컴퓨터교육학회 논문지 2026년 제29권 제7호
https://doi.org/10.32431/kace.2026.29.7.002

정보 교과에서 AI 기반 자동채점 시스템의 오류 유형 및 루브릭 구체화 영향 분석

Analysis of the Error Type and the Impact of Rubric Refinement in AI-Based Automated Grading Systems for Informatics Education

정유하† · 안성진††
Yuha Jeong† · Seongjin Ahn††

요약

본 연구는 고등학교 정보 교과의 서·논술형 평가에서 AI 기반 자동채점 시스템의 현장 적용성을 검증하고 프롬프트 엔지니어링을 통한 루브릭 구체화 방안을 모색하였다. 이를 위해 다중 LLM 기반의 경기도교육청 플랫폼 '하이러닝'을 활용하여 실증 분석을 수행하였다. 연구 결과, AI는 인간 평가자와 유의미한 합치도(QWK 평균 0.73)를 보이며 기존 결과 중심 평가 도구(Online Judge)의 한계를 보완하는 과정 중심 평가의 도구로서 현장 적용 가능성을 확인하였다. 그러나 동시에 AI 채점 과정의 오류 양상을 '유창성 편향'과 '형식적 경직성' 등으로 체계화하여 분석하였으며 루브릭 구체화가 교사의 교과 전문성 및 평가 설계 역량에 기반한 고도의 의사결정 과정임을 규명하였다. 이에 본 연구는 AI를 초안 채점자로 활용하는 '인간-AI 협력(HITL)' 모델의 도입을 제안하며 이를 지원하기 위한 표준 프롬프트 라이브러리 구축 및 교사의 AI 평가 리터러시 강화를 제언한다.

주제어 정보 교육, 대규모 언어 모델, 프롬프트 엔지니어링, 루브릭 구체화, 오류 유형 분석, 인간-AI 협력 평가

ABSTRACT

This study evaluates the field applicability of AI-based automated grading for essay-type assessments in high school informatics and explores rubric refinement via prompt engineering. Using the LLM-based 'Hi-Learning' platform, the study found a significant agreement between AI and human raters (mean QWK=0.73), validating AI as a process-oriented tool that overcomes the limitations of traditional Online Judges. However, qualitative analysis identified specific error patterns, such as 'fluency bias' and 'formal rigidity', revealing that rubric refinement is a sophisticated pedagogical decision-making process. Consequently, this study proposes a 'Human-in-the-Loop (HITL)' model, suggesting the development of standardized prompt libraries and the enhancement of teachers' AI assessment literacy to ensure evaluation validity.

Keywords Informatics Education, Large Language Models(LLM), Prompt Engineering, Rubric Refinement, Error Analysis, Human-in-the-Loop(HITL) Assessment

†정회원 성균관대학교 대학원 컴퓨터교육학과 박사과정
††종신회원 성균관대학교 컴퓨터교육학과 교수 (교신저자)

논문투고 2026년 02월 10일
심사완료 2026년 04월 28일
게재확정 2026년 05월 17일
발행일자 2026년 07월 13일

1. 서론

디지털 대전환에 발맞추어 교육 현장은 비판적 사고력과 문제 해결력을 측정하는 서·논술형 평가로 전환되고 있다[1, 2]. 이에 대응하여 평가의 효율성 및 피드백의 즉시성을 높이기 위해 최근 시도교육청을 중심으로 'AI 기반 자동채점 시스템' 도입이 적극 추진되고 있다[1, 3].

그러나 학생의 성적을 결정짓는 고부담 시험(High-stakes testing) 환경에서 AI의 오채점이나 비일관성은 심각한 공정성 논란을 야기할 수 있다. 따라서 AI 채점 시스템의 교육 현장 안착을 위해서는 효율성에 앞서 교사 평가 기준과의 부합도 및 신뢰도 검증이 선결되어야 한다[1].

특히 2022 개정 교육과정의 정보(Informatics) 교과는 문제 해결 과정과 논리를 중시하는 과정 중심 평가를 지향한다[4]. 그러나 기존의 온라인 저지(Online Judge) 시스템은 결과의 정오(Correctness)만을 판별하거나 사소한 문법 오류에도 0점을 부여하는 등 결과 중심의 이분법적 평가로 학습자의 논리적 사고 과정을 온전히 평가하지 못한다는 구조적 한계를 지닌다[5]. 최근 거대언어모델(LLM)이 대안으로 주목받고 있으나 환각(Hallucination) 오류와 문맥 의존성 문제는 평가 타당성을 위협하는 요인이 된다[6, 7]. 더욱이 기존 연구들은 주로 영미권 데이터나 범용 교과에 편중되어 있어 한국어 기반 정보 교육 현장의 정교한 서술형 답안을 평가하기에는 부족함이 있다[8-12].

이에 본 연구는 정보 교과 서·논술형 평가에 최적화된 AI 자동채점 설계 원리를 규명하고자 한다. 구체적으로 채점 과정에서 발생하는 오류 유형을 분석하고 이를 개선하기 위한 루브릭 구체화 방안을 실증적으로 검증하여 교사가 평가 타당도를 확보할 수 있는 전략을 제시하는 데 목적이 있다. 연구 문제는 다음과 같다.

RQ1. 고등학교 정보 교과 서·논술형 평가에서 AI 기반 자동채점 결과는 인간 교사 채점 결과와 어느 정도의 일치도를 보이는가?

RQ2. 평가 문항의 세부 영역에 따라 AI 기반 자동채점의 신뢰도에 유의미한 차이가 존재하는가?

RQ3. 고등학교 정보 교과 서·논술형 평가 과정에서 식별된 AI 기반 자동채점 오류 유형과 그 특성은 무엇인가?

RQ4. 교사의 루브릭 정교화는 AI 기반 자동채점의 오류 개선과 신뢰도 향상에 어떠한 영향을 미치는가?

