JPEE 실천공학교육논문지
JOURNAL OF PRACTICAL ENGINEERING EDUCATION

에듀테크
J. Pract. Eng. Educ. 16(5), 663-676, 2024

대학 글쓰기 평가에서 인공지능 도구의 활용 가능성 탐색: 인간과 생성형 AI 간 평가 기준 비교

Exploring the Potential of AI Tools in University Writing Assessment: Comparing Evaluation Criteria between Humans and Generative AI

박소영[1], 이병윤[2]*

[1]숙명여자대학교 교육학부, [2]숙명여자대학교 교육연구소

So-Young Park[1], ByungYoon Lee[2]*

[1]Division of Education, Sookmyung Women's University, Seoul 04310, Korea
[2]Education Research Institute, Sookmyung Women's University, Sookmyung Women's University, Seoul 04310, Korea

[요 약]

본 연구는 Learning with AI 관점에서 출발하여, 인공지능이 생성한 글쓰기 평가 기준의 교육적 활용 가능성을 탐색하고자 하였다. 구체적으로, 인공지능이 생성한 평가 기준과 인간이 개발한 기준 사이의 공통점과 차이점을 체계적으로 분석하고자 하였다. 이를 위한 연구 문제는 1) 인공지능 도구가 생성한 글쓰기 평가 기준은 어떤 특성을 가지는가? 2) 인간과 인공지능 도구가 생성한 글쓰기 평가 기준은 서로 어떠한 공통점과 차이점을 갖는가?로 설정하였다. GPT와 Claude를 대표적인 인공지능 도구로 선정하여 대학생 글쓰기 평가 기준을 생성하게 한 후, 그 결과물을 인간이 만든 글쓰기 평가 기준과 대조하였다. 연구 결과, 인간과 인공지능 도구 모두 글의 내용과 관련한 평가 범주에 가장 높은 중요도를 부여한다는 공통점을 보였다. 그러나, 인간은 내용, 조직, 어법 등 세 개의 주요 범주로 평가하였으나, 인공지능 도구들은 형식 및 인용, 독창적(비판적) 사고, 전체적 인상 등의 추가 범주를 포함하여 평가 기준을 제시하였다. 전반적으로 인간은 각 평가 범주 내에서 상세한 항목을 포함하는 반면, 인공지능 도구들은 간결하게 항목을 설정하였다. 특히, 인공지능 도구가 영어를 기반으로 개발되었기 때문에 발생하는 언어적 차이점과 각 항목별 배점 체계와 관련한 차이점이 발견되었다. 이를 통해, 인간과 인공지능의 협력적 평가 모델 개발에 대한 중요한 시사점을 제시하였으며, 향후 교육평가 장면에서 인공지능의 보완적 도구로서의 역할을 탐색하였다.

[Abstract]

This study, from the perspective of Learning with AI, aimed to explore the educational applicability of writing evaluation criteria generated by artificial intelligence. Specifically, it sought to systematically analyze the similarities and differences between AI-generated criteria and those developed by humans. The research questions for this study were set as follows: 1) What characteristics do the writing evaluation criteria generated by AI tools have? 2) What similarities and differences exist between the writing evaluation criteria generated by humans and AI tools? GPT and Claude were selected as representative AI tools, and they were tasked

http://dx.doi.org/10.14702/JPEE.2024.663

Received 19 September 2024; **Revised** 7 October 2024
Accepted 15 October 2024

***Corresponding Author**
E-mail: lee.byungyoon12@gmail.com

with generating writing evaluation criteria for undergraduate students. These AI-generated criteria were then compared with human-created criteria. The results showed a commonality: Both humans and AI-tools placed the highest importance on categories related to content. However, while humans evaluated based on three main categories – content, organization, and language usage – the AI tools included additional categories such as format and citations, original thinking, and overall impression. In general, human tended to include more detailed items within each evaluation category, while AI tools presented more concise items. Notably, differences were observed in language-related aspects and scoring systems, which were influenced by the AI tools being developed based on English. This study offers important insights into the development of collaborative evaluation models between humans and AI, and it explores the potential role of AI as a complementary tool in educational assessment in the future.

Key Words: Writing assessment, Evaluation criteria, Comparisons between humans and AI tools, GPT, Claude

I. 서 론

최근 교육학계에서는 GPT(Generative Pre-trained Transformer) 등 대화형 딥러닝 기반 인공지능 도구를 활용한 “Learning with AI” 관점[1,2]의 연구가 증가하고 있다. 이는 인공지능을 통해 교수·학습을 지원하고, 과제 및 산출물을 평가하는 방안을 탐색하고 개발하는 접근법을 의미한다[3]. GPT의 뛰어난 자연어 처리 기술과 빅데이터를 활용한 사전 학습은 서술형 과제 평가에 있어, 인간평가자의 한계를 극복할 수 있는 대안으로 주목받아 왔다. 특히, 채점자 내 신뢰도 확보와 같은 문제를 해결할 수 있을 것이라는 기대 속에 여러 연구가 진행되어 왔다(예: [4,5,6,7,8,9]).

평가 장면에서 GPT를 활용하는 선행연구들은 대체로 인간평가자의 결과를 기준으로 GPT의 채점 결과가 얼마나 부합하는지를 분석하는 데 초점을 맞추었다. 이는 대체로 GPT를 인간평가자의 보조도구로서 활용하는 가능성에 대해 탐색하는 관점을 취한다. 그러나 인간 채점과 기계 채점 간 신뢰도는 채점 자료 또는 채점 영역에 따라 다른 양상을 나타내는 등 둘의 채점 일치도는 일관성 있다고 보기 힘든 측면이 있다[5-8]. 또한 기계 채점의 결과를 인간 채점 결과에 기대어 일치도를 분석하는 것은 인간 채점의 일관성을 가정하고 있으나, 실제로 인간 채점 결과가 신뢰로운가에 대해서도 확신하기 어렵다. 이에 본 연구에서는 기계 채점을 보다 독립적으로 상정하여 Learning with AI 관점에서 GPT가 직접 생성한 결과물의 실제 활용 가능성을 탐구하는 것이 필요하다고 보았다. 이에, 본 연구는 GPT 등 인공지능 도구가 서술형 과제에 필요한 글쓰기 역량을 평가하기 위해 생성하는 기준의 특성을 분석하고, 이를 인간이 개발한 평가 기준과 비교하고자 한다.

또한 본 연구는 인공지능에 대한 관심이 높아짐에 따라, 널리 알려진 GPT뿐만 아니라 다른 인공지능 언어 모델이 생성하는 채점 기준을 종합적으로 분석함으로써 평가에서 인공지능 활용의 일반성을 높이고자 하였다. 연구진은 여러 인공지능 언어 모델 중 미국 Anthropic사가 개발한 Claude라는 인공지능 언어 모델에 주목하였다. 본 연구에서 GPT와 Claude를 선택한 이유는 다음과 같다. 첫째, 두 모델은 모두 텍스트 기반의 생성형 인공지능으로, 방대한 텍스트 데이터를 학습하여 사용자와의 대화형 상호작용을 통해 맞춤형 결과물을 제공한다는 공통점을 지닌다. 둘째, 현재 접근 가능한 생성형 인공지능 중 GPT와 Claude가 다양한 측면에서 최고 수준의 성능을 보이는 것으로 평가되고 있다[10,11]. 이러한 배경을 바탕으로, 본 연구에서는 이 두 인공지능 도구가 생성해 내는 평가 기준을 비교 분석함으로써, 평가 분야에서의 가능성과 한계를 탐색하고자 하였다.

이에, 본 연구는 각 모델의 고유한 특성과 기능적 차이를 고려하여 인간의 채점 기준과 비교 분석을 통해, 다음과 같은 중요한 시사점을 도출할 것으로 기대한다. 첫째, 서로 다른 인공지능 언어 모델을 비교함으로써, 각 모델의 교육적 활용에 있어 장단점을 파악할 수 있고, 학습자와 교수자가 특정 교육 목표나 과제에 적합한 인공지능 도구를 선택하는 데 실질적인 도움을 줄 수 있다. 뿐만 아니라, 각 모델의 한계점이나 편향성을 파악할 수 있어, 인공지능 도구를 교육 현장에 도입할 때 주의할 점을 인지할 수 있다. 둘째, 두 도구가 생성해 낸 평가 기준을 비교함으로써, 인공지능 도구가 생성한 평가 기준의 일반적 특성과 각 도구별 특수성을 구분할 수 있다. 이는 인공지능 도구가 제시한 평가 기준의 신뢰성과 일반화 가능성을 판단하는 데 중요한 근거가 될 수 있다. 마지막으로, 이러한 비교 연구는 향후 교육용 인공지능 모델 개발에 있어 방향성에 대한 통찰을 제공하고, 평가 영역에서의 인공지능 활용을 확대하고, 사용자에게 더 풍부하고 유용한 정보를 줄 수 있을 것으로 기대된다.

