절성이나 유용성을 평가하는가?	미션3에서 학생은 자신이 분석한 결과에 근거하여 특정 조건의 적절성이나 유용성을 기술하는가?
	탐구적 태도	V20	실험을 통해 자신이 무엇을 배웠는지 기술하는가?	미션3에서 실험을 통해 학생이 무엇을 배웠는지 기술하는가?
	탐구적 태도	V21	본인이 수행한 실험의 강점이나 보완할 점 등을 기술하거나 향후 탐구과제를 제시하는가?	미션3에서 학생이 수행한 실험의 강점이나 보완할 점 등을 기술하거나 향후 탐구과제를 제시하는가?
전반	의사소통	V22	치즈의 특징(관찰 결과)에 대한 언어적 기술이 풍부한가?	학생이 관찰한 치즈의 특징에 대해 기술할 때 언어적 표현이 풍부한가?
	의사소통	V23	참고한 자료의 출처를 제시하는가?	학생이 참고한 자료의 출처를 명확하게 제시하는가?
	의사소통	V24	(V12 외) 계량적 접근이 관찰되는가?	미션 3 이외의 과정에서 계량적 접근이 관찰되는가?
	의사소통	V25	진술이 명료하고, 문단을 구조화하는 등 독자의 가독성을 고려하는가?	진술이 명료하고, 문단을 구조화하는 등 독자의 가독성을 고려하는가?

부록 3. 본 연구에서 활용한 자동채점 플랫폼 화면 예시

[채점항목 및 점수 척도 입력 화면]

Cretia Management

Edit mode

역량 추가하기

역량별 평가기준 관리

과학탐구 역량

역량명

koen

과학탐구 역량Inquiry

평가기준 1

평가기준명(ko)평가기준명(en)

질문 구성/가설 생성Question and hypothesis

세부 평가기준

미션 1에서 학생은 해당 자사에서 학습한 과학적 개념(맛의 종류/세

01

미션 1에서 학생은 치즈의 특징을 분석하기 위하여 해당 자사 학습이

01

미션 2에서 학생은 치즈의 특징을 변화시킬 수 있는 조건을 2개 이상

01

미션 2에서 학생이 제시한 각 조건의 변화가 치즈의 어떤 특징을 어

01

[채점대상(탐구보고서) 업로드 화면]

AI 기반 미래역량 평가 도구

숙명여대 SSK 연구사업 AI-CALI팀 개발

이 점수는 연구진이 개발한 채점기준을 활용하여 GPT-4가 채점을 시행한 결과로, 향후 채점의 신뢰도와 타당도를 평가하고 개선하기 위한 연구자료로 활용됩니다. 현재 단계에서 GPT-4의 채점결과를 실제 해당 역량의 특성을 충분히 반영하고 있지 않을 수 있으므로, 해석과 사용 시 주의가 필요합니다

역량

과학탐구 역량

과제 파일 업로드(.hwp, .docx, pdf, zip)

Drag and drop files here

Limit 200MB per file • HWP, DOCX, PDF, ZIP

Browse files

평가하기

부록 4. 자동채점 플랫폼 채점 결과 출력(엑셀 형식) 예시

[채점항목별로 생성된 점수 및 영역별 합산 점수]

	A	B	C	D	E	F	G	H	I	J
	STU ID	question and hypothesis_1	question and hypothesis_2	question and hypothesis_3	question and hypothesis_4	question and hypothesis_5	question and hypothesis_6	question and hypothesis_7	question and hypothesis_8	question and hypothesis_total
1										
2	240319_143925_1	1	0	1	1	0	0	0	0	3
3	240319_143925_2	1	0	1	1	0	0	0	0	3
4	240319_143925_3	1	0	1	1	0	0	0	0	3
5	240319_143925_4	1	0	1	1	0	0	0	0	3
6	240319_143925_5	1	0	1	1	0	0	0	0	3

[채점영역(절차요소)별로 생성된 서술형 피드백]

	A	AG	AH	AI	AJ	AK
	STU ID	question and hypothesis_descript	experiment_descript	data_descript	reflection_descript	communication_descript
1						
2	240319_143925_1	학생은 치즈의 맛과 질감에 대해 구체적으로 기	학생은 실험 조건을 명확히 기술하고 치즈 제작	학생은 변화된 치즈의 특성을 기본 치즈와 비교	학생은 실험을 통해 배운 점이나 실험의 강점	학생은 치즈의 특징을 언어적으로 풍부하게 기술
3	240319_143925_2	학생은 치즈의 맛, 강도, 촉감 등을 관찰하여 기술	학생은 실험 조건을 명확히 기술하고, 치즈 제작	학생은 변화된 조건에 따른 치즈의 특성을 기	학생은 실험을 통해 배운 점이나 실험의 강점	학생은 치즈의 특징을 언어적으로 풍부하게 기술
4	240319_143925_3	학생은 맛의 종류와 강도, 촉감 등을 언급하며 지	학생은 끓이는 시간이라는 조건을 변화시켜 실험	학생은 변화된 치즈의 특성을 기술하고 비교	학생은 실험을 통해 배운 점이나 실험의 강점	학생은 치즈의 특징을 언어적으로 풍부하게 기술
5	240319_143925_4	학생은 맛의 종류와 세기, 강도, 촉감 등을 구체	학생은 실험 조건을 명확하게 기술하고 치즈 제	학생은 변화된 치즈의 특성을 기본 치즈와 비	학생은 실험을 통해 배운 점이나 실험의 강점	학생은 언어적 표현이 풍부하고 가
6	240319_143925_5	학생은 맛의 종류, 세기, 강도, 촉감 등을 구체	학생은 실험 조건을 명확하게 기술하고, 치즈 제	학생은 변화된 치즈의 특성을 기본 치즈와 비	학생은 실험을 통해 배운 점을 기술하였으나, 실	학생은 언어적 표현이 풍부하고 가
7	240319_143925_6	학생은 치즈의 맛, 촉감 등을 기술하였으나, 고	학생은 실험 조건을 명확히 기술하고 치즈 제	학생은 변화된 치즈의 특성을 기본 치즈와 비	학생은 실험을 통해 배운 점이나 실험의 강점	학생은 치즈의 특징을 언어적으로 풍부하게 기술