

ChatGPT를 활용한 서·논술형 평가 자동 채점 가능성 탐색: 세계지리 서·논술형 평가를 중심으로

성정원* · 신병철**

Exploring the Feasibility of Automatic Scoring of Written Test Using ChatGPT: Focusing on the World Geography Written Test

Jeong-won Seong* · Byoung-chul Shin**

요약 : 본 연구에서는 ChatGPT를 활용하여 서·논술형 평가문항의 채점 가능성을 탐색하기 위해 교사와 ChatGPT의 채점 결과를 비교, 분석하였다. 이를 위하여 세계지리 기후 단원의 서·논술형 평가문항과 평가기준을 개발하고, 15년 이상 교육경력 of 지리 교사 3명이 2차에 걸쳐 채점을 실시하였다. 그리고 ChatGPT는 평가기준만 제시한 경우, 평가기준·예시답안만 제시한 경우, 평가기준·예시답안·비판적으로 채점 명령을 제시한 경우로 나누어 채점을 실시하고 그 결과를 비교, 분석하였다. 비교적 정확한 정답이 있는 1번 문항은 평가기준·예시답안·비판적으로 채점 명령을 제시한 경우, 그리고 평가기준·예시답안을 제시한 경우에 교사의 채점 결과와 높은 상관관계를 보였으며, 이는 서·논술형 평가의 채점에서 교사 간 신뢰도 기준(상관관계 .6 이상)을 충족한 것으로 판단된다. 하지만 채점자의 주관이 많이 개입될 수 있는 2번 문항은 평가기준·예시답안·비판적으로 채점 명령을 제시한 경우, 논술형 평가의 채점에서 요구하는 교사 간 신뢰도 기준에 미치지 못하였다. 하지만 ChatGPT의 채점 결과는 2번 문항에 대한 지리 교사 간 상관관계와 유사한 값이 나왔다. ChatGPT를 활용하여 서·논술형 평가 문항을 채점할 경우, 평가기준·예시답안·비판적으로 채점 명령, 또는 평가기준·예시답안을 제시하는 프롬프트를 활용하면 보조 교사로서 채점자의 역할을 수행할 수 있다.
주요어 : 세계지리, 서·논술형 평가, ChatGPT 채점, 채점자 간 신뢰도

Abstract : In this study, we compared and analyzed the scoring results of teachers and ChatGPT to explore the feasibility of using ChatGPT-4 for scoring written tests. For this study, we developed written test questions and scoring criteria based on the World Geographic Climate Unit. Three geography teachers with over 15 years of educational experience scored the questions twice. Additionally, ChatGPT performed scoring under different conditions: presenting only the rubric, presenting the rubric and example answers, and presenting the rubric, example answers, and critical scoring instructions. We then compared and analyzed the results. Question 1, which had a relatively accurate correct answer, exhibited a high correlation with the teachers' scoring results when the rubric, example answer, and critical scoring instructions were provided. This correlation is believed to meet the criteria for intra-scorer reliability (a correlation of 0.6 or higher) for written tests. However, Question 2, whose scoring can be very subjective, did not meet the reliability criteria for intra-scorer reliability in the scoring involving the rubric, example answer, and critical scoring instructions. Nonetheless, a correlation value similar to that obtained by three teachers for Question 2 was found. When scoring written tests using ChatGPT, it can function as an assistant teacher by presenting prompts with rubrics, example answers, critical scoring instructions, or rubrics and example answers.

Key Words : World geography, Written tests, Scoring by ChatGPT, Intra-scorer reliability

*수원외국어고등학교 교사(Teacher, Suwon Academy of World Languages, seonwoodady@korea.kr)

**수원외국어고등학교 교사(Teacher, Suwon Academy of World Languages, sinbc2004@naver.com)

I. 서론

2022년은 대한민국 교육계에 2가지 큰 화두가 던져졌다. 첫 번째는 학교 교육의 위기와 가능성이라는 양날의 검과 같은 ChatGPT의 등장이고, 두 번째는 학생들의 학습 선택권을 보장하는 고교학점제 도입을 위한 2022 개정 교육과정의 발표된 것이다. ChatGPT와 2022 개정 교육과정은 학교 현장의 교사들에게 많은 변화와 도전을 요구하고 있다. 이러한 변화와 도전은 교사들에게 많은 과업을 요구하면서 학교의 교육 활동을 위축시키는 위기 상황으로 전개될 수도 있다. 하지만 이러한 변화를 슬기롭게 받아들이면 전화위복의 기회가 되어 교육 활동이 활성화될 수도 있을 것이다.

2022년 11월 ChatGPT가 등장하면서 많은 언론에서 ‘학교 위기설’, ‘교육계의 공황’이라는 표현을 쏟아냈다. 학생들이 과제를 할 때 ChatGPT를 이용하여 부정행위를 저지르는 것을 우려하고, 나아가 학습 내용을 교사가 아닌 ChatGPT가 알려줄 수 있다는 점에서 학교와 교사의 무용론이 대두되었다. 하지만 시간이 지나면서 ChatGPT가 학교와 교사를 대체하는 것이 아닌 학생의 상호작용과 개별화 교육을 지원하고, 교사의 업무를 지원하는 수업을 위한 보조 도구로서 가능성에 주목하고 있다(MIT Technology Review, 2023년 8월 7일자).

2025년에 전면 시행되는 고교학점제는 학생의 소질과 적성에 맞는 교육을 위해 다양한 교육 기회를 확대하는 것을 목적으로 한다(김기철 등, 2023). 학생의 진로에 맞는 과목을 선택할 수 있도록 과목 선택권을 최대한 보장하기 위해서는 학교에서 개설되는 과목 수가 많아져야 하기에 2022 개정교육과정은 선택 과목의 수를 확대하였으며, 특히 사회교과(군)의 경우 일반 선택, 융합 선택, 진로 선택으로 나누어 19과목을 제시하였다(교육부, 2022). 학생들의 선택 기회를 보장하기 위해 개설하는 과목 수가 늘어나면 한 과목 당 수강하는 학생 수가 줄어들게 된다. 이는 한 과목에 2명 이상의 교사가 함께 지도하기 보다는 대부분 한 명의 교사가 한 과목을 담당하게 되며, 늘어나는 과목만큼 교사들은 많은 과목을 지도해야 한다는 것을 의미한다. 또한 개정 교육과정에서는 학생들의 다양한 지식과 경험을 융합하며 창의적으로 문제를 해결하는 역량을 함양하고, 이러한 과정에 대한 평가를 위해 서·논술형 평가를 확대할 것을 요구하고 있다(교육부, 2022). 결과적으로 교사들은 교사 단독으로 여러 과

목을 지도해야 하면서 서·논술형 평가를 지금보다 많이 실시해야 한다.

교사들에게 수업을 진행하는 것만큼 준비와 시간이 많이 소요되는 것이 평가이다. 특히, 채점에 시간과 노력이 많이 필요로 하는 서·논술형 평가가 확대되는 상황에서 교사가 지도해야 하는 과목 수가 늘어나면 평가에 대한 부담은 커질 수밖에 없다. 이에 본 연구는 학생들이 작성한 세계지리 기후 단원의 서·논술형 평가 답안을 교사와 ChatGPT가 채점하고 이를 비교, 분석하여 ChatGPT를 활용한 서·논술형 평가 채점 가능성을 탐색해보고자 한다.

II. 선행 연구 고찰

1. 서·논술형 평가

교사는 설정한 교육목표를 달성하기 위해 내용요소와 수업방법을 결정하고 수업을 진행하며, 학습 목표에 도달했는지를 확인하기 위해 평가를 실시한다(오정현, 2006). 학생들의 학습 결과를 확인하기 위한 평가는 선택형을 중심으로 한 지필교사가 큰 비중을 차지하지만, 창의력, 문제해결력, 논리적 사고 등을 평가하기 위한 서·논술형 평가의 비중도 계속해서 높아지고 있다(장유정, 2020). 특히, 서·논술형 평가는 선택형 평가와 비교하면 평가의 내용 타당도가 높고, 과정 평가와 능력에 대한 질적 평가가 가능하다는 장점을 가지고 있어 학교 현장에서 교사나 학생에게 중요한 평가도구로 인식되어 활용되고 있다. 학생들은 이러한 서·논술형 평가에 대해 교과에 대한 자신감과 흥미, 호기심이 향상되고, 문제해결력 및 학습 의욕이 향상되기 때문에 긍정적으로 인식하고 있으며(김남준·배중수, 2006; 황현숙, 2012; 김효선·오영열, 2014), 교사들은 서·논술형 평가를 통해 학생들의 역량과 사고력을 측정할 수 있고 교수학습에 필요한 정보를 수집할 수 있으며 학생에 대한 이해를 높일 수 있다는 점에서 긍정적으로 인식하고 있다(김래영·이민희, 2013).

이러한 서·논술형 평가가 학교 현장에서 자리를 잡기 위해서는 채점의 신뢰도를 높이는 것이 매우 중요하다. 신뢰도를 높이기 위한 채점 방법에 대한 다양한 연구가 진행되고 있으며, 서·논술형 평가의 채점에는 일반적으로 총체적 채점(holistic scoring method) 분석적 채점

(analytic scoring method)이 사용되고 있다. 총체적 채점은 학생들의 응답에 대해 전체적으로 파악하여 하나의 점수를 부여하는 것으로, 이 방법은 학생들의 성취도에 따라 순위를 구분할 때 유리하다. 반면, 분석적 채점은 답안에 포함되어 있는 평가요소들에 점수를 부여하고 그 점수를 종합하는 방법으로, 평가 요소가 구분되어 있어 학생들이 부족한 부분을 파악하고 이에 대한 피드백을 하기에 유리하다(지은림, 2000). 하지만 이러한 서·논술형 평가는 선택형 평가보다 동일한 문항에 대해 채점자 간의 점수 결과가 달라질 수 있고, 채점에 많은 시간과 노력이 필요하다는 해결 과제를 가지고 있다(박도순·홍후조, 2011). 특히, 평가기준이 마련되어 있다고 하더라도 평가자의 주관이 개입될 가능성이 크기 때문에 평가의 신뢰도가 낮아 학교 현장에서는 평가의 공정성을 명분으로 선택형이나 단답형의 평가를 선호하고 있는 것이 현실이다. 이에 채점에 필요한 시간을 줄여줄 수 있고, 사람의 주관이 반영되지 않은 프로그램을 통한 자동 채점의 도입이 요구되고 있다(강원석, 2018).

2. ChatGPT와 같은 생성형 인공지능의 교육적 활용

2015년에 창립된 OpenAI는 인공지능 기술을 연구하고 개발해왔으며 2018년 6월 GPT(Generative Pre-trained Transformer) 모델을 처음 발표하였다. GPT란 Transformer라는 딥러닝 기반의 인공지능경망 모델이 대규모 텍스트 데이터를 통해 사전학습(Pre-trained)되어 새로운 텍스트를 생성(Generative)하는 것을 말한다. OpenAI는 2018년 6월 GPT-1을 기점으로 GPT-2(2019년), GPT-3(2020년), GPT-3.5(2022년), GPT-4(2023년)를 순차적으로 출시했다. 2022년 11월에 공개된 ChatGPT는 GPT-3의 후속 모델인 GPT-3.5 모델을 기반으로 만들어진 인공지능 챗봇으로 사용자와의 자연스러운 대화 생성에 특화되어있으며 다양한 분야의 주제로 대화가 가능하다. 현재 유료 계정 사용자는 GPT-4 모델 기반인 ChatGPT도 사용할 수 있다.