본 연구는 대학 교육에서 날로 그 중요성이 부각되고 있는 글쓰기 평가에서 활용되는 평가 기준에 집중하고자 한다. 대학에서의 글쓰기 교육은 단순한 학문적 글쓰기 기술 습득을 넘어, 의사소통 역량, 비판적 사고력 함양, 그리고 자아 성찰의 기회를 제공한다는 점에서 그 교육적 의미가 크다[12,

13]. 글쓰기 과제 및 활동 자체가 학생들이 긴 시간을 고민하고 여러 내용을 찾아보며 배움의 과정을 촉진할 수 있다는 점에서 그 중요성이 강조되어 왔지만, 다양한 평가 기준으로 인한 일관된 평가의 어려움은 지속적으로 제기되어 왔다[14-16]. 따라서, 본 연구에서는 대학생 글쓰기를 채점을 위해 개발되고 타당화 된 인간의 평가 기준과 GPT가 생성해내는 평가 기준의 특징을 비교하고자 한다. 이에 따른 연구문제는 다음과 같다.

연구문제 1: 인공지능 도구(GPT, Claude)가 생성한 글쓰기 평가 기준은 어떤 특성을 가지는가?

연구문제 2: 인간과 인공지능 도구가 생성한 글쓰기 평가 기준은 서로 어떠한 공통점과 차이점을 갖는가?

이러한 비교 분석을 통해, 인공지능 언어 모델들의 교육적 활용 가능성과 한계를 파악하고, 현장에서의 효과적인 활용 방안을 모색하고자 한다. 특히, 여러 연구를 통해 개발되고 타당화 된 인간의 평가 기준과 인공지능이 생성해 내는 평가 기준의 특징을 비교함으로써, 인공지능 기술이 교육 평가 분야에 미칠 수 있는 영향과 그 잠재적 가치를 심층적으로 탐구할 수 있을 것이다. 이러한 접근은 단순히 인공지능 도구의 성능을 평가하는 데 그치지 않고, 인간과 인공지능의 협력적 평가 모델 개발이나 인공지능을 활용한 새로운 교육 평가 패러다임 구축 등, 교육 혁신을 위한 실질적인 방안을 모색하는 데 기여할 수 있을 것으로 기대한다.

II. 이론적 배경

A. 글쓰기 평가 기준

본 연구는 대학생의 글쓰기를 평가 대상으로, 글쓰기 역량을 정확히 측정하는 것을 목적으로 하여, 평가 기준을 개발하는 데에 초점을 두었다. 다만, 본 연구에서는 특정 글의 유형에 국한하지 않고 일반적인 글쓰기 평가 과정에서 인간과 인공지능 도구가 각각 어떤 기준을 생성하는지를 비교 분석하는 데 집중하였다. 이를 위해 대학생의 글쓰기 능력을 진단하기 위한 평가 기준을 제안한 선행연구들을 중점적으로 검토하였다.

글쓰기 평가 기준은 연구자나 연구 목적에 따라 다양하게 설정되지만, 일반적으로 주로 내용, 조직, 표현의 세 범주로 분류된다[17]. 내용 범주는 글의 주제, 내용의 질, 논리성 등 내용의 실질적 가치를 평가하고, 조직 범주는 글의 전체적 구조와 문단 간 연결성, 논리적 흐름을 다루며, 표현 범주는 어휘 선택의 적절성, 문법, 맞춤법 등 언어 사용의 정확성을 평가한다[18]. 최근 연구들은 이러한 기본 틀을 바탕으로 더욱 세분화되고 구체적인 평가 기준을 제시하고 있다. 예를 들어, 대학 신입생 글쓰기 평가 기준을 다룬 연구에서는[19] 내용 범주에 5개의 하위 준거(주장의 적절성과 명료성, 근거의 적절성, 근거의 충분성, 접근 방법과 아이디어의 참신성, 예상 반론 고려), 조직 범주에 2개의 하위 준거(글 전체 구성, 문단 내 구성), 표현 범주에 2개의 하위 준거(정확한 문장과 적절한 어휘, 맞춤법과 글쓰기 관습)를 포함시켰다. 각 하위 준거는 5점 척도로 평가되며, 1, 3, 5점에 대한 구체적인 채점 기준과 함께 채점자의 주관에 따라 중간 점수인 2, 4점을 부여하도록 채점기준표를 구성하였다.

비슷한 맥락에서, 해외 연구자가 개발한 기준을 한글로 번안하고 내용 전문가의 검토를 거쳐, 내용의 적절성(6개 문항), 조직의 효과성(4개 문항), 어법의 정확성(3개 문항)으로 구성된 평가 기준을 제시한 연구도 있다[20]. 기존 연구에서 '표현'으로 분류되었던 범주를 '어법의 정확성'으로 재정의하여, 문장력과 맞춤법 등 언어 사용의 정확성에 초점을 두었다. 각 문항은 1-5점 척도로 평가되도록 하여 보다 정밀한 평가가 가능하도록 하였다.

대학생 글쓰기 평가 기준을 개발한 두 선행연구를 통해, 글쓰기 평가의 주요 범주들이 서로 다른 가중치를 가지고 있음을 알 수 있다. [19]의 연구에서는 내용, 조직, 표현 범주가 각각 5:2:2의 비율로 가중치가 부여되었고, [20]의 연구에서는 이 세 범주가 6:4:3의 비율로 평가되었다. 이러한 가중치 분배는 연구자의 관점, 연구의 목적, 또는 평가 대상이 되는 글쓰기의 장르와 목적에 따라 유동적으로 조정될 수 있으나, 두 연구에서 공통적으로 관찰되는 중요한 특징은 내용 범주에 가장 높은 가중치가 부여되었다는 점이다. 이는 글쓰기 평가에 있어 내용의 질과 깊이가 가장 중요한 요소로 간주됨을 보여주기도 한다.

B. 인공지능 도구를 활용한 평가

1) 인공지능 도구를 활용한 서술형 과제 평가

교육학 분야에서 GPT와 같은 인공지능 도구의 활용 가능성을 탐색한 최근 연구들은 인간과 인공지능의 평가 결과를 비교하며 흥미로운 결과를 도출하였다. 특히 [7]의 연구는 대학생 인턴 지원자의 자기소개서 평가에 있어 인공지능의 성능을 검증하였다. 84개의 자기소개서를 대상으로 GPT 3.5와 GPT-4.0 모델의 평가 능력을 인간평가자와 비교 분석하였다. 총 19개의 평가 항목에 대해 인간평가자와의 상관계수를 확인한 결과, GPT-3.5 모델과는 .34, GPT-4.0 모델과는 .60의 상관관계를 보였다. 또한, GPT 모델들이 상위 점수를 부여한

지원자와 실제 채용에 합격한 사람들과 일치하는 정도를 확인한 결과, GPT-3.5 모델은 평균 61%, GPT-4.0 모델은 평균 83%의 일치도를 보여, GPT-4.0 모델의 경우 인간평가자와의 평가 일치도가 중상 수준임을 확인하였다. 이는 인공지능 도구들이 실제 채용 결과를 예측하는 능력을 어느 정도 갖춘 것으로 볼 수 있다.

[5]의 연구는 대학생 에세이에 대한 GPT-4.0 모델과 인간평가자의 평가 일치도를 분석하였다. 이 연구에서는 47명의 대학생이 동일한 주제로 작성한 에세이를 대상으로 두 평가자 간의 일치도를 확인했는데, 전반적으로 낮은 수준의 일치도를 보였으며 평가 범주와 세부 평가 항목에 따라 그 정도가 상이했다. 내용 범주의 평가 일치도는 .33의 상관관계를 보였지만, 조직과 표현 범주에서는 통계적으로 유의한 일치도가 나타나지 않았다. 세부 평가 항목을 살펴보면, 내용 범주의 한 문항에서 .46의 상관계수로 가장 높은 일치도를 보인 반면, 표현 범주의 한 문항에서는 .25의 낮은 상관계수를 나타냈다.

[6]의 연구는 세계지리 과목에서 GPT의 서·논술형 평가 능력을 교사와 비교하였다. 이 연구에서는 정답이 명확한 문항과 창의성 및 논리성 등 채점자의 주관적 판단이 필요한 문항에서 GPT와 인간평가자의 평가 일치도가 각각 다르게 나타났다. 정답이 명확한 문항에서는 상관계수 .52의 일치도를, 창의성 및 논리성을 채점하는 문항에서 .30의 일치도를 보였다. 주목할 만한 점은 GPT에게 예시 답안과 비판적 채점 지침을 프롬프트에 제공했을 때, 두 유형의 문항 모두에서 일치도가 크게 향상되어 각각 .67과 .52의 상관계수를 나타냈다는 것이다.

또한, 최근 한 연구[10]는, GPT, Claude, Gemini Advanced와 같은 생성형 인공지능 도구들의 수학적 성능을 수학 서술형 문항을 통해 비교 분석했다. 이는 현존하는 인공지능 도구들의 수학적 능력을 체계적으로 평가하고, 이를 통해 수학 교육에서의 실질적 활용 가능성을 탐색했다는 데에 의의가 있다. 연구 결과, GPT와 Claude가 Gemini Advanced에 비해 다양한 수학 영역에서 더 높은 정답률을 보였다. 그러나, 복잡한 풀이과정을 요구하는 서술형 문제에서는 세 도구 모두 아직 제한적인 것으로 나타나, 교사들이 각 인공지능 모델의 특성을 정확히 이해하고 이를 교육 현장에 적절히 접목시키는 것이 중요하며, 교사와 인공지능 도구의 협력적 교육 모델 개발이 중요함을 시사하였다.