2. 이론적 배경

2.1 2022 개정 정보과 교육과정과 과정 중심 평가

2022 개정 교육과정은 디지털 대전환에 대응하여 학교 교육 전반에서 디지털 기초 소양 함양을 핵심 과제고 명시하고 있다[2]. 이에 정보 교과는 단순한 기능 습득을 넘어 실생활의 복잡한 문제를 컴퓨팅의 관점에서 이해하고 해결하는 컴퓨팅 사고력 함양을 핵심 목표로 한다[5]. 이에 따라 기존 결과 중심 평가의 한계를 극복하고 학습자의 사고 절차를 진단하는 과정 중심 평가로의 전환이 요구되고 있다[2]. 특히 지식·기능·태도를 아우르는 다면적 평가와 함께 서·논술형 평가 확대 및 즉각적인 피드백 제공을 통한 학습자의 사고력 강화가 핵심적인 실천 과제로 강조되고 있다[2, 6].

2.2 정보 교과에서의 서·논술형 평가의 필요성과 기존 자동 평가 시스템(Online Judge)의 한계

학습자의 고차원적 사고 과정을 결과 중심의 평가로 측정하는 데에는 근본적인 한계가 존재한다[12]. 기존 연구에 따르면 대학수학능력시험 국어 영역의 작문 문항과 같은 선다형 평가는 학습자의 실제 작문 수행 능력을 약 21.5%밖에 설명하지 못하는 것으로 나타났다[12]. 이는 객관식 문항이 글쓰기의 사고 과정을 온전히 담아내지 못함을 시사하며 이러한 한계는 정보 교과에도 동일하게 적용된다. 특히 코드의 실행 흐름을 추적하고 이를 논리적으로 기술하는 능력은 프로그래밍 역량과 직결되므로[13] 단순 결과 확인을 넘어 사고의 절차를 진단하는 서술형 평가의 도입이 필요하다.

그러나 기존의 온라인 저지 시스템은 입출력 결과만을 확인하는 블랙박스 테스트(Black-box Testing) 방식에 의존하여[4] 서·논술형 평가를 지원하는 데 구조적 한계를 지닌다. 이는 학습자의 알고리즘 설계 과정이나 코드의 가독성 등 질적 요소를 간과하고[10, 11] 단순 정오 판별에 그쳐 구체적인 오류 원인 분석이나 튜터링 피드백을 제공하지 못한다는 비판을 받고 있다[5]. 따라서 학습 수준이 높아질수록 코드의 내부 논리를 분석하는 정적 평가 비중을 높여야 한다고 주장한다[10, 11]. 이에 현장에서는 서·논술형 평가 확대에 따른 평가의 공정성을 확보하고 교사의 업무를 경감할 수 있는 객관적이고 효율적인 지능형 채점 시스템의 도입을 시급한 과제로 인식하고 있다[14].

2.3 LLM 기반 자동채점의 기술적 타당성과 한계

최신 LLM(GPT-4o, Gemini 1.5 Pro)은 프로그래밍 및 수학 교과 평가에서 인간과 유사한 정확도를 보이며 효율적인 채점 도구로서의 가능성을 보여주었다[1, 8, 9, 15]. 그러나 높은 잠재력에도 불구하고 이를 고부담 평가에 도입하기에는 본질적인 한계가 따른다. LLM은 통계적 확률에 기반한 텍스트 생성 모델로서 통계적 확률에 따라 단어를 예측하는 '확률적 앵무새(Stochastic Parrots)'에 불과하여 학습자의 심층적 의도를 파악하지 못하거나 거짓 정보를 생성하는 환각 오류를 범할 수 있다[6, 16]. 또한 내용의 본질보다 줄바꿈이나 오타 등 사소한 형식적 변인에 민감하게 반응하여 채점 결과가 달라지는 경직성을 보이며[17, 18] 내용의 정확성보다 언어적 유창성에 편향되거나 학습 데이터에 내재된 사회적 편향이 결과에 반영될 위험이 있어 완전 자동화된 채점을 적용하기에는 기술적 신뢰도가 부족한 실정이다[19, 20].

2.4 교육 분야에서의 고위험 AI

AI 기반 자동채점 시스템은 학습자의 성취도 평가와 진로에 미치는 파급력이 크므로 단순한 기술적 도구를 넘어 엄격한 법적 규제의 대상이다. 전우정(2024)에 따르면 EU 인공지능 법(EU AI Act)은 학습자의 입학, 평가, 성취도 측정을 수행하는 AI 시스템을 고위험 AI(High-Risk AI)로 분류하며 국내 제22대 국회 발의안들 역시 성적 산출 등 개인 권익에 영향을 미치는 시스템을 고위험 영역으로 정의하여 관리를 강화하는 추세이다[21]. 이러한 법적 기준은 AI의 편향과 오류 방지를 위해 인간의 개입을 필수 의무로 규정한다[21]. 따라서 교육 평가는 완전 자동화가 아닌 교사가 최종 승인 권한을 갖는 '인간-AI 협업(Human-in-the-Loop)' 구조를 기반으로 설계되어야 한다.

2.5 평가 루브릭의 구체화와 HITL 전략

신뢰할 수 있는 AI 기반 평가를 위해서는 루브릭의 구체화가 선행되어야 한다. AI는 추상적인 지시에 취약하므로[5] 평가 준거를 세분화한 분석적 루브릭과 판단 근거를 단계적으로 추론하는 생각의 사슬(CoT, Chain of Thought) 프롬프트를 결합할 때 채점 타당도가 극대화된다[22]. 또한 정교화된 루브릭을 토대로 인간이 개입하는 전략이 필수적이다. 선행 연구에서는 인간이 채점 과정에서 루브릭을 미세하게 조정하여 성능을 향상시켰으며[23] AI를 평가 협력 파트너로 설정하는 협력적 모델을 제안하였다[24]. 즉 AI의 오류를 방지하기 위해서는 구조화된 루브릭, 추론 유도 프롬프트, 그리고 인간 교사의 최종 검수가 유기적으로 결합되어야 한다.