Transformer 모델은 데이터 내 각 요소의 순서가 있는 시퀀스(sequence) 데이터를 처리하는 동안 모델이 연관성이 있는 부분에 더 집중하도록 하는 어텐션이라는 메커니즘을 사용하고, 다른 언어 모델과는 달리 시퀀스 데이터를 병렬적으로 한번에 처리하여 학습, 추론 시간을

단축할 수 있다(Vaswani *et al.*, 2017). 이러한 병렬 처리 능력을 통해 문장 내에서 멀리 떨어진 단어들 사이의 관계도 잘 파악할 수 있어 번역, 텍스트 요약, 질의 응답 등의 영역에서 뛰어난 성능을 보여 많은 분야에서 활용되고 있다. 현재 이러한 GPT 모델을 응용하여 대화형 영어 인공지능 챗봇, 글쓰기 첨삭 챗봇 등 언어 계열에서 더욱 활발한 사용이 이루어지고 있다.

한편 교과별로 ChatGPT를 교육적으로 활용하거나 가능성을 논하는 여러 연구가 나오고 있다. 국어, 영어 등 언어과목에서는 ChatGPT의 등장 이전부터 인공지능을 활용한 챗봇 연구가 활발하게 이루어졌다. 인공지능 챗봇을 활용했을 때 학생들의 발달과정을 더 정확하게 파악하고, 개별화된 탐구 활동을 수행하는 기회를 제공하며, 언어 수행 능력과 더불어 동기를 향상시키는데 효과적이라는 입장(추성엽·민덕기, 2019; 김인석·김봉규, 2020; 윤여범·박미애, 2020; 최원경, 2020; 장진아 등, 2021; 홍선호 등, 2021)도 있는 반면 인공지능 챗봇이 지닌 가능성과 교육적 효과에도 불구하고 인프라 구축 및 인공지능에 대한 교사 교육의 부족, 오류 가능성, 인공지능이 인간의 미묘한 표현을 이해하지 못하는 등의 한계(이동한, 2018; 양혜진 등, 2019)를 지적하기도 했다. 최근 ChatGPT와 같은 생성형 인공지능의 등장은 영어 교육에서 교수, 학습의 보조 수단으로서 문법 오류 점검, 역할극, 토론, 문제 출제 등에서 활용 가능성이 높아 교육현장의 많은 변화를 불러일으킬 것으로 기대된다(신동광 등, 2023). 수학교육에서는 ChatGPT의 활용 방안 도출을 위한 기초 연구로서 국가수준 학업성취도 평가 및 대학수학능력시험에 대한 응답을 분석하였고(권오남 등, 2023), 예비교사와 학생과의 대화 자료를 사례로 선정하여 ChatGPT를 통해 분석한 결과를 수학교육 전문가의 결과와 비교하여(손태권, 2023) 수학교육에서 ChatGPT의 활용 가능성과 시사점을 도출하였다. 지리교육에서도 ChatGPT는 지리 교과에 대한 심화 문제를 해결할 수 있고, 이는 ChatGPT를 체계적이고 구체적인 지침을 마련하여 사용한다면 지리교육에서도 효과적인 학습 도구로서의 가능성이 있음을 시사하였다(유재진, 2023).

ChatGPT에 대한 교육 분야의 많은 기대에도 불구하고 편향된 정보, 최신 데이터의 학습 부족, 환각 현상, 사회적 혼란, 잘못된 의사결정, 가짜 뉴스 생성, 데이터 유출 등 잘못된 정보 및 보안 문제에 대한 대비가 필요하다(국가정보원, 2023). ChatGPT는 13세 미만의 사용자는

사용을 금하고, 18세 미만의 사용자는 부모 또는 보호자의 감독하에만 사용하도록 하고 있어 교육 분야에서 이를 이용하기 위해서는 각별한 주의가 요구된다. 또한 ChatGPT가 생성하는 결과물을 그대로 과제에 활용하는 표절 문제나 데이터를 오용할 가능성이 있기 때문에 인공지능을 활용할 때 글쓰기에 대한 윤리적 경계 및 교사의 지도가 반드시 필요하다(Gao *et al.*, 2022; Crawford *et al.*, 2023; Qadir, 2023).

3. 인공지능의 자동채점

한국교육과정평가원에서는 서답형 문항 자동채점 프로그램 방안 연구(진경애 등, 2006)를 시작으로 컴퓨터 기반 서·논술형 평가를 위한 자동채점 방안 설계(박종임 등, 2022)까지 다양한 연구보고서를 작성하였다. 최초의 서답형 자동 채점 프로그램 연구에서는 사회과 교육과정 내용을 검토하고, 사회과의 한국어 서답형 문항 자동 채점 프로그램 도입을 위한 기초연구를 실시하여 서답형 문항의 자동 채점 프로그램 개발 방향을 제시하였다(진경애 등, 2006). 2011년 대규모 평가를 위한 서답형 문항 자동 채점 방안 연구에서는 국어·과학의 학업성취도 평가, 한국어능력시험에서 활용된 서답형 문항들의 답안을 분석하여 단어·구 수준의 한국어 서답형 자동 채점 프로그램을 개발하였다(노은희 등, 2012). 이후 2016년에는 2014, 2015 학업성취도 평가 국어·사회·과학 교과의 복문 1문장, 서답형 5문항에 대해 문장 수준의 한국어 자동 채점 프로그램을 적용하여 전년도의 프로그램보다 채점 정확성 및 인간 채점의 결과와 상관을 향상시켜 자동 채점 가능성을 확인하였다(노은희 등, 2016). 하지만 이는 특정 채점 변인과 알고리즘에 의해 자동 채점이 이루어지는 방식이 아니라 답안이 자동 분류되면 답안 범주에 인간이 점수를 부여하고 이를 통해 기계학습을 진행하여 나머지 답안을 채점하는 방식이었으며 문단 이상의 논술형 답안을 채점하기에는 한계가 있었다(박종임 등, 2022).

최근에는 인공지능망 기반의 딥러닝을 이용한 방식이 자동 채점의 새로운 방안으로 떠오르고 있다. 본 연구에서는 기존의 국내 연구에서 서답형 문항들에 대한 자동 채점을 진행한 방법과는 다르게 ChatGPT를 활용하여 고등학생들의 서·논술형 답안의 자동 채점 가능성을 탐색해보고자 하였다. ChatGPT는 대화 기반의 인터페이스

를 가지고 있어 누구나 쉽게 사용할 수 있다는 장점이 있기에 ChatGPT를 채점 도구로써 선택하였고, 인공지능 및 자동 채점에 대한 전문지식이 부족한 교사도 인공지능으로부터 학생들의 답안에 대한 채점 수행을 지원받을 수 있는지 확인해보고자 하였다.

한편, ChatGPT를 활용한 채점의 장점은 첫째, 채점의 영향을 주는 후광효과(Halo effect)를 제거할 수 있다. 채점자는 평소에 가지고 있는 학생의 인상과 느낌에 따라 채점 결과에 영향을 줄 수 있는데(성태제, 2014; 황정규 등, 2016), ChatGPT의 채점은 이러한 후광효과를 원천적으로 배제할 수 있다. 둘째, 채점의 신뢰도를 높이기 위해서는 2명 이상의 채점자들이 채점을 하여 채점자 간 신뢰도를 확인할 필요가 있다(성태제, 2014). 학교 현장에서는 단독으로 수업하는 교사가 많고, 특히, 고교학점제 도입으로 선택 과목 수가 늘어나면서 1인 교사가 가르치고 평가하는 과목 수가 늘어나고 있는 것이 현실이다. 이에 ChatGPT는 1인 채점자로서의 역할을 하며 채점자 간 점수를 비교하여 채점 신뢰도를 높일 수 있다. 셋째, ChatGPT는 정확한 명령어를 입력하면 채점 결과를 짧은 시간 내에 확인할 수 있기 때문에 채점 시간을 획기적으로 줄여줄 수 있다.

ChatGPT와 같은 GPT 모델을 활용하여 자동 채점을 하기 위해서는 다음과 같은 사항을 고려해야 한다. 첫째, 최신의 GPT 모델을 사용해야 한다. GPT-4 모델은 최근의 모델 평가에서 GPT-3.5 모델에 비하여 9개의 카테고리(learning, technology, writing, history, math, science, recommendation, code, business)에서 많은 기능개선이 이루어졌고, 환각(Hallucination) 현상¹⁾을 19% 이상 크게 줄였지만 여전히 사실을 환각하여 추론 오류를 범하기도 한다(OpenAI, 2023). 그러나 OpenAI에서 제공하는 GPT 모델의 올바른 사용법을 통해 환각 현상을 줄이고, 모델의 추론을 개선할 수 있다(OpenAI, 2023년 7월 12일자). 둘째, GPT 모델은 2021년 9월까지의 데이터를 학습했기 때문에 자동 채점하고자 하는 평가가 2021년 9월 이후의 학습을 필요로 한다면 GPT 모델이 올바른 채점을 수행하지 못할 가능성이 있다. 셋째, GPT-4 모델의 경우 한 세션에서 사용할 수 있는 최대 길이가 8,192 토큰이기 때문에 채점을 위해 작성하는 시스템 프롬프트, 학생 답안, GPT-4의 채점 결과 내용이 길어 모델이 지원하는 최대 토큰 수를 넘을 수 있다. 최대 토큰이 초과된 다음 입력되는 프롬프트에 대해서는 GPT-4의 불완전

한 답변을 받을 수 있기 때문에 새로운 세션에서 채점을 수행하거나 시스템 프롬프트를 다시 입력하여 시작해야 한다(OpenAI, 2023년 7월 12일자). 넷째, 모델 출력 결과물의 일관성에 영향을 미치는 변수값을 조절해야 한다. 정확도가 높은 수행이 필요할 때는 값이 커질수록 다양한 결과값을 내는 변수인 Temperature²⁾와 같은 변수값을 작게 설정하는 것이 좋다(OpenAI, 2023년 7월 12일자).

III. ChatGPT 활용 세계지리 서·논술형 평가 채점

1. 연구 절차

ChatGPT의 서·논술형 평가 채점 가능성을 탐색하기 위한 연구 절차는 표 1과 같다. 세계지리 기후 단원에서 서·논술형 평가문항을 개발하고, 이 문항에 대한 평가기준을 마련하였다. 경기도 수원에 소재한 A고등학교에서 세계지리를 수강하는 학생 27명의 수행평가 답안을 활용하여 경력 15년 교사 3명과 ChatGPT의 채점 결과를 비교·분석하였다.