기존 연구들의 결과는 GPT와 같은 인공지능 언어 모델의 평가 영역에서의 잠재력을 보여준다. [21]의 연구에서 지적된 바와 같이, 이러한 모델들은 방대한 사전 정보를 바탕으로 입력된 답변의 적절성을 판단하고, 사용자가 요구하는 형태로 결과를 제공하기 때문에, 인간이 인공지능 도구를 활용할 때 제공하는 평가 기준 자체가 GPT의 평가 성능에 영향을 미칠 수 있다. 그러나 지금까지의 연구들은 주로 인간평가자와의 평가 결과 비교에 초점을 두었다. 이는 인간이 설정한 특정 채점 기준에 GPT가 얼마나 부합하는지를 기준으로 평가한 것이다. 즉, 기존 연구들은 인간의 채점기준을 하나의 기준으로 두고 평가했다는 점에서, GPT의 잠재적 채점 능력을 제한적으로만 탐색했을 가능성이 있다. 이러한 관점에서, 본 연구에서는 인공지능 도구들이 독자적으로 어떤 평가 기준을 생성해낼 수 있는지를 탐구하고자 한다.

2) 인공지능 도구의 채점기준 생성

GPT와 같은 인공지능 도구가 생성한 자료에 대한 연구는 아직 초기 단계에 있지만, [22]의 연구는 이 분야에 중요한 시사점을 제공한다. 이 연구에서는 GPT를 활용하여 평가에 필요한 다양한 자료(채점 기준, 채점 결과, 피드백)를 생성하고, 이에 대한 교사들의 인식을 조사하였다. 고등학교 통합과학 수업을 대상으로 한 이 연구는 평가문항 개발부터 학생 응답 수집, 채점 실시까지의 전 과정에서 GPT의 활용 가능성을 탐색했다. 연구 결과, 교사들은 GPT가 생성한 평가 기준과 피드백에 대해서는 사용 가능성을 높게 평가하였으나, 실제 채점 결과에 대한 신뢰도는 낮게 평가하였다. 이는 GPT가 생성한 결과를 수정하지 않고 그대로 사용하는 것은 평가의 타당성과 신뢰성 측면에서 적절하지 않음을 의미한다. 그러나, 교사들은 GPT를 보조 도구로 활용하는 것에 동의하였고, 특히 평가 자료의 사전 또는 사후 검토 용도로의 사용 가능성을 인정했다.

[22]의 연구는 GPT 등 인공지능 도구가 적절한 평가 기준을 생성할 수 있음을 보여주었지만, 인공지능 도구가 생성한 자료가 어떠한 특징을 보이는지, 인간이 만든 기준과는 어떠한 차이를 보이는지에 대한 연구는 아직 부족한 실정이다. 이에 본 연구는 글쓰기 평가 상황에서 인공지능 도구가 어떤 평가 기준을 생성하는지, 그리고 이러한 기준이 인간이 만든 기준과 어떤 차이를 보이는지를 탐구하고자 한다.

III. 연구 방법

A. 인간이 생성한 글쓰기 평가 기준

본 연구는 [20]에서 개발된 “글쓰기 의사소통 역량 채점기준”을 인간이 생성한 글쓰기 평가 기준으로 채택하여, 인공지능 도구가 생성한 평가 기준과 비교하였다(표 1 참고). 이

표 1. 인간이 생성한 글쓰기 평가 기준[20]

Table 1. Evaluation criteria for writing generated by human [20]

주요 범주	채점 문항
1. 내용의 적절성	1-1. 글의 중심 내용이 명확하다 1-2. 글의 중심 내용이 구체적이다 1-3. 글의 내용이 독창적이다 1-4. 글의 내용을 뒷받침하는 근거나 예시가 다양하게 제시되어 있다 1-5. 저자의 의도를 분명하게 제시한다 1-6. 글의 전체 내용에 일관성이 있다
2. 조직의 효과성	2-1. 글의 구조가 글의 목적에 부합한다 2-2. 문단과 문단 간 역할이 분명하다 2-3. 문단 내 중심 문장과 뒷받침 문장이 명료하게 드러나 있다. 2-4. 필요한 정보가 유기적으로 배열되어 있다
3. 어법의 정확성	3-1. 맞춤법과 띄어쓰기가 지켜지고 있다 3-2. 어법을 잘 지키고 있다 3-3. 이해하기 쉽게 문장을 작성하였다

주1. 문항당 1-5점으로 평가됨.

평가 기준은 한글로 번안된 설명문 글쓰기를 평가하는 원문항을 연구진이 GPT의 채점 오류를 최소화하도록 다각도로 수정하고, 전문가(교육학 전공 교수 3인)의 자문 및 검토를 거쳐 최종 개발되었다.

본 연구에서 이 평가 기준을 선택한 이유는 실제 GPT 활용 평가플랫폼에 탑재되어 신뢰성 있는 채점을 위해 다양한 요소를 고려하여 개발되었으며, 인공지능 도구가 생성한 기준과 비교하기에 가장 적합했다고 판단했기 때문이다. 예를 들어, [20]의 연구진은 평가 기준의 명확성을 높이기 위해 각 문장에 하나의 평가 내용만을 포함시키고 다의적 해석 가능성이 있는 단어를 배제하였다. 이는 GPT가 여러 내용이 포함된 채점 문항(예: 글의 내용이 명확하고 구체적이며, 중심 내용에 부합한다)에 대해 모든 요소를 균형 있게 평가하지 못하는 문제를 해결하고자 한 것이다. 또한, 평가 내용을 더 분명하게 구분하기 위해 원문항의 5개 범주(내용, 조직, 표현, 단어 선택, 형식 및 어법)를 3개 범주(내용의 적절성, 조직의 효과성, 어법의 정확성)로 재구성하였다. 이는 원문항에서 일부 문항이 여러 범주에 걸쳐 나타나는 문제를 해결하여 GPT의 평가 시 범주 간 구분을 명확히 하고 채점의 정확성을 높이기 위함인 것으로 보인다. 마지막으로, 이 평가 기준은 채점 방식에 있어 원문항의 불연속적인 5점, 3점, 1점으로 배점 체계 대신, 각 문항에 1-5점의 연속적 점수를 부여하는 분석적 채점방식을 도입하였다. 이 또한 [20]의 연구진이 GPT의 실제 채점 성능을 향상시키기 위한 과정을 거쳐, 원문항의 배점 방식으로 채점한 결과를 분석하여 연속 점수 부여 방식이 더 적절하다고 판단한 것을 알 수 있다.

B. 인공지능 도구가 생성한 글쓰기 평가 기준

본 연구에서는 현재 가장 널리 사용되는 거대언어모델인 GPT와 Claude를 인공지능 도구로 선정하였다. 두 모델은 방대한 텍스트 기반 데이터를 분석하여 채팅 형식으로 사용자의 요구에 맞는 결과물을 제공한다는 공통점이 있지만, GPT는 실시간 인터넷 검색을 통해 관련 데이터를 찾아 답변하는 반면, Claude는 사전에 학습된 데이터만을 기반으로 응답한다는 차이점이 있다[23]. 따라서, GPT는 훈련된 데이터보다도 최신 데이터를 제공한다는 장점이 있다[24]. 수학 문제 생성 능력에 대해 GPT와 Claude를 비교한 선행연구[10]에 따르면, Claude는 뛰어난 텍스트 처리 능력을 바탕으로 개념 설명이나 이론적 문항 개발에 강점을 보이는 반면, GPT는 복잡한 문제 해결 과정을 요구하는 고난도 문항 개발에 더 적합한 것으로 나타났다. 또한, Claude는 응답 시 창의성보다는 정확성을 우선시하도록 훈련받았는데, 이는 반대로 말하면 GPT는 창의적인 콘텐츠를 생성하는 데 Claude보다 더 능숙하다는 평가를 받는다[25]. 이러한 차이를 바탕으로 본 연구에서는 두 인공지능 도구가 생성해내는 평가 기준을 인간이 만든 기준과 함께 비교해보고자 하였다.

연구의 정확성을 높이기 위해 두 인공지능 도구의 유료 버전인 GPT-4o 모델과 Claude Pro 모델을 사용하였다. 글쓰기 평가 기준 생성을 위해 각 모델의 기본 채팅창에 동일한 프롬프트[1]를 입력하였다("너는 대학교에서 학생들의 글을 평가하는 교수야. 학생들의 글쓰기를 잘 평가할 수 있는 채점 기준을 만들어줘."). 이후 각 인공지능 도구가 최초로 생성한 기준들을 비교 분석의 대상으로 활용하였다.

C. 글쓰기 평가 기준 비교 분석틀

본 연구는 인간이 생성한 글쓰기 평가 기준과 두 인공지능 도구가 생성한 기준을 비교 및 분석하기 위해 선행연구 검토를 통해 도출된 글쓰기 평가 시 주의해야 할 세 가지 주요 사항에 집중하였다. 첫째, 각 기준의 전체적인 구조를 비교에 초점을 두었다. [26]의 연구에 따르면, 인공지능 도구를 활용한 평가 플랫폼 구축 시, 층(layer)의 구분이 중요하다. 이에 따라, 상위 층(상위 개념 혹은 평가 범주)과 하위 층(하위 개념 혹은 세부 평가 항목)의 존재 여부를 명확히 하고, 각 층에

[1]다만 본 연구의 목적은 실제 글쓰기 과제를 평가하는 것이 아니라, 평가 기준 자체의 생성과 그 내용을 분석하는 데 있다. 따라서 본 연구에서는 특정 글쓰기 샘플을 제시하지 않고, 프롬프트를 활용하여 평가 기준을 자체적으로 생성하도록 요청하였다.