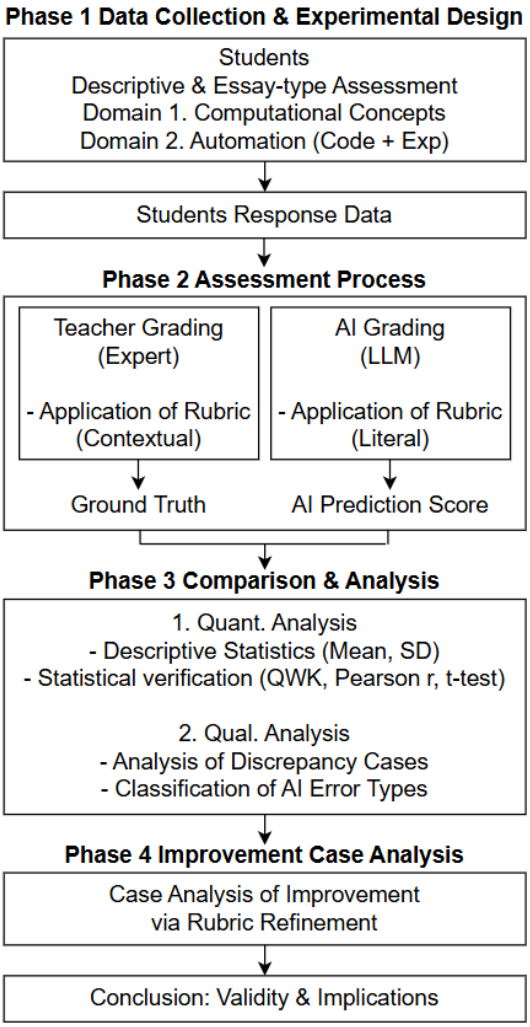


Figure 1. Overall Research Procedure

3. 연구 방법

3.1 연구 설계

본 연구는 Fig. 1과 같은 단계로 진행되었다. 수집된 학습자 답안에 대해 교사와 AI가 1차 평가를 수행한 후, 결과 간 불일치 오류를 유형화하여 이를 토대로 루브릭을 정교화하였다. 이어 개선된 루브릭으로 재평가하여 타당도 변화를 검증하고 AI 기반 평가의 실질적 적용 가능성을 확인하였다.

3.2 연구 대상 및 연구 도구

연구 대상은 2022개정 정보 교과 과정을 이수한 경기도 소재 고1 여학생 총 104명이며 불성실 응답을 제외한 문항별 71~94건의 유효 표본을 최종 분석하였다. 자동채점 도구로는 경기도교육청의 '하이러닝' 플랫폼을 활용하였다. 해당 플랫폼은 특정 상용 업체의 단일 모델에 고정되어 있지 않고 국내 공공망 환경에 자체 구축된 LLM API를 기반으로 운영된다. 교육적 맥락 파악에 최적화된 최신 다중 LLM과 자체 학습 모델을 결합한 유연한 하이브리드 AI 평가 파이프라인을 활용하고 있다. 이를 통해 교사가 설정한 루브릭에 따라 학생의 서술형 답안과 소스코드를 분석하여 자동 점수와 피드백을 생성한다.

3.2.1 평가 방식 및 문항 구성

본 연구는 이미지 압축 및 파이썬 실습 등 51차시의 정규 교육과정을 마친 후 실시된 총괄 평가를 분석 대상으로 한다. 평가 도구는 핵심 성취기준을 반영한 서·논술형 7문항

(NCIC 2문항[25], 자체 개발 5문항)으로 구성하였으며 정보 교사 2인(5년 이상)으로부터 검토를 받아 내용 타당도를 확보하였다. 평가 시행 시 채점 신뢰도를 높이기 위해 수기 답안의 광학 문자 인식 오류를 차단하고자 학생들이 구글 설문지에 직접 입력하는 방식을 채택하였다. 평가 영역은 논리적 서술 능력을 측정하는 '컴퓨팅 개념'과 코드 분석 기반의 '자동화'로 구분하였다(Table 1). Table 2는 각 영역의 특성을 대표하는 문항(Q1, Q6)과 점수대별 실제 학생 답안이다. 이는 전문가 집단이 부여한 실제 점수(Ground Truth)이며 이 데이터는 향후 4.3절에서 다룰 AI 채점 오류 기제 분석의 핵심 기초 자료로 활용된다.

Table 1. Evaluation items

Domain	No	Key Concepts	Assessment Criteria
Computational Concepts	Q1	Data Compression	Understanding principles of data compression
	Q2	Sorting Algorithm	Explaining logical flow and comparing efficiency
	Q3	Modeling	Explaining why modeling is necessary and applying to real-world problems
	Q4	AI Ethics	Critical thinking on AI ethical issues
Automation	Q5	Data Type	Identifying and correcting data type errors
	Q6	Selection Structure	Predicting outputs and modifying conditional code
	Q7	Loop Structure	Tracing nested loops and modifying code

Table 2. Examples of items and answers

No	Item	Student Responses by Score
Q1	Find the incorrect description among the following image compression techniques and explain the reason. GIF: ... (omitted) JPG: Using jpg maintains lossless image quality, so it is used when sending a landscape photo containing various colors and rich details to a friend. PNG: ...(omitted)	[2 pts] "JPG is the wrong explanation. Reduce file size using loss compression. It's wrong to keep the picture quality without loss because it throws away some picture quality."
		[1 pt] "JPG, Lost image quality", "There is a loss in jpg.", "JPG, It's loss compression."
		[0 pts] "GIF", "The storage method is crummy, so the quality is broken and the pixels are visible"
Q6	Read the following code and answer the question. score = 85 if score >= 90: ...(omitted)... print("score:", score, "grade:", grade) [Sub-question 1] Write the execution result of the above code.	[2 pts]"score: 85 grade: B"
		[1 pt] "85 B", "85", "B"
		[0 pts] "score"

Note. Q6 refers only to sub-item 1 (Max score: 2 pts).