2. 세계지리 서술형 평가 문항과 평가기준

본 연구에 활용한 수행평가는 세계지리 2단원 '세계의 자연환경과 인간 생활'의 기후를 주제로 문항을 구성하였다. 지리교육은 다양한 지역의 정보를 제공하여 지역을 편견 없이 바라보게 하며(김기남·성정원, 2023), 지리적 지식과 안목을 바탕으로 습득한 지식을 실생활에 적용할 수 있는 시민을 육성하는 과목이다. 이에 지리교육의 목표 달성을 확인하기 위한 평가는 다양한 지역의 현실 자료를 바탕으로 추론할 수 있는 내용으로 구성하고, 학생들의 지식과 탐구를 통해 합리적인 의사결정과 문제해결을 요구해야 한다(박선미, 1999). 기후는 우리의 일상 생활에서 가장 밀접한 관련이 있는 주제이며, 명확한 근거를 바탕으로 학습하고 평가하는 내용 요소라고 판단된다. 이러한 지리교육의 목적 달성을 고려하여 본 연구의 수행평가는 제시된 지역의 월별 기온과 강수량을 바탕으로 컴퓨터 프로그램(한셀)을 이용하여 기후 그래프를 그리고, 이를 활용하여 해당 지역의 기후를 분석하는 문항과 해당 지역의 기후조건을 활용한 창업 아이디어를 제시하는 문항으로 구성하였다. 수행평가는 컴퓨터실에서 진행되었으며, 학생들이 사용하는 컴퓨터의 인터넷을 통제된 상태에서 한셀과 한글 프로그램을 이용하여 진행하였다. 학생들에게 제시된 기후 값은 표 2

표 1. 연구절차

구분	연구절차	연구내용
서·논술형 평가문항 및 평가기준 개발	서술형 평가 문항 개발	<ul style="list-style-type: none"> • 1번 문항: 해당 지역의 기후를 분석하는 문항 • 2번 문항: 기후를 활용한 창업 계획서를 작성하는 문항
	평가기준 개발	<ul style="list-style-type: none"> • 1번 문항 평가기준표 제작 • 2번 문항 평가기준표 제작
학생 평가	서·논술형 평가 실시	<ul style="list-style-type: none"> • 6월 16일 실시 • 고등학교 2학년 세계지리 이수 학생 27명 • 컴퓨터실에서 한셀과 한글 프로그램을 활용하여 진행
서·논술형 평가 채점	프롬프트 개발 및 ChatGPT 채점	<ul style="list-style-type: none"> • 7월 12일~8월 12일 GPT-4 모델로 실시 • ChatGPT의 선행지식 확인 (기후 자료 및 수행평가 문항을 제시하고 ChatGPT가 수행할 수 있는지 파악) • ChatGPT의 채점 수행을 위한 프롬프트 작성 • 10명의 샘플 답안으로 파일럿 테스트 • 한 세션당 1명씩 채점하고 반복 수행
	지리교사 채점	<ul style="list-style-type: none"> • 1차 채점: 7월 10일~7월 14일 실시, 2차 채점: 7월 17일~21일 실시 • 1, 2번 문항에 대해 채점과 채점 근거 작성
채점 결과 분석	채점 결과 분석	<ul style="list-style-type: none"> • 지리교사와 ChatGPT의 점수 결과 비교 분석 • 1번 문항과 2번 문항 결과 비교분석
	ChatGPT 피드백 분석	<ul style="list-style-type: none"> • 학생 답안에 대한 ChatGPT 피드백 경향 분석

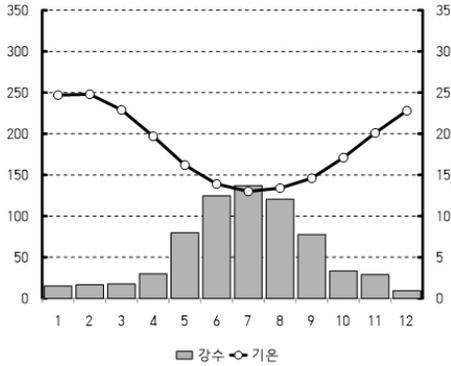


그림 1. 학생들이 제작한 기후 그래프

와 같은 형태이며, 학생들은 한셀 프로그램을 이용하여 그림 1과 같은 그래프를 제작하였다. 이러한 활동을 평가에 포함한 이유는 지리교과에서 차트, 도표, 지도 등을 평가문항에 활용하여 도해력을 평가할 수 있고(조경철, 2012), 학생들에게 지역의 정보를 활용하여 다양하게 시각화하여 표현하는 활동은 지리 학습에서 중요하다고 인식하고 있기 때문이다(이진희·조인정, 2016). 하지만, 인공지능 채점을 수행하였던 2023년 7월 12일~8월 12일에는 ChatGPT의 이미지 분석기능을 이용할 수 없었기에 본 연구는 학생들이 서술형으로 작성한 2개 문항에 대해서만 분석을 진행하였다.

1) 해당 지역 기후 분석

첫 번째 문항은 정답과 오답을 구분할 수 있는 명확한 기준이 있는 문항으로 제작하였다. 제시된 월별 기온과 강수량 값을 바탕으로 쾨펜의 기후구분 원리를 적용하여 해당 지역의 기후를 분석하고, 해당 지역의 위치를 추론하도록 하였다. 쾨펜의 기후구분에서 1차 구분은 최한월과 최난월의 기온 값을 통해 열대기후(A), 온대기후

(C), 냉대기후(D), 한대기후(E)를 구분하고, 2차 구분은 강수량의 계절적 변동을 통해 겨울 건조(w), 여름 건조(s), 연중 습윤(f)로 구분한다. 이 원리를 바탕으로 해당 지역의 기후를 분석하고, 최난월과 최한월의 시기를 통해 남반구 또는 북반구에 위치한 지역임을 추론할 수 있다. 실제 문항에서는 남반구 오스트레일리아의 퍼스(Cs), 남반구 오스트레일리아의 다윈(Aw), 북반구 싱가포르(Af), 북반구 홍콩(Cw), 북반구 알제리의 다르엘베이다(Cs)의 기후 값을 제시하였다. 특히, 남반구와 북반구의 계절 차이를 이해하고 있는지 확인하기 위하여 2차 구분에서 여름 건조(s)와 겨울 건조(w)가 나타나는 지역을 고르게 선정하였다. 수행평가에서는 학생들이 무작위로 해당 지역의 자료를 뽑아가게 하고 평가를 수행하였다.

첫 번째 문항은 5점 만점이고, 답안을 작성하면 기본 점수가 2점이며, 평가기준에 해당하는 내용을 정확히 서술하면 점수를 부여하도록 표 3과 같이 평가기준을 마련하고 채점을 진행하였다.

2) 창업 계획서 작성

두 번째 문항은 창의성과 논리적 사고를 측정하는 문항으로 제작하였다. 제시된 지역의 기후 특성을 활용한 창업 아이템을 개발하는 것으로, 학생들은 해당 지역에 거주하는 사람의 입장이 되어 그 지역에서 필요로 하는 아이템을 고안해야 한다. 특히, 창업하는 아이템과 해당 지역의 기후조건과의 관련성을 논리적으로 서술하도록 문항을 제시하였다.

두 번째 문항은 창의성과 논리성을 평가하는 문항으로, 완성된 문장의 수준에 따라 점수를 차등적으로 부여하도록 표 4와 같이 평가기준을 마련하고 채점을 진행하였다.

표 2. 수행평가에 제시된 기후 자료 예시(오스트레일리아 퍼스)

월	1	2	3	4	5	6	7	8	9	10	11	12	전년
기온(°C)	24.7	24.8	22.9	19.7	16.2	13.9	13.0	13.4	14.6	17.1	20.1	22.8	18.6
강수(mm)	15.2	16.6	17.6	30.0	79.8	124.7	137.1	120.6	77.7	33.3	28.9	9.3	690.8

표 3. 첫 번째 문항의 평가기준

배점	평가기준
1	쾨펜의 기후구분에 따라 기온 조건을 통해 1차 구분을 정확하게 하였다. (열대, 온대, 냉대, 한대 등)
1	쾨펜의 기후구분에 따라 강수 조건을 바탕으로 2차 구분을 정확하게 하였다. (겨울건조, 여름건조, 연중습윤 등)
1	최난월과 최한월 조건을 바탕으로 해당 지역의 위치를 추론하였다. (남반구 혹은 북반구 수준에서 구분하면 정답 인정)

3. GPT-4 모델 기반인 ChatGPT의 세계 지리 수행평가 채점

프롬프트 엔지니어링(Prompt engineering)이란 사용자가 원하는 출력값을 얻기 위해 인공지능 모델의 입력값을 조작하는 기법으로 ChatGPT로부터 좋은 답변을 얻기 위한 8가지 규칙 및 6가지 전략은 표 5, 표 6과 같다(OpenAI, 2023년 7월 12일자). 8가지 규칙 및 6가지 전략에서 ChatGPT의 세계지리 수행평가 채점을 위해 채점 수행과 관련이 있다고 판단되는 항목을 선정하여 채점

시스템 프롬프트(System Prompt) 작성을 준비하였다. 채점 수행은 최신 모델인 GPT-4를 사용하였으며(규칙 1), 컨텍스트를 ‘###’, ‘<’ 기호를 이용하여 구분하였고(규칙2, 전략1), 수행평가 문항 및 평가기준을 포함하여 채점 시스템 프롬프트를 구체적이고 명확하게 작성하였다(규칙3, 규칙6, 규칙7, 전략2).

또한 프롬프트 내용 및 표현을 바꾸어가며 임의로 뽑은 10명의 학생 답안에 대해 채점을 진행하여 최적의 채점 시스템 프롬프트의 찾는 과정을 거쳤다(전략6). 프롬프트의 시작과 끝에 입력되어있는 정보에 상대적으로

표 4. 두 번째 문항의 평가기준

점수	평가기준
1	해당 지역에서 창업하고자 하는 업종을 제시하고 설명하였다.
2	해당 지역에서 창업하고자 하는 업종을 제시하고, 창업 업종을 기후조건과 관련하여 설명하였다.
3	해당 지역에서 창업하고자 하는 업종을 제시하고, 창업 업종을 기후조건과 관련하여 논리적으로 설명하였다.

표 5. 프롬프트 엔지니어링을 통해 좋은 답변을 받기 위한 8가지 규칙

규칙
(규칙1) 최신 모델을 사용한다.
(규칙2) 프롬프트 시작 부분에 명령을 입력하고, ### 또는 ""를 사용하여 명령과 컨텍스트를 구분한다. ³⁾
(규칙3) 원하는 컨텍스트, 결과, 길이, 형식, 스타일 등에 대해 가능한 구체적이고 상세하게 작성한다.
(규칙4) 예제를 통해 원하는 출력 형식을 표현한다.
(규칙5) 제로샷으로 시작한 다음 퓨샷을 적용하고, 둘 다 작동하지 않는다면 파인튜닝 한다. ⁴⁾
(규칙6) 모호하고 불명확한 설명을 줄인다.
(규칙7) 하지 말아야 할 것을 서술하지 말고, 해야 할 것을 말한다.
(규칙8) (코드 생성 관련) 리딩 어휘를 사용하여 특정 패턴이 나오도록 유도한다.