포함된 요소들의 수를 파악하였다. 즉, 인공지능 도구가 생성한 평가 기준에서 상·하위 개념의 구분 여부와 각 층에 포함된 요소(예: 주요 범주 개수, 세부 문항 개수)의 형태를 비교분석하는 것이 중요하다고 판단되었다. 둘째, 각 평가 기준의 구체적인 내용을 비교하였다. 인간이 생성한 글쓰기 평가 기준도 내용적 측면에서 상당한 편차를 보이므로[14,19], 인간과 각 인공지능 도구가 글쓰기 평가 시 어떤 부분에 더 중점을 두는지 파악하고자 하였다. 셋째, 배점 방식의 차이를 비교하였다. 글쓰기 평가에서 채점 기준별 배점 및 가중치 설정은 여전히 중요한 문제로 꼽힌다[26]. 따라서, 본 연구에서는 인공지능 도구들이 평가 기준 생성 시 어떤 방식으로 점수를 부여하는지 확인하고, 이를 인간의 배점 기준과 비교분석하고자 하였다.

IV. 결과 및 논의

A. 인공지능 도구가 생성한 평가 기준

GPT와 Claude가 생성한 평가 기준은 표 2와 같다. GPT는 평가 기준을 6개 범주(내용의 질, 구조 및 전개, 문장력 및 언어 사용, 형식 및 인용, 독창성 및 비판적 사고, 전체적 인상)로 구분하였다. Claude는 총 5개 범주(내용, 구조, 언어 사용, 형식과 스타일, 비판적 사고)로 나누었다. 두 도구 모두 각 범주별로 2-3개의 세부 평가 항목을 포함하였다. GPT는 각 범주에서 세부 범주명을 제시하고 그에 해당하는 세부 항목을 명시하였다(예: '내용의 질' 범주 – '주제 명확성' 세부 범주 – 글이 명확한 주제를 가지고 있으며, 그 주제가 일관되게 유지되는가?). Claude는 각 범주 아래 세부 범주를 두되, 이를

표 2. GPT와 Claude가 생성한 글쓰기 평가 기준

Table 2. Evaluation criteria for writing generated by GPT and Claude

GPT		Claude	
주요 범주(배점)	채점 문항	주요 범주(배점)	채점 문항
1. 내용의 질 (30점)	1-1. 주제 명확성 (10점): 글이 명확한 주제를 가지고 있으며, 그 주제가 일관되게 유지되는가? 1-2. 논리적 타당성 (10점): 제시된 주장이 논리적이고 타당한가? 논거가 충분하고 설득력 있는가? 1-3. 창의성 및 독창성 (10점): 글이 창의적이며, 새로운 관점이나 독창적인 아이디어를 포함하고 있는가?	1. 내용 (30점)	1-1. 주제의 명확성과 적절성(10점) 1-2. 논리의 일관성과 설득력(10점) 1-3. 아이디어의 독창성과 깊이(10점)
2. 구조 및 전개 (25점)	2-1. 논리적 전개 (10점): 글의 전개가 논리적이고 일관성이 있는가? 서론, 본론, 결론이 잘 연결되어 있는가? 2-2. 문단 구분 (10점): 문단이 잘 구분되어 있으며, 각 문단이 하나의 핵심 아이디어를 명확하게 전달하는가? 2-3. 서론과 결론 (5점): 서론이 주제를 잘 도입하고, 결론이 글을 효과적으로 마무리하는가?	2. 구조 (20점)	2-1. 서론, 본론, 결론의 균형(6점) 2-2. 문단 간 연결성과 흐름(7점) 2-3. 전체적인 글의 조직력(7점)
3. 문장력 및 언어 사용 (25점)	3-1. 어휘 선택 (10점): 어휘가 적절하고, 다양한 어휘가 사용되었는가? 단어 선택이 글의 주제와 톤에 맞는가? 3-2. 문장 구조 (10점): 문장이 문법적으로 올바르고, 다양한 문장 구조가 사용되었는가? 3-3. 명확성 및 일관성 (5점): 문장이 명확하고 일관되게 전달되는가? 모호한 표현이 없는가?	3. 언어 사용 (20점)	3-1. 문법과 맞춤법의 정확성(8점) 3-2. 어휘의 다양성과 적절성(6점) 3-3. 문장 구조의 다양성과 명확성(6점)
4. 형식 및 인용 (10점)	4-1. 지침 준수 (5점): 과제의 형식적 지침(글자 크기, 줄 간격, 페이지 설정 등)을 정확히 준수하였는가? 4-2. 참고문헌 및 인용 (5점): 인용과 참고문헌이 정확하게 표기되었는가? 표절 없이 학문적 윤리를 준수하였는가?	4. 형식과 스타일 (15점)	4-1. 학술적 글쓰기 규범 준수(5점) 4-2. 인용과 참고문헌의 적절한 사용(5점) 4-3. 글의 톤과 스타일의 일관성(5점)
5. 독창성 및 비판적 사고 (10점)	5-1. 독창성 (5점): 글이 학생 자신의 독창적인 생각을 반영하고 있는가? 5-2. 비판적 사고 (5점): 학생이 주제에 대해 비판적으로 사고하고, 다양한 관점을 고려하였는가?	5. 비판적 사고 (15점)	5-1. 다양한 관점의 고려(5점) 5-2. 증거 기반 논증(5점) 5-3. 반론 예상 및 대응(5점)
6. 전체적 인상 (10점)	6-1. 완성도 (5점): 글이 전반적으로 잘 완성되었고, 전체적으로 하나의 일관된 작품으로 느껴지는가? 6-2. 흥미 유발 (5점): 글이 독자의 흥미를 끌고, 끝까지 집중하게 만드는가?	n/a	

표 3. 평가 기준 비교 요약

Table 3. Summary of evaluation criteria by human and AI tools

비교 사항		인간이 생성한 기준[20]	인공지능 도구가 생성한 기준	
			GPT	Claude
상위 개념과 하위 개념의 구성 형태		– 3개의 상위 개념: • 내용의 적절성(6개 하위 개념) • 조직의 효과성(4개 하위 개념) • 어법의 정확성(3개 하위 개념) – 총 13개 평가 항목으로 구성	– 6개의 상위 개념 • 내용의 질(3개 하위 개념) • 구조 및 전개(3개 하위 개념) • 문장력 및 언어 사용(3개 하위 개념) • 형식 및 인용(2개 하위 개념) • 독창성 및 비판적 사고(2개 하위 개념) • 전체적 인상(2개 하위 개념) – 총 15개 평가 항목으로 구성	– 5개의 상위 개념 • 내용(3개 하위 개념) • 구조(3개 하위 개념) • 언어 사용(3개 하위 개념) • 형식과 스타일(3개 하위 개념) • 비판적 사고(3개 하위 개념) – 총 15개 평가 항목으로 구성
평가 범주와 세부 항목의 특징	진술 방식	– 평서문 예: 글의 중심 내용이 명확하다	– 의문문 예: 글이 명확한 주제를 가지고 있으며, 그 주제가 일관되게 유지되는가?	– 명사형 예: 주제의 명확성과 적절성
	주요 범주 구성	– 내용의 적절성, 조직의 효과성, 어법의 정확성	– 내용의 질, 구조 및 전개, 문장력 및 언어 사용, 형식 및 인용, 독창성 및 비판적 사고, 전체적 인상	– 내용, 구조, 언어 사용, 형식과 스타일, 비판적 사고
배점 방식		– 1–5점 척도 – 총 65점 만점	– 차등 배점 – 총 100점 만점	– 차등 배점 – 총 100점 만점

곧바로 평가 항목으로 활용하였다(예: '내용' 범주 – '주제의 명확성과 적절성' 세부 범주이자 평가 항목).

또한, 두 도구는 범주별로 상이한 배점을 부여하였다. GPT는 각 범주의 배점과 함께 세부 항목별 배점을 상세히 제시한 반면(예: 내용의 질 범주에 포함된 주제 명확성, 논리적 타당성, 창의성 및 독창성은 각각 10점으로 배점), Claude는 범주별 총점만을 제시하였다(예: 내용 범주는 30점, 구조 범주는 20점). 그러나, 두 인공지능 도구 모두 전체 배점의 합계를 100점으로 일치시켰다는 공통점을 보였다. 범주별 점수 비율에도 차이를 보였다. 두 도구 모두 내용 관련 범주에 30%로 가장 높은 비중을 두었지만, GPT는 구조 및 전개, 문장력 및 언어 사용 범주에 각각 25%를 배정했고, 형식 및 인용, 독창성 및 비판적 사고, 전체적 인상 범주에 각각 10%를 할당했다. Claude는 구조와 어휘 사용 범주에 각각 20%, 형식과 스타일, 비판적 사고 범주에 각각 15%를 두었는데, 이러한 배점 차이는 각 도구가 글쓰기 평가에서 중요하게 여기는 요소들의 우선순위를 반영한다고 볼 수 있다.