3.2.2 채점 기준

채점 기준은 2022 개정 정보과 교육과정의 성취기준에 근거하여 수립하였다. NCIC 제공 2문항은 원안을 준용하였으며 자체 개발 5문항은 지도서의 핵심 개념을 토대로 정답 범위를 구체화하여 평가의 객관성을 확보하였다. 영역별 세부 채점 루브릭은 Table 3 및 Table 4와 같다.

Table 3. Rubric for Computational Concepts

No	Criteria	Pts
Q1	Identifying errors in JPG compression principles (1)	8
	Linking lossy compression to quality degradation (1)	
Q2	Comparison/swapping principles (3)	9
	Tracing sorting steps (3)	
	Analyzing the causes of inefficiency (3)	
Q3	Concept: Explaining the necessity of modeling (3)	12
	Real-world Scenarios: Selecting and justifying appropriate modeling types (3)	
	Selection and justification of modeling types (3)	
	Identifying modeling cases in daily life (3)	
Q4	Explaining characteristics of VR and AI(3)	15
	Clearly stating personal position (3)	
	Providing ethical/social justification (3)	
	Addressing and refuting counter-arguments (3)	
	Structure and clarity of writing (3)	

Table 4. Rubric for Automation

No	Criteria	Pts
Q5	Explaining concepts of data types (1)	3
	Converting strings to integers (2)	
Q6	Providing accurate execution results (2)	7
	Explaining 'if-elif-else' structures (3)	
	Inferring outputs from changed inputs (2)	
Q7	Providing accurate execution results (1)	5
	Explaining compound assignment(+=) (2)	
	Updating code for new conditions (2)	

3.3 자료 수집 및 준거(Ground Truth) 설정

본 연구는 AI 자동채점 결과의 신뢰도 검증 및 타당성 확보를 위해 현직 정보 교사 집단의 채점 결과를 준거(Ground Truth)로 설정하였다. 평가의 일관성을 확보하기 위해 정보교육 전문가인 연구자가 전체 답안(590개)을 대상으로 1차 채점 후 3주의 간격을 두고 2차 채점을 실시한 결과 채점자 내 신뢰도는 0.85로 높게 나타났다.

또한 객관성 확보를 위해 현직 정보 교사 2인(7년 이상)과 교차 검증을 실시하였다. 채점 결과를 분석한 결과 QWK는 각각 0.80, 0.86으로 도출되었으며 이는 Landis & Koch(1977)의 기준에 따라 거의 완벽한 일치에 해당한다. 설정한 채점 루브릭과 준거 기준이 전문적 판단 기준과 높은 수준에서 부합함을 입증하여 전문가 집단에 의해 객관성이

검증된 연구자의 1차 채점 결과를 최종적인 참값으로 정의하고 AI 자동채점 결과와의 비교 분석에 적용하였다.

3.4 자료 분석

3.4.1 양적 분석 방법

본 연구는 Python 환경에서 라이브러리를 활용하여 Table 5와 같은 검증 절차를 구현하였다. 교사와 AI의 채점 데이터를 입력받아 혼동 행렬 및 히트맵을 생성하고 채점 경향을 시각적으로 분석하였다(Line 1-4). AI 자동채점 결과가 인간 교사와 어느 정도의 일치도를 보이는지 확인하기 위해(RQ1) 오차에 엄격한 페널티를 부여하는 이차 가중 카파(Quadratic Weighted Kappa; QWK, Cohen, 1968)를 산출하였다(Line 5-7). 또한 피어슨 상관계수()와 대응 표본 t-검정()을 통해 두 평가자 간의 선형적 상관관계와 시스템적 편향을 통계적으로 판별하였다(Line 8). 최종적으로 이러한 통계 지표들을 평가 문항 영역(컴퓨팅 개념, 자동화)별로 교차 분석함으로써 문항 특성에 따른 AI 자동채점 신뢰도의 유의미한 차이를 규명하였다(RQ2; Line 9).

Table 5. Algorithm for Calculating Validity and Reliability (Based on Cohen, 1968)[26]

Pseudocode	
Input	H, A : Vector of Human/AI scores; : Number of grading scales
Output	O(Confusion Matrix), k(QWK), r(Pearson), p (P-value)
Procedure	1: function EVALUATE_MODEL(H, A, N) 2: O←Initialize $N\times X$ matrix with zeros 3: for each $h\in H, a\in A$ do $O_{h_i,a_k}\leftarrow O_{h_i,a_k}+1$ 4: Generate Heatmap from O 5: $W_{i,j}\leftarrow\frac{(i-j)^2}{(N-1)^2}$ for $i,j\in 0,\ldots,N-1$ 6: E←Outer product of marginal totals 7: $k\leftarrow 1-\frac{\sum(W\cdot O)}{\sum(W\cdot E)}$ 8: r←Pearson(H, A); p←Paired_T_Test(H, A) 9: return k, r, p, O

3.4.2 질적 분석 방법

양적 분석 결과를 바탕으로 AI 자동채점의 오류 기제를 규명하기 위해 Table 6과 같은 '오류 사례 식별-원인 진단-유형 범주화-루브릭 개선 및 최종 성능 검증'의 질적 분석 절차를 수행하였다

Table 6. Procedure for Qualitative Analysis

Step	Process	Key Activities	RQ
1	Error case Identification	Extract individual error cases by comparing teacher and AI scores on a 1:1 basis.	–
2	Coding	Analyze underlying causes of AI misjudgment (e.g., fluency bias, overlooking logical errors).	–
3	Categorization	Group diagnosed causes by common attributes to establish an AI scoring error taxonomy.	RQ3
4	Refinement	Refine scoring rubrics by specifying detailed criteria based on identified error types.	–
5	Validation	Re-score with refined rubrics and quantitatively validate improvements in reliability.	RQ4

AI와 교사의 채점 결과가 불일치하는 답안을 분석 대상으로 확정하였으며 하나의 답안에서 복수의 오류가 발견되는 특성을 반영하여 중복 코딩(Doouble Coding) 방식을 채택하였다. 이를 통해 도출된 오류 유형별 분석 결과는 4.2절에서 기술하였다.