표 6. 프롬프트 엔지니어링을 통해 좋은 답변을 받기 위한 6가지 전략

전략	세부 전략
(전략1) 명확한 지침을 작성한다.	<ul style="list-style-type: none"> • 관련성 높은 답변을 얻으려면 요청사항에 세부 정보를 포함한다. • 모델에게 페르소나⁵⁾ 채택을 요청한다. • 구분 기호(< >)를 사용하여 입력의 고유한 부분을 명확하게 표시한다.⁶⁾ • 예시를 제공한다. • 원하는 출력 길이를 지정한다.
(전략2) 참조 텍스트를 제공한다.	<ul style="list-style-type: none"> • 모델이 참조 텍스트를 사용하여 대답하도록 지시한다. • 참조 텍스트에서 인용하여 대답하도록 모델에 지시한다.
(전략3) 복잡한 작업을 더 간단한 하위 작업으로 분할한다.	<ul style="list-style-type: none"> • 의도 분류를 사용하여 사용자 요청과 가장 관련성이 높은 지침을 식별한다.⁷⁾ • 대화가 매우 길 경우 이전 대화를 요약하거나 필터링한다. • 긴 문서를 부분별로 요약하고 다시 합쳐서 전체 이야기를 이해한다.
(전략4) ChatGPT에게 생각할 시간을 준다.	<ul style="list-style-type: none"> • 성급하게 결론을 내리기 전에 모델이 스스로 해결할 수 있도록 지시한다. • 내면의 독백 또는 일련의 쿼리를 사용하여 모델의 추론 과정을 숨긴다.⁸⁾ • 모델이 이전 과정에서 놓친 것이 있는지 물어본다.
(전략5) 외부 도구를 사용하여 ChatGPT의 약점을 보완한다.	<ul style="list-style-type: none"> • 임베딩⁹⁾ 기반 검색을 사용하여 효율적인 지식 검색을 구현한다. • 코드 실행을 사용하여 보다 정확한 계산을 수행하거나 외부 API¹⁰⁾를 호출한다. • 모델이 특정 기능에 접근할 수 있는 액세스 권한을 부여한다.
(전략6) 변경 사항을 체계적으로 테스트한다.	<ul style="list-style-type: none"> • 표준 답변을 참조하여 모델 결과를 평가한다.

높은 성능을 보이고, 프롬프트가 길어지면 언어모델의 성능이 저하되는 경향이 있어(Liu *et al.*, 2023) 각 문항을 채점하는 채점 시스템 프롬프트를 분리하였다. 즉, 문항별 채점 시스템 프롬프트를 바탕으로 1번, 2번 문항을 서로 다른 세션에서 채점하여 한 세션에 입력되는 채점 시스템 프롬프트의 길이를 최대한 간결하게 하였다. ChatGPT가 소모하는 토큰 수¹¹⁾를 줄이기 위해 예시 답안 및 학생 답안을 제외한 모든 프롬프트는 영어로 작성하였으며 ChatGPT의 환각 현상을 줄이고 답변의 일관성 확보를 위해 채점 시스템 프롬프트를 수차례 수정한 결과 '역할, 문제, 평가기준(Rubric)' 순서로 채점 시스템 프롬프트를 구성하였다. 이때, 채점 시스템 프롬프트를 한글로 먼저 작성한 뒤 deepl.com에서 영어로 번역하고 영문 표현을 두 연구자가 검토하였다. 한국에서 사용하는 교과 지식의 언어 표현과 영어에서의 표현이 다를 수 있기 때문에 영문 채점 시스템 프롬프트의 번역 표현에 대한 검토 과정을 거쳐야 했다.

한편 ChatGPT는 출력물에 대한 설명이나 추가 텍스트를 요청할 경우 단순히 결과만 제공하는 요청에 비해 극적으로 향상된다(Shakarjian *et al.*, 2023). 따라서 채점 이유 및 부족한 부분에 대한 피드백을 작성하도록 '출력 규칙'을 추가하여 최종적으로 표 7과 같은 시스템 프롬프트를 도출하였다.

OpenAI는 각 세부 전략에 대해 Playground에서 수행해 볼 수 있도록 예시를 제공하고 있다. (전략 4)에서는

ChatGPT의 답변에 영향을 미치는 변수인 Temperature 값을 0으로 설정하여 학생 답안을 채점하는 예시가 있었고, 본 연구는 이 예시와 동일하게 변수값을 설정하였으며 모델은 GPT-4를 선택하였다. 이를 위해 본 연구의 채점 수행은 GPT-4 모델의 다양한 변수값을 조절할 수 있는 OpenAI의 Playground (<https://platform.openai.com/playground>)에서 진행되었으며 사전에 GPT-4의 API를 사용할 수 있도록 권한을 부여받은 유료 계정으로 실시하였다.

1번 문항을 채점할 때 Playground의 'SYSTEM'란에 표 7의 1번 문항 시스템 프롬프트 내용을 입력하였고, 'USER'란에 '<student's answer> [채점하고자 하는 학생의 1번 문항 답안] </student's answer>'를 입력한 뒤 GPT-4 모델의 답변을 출력하였으며 그 예시는 그림 2¹²⁾와 같다. 다음 학생을 채점할 때는 이전 문항에 대한 GPT-4의 채점 결과를 삭제하고, 'USER', 'SYSTEM'란에 채점하고자 하는 대상 학생이 받은 수행평가 문제 및 학생의 답안으로 바꾼 다음 다시 GPT-4 모델의 답변을 출력하였다. 이를 반복하여 표 8의 시스템 프롬프트를 입력했을 때 1번 문항에 대한 27명의 채점 결과를 수집하였고, 2번 문항도 같은 방식으로 진행하였다. SYSTEM란에 작성하는 프롬프트 내용, USER란에 작성하는 학생의 답안을 모두 한글로 작성한 경우 GPT-4 모델의 채점 결과까지 고려하였을 때 토큰수가 너무 길어져 부정확한 채점 결과를 제공할 수 있고, 영문 프롬프트에서의 GPT-4 모델의 기능이 우수하기 때문에 '예시 답안', '학생 답안'



그림 2. Playgorund에서 채점 예시

표 7. 각 문항의 채점 시스템 프롬프트 구성 및 내용

구성	1번 문항 채점 시스템 프롬프트 내용	2번 문항 채점 시스템 프롬프트 내용
역할	###You are a geography teacher. Please grade your (student's answer) to (Question) using (scoring criteria)and output them as (output). answer in korean.###	
문제	<p><Question> [Temperature and Precipitation Data]</p> <p>The temperature in January is 16,1°C with a precipitation of 32,7mm. The temperature in February is 16,8°C with a precipitation of 37,0mm. The temperature in March is 19,1°C with a precipitation of 68,9mm. The temperature in April is 22,7°C with a precipitation of 138,5mm. The temperature in May is 26,0°C with a precipitation of 284,8mm. The temperature in June is 28,0°C with a precipitation of 453,7mm. The temperature in July is 28,6°C with a precipitation of 382,0mm. The temperature in August is 28,4°C with a precipitation of 456,1mm. The temperature in September is 27,6°C with a precipitation of 320,6mm. The temperature in October is 25,3°C with a precipitation of 116,6mm. The temperature in November is 21,9°C with a precipitation of 39,2mm. The temperature in December is 17,8°C with a precipitation of 29,2mm. The average temperature for the year is 23,2°C with a total precipitation of 2359,3mm.</p>	
	<p>Question, Describe the location and climate characteristics of this area. [3 points]</p> <p></Question></p>	<p>Question, Write a business plan utilizing the climate of the area. [case1 : 1point, case 2: 2points, case3 : 3points]</p> <p></Question></p>
평가 기준	<p><scoring criteria></p> <p>### The student's answer to question is scored by adding up the scores from the following three items.###</p> <p>* Correctly identifies the primary distinction based on Köppen's climate classification using temperature conditions (equatorial, warm temperate, snow, polar, etc.) (0 or 1 point)</p> <p>* Correctly identifies the secondary distinction based on Köppen's climate classification using precipitation conditions (winter dry, summer dry, fully humid, etc.) (0 or 1 point)</p> <p>* Infer the location of a region (if student state Southern Hemisphere/Northern Hemisphere or specifically state a city located in the Southern Hemisphere/Northern Hemisphere) based on the conditions of the hottest and coldest months. (0 or 1 point)</p> <p></scoring criteria></p>	<p><scoring Criteria></p> <p>- CASE 1: If the student presents and explains the type of business he/she wants to start in a specific region, award 'just 1 point'.</p> <p>- CASE 2: Award 'just 2 points' if the student presents and explains the type of business he/she wants to start in a specific region and relates it to the climate conditions.</p> <p>- Case 3 : 'just 3 points' will be awarded if the student presents a business that he/she wants to start in a specific region and logically explains the business in relation to the climate conditions.</p> <p></scoring Criteria></p>
출력 규칙	<p><Output></p> <p>//¹³⁾In square brackets []¹⁴⁾ below, 'Evaluate the content of students' answers and provide specific feedback on what they need to do to make their answers appropriate, along with reasons if they're not,' will output ChatGPT's feedback on this content.</p> <p>*Primary distinction through temperature conditions according to Köppen's climate classification: []</p> <p>*Secondary distinction based on precipitation conditions according to Köppen's climate classification: []</p> <p>*Inference of the region's location considering the conditions of the coldest and hottest months: []</p> <p>*Score: [Total score]</p> <p>*Overall feedback on the student's answer [strengths, weaknesses, areas to improve, etc.]</p> <p></Output></p>	<p><Output></p> <p>//Evaluate your student's answers to the three items below, and find the nearest case. Don't award points for any of the three items. In square brackets [] below, 'Evaluate the content of students' answers and provide specific feedback on what they need to do to make their answers appropriate, along with reasons if they're not,' will output ChatGPT's feedback on this content.</p> <p>*Choosing the industry: []</p> <p>*Explanation related to climate: []</p> <p>*Logical explanation: []</p> <p>*Nearest case : [score of that case]</p> <p>*If not Case3: [specific reason for not being categorized as case3]</p> <p></Output></p>