B. 인간이 생성한 평가 기준과 인공지능 도구가 생성한 평가 기준과의 비교

대학생 글쓰기 평가를 위해 인간이 개발한 평가 기준과 두 인공지능 도구가 도출해낸 평가 기준에는 어떠한 차이가 있는지를 세부적으로 살펴보았다(표 3 참고). 전술하였듯이, 본 연구에서는 1) 평가 기준 제시 시 상위 개념(평가 범주)과 하위 개념(세부 평가 항목)의 구성 형태, 2) 세부 항목의 구체적인 내용 및 특징, 3) 평가 범주와 세부 항목별 배점 방식 등에 집중하여 비교 분석하였다.

1) 상위 개념(평가 범주)과 하위 개념(세부 평가 항목)의 구성 형태

인간이 개발한 평가 기준은 상위 개념(평가 범주) 3개(내용의 적절성, 조직의 효과성, 어법의 정확성)를 중심으로 구성되었으며, 각 범주 아래 6개, 4개, 3개의 세부 항목을 포함하여 총 13개로 구성된 평가 기준을 제시하였다. 두 인공지능 도구 또한 상위 개념과 하위 개념을 함께 제시하였다. GPT는 총 6개의 상위 개념과 그에 따른 하위 개념을 2-3개씩 제안하여, 총 15개 항목으로 구성된 평가 기준을 제시하였다. Claude는 총 5개의 상위 개념과 그에 따른 하위 개념을 3개씩 제시하여 총 15개 항목으로 구성된 평가 기준을 제안하였다. 세 평가 기준 모두 내용, 조직(구조), 어법(문법) 관련 평가 범주를 공통적으로 포함하고 있었으나, GPT는 형식 및 인용, 독창성 및 비판적 사고, 전체적 인상 등 3개의 평가 범주를 더 제시하였고, Claude는 형식과 스타일, 비판적 사고 등 2개의 평가 범주를 더 포함시켰다. 이는 인공지능이 글쓰기 평가에 있어 더 폭넓은 관점을 제시함을 확인할 수 있었다.

2) 평가 범주와 세부 항목의 구체적인 내용 및 특징

a) 세부 평가 항목 진술 방식

각 평가 범주와 세부 평가 항목에 대한 인간과 인공지능의 접근 방식을 비교한 결과, 각각의 특징과 차이점이 확인되었다. 먼저, 세부 평가 항목의 진술 방식에서 가장 두드러진 차

표 4. 주요 범주별 평가 내용 비교

Table 4. Comparison of evaluation criteria by major categories

주요 범주		인간이 생성한 기준[20]	인공지능 도구가 생성한 기준 (GPT, Claude)
내용 범주		– 내용의 명확성 – 내용의 구체성 – 내용의 독창성 – 근거와 예시의 다양성 – 내용의 일관성	– 주제의 명확성/적절성 – 논리의 타당성/일관성/설득력 – 아이디어의 독창성
조직(구조) 범주		– 구조와 글의 부합성 여부 – 문단 간 역할의 명확성 – 중심 문장과 뒷받침 문장의 명료성 – 정보의 유기적 배열 여부	– 서론, 본론, 결론의 연결성 – 문단 간 연결성 – 전체적인 글의 조직력(서론과 결론의 역할 분명)
어법 범주		– 맞춤법, 띄어쓰기 – 어법 – 이해하기 쉬운 문장 사용	– 어휘의 다양성 및 적절성 – 문장 구조의 다양성 및 명확성 – 문법의 정확성
추가 범주	형식 및 인용	n/a	– 인용과 참고문헌의 적절한 사용 – 학술적 글쓰기 규범 준수 – 과제의 형식적 지침 준수
	비판적 사고		– (GPT) 독창성 – (GPT) 비판적 사고 – (Claude) 다양한 관점 고려 – (Claude) 증거 기반 논증 – (Claude) 반론 예상 및 대응
	(GPT만) 전체적 인상		– 완성도 – 흥미 유발

이가 관찰되었다. 인간이 개발한 평가 기준은 '글의 중심 내용이 명확하다'와 같은 평서문 형태를 사용한 반면, GPT는 "제시된 주장이 논리적이고 타당한가?"와 같은 의문문 형태를, Claude는 '주제의 명확성과 적절성'과 같은 핵심 키워드(명사형) 나열 방식을 채택하였다. 이러한 진술 방식의 차이는 평가자의 접근 방법이나 평가 과정에 영향을 미칠 수 있다는 점에서, 진술 형식에 따른 평가 결과 차이 등에 대한 추후 연구가 필요할 것을 보여준다.

b) 주요 범주별 평가 내용

주요 범주별 평가 내용에 대한 비교는 표 4와 같다. 내용 범주와 관련하여서는, 인간은 글의 전반적인 내용의 적절성에 초점을 두어, 명확성, 구체성, 독창성, 일관성 등 다양한 측면을 평가하는 것으로 보인다. 특히 주요 내용을 뒷받침하는 근거나 예시의 다양성, 저자 의도의 명확성 등을 포함하여 글의 내용을 다각도로 평가하는 특징을 보인다. 반면, GPT는 '내용의 질'이라는 범주 하에, 주제의 명확성과 일관성, 주장의 논리성과 타당성, 그리고 아이디어의 창의성과 독창성을 중점적으로 평가한다. Claude는 비교적 간결하게 '내용'이라는 범주 안에 주제의 명확성과 적절성, 논리의 일관성과 설득력, 아이디어의 독창성과 깊이를 평가 항목으로 제시하였다.

이러한 비교를 통해, 인간이 개발한 기준이 두 인공지능 도구의 기준보다 더 구체적이고 세분화 되어 있음을 알 수 있다. 이는 인간의 평가 기준이 각 평가 항목에 단일 내용만을 포함시키려는 의도적인 접근 방식[20]에서 기인한다. 특히, 인간의 평가 기준은 글 내용의 구체성, 저자 의도의 명확성, 제시된 근거의 다양성 등을 개별적으로 평가하는 항목을 포함하고 있어, 내용 범주에 더 큰 비중을 두고 있다. 반면, GPT와 Claude는 '글의 내용'이라는 포괄적인 표현 대신, '주제', '아이디어', '논리', '논거' 등의 용어를 사용하여 평가 항목을 구성하였다. 이는 글의 내용을 전반적인 주제, 저자의 아이디어, 주장 및 논리 등으로 세분화하여 평가하려는 의도로 해석할 수 있으나, 해당 용어 사이에 뚜렷한 구분이 어려워 평가 시 혼란을 야기할 수 있다는 단점도 있다. 이러한 차이점은 인간과 인공지능의 글쓰기 평가에 대한 접근 방식의 차이를 반영한다. 인간의 기준이 더 세밀하고 다각적인 평가를 지향하는 경향을 보이며, 이는 글의 질적 측면을 다각도로 평가하고자 하는 인간 평가 기준의 의도를 반영하는 것으로 보인다. 반면, 인공지능의 기준은 보다 간결하면서도 핵심적인 요소에 집중하는 경향을 보인다.

글의 조직(구조) 범주에 대한 인간과 인공지능의 평가 기준을 비교해보면, 인간이 개발한 평가 기준은 문장-문단-전체 글의 구조를 모두 아우르는 포괄적인 접근을 취하고 있

다. 특히 문단 간의 역할과 문단 내 문장이나 정보의 조직까지 세밀하게 평가하여, 글의 미시적 구조와 거시적 구조를 균형 있게 평가하는 것으로 보인다. 반면, GPT와 Claude는 글의 전체적인 구조에 더 초점을 맞추는 경향을 보인다. GPT는 '구조 및 전개'라는 범주 하에 논리적 전개, 문단 구분, 시작(서론)과 마무리(결론)의 적절성을 주요 평가 항목으로 삼고 있다. Claude 역시 '구조'라는 범주에서 서론, 본론, 결론의 균형, 문단 간 연결성과 흐름, 전체적인 글의 조직력을 가지고 평가하고자 하였다.

이러한 차이점을 통해, 인간의 평가 기준이 두 인공지능 도구보다 세부적이고 미시적인 접근을 취하고 있음을 알 수 있다. 인간의 조직(구조) 범주는 글의 전체적인 구조 뿐만 아니라 문단 내부의 문장이나 정보의 유기적 구성까지 평가 대상으로 삼아, 보다 깊이 있는 평가를 하는 것으로 보인다. 반면, GPT와 Claude는 공통적으로 '서론, 본론, 결론'의 연결성, 문단의 구분, 전체 글의 효과적인 조직력 등 거시적 구조에 더 주목하는 경향을 보인다. 이는 인공지능이 글의 구조를 평가할 때, 일반적으로 통용되는 글의 기본 구조에 얼마나 잘 부합하는지, 그리고 그 흐름에 따라 얼마나 효과적으로 작성되었는지에 더 큰 관심을 두고 있음을 시사한다. 이러한 접근은 전체적인 구조와 흐름을 빠르게 파악하는 데 유용할 수 있지만, 세부적인 구조적 요소를 놓칠 가능성도 배제할 수 없다.