4. 연구 결과

4.1 AI 자동채점 결과의 영역별 타당성 검증

본 절에서는 질적 분석을 통해 개선된 루브릭(4.3절 참조)을 적용한 AI 자동채점 최종 결과의 타당성을 분석하였다. 교사 채점과의 일치도를 기반으로 정량적 검증을 수행한 결과는 Table 7 및 Table 8과 같다.

Table 7. Descriptive Statistics

Domain	No	N	Score	AI		Teacher	
				M	SD	M	SD
Computational Concept	Q1	94	2	1.48	0.60	1.44	0.59
	Q2	90	9	6.83	1.13	6.97	1.17
	Q3	89	12	10.12	1.65	10.04	1.90
	Q4	85	15	12.26	1.50	12.08	2.31
Automation	Q5	83	3	2.1	1.19	2.11	1.10
	Q6	78	7	4.03	2.06	4.41	1.85
	Q7	71	5	2.89	1.52	2.79	1.32

Table 8. Statistical Test

Domain	No	N	Score	QWK	Pearson r	p-value
Computational Concept	Q1	94	2	0.70	0.70	0.37
	Q2	90	9	0.67	0.67	0.18
	Q3	89	12	0.71	0.72	0.58
	Q4	85	15	0.51	0.56	0.40
Average				0.65	0.66	0.38
Automation	Q5	83	3	0.90	0.91	0.83
	Q6	78	7	0.79	0.80	0.11
	Q7	71	5	0.86	0.87	0.28
Average				0.85	0.86	0.41

분석 결과, Table 7, 8과 같이 교사와 AI 채점 간 통계적으로 유의미한 차이는 없었으며($p>.05$) 평균의 차이도 최소 0.04로 큰 차이를 보이지 않았다. 표준편차 역시 유사한 분

포를 보여 AI가 학습자의 성취 수준을 변별하고 있음을 확인하였다. 또한 QWK(0.51~0.90)와 상관계수(0.56~0.91)를 통해 AI가 시스템적 편향 없이 교사의 평가 패턴을 정밀하게 모사함을 입증하였다.

다만 문항의 성격에 따라 정확도의 차이가 관찰되었다. 논리 구조가 명확한 자동화 영역(Q5~Q7)은 높은 일치도(평균 QWK=0.85)를 보였다. 반면 비정형적 응답이 포함된 개방형 문항(Q4)은 QWK 0.51로 상대적으로 낮은 일치도를 보였다. 이는 문항의 개방성이 높을수록 AI 자동채점의 난도가 상승하며 루브릭의 정교화 정도에 따라 신뢰도가 민감하게 반응함을 시사한다.

4.2 채점 오류 유형 및 특성에 대한 질적 분석

교사와 AI의 채점 불일치가 발생한 308건의 답안을 대상으로 중복 코딩을 실시한 결과 총 381건의 오류 현상이 식별되었다. 오류 유형별 빈도와 비율은 Table 9와 같다. 루브릭의 추상적 지시어 미인식(36.5%)이 가장 높은 비중을 차지하였으며 학습자 언어의 비유연성(15.5%)과 형식적 경직성(15.2%)이 뒤를 이었다.

Table 9. Frequency of AI Scoring Error Types

Error Type	Freq.	%
Undefined Abstract Instructions	139	36.5
Inflexibility to Learner Language	59	15.5
Formal Rigidity	58	15.2
Disregard for Fragmented Answers	45	11.8
Fluency Bias	33	8.7
Simple Keyword Matching	33	8.7
Overlooking Logical Errors	14	3.6
Total	381	100.0

Note. N=381 instances from 308 cases via double coding.

AI 채점 오류의 발생 기제를 체계화하기 위해 Table 11과 같이 모델의 처리 과정에서 발생할 수 있는 오류, 루브릭 설계의 측면, 모델의 역량 세 가지 차원으로 구조화하여 분석하였다.

AI는 문장의 형태적 완성도가 높을 경우 기술적 오개념을 간과하는 유창성 편향을 보였다. Table 10과 같이 학생의 답안에 명백한 오개념을 포함하고 있더라도 문장 구조가 유려하고 인과관계의 접속어('~때문에' 등)가 적절히 사용된 경우 AI는 이를 논리적인 답안으로 판단하여 만점을 부여하였다.

Table 10. Case Analysis 1

Item	Description
Question	Identify the incorrect description among image compression techniques (GIF, JPG, PNG) and describe the reason.
Student Response	"JPG. JPG has no loss when compressing 256 colors, but images with more than 256 colors are converted to supported colors, resulting in image loss. Therefore, it is an incorrect explanation."
Human Scoring	1 point (Awarded partial credit for identifying the correct keyword 'JPG', but penalized for the clear technical misconception confusing it with GIF)
AI Scoring	2 points (Full score)
AI Inference Analysis	AI failed to detect a conceptual error due to fluency bias.

반면 정답의 핵심을 명확히 제시했음에도 문장 성분이 생략되거나 구어체로 작성된 단답형 답안에 대해서는 과소평가 결과를 초래하였다. 예를 들면, '손실 압축이고 이미지의 세밀한 부분의 정보를 제거하기 때문'과 같이 정답을 정확히 언급했지만 문장 성분이 일부 생략된 비정형적 구조인 답안을 AI는 미완성 답안으로 간주하여 감점하였다. 이는 내용의 질적 수준보다 서술의 유창성에 따라 평가 결과가 달라지는 경향이 있음을 시사한다.