표 8. 예시답안, 비판적 채점 프롬프트로 추가된 내용

구성	1번 문항 시스템 프롬프트 내용 추가	2번 문항 시스템 프롬프트 내용 추가
역할	###You are a geography teacher. Please grade your <students' answer> to <Question> using <scoring criteria> and <example answer> and output them as <output>. answer in korean,###	
문제	... (나머지는 동일) ...	
평가 기준	<scoring criteria> // The student's answer to question is scored by adding up the scores from the following three items. Please grade critically and adhere strictly to the grading criteria below, keeping in mind that the student's answer may be incorrect. ... (나머지는 동일) ... * Incorrect or vague words should also be evaluated. If there are any incorrect terms or ambiguous expressions related to geography, deduct one point accordingly. </scoring criteria>	<scoring Criteria> // Compare the student's answer to the assistant's answer, assigning points to the closest of the three cases below, and grade very carefully and rigorously. Please grade critically and adhere strictly to the grading criteria below, keeping in mind that the student's answer may be incorrect. In particular, "Case 2 is a case in which the student explains the business differently from Case 1, and Case 3 is a case in which the student logically explains the business differently from Case 2." ... (나머지는 동일) ... * Incorrect or vague words should also be evaluated. If there are any incorrect terms or ambiguous expressions related to geography, deduct one point accordingly. </scoring Criteria>
예시 답안	<example answer> 이 지역은 7월과 8월의 기온이 1월과 2월의 기온보다 높게 나타나는 것으로 보아 북반구에 위치하고 있으며, 최한월 평균기온이 -3도에서 18도 사이에 위치하여 온대 기후이며, 강수량을 보면 겨울보다 여름 강수량이 적은 여름건조기후입니다. 즉, 이 지역은 온대여름건조기후(지중해성기후)입니다. </example answer>	<example answer> 이 지역은 Cs 기후로, 겨울에 비가 많이 오고 여름이 건조하며, 온도 변화가 적습니다. 이러한 기후 조건을 활용한 사업 아이디어는 와인 양조장입니다. 여름철 건조함은 포도 재배에 좋은 환경을 제공하며, 겨울에 내리는 비는 포도를 키우는데 필요한 충분한 물을 제공합니다. 또한, 기온 변화가 적은 것이 와인의 질을 유지하는 데 도움이 됩니다. Cs 기후는 미지근하고 적절한 강수량으로 포도 재배에 이상적인 조건을 제공합니다. 와인 양조장은 이러한 기후 조건을 통해 안정적인 수확량과 고품질의 와인을 제공할 수 있어 경쟁력을 유지할 수 있습니다. </example answer>
출력 규칙	<Output> ... (나머지는 동일) ... *Deduction for incorrect or vague terminology: [] </Output>	<Output> *Deduction for incorrect or vague terminology: [] </Output>
비판적 채점 강조	###Please grade critically and adhere strictly to the <scoring Criteria>, keeping in mind that the student's answer may be incorrect. Be very rigorous and demanding in grading students' responses,###	###Please grade critically and adhere strictly to the <scoring Criteria>, keeping in mind that the student's answer may be incorrect. Be very rigorous and demanding in grading students' responses,###

만 한글로 입력하고, 나머지 내용은 영어로 번역하여 작성하였다. 또한, GPT-4 모델의 채점 결과는 한글로 출력하도록 '역할' 프롬프트에서 'answer in korean,'을 입력하였다.

ChatGPT를 비롯한 다수의 생성형 인공지능은 같은 질문에 대한 답을 다르게 생성하기 때문에 한 번은 올바르게 응답해도 같은 질문을 다시 입력했을 때 틀린 답이 나올 수 있다(Plevris et al., 2023). 따라서 각 문항에 대한 27명 학생의 답안을 동일한 방식으로 3회 반복하였고, 이에 대한 답변의 일관성 및 교사 3명의 채점 결과와

비교하고자 하였다.

또한 ChatGPT에게 제공되는 시스템 프롬프트의 정보가 달라짐에 따른 채점 결과의 변화를 관찰하기 위해 채점 시스템 프롬프트에 '예시답안(Example)' 프롬프트만 추가한 경우와 '예시답안, 비판적(Critical) 채점 명령' 프롬프트를 추가하여 표 8과 같이 표현하였고, 각각의 경우에 대해 위와 동일한 방식으로 문항별 채점을 3회씩 반복하였다. 예시답안은 ChatGPT에게 문제를 주고 평가기준에 모두 부합하는 답안을 생성하게 한 뒤 본 연구진이 검토, 수정하여 만들었으며 비판적 채점 명령의 경

우 시스템 프롬프트 마지막에 한 번 더 강조하여 채점을 더 까다롭게 수행할 수 있게 하였다. 정리하면, ChatGPT에게 역할과 문제를 부여하고, 3가지 유형의 프롬프트 즉, '평가기준·예시답안·비판적 채점', '평가기준·예시답안', '평가기준'에 대해 각각 3번씩 채점하여 총 27명의 학생들을 대상으로 9번씩 채점하였다.

4. GPT-4 모델 기반인 ChatGPT의 세계 지리 수행평가 채점

채점자 간 신뢰도(intra-scorer reliability)는 서로 다른 채점자들의 채점 결과가 얼마나 유사한가를 의미하며, 논술형 문항의 채점에서 흔히 사용하는데 최근 학교 현장에서 수행평가의 비중이 높아지면서 채점자 간 신뢰도가 매우 중요하게 인식되고 있다(성태제, 2014; 최진영·하민수, 2023). 채점자 간 신뢰도를 추정하기 위해서는 채점자가 동일한 문항에 응답해야 하고, 상호 독립적이어야 하며, 동일 대상을 평가해야 한다(성태제, 2014). 본 연구에서는 이러한 조건을 충족하여 ChatGPT와 지리 교사의 채점 결과를 분석하였다.

1) 지리교사의 채점

지리 교사의 채점은 2023년 7월 10일~7월 14일까지 교육경력 15년 이상의 지리 교사 3명을 대상으로 1차 채점을 진행하였다. 1차 채점을 위해 27명의 수행평가 답

안지와 평가기준표를 제공하였다. 채점자 간 신뢰도를 높이기 위하여 교사들의 채점 결과를 평균, 표준편차, 최하점, 최고점의 값으로 정리하여 제공하고 2023년 7월 17일~7월 21일까지 최종 점수를 부여할 수 있도록 안내하였다. 평가문항에 대한 교사들의 채점 결과는 표 9와 같으며, 1차 채점과 2차 채점의 결과를 비교해보면 1번 문항보다 주관적 판단이 많이 들어가는 2번 문항에서 채점 결과의 변화가 상대적으로 큰 것을 확인할 수 있다. 특히, 교사3은 2번 문항 2차 채점에서 기후조건과 창업아이템의 논리적 연관성에 대한 근거를 1차보다 유연하게 적용하여 평균 점수가 크게 상승하였다. 즉, 교사들의 채점에서도 주관적 판단이 들어가는 채점은 수정 가능성이 크다는 것을 확인할 수 있다.

3명의 교사 간 1차 채점과 2차 채점에 대한 상관관계는 표 10과 같다. 1번 문항에 대한 1차 채점 상관관계는 .782~.838이고, 2차 채점 상관관계는 .838~.934로 향상되었다. 또한, 2번 문항에 대한 1차 채점 상관관계는 .218~.499이고, 2차 채점 상관관계는 .393~.588로 향상되었다. 1번 문항은 정답에 대한 기준이 명확하여 채점자 간 상관관계가 높게 나타났지만 2번 문항은 창의성과 논리적 타당성에 대해 옳고 그름을 주관적으로 판단해야 하기 때문에 채점자 간 신뢰도가 높지 않게 나타났다. 하지만 두 문항 모두 채점 결과를 공유하여 재채점 한 결과 상관관계가 높아진 것으로 확인되었다. 이는 서술형 평가에서 동료 교사와 채점 결과를 공유하면 채점자 간

표 9. 교사 채점 결과 기술통계

구분	N	1번 문항		2번 문항		
		M	SD	M	SD	
1차	교사1	27	4.48	0.70	4.59	0.75
	교사2	27	4.15	0.99	4.48	0.64
	교사3	27	4.52	0.64	4.48	0.75
2차	교사1	27	4.48	0.70	4.59	0.75
	교사2	27	4.15	0.99	4.44	0.64
	교사3	27	4.41	0.75	4.67	0.55

표 10. 지리교사의 문항 채점 결과

구분	1번 문항		2번 문항	
	1차	2차	1차	2차
교사1-교사2	.838**	.838**	.424*	.393*
교사2-교사3	.782**	.934**	.218	.588**
교사1-교사3	.792**	.852**	.499**	.541**

신뢰도를 높일 수 있다는 주장(성태제, 2014)과 일치한다.

2) GPT-4 모델 기반인 ChatGPT의 채점

채점에 활용한 프롬프트는 3종류로 나누어 실시하였다. 첫 번째 프롬프트는 ‘평가기준(Rubric)’만 제공하였고, 두 번째 기준은 ‘평가기준과 예시답안(Example)’을 함께 제시하였다. ChatGPT의 채점 결과 점수가 대부분 높게 나타나는 것을 확인하고, 마지막 프롬프트는 ‘평가기준, 예시답안, 비판적(Critical) 채점이라는 명령’을 함께 제공하였다. 동일한 질문에 동일한 응답을 하는 것이 아닌 ChatGPT의 특성(Plevris *et al.*, 2023)을 고려하여 동일한 프롬프트에 대해 3번의 채점을 실시하고 그 평균값을 활용하였으며 그 결과는 표 11과 같다. 1번 문항은 ‘평가기준·예시답안’을 제시한 경우(M=4.78), ‘평가기준’만 제시한 경우(M=4.68), ‘평가기준·예시답안·비판적 채점 명령’을 함께 제시한 경우(M=4.65) 순으로 평균 점수가 높게 나타났고, 2번 문항은 ‘평가기준·예시답안·비판적 채점 명령’을 함께 제시한 경우(M=4.89), ‘평가기준’만 제시한 경우(M=4.84), ‘평가기준·예시답안’을 제시한 경우(M=4.77) 순으로 평균값이 높게 나타났다.

3) 교사와 GPT-4 모델 기반인 ChatGPT의 채점 결과 비교

교사의 2차 채점 결과와 ChatGPT의 3가지 유형 프롬프트 결과에 대해 상관관계 분석을 실시하였으며 그 결과는 표 11과 같다.

1번 문항은 평가기준을 활용하여 학생의 정답에 대해 비교적 정확하게 옳고 그름을 판단할 수 있는 문항이다. 교사의 2차 채점 결과와 상관관계 결과를 보면 ‘평가기준·예시답안·비판적 채점 명령’을 함께 제시한 경우($r=.672$), ‘평가기준·예시답안’을 함께 제시한 경우($r=.610$), ‘평가기준’만 제시한 경우($r=.516$) 순으로 상관관계가 높게 나타났다. 채점자 간 상관성이 .6 이상이면 신뢰도

가 확보된다고 보기 때문에(성태제, 2014), 정답이 명확하게 있는 서·논술형 문항의 경우 평가기준·예시답안 또는 비판적 채점 명령을 추가하여 ChatGPT로 채점하면 1명의 동료 교사와 함께 채점하는 역할을 수행할 수 있다고 판단된다.