어법 평가에 있어서는 인간과 인공지능의 접근 방식은 유사하면서도 차이점을 보였다. 모두 어법을 평가하고자 하지만, 그 구체적인 내용에서 차이가 있었고 이는 범주명에서도 드러난다. 인간은 '어법의 정확성'이라는 명칭 하에 맞춤법, 띄어쓰기, 문법 규칙 준수에 중점을 둔다. 반면, GPT는 '문장력 및 언어 사용'이라는 더 넓은 범주를 사용하며, 어휘 선택, 문장 구조, 사용된 문장의 명확성 및 일관성을 종합적으로 평가한다. Claude 또한 '언어 사용'이라는 범주 아래 맞춤법의 정확성, 어휘의 다양성과 적절성, 문장 구조의 다양성과 명확성을 평가한다. 또한, 인간이 만든 어법 범주에서는 '이해하기 쉽게 문장을 작성하였다'와 같이 문장의 간결성을 중요시하는 반면, GPT와 Claude는 문장 구조의 다양성에 더 주목한다. 마지막으로, 인간의 어법 범주에는 어휘 관련 평가 항목이 없지만, 두 인공지능 도구는 모두 어휘의 적절성과 다양성을 중요한 평가 요소로 간주하는 것으로 보인다.

이러한 차이점들을 종합해 볼 때, 인간의 어법 평가는 정확성과 간결성에 초점을 맞추는 반면, 인공지능은 정확성과 더불어 언어 사용의 수려함과 다양성까지 포괄적으로 평가하는 경향이 있음을 알 수 있다. 이러한 차이는 언어적 배경에서 기인할 수 있다. GPT와 Claude가 주로 영어를 기반으로 개발되었다는 점을 추측해 볼 수 있다. 한글에 중점을 두어 인간의 어법 범주에는 띄어쓰기에 대한 세부 평가가 포함되어 있지만, GPT와 Claude의 어법 범주에는 이러한 항목이 명시되어 있지 않다. 최근 인공지능 도구들의 한글 처리 능력이 크게 향상되었음을 확인한 연구 결과[4]가 있음에도 불구하고, 기본적으로 설계된 언어 체계의 차이가 어법 평가 기준의 차이로 이어졌을 가능성이 있다. 이는 인공지능의 언어 모델이 주로 학습된 언어와 그 특성에 따라 세부 평가 항목을 형성할 수 있음을 보여주기도 한다. 따라서 인공지능 도구를 활용하여 한글로 된 글을 평가를 위한 기준 설정 시, 이러한 언어적 특성을 미리 고려할 필요가 있다.

c) 인공지능 도구가 제시한 추가 범주별 평가 내용

인간의 글쓰기 평가 기준은 3개 범주로 한정되어 있는 반면, GPT와 Claude는 그 외에 추가적인 평가 범주를 제시하였다. 특히 주목할 점은 두 인공지능 도구가 공통적으로 글쓰기의 형식적 측면과 인용 방식에 대한 평가를 다룬다는 점이다. GPT는 '형식 및 인용'으로, Claude는 '형식과 스타일'이라는 범주로 명명하여 평가 기준에 포함시켰다. 이러한 접근은 대학생 수준의 학술적 글쓰기에서 특히 중요한 의미를 가진다. 대학생들은 자신의 주장을 뒷받침하기 위해 다양한 문헌과 자료를 활용하게 되므로, 적절한 인용 방식과 참고문헌 표기는 학술적 정직성과 글의 신뢰도를 높이는 데 필수적이다. 따라서 이러한 요소들을 평가 기준으로 설정한 인공지능의 접근은 타당하다고 볼 수 있으며, 향후 인간의 글쓰기 평가 기준에도 이 같은 요소들이 포함되어야 할 필요성을 시사한다.

또한, 두 인공지능 도구는 공통적으로 비판적 사고와 관련한 범주를 포함하고 있으며, 특히 글이 다양한 관점을 고려하는지를 평가하는 세부 항목을 포함하였다. GPT의 경우, '독창성 및 비판적 사고' 범주에서 글쓴이의 독창적인 생각이 반영되었는지를 평가하는 세부 항목이 있지만, 이는 '내용의 질' 범주의 세부 항목("글이 창의적이며, 새로운 관점이나 독창적인 아이디어를 포함하고 있는가?")과 중복되는 면이 있다. 반면, Claude는 '비판적 사고' 범주에서 증거 기반 논증과 반대 의견에 대한 예상 및 대응을 확인하는 세부 항목을 포함하여, GPT보다 글의 비판적 내용 측면에 더 중점을 두는 것으로 보인다. Claude가 제시한 이러한 평가 항목은 논증적 글쓰기 평가 기준[12]과 상당 부분 일치한다. 그러나, 세 가지 평가 기준 모두 글의 장르를 명시하지 않고 단순히 글쓰기 평가 기준을 생성했다는 점에서, 이러한 차이가 발생한 원인에 대해서는 추가적인 분석이 필요하다.

마지막으로, GPT는 '전체적 인상'이라는 독특한 범주를

포함하고 있다. 이 평가 범주는 글의 전반적인 완성도와 독자의 흥미 유발 및 집중도를 평가한다. 완성도 평가 항목은 '구조 및 전개' 범주의 일관성 평가 항목과 일부 중복되는 듯하나, 전체적인 작품성을 더 포괄적으로 다루는 것으로 보인다. 특히 독자의 흥미와 집중도를 평가하는 부분은 새로운 시도로, 글의 효과성을 독자 관점에서 측정한다는 점에서 의의가 있다. 다만, 일부 평가 항목의 중복성과 모호성(예: 한 항목 안에 두 개의 내용을 평가)으로 인해 실제 적용 시 혼란이 발생할 수 있어, 이 범주에 대한 더 명확한 기준 설정과 추가적인 연구가 필요하다.

3) 평가 범주와 세부 항목별 배점 방식

GPT와 Claude가 제안한 평가 기준은 인간이 설계한 기준과 비교하여 평가 범주와 세부 항목별 배점에서 주목할 만한 차이를 보였다. 인간의 평가 기준이 모든 세부 항목에 대해 동일하게 1-5점 척도를 적용한 반면, 두 인공지능 도구들은 평가 범주와 세부 항목에 따라 차등화된 배점 체계를 도입했다. 특히 GPT의 경우, '내용의 질' 범주에 30점이라는 최고 배점을 부여하고 그 아래 3개 세부 항목에 각 10점을 할당하였다. '구조 및 전개'와 '문장력 및 언어 사용' 범주는 각각 25점을 배정하였고, 세부 항목들은 10점 또는 5점으로 세분화되었다. 나머지 '형식 및 인용', '독창성 및 비판적 사고', '전체적 인상'은 각각 10점씩 배점되었고, 그 아래 2개씩의 세부 항목에 5점씩 균등하게 분배되었다.

Claude는 앞서 언급했듯이 범주별 총점만 제시하고 세부 항목별 점수는 초기에 제공하지 않았다. 그러나 추가 요청을 통해 세부 항목별 점수도 확인할 수 있었다. GPT와 마찬가지로 Claude도 '내용' 범주에 가장 높은 점수(30점)를 배정하고, 3개 세부 항목에 각 10점씩 할당하였다. '구조'와 '언어 사용' 범주는 각각 20점씩 배점되었으나, 세부 항목별로 점수 분배가 달랐다. Claude는 서론, 본론, 결론의 균형을 6점, 문단 간 연결성과 흐름과 전체적인 글의 조직력에 각 7점을 부여하였다. 이는 GPT가 서론, 본론, 결론의 연결성에 더 높은 점수를 할당한 것과는 대조적이다. '언어 사용' 범주에서 Claude는 문법과 맞춤법의 정확성에 8점, 어휘의 다양성과 적절성 그리고 문장 구조의 다양성과 명확성에 각 7점을 배정하였다. 이 또한 GPT가 어휘 선택에 더 높은 비중을 둔 것과는 차이를 보인다. '형식과 스타일', '비판적 사고' 범주는 각각 15점씩 배점되었으며, 각 범주 내 3개의 세부 평가 항목에는 동일하게 5점씩 할당되었다.

이렇게 차등화된 배점 체계는 각 평가 범주 및 항목의 상대적 중요도를 반영하려는 인공지능의 시도로 해석될 수 있으며, 인간의 일률적 평가 방식과는 대조를 이룬다. 인간과 인공지능 도구 간의 배점 체계의 차이뿐만 아니라 인공지능 도구들 사이의 차이도 두드러졌다. 인간이 설계한 평가 기준은 모든 항목에 동일한 배점을 적용하되, 글의 내용과 관련된 범주에 더 많은 항목을 둠으로써 간접적으로 가중치를 부여한 것으로 보인다. GPT와 Claude 역시 내용 관련 범주에 최고 배점을 할당하여 이 영역의 중요성을 강조했다. 이는 인간과 인공지능 모두 글 평가에 있어 내용을 최우선으로 고려한다는 점을 시사한다. 또한, GPT와 Claude가 기존의 내용, 구조, 어법 범주 외에 추가한 범주들에는 상대적으로 낮은 배점을 부여했다. 이는 이러한 추가 범주들이 평가에 필요하다고 판단되지만, 핵심적인 3개의 범주보다는 중요도가 낮다고 인식함을 보여준다. 이러한 배점 체계의 차이는 인간과 인공지능, 그리고 인공지능 도구들 간의 평가 우선순위와 중요도 인식의 차이를 반영하고 있으며, 동시에 평가 기준 설계에 있어 각각의 독특한 접근 방식을 보여준다.