루브릭 설계 측면의 오류는 미인식된 추상적 용어에 대한 경직성이다. 교사가 단계별 점수 부여를 위해 루브릭에 관습적으로 명시하는 '부분적으로 서술함.'과 같은 표현이 이에 해당한다. 교사는 해당 지표를 바탕으로 학생의 성취 수준에 따라 유연하게 중간 점수를 부여한다. 그러나 AI는 이러한 추상적 용어를 해석하지 못하고 Table 12의 사례와 같이 관련 용어가 단순 언급되면 관대하게 점수를 부여하거나 아예 점수를 부여하지 않는 이분법적 판단을 내리는 양상을 보였다. 이는 AI 자동채점의 최적화를 위해서는 루브릭의 기술적 구체화가 필수적임을 시사한다. 또한 입력 양식 처리에 있어서도 한계가 관찰되었다. 학생 답안에 줄바꿈이 있는 등 입력 형식이 학습 데이터와 상이할 경우 내용의 정오와 무관하게 오답 처리하는 경향을 보였다.

Table 11에 제시된 각 오류 원인들은 상호 배타적이지 않으며 실제 채점 과정에서는 복합적인 상호작용을 일으키는 것으로 분석되었다. 모델의 처리 기제와 루브릭 설계의 한계가 결합하여 나타나는 혼재적 양상은 AI가 오채점하게 만드는 요인이 된다. 이는 AI의 특성을 고려한 루브릭의 정교한 설계가 선행되어야 하며 AI의 강점은 활용하되 약점은 인간 평가자가 보완하는 인간-AI 협업 기반의 채점 설계가 필수적임을 뒷받침한다.

반면 다양한 한계에도 불구하고 AI는 기존의 자동 채점 시스템이 갖지 못한 구문 독립적 논리 평가 역량을 보여주었다. AI는 표면적 문법 오류가 있더라도 알고리즘의 전체적인 논리 구조가 타당할 경우 부분 점수를 부여하는 유연성을 보였다. 이는 결과값의 일치를 넘어 문제 해결 과정의 논리성을 평가하는 과정 중심 평가 도구로서의 교육적 가능성을 보여준다.

Table 11. Qualitative Analysis of AI Assessment Characteristics: Limitations and Strengths

Primary Source	Root Cause (Bias)	Error Type	Sub-factors	Specific Observed Cases
1. Limitations				
Model Processing	Bias towards Fluency & Completeness	OE	Fluency Bias	High scores are awarded for perfect sentence structures even if the content is incorrect.
			Overlooking Logical Errors	The model fails to detect logical contradictions within lengthy, fluent sentences.
		UE	Disregard for Fragmented Answers	Correct standalone keywords are penalized as the model favors complete sentences.
			Inflexibility to Learner Language	Rejects simplified student explanations that lack academic terminology.
	Simple Keyword Matching	OE	Context-free Keyword Acceptance	Points are awarded for the mere presence of keywords even when the logical/causal link is missing.
	Formal Rigidity	UE	Input Formatting	Zero scores given for minor formatting deviations such as line breaks.
Rubric Design	Undefined Abstract Instructions	UE	Unrecognized Abstract Terms	The model failed to interpret "partially" without concrete examples, denying intermediate credit.
2. Strengths				
Model Capability	Syntex-Independent Logic Evaluation	N/A	Algorithm-Oriented Assessment	Validation of potential as a process-oriented assessment tool.
			Differentiation from Existing Systems	Partial credit is granted for correct logical flow despite syntax errors.

Note. OE = Overestimation; UE = Underestimation; N/A = Not

Table 12. Case Analysis 2

Item	Description
Question	Write an essay on the introduction of VR/AI restoration technology for deceased family members (Condition: Include technical characteristics, etc.).
Student Response (Excerpt)	"... VR technology is an emerging technology these days. (Omitted) There are advantages, but also disadvantages. ... Therefore, I am against its social introduction."
Human Scoring	9/15 points (Strictly penalized as the required sub-criterion of 'describing technical characteristics' was missing)
AI Scoring	11/15 points (Overestimated)
AI Inference Analysis	AI's lenient interpretation of abstract criteria led to an overestimation of stylistically superior but technically incomplete responses.

4.3 루브릭 구체화를 통한 개선 사례 분석

본 절에서는 앞서 분석한 AI의 채점 오류를 개선하기 위해 루브릭을 구체화한 과정을 기술한다. 이를 위해 전체 7개 문항 중 평가 영역별 특성을 대표하면서도 교사 채점자와 AI 간의 채점 불일치가 높았던 Q1 문항(QWK=0.29)과 Q6 문항($p < .001$)을 심층 분석 대상으로 선정하였다. 이처럼 줄글로 서술하는 답안과 코드를 작성하는 방식의 상이한 답안 형태에 따라 AI의 대표적인 채점 한계가 뚜렷하게 나타난 두 문항의 사례를 중심으로 초기 루브릭의 한계와 이를 극복하기 위한 개선 전략을 분석하였다.

4.3.1 초기 루브릭의 한계 및 오류 원인 분석

앞서 4.2절에서 식별된 오류 양상을 바탕으로 초기 루브릭의 한계를 분석한 결과는 다음과 같다. Q1 문항에서는 학습자의 비정형 응답에 대한 AI의 낮은 수용성이 한계로 확인되었다. 정답의 핵심 키워드인 'jpg'를 포함하더라도 설명이 간결하거나 구어체 조사(예: '~요')가 결합된 경우 이를 오답으로 처리하는 비유연성을 보였다. 반면 기술적 오개념이 포함되더라도 문장 구조가 유려하면 정답으로 오판하는 유창성 편향이 동시에 관찰되었다. 또한 논리적 인과관계가 빈약하더라도 '손실 압축', '화질 저하'등 핵심어의 단순 나열만으로 정답을 인정하는 등 표면적 텍스트 매칭에 그치는 한계를 보였다. 이러한 오채점의 원인은 Q1 초기 루브릭(Table 13)에서 '명확하게 설명함', '부정확하게 설명함'과 같은 정성적 지표를 사용한 점과 모델 처리과정에서 발생하는 기제가 결합하여 나타난 복합적 오류의 결과로 분석된다.

Table 13. Before changing the Q1 rubric

Concept	Pts	Rubric
Understanding Image Data Compression Techniques	2	The image data compression principle finds the wrong explanation and clearly explains why it is wrong.
	1	The image data compression principle may find an incorrect explanation, but it may inaccurately describe why it is wrong.
	0	All responses are inappropriate.