2번 문항은 학생들의 답안을 통해 창의성과 논리성을 판단하여 채점을 해야 하는 문항이다. 교사의 2차 채점 결과와 상관 관계 결과를 보면, ‘평가기준·예시답안’을 제시한 경우($r=.521$), ‘평가기준·예시 답안·비판적 채점 명령’을 함께 제시한 경우($r=.517$), ‘평가기준’만 제시한 경우($r=.304$) 순으로 높게 나타났다. 학생 답변의 논리성과 창의성을 묻는 경우 ‘비판적 채점’에 대한 명령문은 큰 역할을 하지 못하는 것으로 판단되며, 오히려 상관관계가 낮아지는 결과를 보였다. 논리성과 창의성을 물어보는 문항에서 비판적 채점 명령은 적합하지 않은 명령문이라고 판단된다. 그리고 주관적 판단이 많이 포함되는 문항의 경우 채점자 간 신뢰도를 확보하는 상관관계 .6보다 낮게 나타났지만, 교사 3인의 2번 문항 채점에 대한 상관 계수인 .398~.588과 비교해보면 평가기준·예시답안을 제시한 프롬프트 또는 비판적 채점 명령을 추가하는 프롬프트를 활용한 경우 교사들과 비슷한 상관관계를 보이는 것을 확인할 수 있다.

5. GPT-4 모델 기반인 ChatGPT의 피드백 결과 분석

ChatGPT가 제시한 프롬프트를 활용하여 채점을 진행하면 ChatGPT는 표 12와 같이 판단 근거와 함께 점수를 제시한다. 각각의 평가 요소별로 평가기준을 제시하는 분석적 채점방식을 사용하였기 때문에 ChatGPT는 각각의 요소에 따른 점수와 점수를 부여한 근거를 제시한다. 점수를 부여한 근거를 바탕으로 ChatGPT의 채점 오류를 판단할 수 있다.

표 11. GPT-4 채점 결과 및 GPT-4 채점과 교사 채점과의 상관관계

구분	N	1번 문항		2번 문항	
		평균	교사 채점과의 상관 계수	평균	교사 채점과의 상관 계수
R	27	4.68	.516**	4.84	.304
R·E	27	4.78	.610**	4.77	.521**
R·E·C	27	4.65	.672**	4.89	.517**

* R: 평가기준 제시, E: 예시답안 제시, C: 비판적 채점 명령 제시

표 12. ChatGPT의 채점 결과 예시

(Output)

*Primary distinction through temperature conditions according to Köppen's climate classification: [학생의 답변은 온대 기후의 특징을 잘 설명하였습니다. 매월의 기온이 18도 이상이지 않고, 최한월의 평균 기온이 -3도 이상이라는 점을 통해 온대 기후임을 정확하게 파악하였습니다. (1점)]

*Secondary distinction based on precipitation conditions according to Köppen's climate classification: [학생은 겨울철에 강수량이 여름철보다 비교적 많다는 점을 통해 여름이 건조한 지중해성 기후임을 정확하게 파악하였습니다. (1점)]

*Inference of the region's location considering the conditions of the coldest and hottest months: [학생은 지역의 위치를 위도상으로 남북위 30도에 주로 분포한다고 했지만, 가장 더운 달과 가장 추운 달의 조건을 고려하여 지역의 위치를 추론하는 것이 문제의 요구사항이었습니다. 따라서 이 부분에 대한 점수는 부여할 수 없습니다. (0점)]

*Score: [2점]

*Overall feedback on the student's answer: [학생의 답변은 기후 분류에 대한 이해를 잘 보여줍니다. 특히, 온대 기후와 지중해성 기후의 특징을 잘 설명하였습니다. 그러나, 가장 더운 달과 가장 추운 달의 조건을 고려하여 지역의 위치를 추론하는 것이 문제의 요구사항이었으므로, 이 부분에 대한 이해와 답변이 필요합니다.]

(/Output)

ChatGPT를 활용하여 평가기준·예시답안·비판적 채점 명령을 제시한 프롬프트로 3번 채점한 평균값과 교사 3명의 평균값을 비교하여 1점 이상 점수 차이가 나는 채점 결과를 비교한 결과 ChatGPT의 오류 경향을 확인할 수 있었다. 첫째, 영어를 번역하는 과정에서 나타나는 오류이다. ChatGPT는 최한월(the temperature of the coldest month)을 '최난월'로 표현한 경우도 옳은 답이라고 채점하거나, '겨울 기온'이라는 정확하지 않은 표현을 사용한 경우에도 정답으로 인식하였다. 둘째, 축약해서 표현한 단어를 사용한 경우 인식하지 못하는 오류이다. 기후구분에서 지중해성 기후는 Cs, 사바나 기후는 Aw 등으로 표현하는데, ChatGPT가 채점한 경우 이러한 약어를 인식하지 못하고 옳지 않은 답으로 인식하였다. 학생이 'Cs'로 표현한 답을 '지중해성 기후'로 변경해서 채점한 결과 정답으로 인식하였다. 셋째, 논리적 판단의 오류이다. 특히, 2번 문항에서 창업 아이템과 기후를 논리적으로 연계했는지를 판단하는 채점에서 사바나 기후의 장소와 단목이 형성되는 것이 기후와 관련이 없다고 판단하고 감점 처리 하는 등 전체적인 맥락에서 판단하지 못하고 서술되어 있는 일부분에 대해서만 논리성을 판단하였다. 넷째, 우리나라에서 가르치는 내용과 영어권 자료의 차이에서 오는 오류이다. ChatGPT는 2021년 9월까지의 온라인 정보를 바탕으로 질문에 응답을 하며, 기본적으로 영어권 자료가 더 많이 활용된다. 우리나라 교육과정에서 온대기후와 냉대기후를 구분하는 기준을 최한월 평균기온 -3°C로 하지만 영문판 Wikipedia(2023년 8월 8일자)나 일부 영어권 논문에서 쿼펜의 기후구분을

일부 수정하여 온대와 냉대기후를 구분하는 기준을 -3°C가 아닌 0°C를 기준을 사용하고 있다(Peel, *et al.*, 2007). 이에 최한월 기준을 틀리게 쓴 학생들의 답안을 정답으로 인식하는 오류를 많이 범하였다. 가르치는 내용에 있어 국가별로 차이가 있는 경우, 영어권 온라인 자료를 기반으로 하는 ChatGPT가 채점의 오류를 일으킬 가능성이 있다.

그리고 학생 답안에 대한 ChatGPT의 피드백 결과는 표 13과 같다. ChatGPT의 피드백 내용을 보면, 정답을 적은 학생에게는 답안이 평가기준에 어떻게 부합하였는지를 설명하고, 학생들에게 칭찬을 포함한 긍정적인 표현을 많이 제시하였고, 틀린 답을 작성한 학생들에게는 답안의 오류를 지적하고, 학습의 방향을 안내하고 있다. 특히 비교적 명확한 평가기준이 있는 1번 문항의 경우 학생들이 틀리게 진술한 부분을 구체적으로 설명하고 있어 실제 학교 현장에서 교사가 채점 후 학생들에게 점수를 확인해줄 때 중요한 정보를 제공할 수 있을 것으로 판단된다. 창의성과 논리성에 대한 주관적 판단이 필요한 2번 문항의 경우 학생들 진술의 논리적 부족을 제시하고 있지만, 그 내용이 구체적이지 않아 1번 문항과 같이 비교적 명확한 평가기준이 있는 문항보다 교사의 채점에 대한 세심한 확인이 필요하다고 판단된다.

김승주(2022)의 연구에 의하면 글쓰기 자동채점 연구의 주요 키워드를 분석한 결과, 피드백이 두 번째로 높은 빈도를 보였고, 이는 채점의 자동화와 관련된 연구들에서 학생들에게 제공하는 피드백이 매우 중요한 연구 주제로 다루어졌음을 시사한다. 이처럼 ChatGPT를 이용

표 13. 학생 답안에 대한 ChatGPT의 피드백 예시

문항	피드백 내용
1번	학생의 답변은 매우 잘 작성되었습니다. 기후 분류를 올바르게 이해하고, 그에 따른 지역의 위치를 정확하게 추론하였습니다. 이러한 분석력은 지리학적 사고력을 향상시키는 데 큰 도움이 됩니다. 계속해서 이런 논리적인 사고를 유지하면서 학습해 나가기를 바랍니다.
	학생의 답변은 대체로 잘 작성되었습니다. 궤선의 기후 분류에 따른 기온과 강수량 조건을 올바르게 사용하여 이 지역의 기후 특성을 잘 설명하였습니다. 그러나 이 지역이 북반구인지 남반구인지에 대한 구체적인 언급이 없어, 이 부분에 대한 피드백을 제공합니다. 북반구와 남반구의 기후 특성은 다르기 때문에, 이를 구분하는 것이 중요합니다.
	학생의 답변은 기후 분류와 지역의 위치를 대체로 잘 이해하고 있습니다. 그러나 강수량에 따른 기후 분류에 대한 이해가 부족한 것으로 보입니다. 이 부분에 대해 더 학습하고 이해하는 것이 필요합니다. 또한, 기후에 따른 생활 방식에 대한 설명은 문제의 요구사항에 포함되지 않았으므로, 이 부분은 점수에 반영되지 않았습니다.
2번	기후와 강수량이 토착민의 생활에 어떻게 영향을 미치는지, 그리고 이를 어떻게 관광 산업에 활용할 수 있는지에 대한 논리적인 설명이 필요합니다
	사업 아이디어와 기후 조건 사이의 논리적인 연결성을 충분히 설명하지 않았기 때문입니다. 기후 조건이 사업 아이디어에 어떻게 영향을 미치는지에 대한 논리적인 설명이 필요합니다.
	사업의 경제적 가치와 고객들이 이 제품을 선택해야 하는 이유에 대한 논리적인 설명이 부족합니다.

한 서술형 평가의 자동 채점 가능성을 논하는 것은 미래의 각종 평가에서 학습자에게 개별화된 피드백을 수시로 제공할 수 있다는 점에서 의미가 있다.

IV. 결론 및 제언

2022 개정교육과정 총론(교육부, 2022)에서는 다양한 경험과 지식을 바탕으로 창의적으로 문제를 해결하고 문제상황에 대처하는 능력을 길러주며, 이러한 학습의 과정을 평가하기 위하여 수행평가 및 서·논술형 평가 비중을 확대할 필요성을 제시하였다. 서·논술형 평가는 흥미, 호기심, 문제해결력을 향상시켜주는 면에서 학생들에게 긍정적인 평가를 받고 있으며(김남준·배종수, 2006; 황현숙, 2012), 교수학습에 필요한 정보 및 학생 이해 자료를 수집할 수 있다는 점에서 교사들도 긍정적으로 인식하고 있다(김래영·이민희, 2013). 이렇듯 학교 현장에서는 서·논술형 평가를 긍정적으로 생각하지만, 답안을 채점하는 데 시간이 많이 소요되고, 평가에 채점자의 주관적 판단이 개입될 가능성이 크기 때문에 채점의 신뢰도 확보에 문제가 있어 선택형이나 단답형 평가에 비해 외면받고 있는 것이 현실이다(박도순·홍후조, 2011). 이에 교사의 주관적 판단 개입이 없고, 채점 시간이 많이 필요하지 않은 자동채점을 도입해야 한다는 목소리가 커지고 있어 본 연구에서는 최근 등장한 ChatGPT를 활용하여 서·논술형 평가를 채점하고 그 가

능성에 대해 논하고자 한다.