C. 논의 및 제언

본 연구의 목적은 인공지능 도구(GPT와 Claude)가 생성한 글쓰기 평가 기준의 특성을 분석하고, 이를 인간이 개발한 평가 기준과 비교함으로써, 인공지능 활용의 교육적 가능성과 한계를 탐구하는 데 있다. 이를 위해 설정한 두 가지 연구 문제인 인공지능 도구가 생성한 글쓰기 평가 기준의 특성과 인간이 생성한 평가 기준을 비교하여 논의하면 다음과 같다.

1) 인공지능 도구가 생성한 평가 기준의 특성

GPT와 Claude가 생성한 글쓰기 평가 기준은 인간이 개발한 기준과 비교했을 때 몇 가지 독특한 특성을 가지고 있다. 먼저, 두 인공지능 도구는 평가 기준에서 더 다양한 범주를 도입하여, 글쓰기의 포괄적인 평가를 시도하고 있다. 예를 들어, GPT는 '형식 및 인용', '독창성 및 비판적 사고', '전체적 인상' 등의 범주를 추가함으로써, 글쓰기 평가에서 더 넓은 시각을 제공한다. 이는 [2]의 연구에서 인공지능이 포괄적인 평가를 제공할 수 있는 잠재력을 언급한 바와 일치한다. 이러한 특성은 인간 평가자가 간과할 수 있는 측면을 포착할 가능성을 시사하며, 인공지능 도구가 글쓰기의 다양한 측면을 고려하는 데 유용할 수 있음을 보여준다.

또한, GPT와 Claude는 각각 '내용의 질', '언어 사용', '구조'와 같은 범주에서 다양한 세부 항목을 포함하고 있으며, 이러한 항목들이 글쓰기의 각 측면을 구체적으로 평가할 수 있도록 구성되어 있다. 이는 인공지능 도구가 평가 기준을 설정하는 데 있어 일관된 구조를 가지고 있으며, 사용자가 요구하는 평가 지침에 맞춰 다양한 평가 요소를 포괄할 수 있

음을 나타낸다.

2) 인간과 인공지능 도구가 생성한 평가 기준 간 비교

연구 결과, 인간과 인공지능 도구가 생성한 평가 기준 간 유사성과 차이점이 있다. 인간과 인공지능 도구가 생성한 평가 기준 간 가장 큰 유사점은 인간과 인공지능 도구 모두 글의 내용과 관련한 범주에 가장 높은 중요도를 부여한다는 점에 있다. 이는 글쓰기 평가에서 내용의 중요성을 강조한 선행 연구들[17,19]과 일치하는 결과이다. 인간이 개발한 평가 기준은 내용 관련 범주에 가장 많은 세부 항목을 두었고, 인공지능 도구들은 내용 범주에 더 높은 점수 비중을 두어 그 중요성을 강조하고 있다. 이는 인간과 인공지능 평가 도구 간 협업 가능성이 높다는 것을 보여준다.

그러나 인간과 인공지능 도구가 생성한 평가 기준 사이에는 몇 가지 중요한 차이가 존재한다. 첫째, 인간이 개발한 평가 기준은 각 평가 범주 내에서 더 세부적인 항목을 포함하고 있으며, 특히 내용 평가에서 구체성과 깊이를 강조한다. 이는 인간 평가자가 글의 미묘한 뉘앙스를 파악하고, 각 요소를 정교하게 평가하려는 의도를 반영한다. 반면, GPT와 Claude는 평가 기준을 더 간결하게 설정하는 경향이 있다. 즉, 인간의 평가 기준은 미시적 구조(문장-문단)와 거시적 구조(전체 글)를 균형 있게 평가하는 반면, 인공지능 도구들은 주로 거시적 구조에 초점을 맞추는 경향을 보였다. 예를 들어, GPT는 '주제 명확성', '논리적 타당성', '창의성 및 독창성'과 같은 항목을 포함하여 글의 핵심 요소에 집중하고, Claude는 '주제의 명확성과 적절성', '논리의 일관성과 설득력' 등을 통해 논리적 구조를 강조한다. 이는 인공지능 도구가 평가 과정에서 효율성을 추구하는 반면, 인간 평가자는 더 깊이 있는 평가를 시도한다고 해석할 수 있으며, 인공지능 모델들이 방대한 사전 정보를 바탕으로 입력된 답변의 적절성을 판단하는 특성[21]에서 기인한 것으로 볼 수 있다. 이러한 차이는 인간과 인공지능의 과학 탐구 보고서 채점 능력을 비교한 연구[4]의 결과와 맥을 같이 한다고 볼 수 있다. 인간 평가자는 해당 과제에 대한 폭넓은 이해를 바탕으로 일부 문항에서 GPT보다 더 엄격한 기준을 적용한 반면 GPT는 주어진 학생 응답만을 토대로 평가를 진행하여 상대적으로 관대한 채점 경향을 보였을 가능성과 유사한 것으로 보인다. [5]의 연구에서도 이러한 효율성과 뉘앙스 사이의 상충관계가 언급된 바 있으며, 이는 향후 연구에서 인공지능 도구의 보조 평가가 글쓰기 평가에서 더 나은 균형을 찾을 수 있도록 하는 과제가 될 것이다.

또한, 언어적 및 문화적 고려사항에서도 차이가 나타났다. GPT와 Claude는 주로 영어 기반으로 개발되었기 때문에, 한국어 글쓰기 평가에서 띄어쓰기와 같은 세부 요소를 충분히 반영하지 못할 가능성이 있다. 연구 결과에 따르면, 어법 평가에 있어서 인간과 인공지능 도구 간의 차이점이 두드러졌다. 인간의 평가 기준이 맞춤법, 띄어쓰기, 문법 규칙 준수 등 정확성에 초점을 맞춘 반면, 인공지능 도구들은 어휘의 다양성과 적절성, 문장 구조의 다양성 등을 추가적으로 고려하였다. 이러한 차이는 [4]의 연구에서 언급된 바와 같이, 인공지능 도구들의 한글 처리 능력 향상에도 불구하고 여전히 존재하는 언어적 특성의 차이에서 기인한 것으로 볼 수 있으며, [1]의 연구에서 지적된 바와 같이, 인공지능 도구의 언어적 및 문화적 적응이 필요함을 시사한다. 따라서 한국어 글쓰기 평가를 위한 인공지능 도구 개발 시, 한국어의 특성을 충분히 반영할 수 있는 방안을 모색해야 할 것이다.

마지막으로, 인공지능 도구들은 평가 범주와 세부 항목에 따라 차등화된 배점 체계를 도입하였다. 이는 인간 평가 기준의 균일한 배점 체계와 대조적으로, 글의 다양한 측면에 상대적 중요도를 부여하려는 시도를 반영한다. 그럼에도 불구하고 GPT와 Claude 모두 '내용' 범주에 가장 높은 배점을 할당하여, 글의 내용적 요소를 가장 중요하게 여기는 경향을 보였다. 이는 앞서 언급한 바와 같이 인간과 인공지능 도구 모두 글쓰기 평가에서 내용의 중요성을 인식하고 있음을 보여준다.

3) 제언

본 연구는 인공지능 도구가 생성한 평가 기준이 포괄적이고 효율적인 평가 도구로서의 잠재성을 가지고 있음을 보여주었다. 그러나, 인공지능 도구가 평가 기준을 생성할 때 나타날 수 있는 중복성 및 모호성 문제는 여전히 해결해야 할 과제로 남아 있다. 본 연구의 제한점 및 후속 연구 제언은 다음과 같다. 첫째, 본 연구에서 비교 대상으로 삼은 인간이 생성한 평가 기준은 '글쓰기 의사소통 역량 채점 기준'에 특화되어 있었던 반면, 인공지능 도구에 요청한 평가 기준은 일반적인 글쓰기 평가에 초점을 맞추었다. 이러한 프롬프트 내용의 차이로 인해 평가 기준의 양적, 질적 차이가 발생했을 가능성이 있다. GPT의 경우 프롬프트에 따라 결과가 달라질 수 있다는 점[5]을 고려할 때, 향후 연구에서는 프롬프트 내용의 변화에 따른 평가 기준의 비교 분석이 필요할 것이다. 둘째, 본 연구에서는 인공지능 도구가 최초로 제시한 채점 기준만을 활용하였다. 그러나 반복적으로 기준 생성을 요청했을 경우 결과가 달라질 수 있으므로, 다양한 시도를 통해 생성된 평가 기준들의 차이를 분석하는 후속 연구가 요구된다. 이를 통해 인공지능 도구를 활용한 평가 기준 생성의 질적 향상을 도모할 수 있을 것이다. 셋째, 본 연구에서 도출된 평가 기준을 실제 채점에 적용하여 비교하는 연구가 필요하

다. 각 기준의 신뢰도와 일치도를 검증하고, 실제 교수자들의 선호도를 조사하는 등 실용적인 관점에서의 후속 연구가 이루어져야 할 것이다. 마지막으로, 본 연구에서는 인간과 인공지능 도구의 평가 기준을 비교하기 위해 특정 선행연구[20]에서 제안한 단일 평가 기준만을 활용하였다. 이 평가 기준을 선택한 구체적인 근거는 앞서 설명하였으나, 다양한 평가 기준을 적용했을 때 나타날 수 있는 차이점을 고려하지 못한 한계가 있다. 따라서 향후 연구에서는 다양한 평가 기준을 비교 분석함으로써, 인공지능 도구의 평가 방법을 더욱 정교화하고 발전시킬 필요가 있다. 이러한 다각적인 접근은 인공지능 도구의 성능을 더욱 정확하게 파악할 수 있게 할 뿐만 아니라, 인간과 인공지능의 상호보완적 협력 방안을 모색하는 데에도 중요한 통찰을 제공할 것이다.