Table 14. Before changing the Q6 rubric

Concept	Pts	Rubric
Program Results Output	2	Accurately write the results of the program
	1	Partially write the results of the program.
Changing Program Conditions	2	Accurately write the results of the program
	1	Partially write the results of the program.

Note. Items without change were excluded from the table

Q6 문항에서도 마찬가지로 관습적 평가 용어에 대한 AI의 해석 실패가 주된 원인으로 나타났다. 초기 루브릭(Table 14)의 '결과를 정확하게 작성함' 또는 '부분적으로 작성함'과 같은 정성적 표현에 대해 AI는 부분적인 허용범위를 정의하지 못했다. 특히 학습자가 정답을 정확히 작성했음에도 불구하고 AI가 이를 오답으로 처리하는 사례가 빈번하게 관찰되었다. 이러한 결과는 AI가 맥락적·객관적 판단을 수행할 수 있도록 부분 점수의 범위를 명시적 수치로 구체화하고 다양한 정답 케이스를 예시 답안으로 명시해야 할 필요성을 시사한다.

4.3.2 프롬프트 엔지니어링 기반 루브릭 구체화

분석된 오류를 해결하기 위해 본 연구에서는 초기 루브릭의 정성적 표현을 배제하고 AI의 채점 과정을 제어할 수 있는 분석적 루브릭(Table 15, 16)을 설계하였다. 논리적 추론 능력을 평가하는 Q1 문항은 단순 예측이 아닌 논리적인 추론 사슬(CoT) 프롬프팅 구조를 적용하여 루브릭을 분석적 형태로 재설계하였다.

Table 15. After changing the Q1 rubric

Concept		Rubric
Find incorrect description of image data compression techniques	1	There are references to the words 'JPG', 'JPEG', 'jpg', and 'jpeg'.
	0	No mention of the words 'JPG', 'JPEG', 'jpg', and 'jpeg'.
Explain why the description of the image data compression technique is incorrect	1	The answer should explicitly reveal both of the following concepts (A, B). Concept A (cause): Words such as 'loss', 'compression', 'data deletion', 'information abandonment' Concept B (Result): Referring to words such as 'deteriorated quality', 'breaking', 'unrecoverable', and 'blurred'
	0	Failed to explain the features of jpg. [Example of incorrect answer – Never give 1 point] 'JPG has poor image quality' (zero point because there is only a result and no mention of the word 'loss/compression' which is the cause) 'JPG is loss compression' (zero point if there is only cause and no mention of the resulting word 'quality/recovery') "JPG is small." (It's a misconception, so 0 point)

기존의 통합적 평가 기준을 정답 식별, 원인, 결과라는 인과적 논리 단계로 분해하였다. 이는 AI가 수행해야 할 중간 추론 단계의 이정표 역할을 수행한다. 구체화된 루브릭은 답안 내 핵심 용어(예: jpg)를 식별하도록 작성하였고 단순 키워드 매칭으로 인한 위음성을 방지하기 위해 유의어 리스트를 함께 제공하여 문맥적 의미를 파악하도록 하였다. 또한 원인과 결과가 모두 충족될 때만 정답으로 인정하는 AND 조건을 적용하였으며 오답 사례를 구체화하였다.

이 과정에서 연구자는 정답 여부를 판별하기 위한 해당 개념을 정의하는 가장 정확한 어휘가 무엇인지, 그리고 학생 수준에서 허용 가능한 표현의 범위가 어디까지인지를 규명하는 과정을 거쳤다. 이는 루브릭 구체화 과정이 단순한 프롬프트 작성 기술을 넘어 교사의 교과 전문성 및 평가 설계 역량을 정교화하는 과정임을 발견하였다. 결과적으로 Table 15는 LLM의 블랙박스 문제를 방지하고 구체적인 감점 사유를 제시하는 설명 가능성의 기반이 되었다.

프로그래밍 실행 결과를 묻는 Q6 문항은 출력값의 정확성을 검증하는 결정론적 특성을 갖는다. 이에 따라 Table 16과 같이 정성적 표현을 배제하고 엄격한 문자열 매칭 방식을 적용하였다. 초기 루브릭(Table 14)의 모호한 부분적 서술 기준을 '85 또는 B 포함'이라는 객관적 수치와 텍스트 조건으로 전환하여 구체화하였다. 특히 다양한 정답 케이스를 예시 답안으로 추가 명시하여 AI가 텍스트 매칭 조건에 따라 객관적인 채점을 수행하도록 수정하였다.

Table 16. After changing the Q6 rubric

Concept	Pts	Rubric
Program Results Output	2	Correctly write the result of the program execution, 'Score: 85 Grade: B'.
	1	Write part of the results of the program execution (e.g., 'B' or '85 B').
Changing Program Conditions	2	Correctly write the result of the program execution, 'Score: 75 Grade: C'.
	1	Write part of the results of the program execution (e.g., 'C' or '75 C').

Note. Items without change were excluded from the table

4.3.3 루브릭 개선 전후 성능 비교 및 효과 검증

프롬프트 엔지니어링 기반의 루브릭 구체화를 적용한 결과, 문항의 유형에 따라 교사와 AI 간 채점 결과 합치도(QWK)의 변화 양상이 상이하게 나타났다. Table 17에 제시된 바와 같이 개념 설명을 요하는 컴퓨팅 개념 문항(Q1)에서는 QWK가 대폭 향상된 반면 자동화 문항(Q6)에서는 소폭 하락하는 결과가 도출되었다.