본 연구는 세계지리 정규 수업 시간에 실시한 기후 단원의 서·논술형 평가문항을 활용하여 교사의 채점과 ChatGPT 채점 결과를 비교, 분석하였다. 1번 평가문항은 해당 지역의 기온과 강수 자료를 활용하여 해당 지역의 기후를 분석하고, 위치를 추론하는 문항으로 비교적 정답이 명확한 문항이고, 2번 평가문항은 해당 지역의 기후 특성을 활용한 창업 아이디어를 제안하는 것으로 창의력과 논리적 사고를 판단해야 하기에 채점에 주관적 판단이 더 많이 들어가는 문항이다. 이 문항에 대한 학생 27명의 답안을 15년 이상 경력의 지리 교사 3명이 7월 10일~14일에 1차 채점을 진행하고, 1차 채점 결과의 평균, 표준편차 등의 자료를 공유한 후 7월 17일~21일에 최종 채점을 진행하였다. 프롬프트 개발 및 ChatGPT의 채점은 7월 12일~8월 12일에 GPT-4 모델을 이용하여 수행하였다. OpenAI에서 제안한 좋은 답변을 얻기 위한 8가지 원칙과 6가지 전략을 활용하여 채점 프롬프트를 제작하고, 평가기준만 제시한 경우, 평가기준·예시답안을 제시한 경우, 평가기준·예시답안·비판적 채점 명령을 제시한 경우로 나누어 채점을 진행하고, 각각의 경우에 대해 3번씩 채점하여 평균한 값을 활용하였다. 이렇게 얻어진 지리 교사 3명의 2차 채점 결과의 평균값과 ChatGPT 채점 결과의 평균값을 활용하여 상관관계를 분석하였다.

분석 결과를 보면, 정답이 비교적 명확한 1번 문항은 평가기준·예시답안·비판적 채점 명령을 제시한 경우

($r=.672$), 평가기준·예시답안을 제시한 경우($r=.610$), 평가기준만 제시한 경우($r=.516$) 순으로 상관관계가 높게 나타났다. 채점자 간 상관 관계가 .6 이상의 경우 채점자 간 신뢰도가 확보된다고 보는데, 정답이 비교적 명확한 문항은 평가기준·예시답안·비판적 채점 명령을 제시하거나 평가기준·예시답안을 제시하여 GPT로 채점한 경우 교사 1인과 함께 채점하는 효과를 얻을 수 있다고 판단된다. 상상력, 논리적 사고 등을 평가하는 2번 문항은 평가기준·예시답안을 제시한 경우($r=.521$), 평가기준·예시답안·비판적 채점 명령을 제시한 경우($r=.517$), 평가기준만 제시한 경우($r=.304$)순으로 상관관계가 높게 나타났다. 채점자 간 상관 관계가 .6미만으로 나왔지만, 2번 문항에 대한 지리 교사 간 상관 관계가 $r=.398\sim.588$ 인 것을 감안하면, 평가기준·예시답안·비판적 채점 명령을 제시한 경우와 평가기준·예시답안을 제시한 경우 교사들 간의 상관관계와 유사한 상관을 보이는 것을 확인할 수 있다. 즉, ChatGPT를 활용하여 서·논술형 평가 문항을 채점할 경우 OpenAI에서 제안한 좋은 답변을 얻기 위한 8가지 원칙과 6가지 전략을 바탕으로 평가기준·예시답안·비판적 채점 명령을 제시하거나, 평가기준·예시답안을 제시하면 교사 1인이 채점을 하는 결과를 확보할 수 있다. 서·논술형 채점은 2인 이상의 교사가 채점하고 그 결과를 공유하여 점수를 조정하는 절차가 필요하다. 하지만 2022 개정교과과정에서 도입되는 고교학점제에서는 학생들의 다양한 과목 선택을 확보하기 위해 여러 과목이 개설되고 그 결과 과목당 학생 수가 많지 않으며, 이는 한 과목에 여러 명의 교사가 함께 가르치는 경우가 많음을 의미한다. 이러한 상황에서 서·논술형 평가의 확대가 요구되고 있기 때문에 ChatGPT를 활용한 채점은 학교 현장에서 평가의 신뢰도를 향상시키고, 교사의 채점에 들어가는 시간을 줄여줄 수 있는 역할을 해줄 것으로 기대한다.

그리고 ChatGPT의 채점에 있어서 발생하는 오류의 원인을 보면 영어를 번역하는 과정에서 발생하는 오류, 축약해서 표현하는 단어를 이해하지 못하는 오류, 논리적 판단의 오류, 우리나라의 교육과정 내용과 영어권 온라인 자료의 차이에서 생기는 오류 등을 확인할 수 있었다. 이러한 오류는 ChatGPT에게 제공하는 프롬프트를 수정하는 등의 방법을 통해 오류의 가능성을 줄일 수 있다고 판단되며, 이를 후속 연구로 제안하고자 한다.

본 연구는 ChatGPT가 단독으로 서·논술형 채점을 할

수 있는 가능성에 대한 탐색이 아니라, 교사와 함께 채점을 할 경우 보조 교사로서의 역할에 대한 가능성을 제안한 것이다. 교사의 다교과 지도에 따른 업무 과중을 줄여주고 평가의 신뢰도 확보를 위한 ChatGPT의 활용을 위해 후속 연구를 위한 제언은 다음과 같다. 첫째, 인공지능이 인간 교사 수준으로 채점 수행 결과를 출력할 수 있도록 프롬프트 엔지니어링에 대한 지속적인 연구가 필요하다. 인공지능 모델 및 인공지능 모델의 업데이트 시기에 따라 새로운 프롬프트 엔지니어링이 필요할 가능성이 높고, 본 연구에서 나타난 ChatGPT 오류 경향에 대해 프롬프트를 수정하여 보다 정교한 채점이 가능할 수 있는 프롬프트 개발 연구가 필요하다. 둘째, 인공지능을 활용하여 채점하고자 하는 다양한 유형의 문항들에 대한 자동 채점 시도가 필요하다. 본 연구에서는 주어진 자료를 바탕으로 교과 지식을 활용하여 답하는 닫힌 문제와 교과 지식을 바탕으로 해당 기호와 연관된 창업 아이템을 논리적으로 서술하는 열린 문제를 개발하여 진행했지만, 보다 다양한 교과에서 다양한 유형의 평가를 시도하고 연구 결과를 정리할 필요가 있다. 특히, 채점의 오류 경향을 고려하여 자동 채점이 유용한 평가 내용과 방법에 대해 탐색하고, 그 구체적 사례를 찾아보는 연구가 필요하다. 셋째, 자동 채점과 함께 학생에게 개별화된 피드백을 제공할 수 있는 방안에 대한 연구가 함께 진행되어야 한다. 자동 채점의 목적은 단순히 교사의 채점 수행을 돕기 위함이 아니라 학생들의 학습 상태를 진단하고 진단 결과에 따른 피드백을 제공하여 학생들의 학습을 돕기 위함이기 때문이다. 이러한 제언을 바탕으로 학교 현장에서 ChatGPT가 교사의 보조 수단으로 활용될 수 있도록 활발한 후속 연구가 진행되기를 기대한다.

註

- 1) 환각(Hallucination)현상이란 생성형 인공지능 언어모델이 사용자의 질문에 대해 부정확하거나 허구의 정보를 제공하는 것을 말한다. 일례로 사용자의 질문에 대해 존재하지 않는 논문을 인용하여 답하는 경우, 조선시대에 컴퓨터를 사용하였다는 등의 역사적 사실 왜곡이 있다. 현재 2023년 9월에는 GPT 모델이 업데이트 되면서 환각 현상이 많이 감소하였지만 인공지능의 답변에 대한 사

용자의 확인이 반드시 필요하다.

- 2) Temperature와 같이 모델의 답변에 영향을 미치는 변수 값들은 ChatGPT 홈페이지인 <https://chat.openai.com/>에서 사용자가 상황에 맞게 변경하는 것이 불가능하며 ChatGPT의 기반이 되는 GPT 모델을 직접 사용하는 곳인 Playground(<https://platform.openai.com/playground>)에서만 설정이 가능하다. Temperature는 0에서부터 2까지의 값을 설정할 수 있으며 값이 클수록 다양한 결과물을 출력한다.
- 3) ###, "" 기호는 프롬프트의 내용이 많은 경우 GPT 모델이 이를 잘 이해하게 만들기 위해 프롬프트 내용을 문단별로 구분하는데 사용한다.
- 4) 제로샷(Zero-shot)이란 GPT 모델에게 예시를 전혀 제공하지 않고 사용자의 명령을 수행하도록 하는 방식이고, 퓨샷(Few shot)은 GPT 모델이 수행해야 하는 것에 대한 몇 가지 예시를 제공하여 사용자의 명령을 수행하도록 하는 방식이다. 파인튜닝(Fine-tuning)이란 사전 학습된 모델을 새로운 데이터 집합에 대해 답할 수 있도록 모델을 업데이트하는 방식이다.
- 5) 페르소나란 ChatGPT에게 부여하는 역할을 말한다. 예컨대 ‘당신은 고등학교 2학년 학생들을 가르치는 지리교사이다.’와 같이 ChatGPT의 역할을 부여하고 구체적인 지침을 작성한다.
- 6) (규칙2)와 유사하게 (전략1)에서도 구분기호 ◇를 이용하여 프롬프트 내용의 문단을 구분할 수 있다. <https://platform.openai.com/docs/guides/gpt-best-practices/strategy-write-clear-instructions>에서 제시한 예를 보면 HTML과 같이 웹 페이지의 구조와 내용을 나타내는 방법과 유사하게 문단의 시작과 끝을 각각 ◇, </>를 이용하여 구분하고 있다.
- 7) 사용자의 요청이 어떤 범주와 가장 관련성이 높은지를 파악하고, 이에 대한 답변을 출력하게 하는 방식이다. <https://platform.openai.com/docs/guides/gpt-best-practices/strategy-split-complex-tasks-into-simpler-subtasks>
- 8) 사용자의 요청에 대한 답을 출력하기 위해 GPT 모델이 수행하는 추론 과정을 "" 기호 안에 작성하게 하여 사용자에게 최종적으로 제공해야 할 결과물과 추론 과정을 구분한다. 즉, GPT 모델의 내

부 독백을 "" 기호 안에 작성하여 사용자에게 보여지지 않게 하고, 독백을 바탕으로 한 결과물만 출력하게 한다.