이 연구의 결과는 인공지능 도구의 평가 결과가 인간이 의도한 평가 결과로 도출하기 위해 사전적인 조건을 제시하는 등의 작업이 필요하다는 점을 보여주고 있다. 인공지능 도구는 포괄적이고 구조적인 평가 기준을 도출하는 반면, 인간은 상대적으로 미시적 평가 기준을 도출하는 경향이 있으므로 이런 특징을 고려하여 인공지능 평가 도구를 활용할 필요가 있다. 즉, 인공지능 도구의 평가 결과에 대한 인간의 감독과 개선이 필요하며, 인공지능과 인간 평가의 협력적 접근이 중요하다[10,22]. 향후 연구에서는 인공지능 도구가 다양한 글쓰기 장르에 적응할 수 있도록 훈련하는 방법을 탐구해야 하며, 인공지능 도구가 생성한 평가 기준이 교육 현장에서 어떻게 활용될 수 있을지를 검토해야 할 것이다. 또한, 인공지능 도구의 언어적, 문화적 적응을 통해 더 정확하고 신뢰할 수 있는 평가 기준을 개발하는 것이 필요하다. 인공지능 도구의 글쓰기 평가 기준은 포괄적이고 효율적인 평가 도구로 활용될 수 있지만, 인간 감독과의 협력적 접근이 여전히 중요하다. 여전히 현재 시점에서 인공지능 도구는 인간 평가의 대체가 아닌 보완적 도구로 활용될 때의 가치가 높다고 볼 수 있으며, 어떻게 협력해야 할 것인가에 대한 지속적인 연구가 필요하다.

감사의 글

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

참고문헌

[1] W. Holmes, M. Bialik, and C. Fadel, “Artificial intelligence in education: Promise and implications for teaching and learning,” Boston: Center for Curriculum Redesign, 2019.

[2] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, “Systematic review of research on artificial intelligence applications in higher education – where are the educators?” *International Journal of Educational Technology in Higher Education*, vol. 16, no. 1, pp. 39-66, October, 2019.

[3] S. Hong, B. Cho, I. Choi, K. Park, H. Kim, Y. Park, and J. Park, “Artificial intelligence and edutech in school education,” Jinchoen: Chungbuk, Korea Institute for Curriculum and Evaluation, 2020.

[4] S. Park, B. Lee, Y. Lee, E. Ham, and S. Lee, “Exploring the possibility of science-inquiry competence assessment by ChatGPT-4: Comparisons with human evaluators,” *Korean Journal of Educational Research*, vol. 61, no. 4, pp. 299-332, June, 2023.

[5] S. Park, Y. Hong, and B. Lee, “A study on exploring the potential of ChatGPT in writing skills assessment : Focusing on essay writing,” *Korean Journal of Educational Research*, vol. 62, no. 5, pp. 219-248, August, 2024.

[6] J. W. Sung and B. C. Shin, “Exploring the feasibility of automatic scoring of written test using ChatGPT: Focusing on the world geography written test,” *Journal of the Association of Korean Geographers*, vol. 12, no. 3, pp. 415-432, September, 2023.

[7] W. J. Yoon, Y. H. Koo, and H. Jang, “A study on the performance of ChatGPT in evaluating coverletters and its potential to robotic process automation,” *Journal of Organization and Management*, vol. 47, no. 4, pp. 27-51, November, 2023.

[8] E. H. Ham, S. Park, B. Lee, S. Lee, Y. Lee, and Y. Hong, “Characteristics of GPT-4 automated scoring of scientific inquiry competency,” *The Journal of Educational Information and Media*, vol. 30, no. 3, pp. 713-742, June, 2024.

[9] R. H. Nehm, M. Ha, and E. Mayfield, “Transforming biology assessment with machine learning: automated scoring of written evolutionary explanations,” *Journal of Science Education and Technology*, vol. 21, no. 1, pp. 183-196, April, 2011.

[10] S. Oh, J. Yoon, Y. Chung, Y, Cho, H. Shim, and O.

N. Kwon, “Analysis of generative AI's mathematical problem-solving performance: Focusing on ChatGPT 4, Claude 3 Opus, and Gemini Advanced,” *The Mathematical Education*, vol. 63, no. 3, pp. 549-571, August, 2024.

[11] H. Kim and S. Sul, “Service design-focused comparative analysis of intelligent expert assistant (IEA) and task-oriented rule-based chatbot,” *Journal of Digital Contents Society,* vol. 25, no. 6, pp. 1443-1452, June, 2024.

[12] J. S. Kim, “The current status and future direction of communication competency subject according to changes in university education,” *Journal of Humanities*, vol. 35, pp. 185-217, July, 2023.

[13] S. Y. Lee, and J. H. Kim, “A study of the instructors’ perceptions on university writing and key competencies,” *Research on Writing*, vol. 20, pp. 135-163, March, 2014.

[14] S. Kim, “A study on automated essay evaluation method for argumentative writing task using deep learning based natural language processing technique - Based on model of collaborative scoring between teacher scorer and machine scorer,” Ph. D. dissertation, Korean National University of Education, Chungbuk, 2022.

[15] H. Y. Park, S. S. Kim, K. H. Kim, M. J. Lee, K. G. Kim, and J. Y. Kim, “Substanitializing methods of restricted and extended response essay assessment through enforcing the instruction-assessment alignment,” Jincheon, Chungbuk: Korea Institute for Curriculum and Evaluation, December, 2019.

[16] Y. S. Lee, S. G. Gu, and M. B. Lee, “Currents and suggestions of scoring reliability and validity,” Seoul: Korea Institute for Curriculum and Evaluation, 2013.

[17] Y. M. Park, “The method and procedure of evaluating writing ability,” *The Education of Korean Language*, vol. 99, pp. 1-29, June, 1999.

[18] B. G. Min, K. Y. Nam, S. H. Kim, S. M. Jang, S. J. Lee, and E. S. Kwon, “Development of a national writing skills assessment system in 2023,” Seoul: National Institute of Korean Language, 2023.

[19] B. G. Min, Y. Oh, S. Lee, S. Ahn, K. Kim, M. Kim, J. Son, J. Lee, and S. Chang, “Development of a writing ability diagnosis system for university first-year students – The Seoul National University Essay Examination,” *The Korean Journal of Literacy Research*, vol. 13, no. 6, pp. 45-76, December, 2022.

[20] S. Park, B. Lee, and Y. Hong, “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*, vol. 16, no. 1, pp. 21-31, February, 2024.

[21] C. K. Lo, “What is the impact of ChatGPT on education? A rapid review of the literature,” *Education Sciences*, vol. 13, no. 4, pp. 410-425, April, 2023.

[22] D. Lee, H. Shim, and J. Baek, “Exploration on the feasibility of utilization and teacher perceptions of using ChatGPT for student assessment in Science,” *Journal of the Korean Association for Science Education*, vol. 44, no. 1, pp. 119-130, February, 2024.

[23] Chsosun Biz, “Human-level comprehension” Claude 3 unveiled: ChatGPT and Gemini on edge... Anthropic shaking up the generative AI landscape,” March 12, 2024. [Online]. Available: https://biz.chosun.com/it-science/ict/2024/03/12/XDRFR5TOYVFWRCGJKGIBCCYECQ/

[24] News2day, “Generative AI, Encroaching on human creativity! (62) How far ahead is Anthropic’s Claude compared to ChatGPT-4? (part 2),” July 16, 2024 [Online]. Available: https://www.news2day.co.kr/article/ 20240715500210

[25] News2day, “Generative AI, Encroaching on human creativity! (61) How far ahead is Anthropic’s Claude compared to ChatGPT-4? (part 1),” July 9, 2024 [Online]. Available: https://www.news2day.co.kr/article/ 20240708500134

[26] S. Lee, J. Kim, and H. Cheon, “A study on the evaluation of expository writing skills of middle school students,” *Korean Literature & Language Education*, vol. 40, pp. 71-99, August, 2022.

박 소 영 (So-Young Park)_정회원

1998년 2월 : 서울대학교 교육학과 졸업
2000년 8월 : 서울대학교 대학원 교육학과 석사(교육행정)
2003년 5월 : University of Wisconsin, Madison 교육행정 박사
2009년 9월 ~ 현재 : 숙명여자대학교 교육학부 교수
〈관심분야〉 교육정책, AI 기반 교육 평가, 교사교육

이 병 윤 (ByungYoon Lee)_정회원

2012년 8월 : University of Minnesota, Twin Cities 심리학과 졸업
2016년 8월 : 서울대학교 교육학과 석사
2023년 2월 : 서울대학교 교육학과 박사
2023년 3월 ~ 현재 : 숙명여자대학교 교육연구소 전임연구원
〈관심분야〉 교육심리, AI 기반 교육 평가