- 9) 임베딩은 데이터의 내용 및 의미를 보존하기 위한 데이터 조각을 숫자 형태인 벡터로 표현한 것이다. 벡터의 형태로 데이터가 표현되면 단어들 사이의 유사도를 계산할 수 있어 단어 사이의 관계를 도출할 수 있다.
- 10) API란 다양한 소프트웨어 간의 상호작용을 가능하게 하는 도구로 ChatGPT와 외부 서비스의 기능을 함께 이용할 때 사용한다.
- 11) <https://platform.openai.com/tokenizer>에서 토큰 수를 계산할 수 있다. ‘Hello, I am a geography teacher’의 경우 7 토큰, ‘안녕하세요. 저는 지리교사입니다.’의 경우 40 토큰이 소모된다. 두 문장은 같은 의미를 담고 있지만 영어로 작성한 경우 토큰 수가 더 적다. GPT-4 모델은 8,192 토큰, GPT-3.5-turbo 모델은 4,096 토큰까지 대화를 기억할 수 있기 때문에 프롬프트가 길어지는 경우 영어로 프롬프트를 입력해야 한다.
- 12) ①SYSTEM란에 <역할>, <문제>, <평가기준>, <예시답안> 등의 프롬프트를 작성하고, ②USER란에 학생이 작성한 <학생 답안>과 시스템 프롬프트에 대한 요약문을 넣었다. submit 버튼을 누르면 ③ASSISTANT란에 GPT-4 모델의 답변이 생성된다. 한 명의 학생에 대한 채점이 끝나면 방금 출력된 ASSISTANT, USER란을 삭제한 뒤, 그 다음 순서 학생의 <예시답안>을 SYSTEM란에 수정하고, <학생 답안>을 USER란에 입력하여 위 순서를 반복하였다. 그림 오른쪽에서 ④GPT 모델, ⑤temperature 등의 변수값 설정이 가능하다.
- 13) ‘//’, ‘#’는 프로그래밍 언어에서 주석 기호로 사용된다. <output>에 있는 내용을 부연 설명하는 것으로 ChatGPT가 결과물을 출력할 때 ‘//’이 있는 내용은 출력하지 않고, 참고하게 하고자 해당 기호를 사용하였다.
- 14) ‘[]’는 ChatGPT의 채점 결과를 ‘[]’안에 작성하도록 하기 위해 사용하였다. ChatGPT가 출력하는 모든 학생들의 채점 결과가 같은 구조로 출력되어야 일관된 분석을 할 수 있다고 판단하였고, 모든 학생의 출력 결과는 단 한 명도 빠짐없이 해당 구조로 채점 결과를 정확히 출력하였다.

참고문헌

- 강원석 2018, “서술형 문항 채점을 위한 복합문 구문 의미분석 시스템에 대한 연구” 컴퓨터교육학회 논문지, 105-115.
- 교육부, 2022, 「초·중등학교 교육과정 총론(교육부 고시 제 2022-33호 별책 1)」.
- 국가정보원 2023, 「챗GPT 등 생성형 AI활용 보안 가이드라인」, 국가사이버안보센터.
- 권오남·오세준·윤정은·이경원·신병철·정원 2023, “ChatGPT의 수학적 성능 분석: 국가수준 학업성취도 평가 및 대수학능력시험 수학 문제 풀이를 중심으로” 수학교육 논문집, 37(2), 233-256.
- 김기남·성정원 2023, “세계지리 평가가 학생들의 지역 편견 형성에 미치는 영향” 한국지리학회지, 12(2), 335-346.
- 김기철·김상범·서지영·백경선·이은경·이미영·이수정·이승미·조기희, 2023. 「고교학점제 도입·운영 안내서(연구자료 ORM 2023-25)」, 교육부·교육과정평가원.
- 김남준·배종수, 2006, “서술형 평가가 초등학생의 수학적 성향에 미치는 영향 연구” 한국초등수학교육학회지, 10(2), 195-219.
- 김래영·이민희, 2013, “수학과 서술형 평가에 대한 중학교 교사들의 인식연구” 수학교육학연구, 23(4), 533-551.
- 김승주, 2022, “키워드 네트워크 분석 방법을 활용한 국외 글쓰기 자동채점 연구 동향 분석-인문학 및 사회과학 영역의 학술지 논문을 중심으로” 국어교육연구, 80, 45-93.
- 김인석·김봉규, 2020, “인공지능형 대화형 챗봇 현황과 영어 교육용 챗봇 개발 방안” 영어영문학 21, 33(2), 73-92.
- 김효선·오영열, 2014, “서술형 수학 쓰기 수업이 초등학생의 문제해결 및 수학적 성향에 미치는 효과” 수학교육 논문집, 28(1), 131-154.
- 노은희·심재호·김명화·김재훈, 2012, 「대규모 평가를 위한 서답형 문항 자동채점 방안 연구 (연구보고 RRE 2012-6)」, 서울: 한국교육과정평가원.
- 노은희·송미영·박종임·김유향·이도길, 2016, 「한국어 문장 수준 서답형 문항 자동채점 프로그램 고도화 개발 및 적용 (연구보고 RRE 2016-11)」, 서울: 한국교육과정평가원.
- 박도순·홍후조, 2011, 「교육과정과 교육평가」, 서울: 문음사.
- 박선미, 1999, “지리교육의 평가방법 재구조화” 대한지리학회 학술대회논문집, 155-162.
- 박종임·이상하·송민호·이문복·이민정·최속기, 2022, 「컴퓨터 기반 서술형 평가를 위한 자동채점 방안 설계(I) (연구보고 RRE 2022-6)」, 진천: 한국교육과정평가원.
- 성태제, 2014, 「문항제작 및 분석의 이론과 실제」, 서울: 학지사.
- 손태권, 2023, “ChatGPT의 수학교육 활용 가능성 탐색: 분수 문제에 관한 학생의 산출물과 예비교사의 담화 사례를 중심으로” 초등수학교육, 26(2), 99-113.
- 신동광·정혜경·이용상, 2023, “내용중심 영어 교수 학습의 도구로서 ChatGPT의 활용 가능성 탐색” 영어교육교육, 22(1), 171-192.
- 양혜진·김혜영·신동광·이장호, 2019, “인공지능 음성챗봇 기반 초등학교 영어 말하기 수업 연구” Multimedia-Assisted Language Learning, 22(4), 184-205.
- 오정현, 2006, “고등학교 역사과 서술형·논술형 평가의 특징과 개선 방안” 역사교육연구, 4, 109-157.
- 유재진, 2023, “인공지능을 활용한 지리교육 연구 -ChatGPT 기반 질의·응답을 중심으로-” 한국사지리학회지, 33(1), 162-173.
- 윤여범·박미애, 2020, “인공지능과 초등영어교육: 챗봇의 현황과 발전 방향을 중심으로” 한국초등교육, 31(특별호), 77-90.
- 이동한, 2018, “인공지능을 활용한 영어 학습용 챗봇 시스템 개발 방안 연구” 중등영어교육, 11(1), 45-68.
- 이진화·조인정, 2016, “지리공간기술의 지리교육현장 적용에 대한 예비교사들의 태도 연구” 한국지리학회지, 5(2), 85-97.
- 장유정, 2020, “[윤리와 사상] 서술형 평가에 대한 학생 인식 조사” 도덕윤리교육, 68, 195-222.
- 장진아·박준형·박지선, 2021, “인공지능 챗봇 관련 국내 연구 동향 및 챗봇 활용 현황 분석: 과학 교육에서의 활용을 위한 시사점을 중심으로” 학습자중심교과교육연구, 21(13), 729-743.
- 조경철, 2012, “지리교육과정의 성취목표와 국가수준학업성취도 평가의 평가목표의 비교분석: Bloom의 신교육 목표분류학에 근거하여” 한국지리학회지, 1(1), 19-31.
- 지은림, 2000, “논술형 수행평가를 위한 채점방법들의 비교” 경희대학교 교육문제연구소 논문집, 16, 235-246.
- 진경애·남명호·김명화·오상철·김민정·주형미, 2006, 「서답형 문항 자동채점 프로그램 도입 방안 연구(I) (연구보고 RRI 2006-6)」, 서울: 한국교육과정평가원.

- 추성엽·민덕기, 2019, “영어 상호작용 촉진을 위한 과업 기반 AI 챗봇 활용 및 학생 발화 분석” *초등영어교육*, 25(2), 27-52.
- 최원경, 2020, “AI 챗봇을 활용한 초등영어 과업중심 말하기 평가” *초등영어교육*, 26(1), 131-152.
- 최진영·하민수, “2023, 국어과 읽기 영역 서술형 평가를 위한 비지도 기반 인공지능 채점 보고 프로그램(SAAI)의 성과와 활용도 탐색”, *청람어문교육*, 92, 7-48.
- 홍선호·윤택남·이삭·오은진, 2021, “초등영어교육에서 음성 대화형 챗봇 활용 방법론 탐색 및 효과”, *한국초등교육*, 31(5), 31-55.
- 황정규·서민원·최종근·김민성·양명희·김재철·강태훈·이대식·김준엽·신종호·김동일, 2016, 「교육평가의 이해」, 서울: 학지사.
- 황현숙, 2012, “중학교 1학년 과학에서 서술형 수행평가가 자기주도적 학습능력과 과학적 학습태도에 미치는 효과”, *한국교원대학교 석사학위논문*.
- Crawford, J., Cowling, M., and Allen, K.A., 2023, Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI), *Journal of University Teaching & Learning Practice*, 20(3), 1-19.
- Gao, C.A., Howard, F.M., Markov, N.S., Dyer, E.C., Ramesh, S., Luo, Y., and Pearson, A.T., 2022, Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers, *BioRxiv*, 1-18.
- Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P., 2023, Lost in the middle: How language models use long contexts, *arXiv preprint arXiv:2307.03172*, 1-19.
- OpenAI, 2023, GPT-4 technical report, *arXiv:2303.08774*, 1-100.
- Peel, M.C., Finlayson, B.L., and McMahon, T.A., 2007, Updated world map of the Köppen-Geiger climate classification, *Hydrology and earth system sciences*, 11(5), 1633-1644.
- Plevris, V., Papazafeiropoulos, G., and Rios, A. J., 2023, Chatbots put to the test in math and logic problems: A preliminary comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard. *arXiv preprint arXiv:2305.18618*. 1-20.
- Qadir, J., 2023, Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education, *In 2023 IEEE Global Engineering Education Conference (EDUCON)*, 1-9.
- Shakarian, P., Koyyalamudi, A., Ngu, N., and Mareedu, L., 2023, An Independent Evaluation of ChatGPT on Mathematical Word Problems (MWP), *arXiv preprint arXiv:2302.13814*, 1-9.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I., 2017, *Advances in neural information processing systems*, 1-15.
- MIT Technology Review, 2023년 8월 7일자, “ChatGPT is going to change education, not destroy it,” <https://www.technologyreview.kr/chatgpt-is-going-to-change-education-not-destroy-it>
- OpenAI, 2023년 7월 12일자, “Documentation”, <https://platform.openai.com/docs/introduction>
- OpenAI Playground, <https://platform.openai.com/playground>
- Wikipedia, 2023년 8월 8일자, https://en.wikipedia.org/wiki/K%C3%B6ppen_climate_classification
- 교신 : 성정원, 16227, 경기도 수원시 영통구 창룡대로 263, 수원외국어고등학교(이메일: seonwoodady@korea.kr)
- Correspondence: Jeongwon-Seong, 16227, 263 Changryong-daero, Yeongtong-gu, Suwon-si, Gyeonggi-do, Republic of Korea, Suwon Academy of World Languages (Email: seonwoodady@korea.kr)

투고접수일: 2023년 8월 21일
심사완료일: 2023년 9월 5일
게재확정일: 2023년 9월 18일