Table 17. Comparison of results before and after rubric modification

No	Before				After			
	QWK	Pearson r	p-value	MAE	QWK	Pearson r	p-value	MAE
Q1	0.29	0.29	0.18	0.48	0.70	0.70	0.37	0.21
Q6	0.88	0.90	0.00	0.59	0.79	0.80	0.11	0.77

jpg 압축 개념을 서술하는 Q1 문항의 경우 루브릭 최적화 전 QWK는 0.29로 매우 낮은 수준이었으나 루브릭 최적화 이후 0.70으로 뛰어난 성능 향상을 보였다. 기존의 포괄적 루브릭에서 평가 기준을 구체적이고 논리적 단계로 분해하여 적용한 결과 인간 채점자와의 합치도가 2배 이상 상승하였다. 또한 Fig. 2에서 개선 후 데이터의 밀도가 대각선을 중심으로 뚜렷하게 수렴하는 양상을 통해 채점 신뢰도의 개

선이 입증되었다.

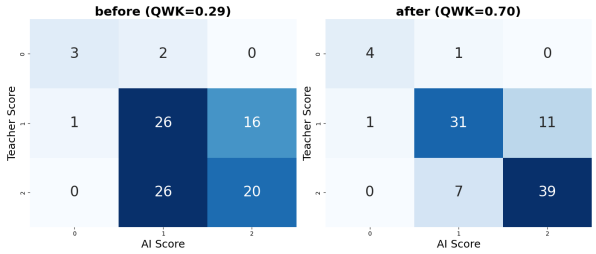


Figure 2. Q1 before and after rubric modification

반면 프로그래밍 코드에 관한 문항 Q6에서는 대응 표본 t-검정 결과, 루브릭 구체화 전에는 유의확률(p)이 .00으로 나타나 교사와 AI간에 통계적으로 유의미한 점수 차이가 존재하였다. 이는 초기 루브릭으로 AI가 채점했을 때 교사의 채점 기준을 완벽히 반영하지 못하고 과소평가하는 경향이 있었다는 것이다. 그러나 루브릭을 정교화한 후에는 유의확률이 .11로 상승하여 Fig. 3처럼 AI가 교사의 채점 기준을 학습하여 시스템적 편향을 해소하고 전체적인 채점 난이도를 교사 수준으로 맞추는 데 성공했음을 보여준다.

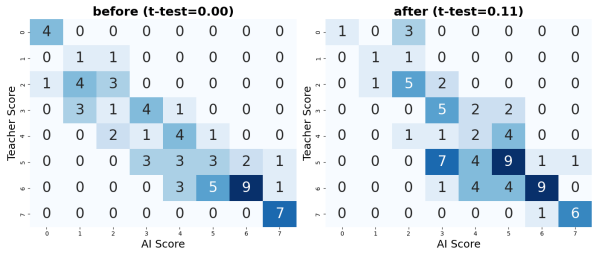


Figure3. Q6 before and after rubric modification

QWK와 상관계수의 소폭 하락 및 MAE 증가 현상은 평가의 유연성과 객관성 간 상충 관계로 해석된다. 정성적 기준을 엄격한 키워드 매칭으로 전환하며 집단 전체의 공정성과 일치도는 확보했으나 LLM의 문맥 추론 제한으로 인해 미세한 변별력은 다소 감소한 것이다. 이는 객관적 채점 신뢰도를 담보하기 위해 수반되는 기계적 채점 방식의 구조적 특성으로 분석된다.

5. 결론 및 제언

본 연구는 고등학교 정보 교과 서·논술형 평가에서 AI 자동채점 시스템의 타당성을 검증하고 프롬프트 엔지니어링 기반의 루브릭 구체화 방안을 모색하였다. 연구 결과를 바탕으로 도출한 결론은 다음과 같다.

첫째, 본 연구는 하이러닝의 AI 자동채점 기능은 기존의 규칙 기반 프로그래밍 평가 도구(Online Judge)가 지닌 한계를 보완하고 과정 중심 평가 도구로서의 활용 가능성을 확인하였다. AI는 구문 오류가 있더라도 알고리즘의 논리적 타당성을 평가하는 질적 유연성을 보였으며 이는 컴퓨팅 사고력의 절차적 요소를 진단하는 데 유효하게 활용될 수 있음을 확인하였다.

둘째, 루브릭 구체화는 교사의 고도화된 전문성이 요구되는 교육적 의사결정 과정임을 확인하였다. 추상적 평가 기준을 논리 구조로 변환하고 공정성과 유연성 사이를 조율하는 것이 프롬프트 기술을 넘어선 교수학적 판단의 영역임을 시사한다. 다만 이는 AI 채점의 신뢰도가 설계 주체인 교사의 평가 설계 전문성에 의존할 수 있으며 이에 따른 평가 품질의 가변성을 최소화하기 위한 시스템적 지원이 필요하다.

셋째, 본 연구는 AI의 채점 오류 유형을 체계화하여 AI 단독 평가 시 발생할 수 있는 공정성 훼손의 위험을 규명하였다. 이를 보완하기 위해 인간-AI 협력(HITL) 체계가 필수적이며 교사는 단순 채점자에서 AI의 판단 근거를 검토하고 최종 의사결정을 내리는 평가설계자로 역할이 변화되어야 한다.

위 결론을 바탕으로 AI 자동채점 시스템의 교육 현장 안착을 위해 다음과 같이 제언한다.

첫째, 본 연구에서 확인된 AI의 과정 중심 평가 역량을 다양한 교육적 맥락으로 확장하는 후속 연구가 필요하다. 특정 지역 및 단일 플랫폼의 한계를 넘어 다양성을 확보하고 타 교과로 적용 범위를 넓힘으로써 AI를 보편적 평가 도구로서의 타당성을 검증해야 한다.

둘째, 교사 개인의 루브릭 구체화 역량에 따른 AI 채점 품질의 가변성 리스크를 보완해야 한다. 표준화된 프롬프트 라이브러리와 오류 유형별 검수 가이드라인을 개발·보급해야 한다.

셋째, 교사가 평가 설계자로서 전문성을 발휘할 수 있도록 AI 평가 리터러시 강화 연수를 지원해야 한다. 특히 AI의 블랙박스 특성과 오류 유형을 이해하고 루브릭을 재설계하는 과정을 포함하여 교사가 기술을 주도적으로 활용할 수 있도록 지원해야 한다.

넷째, 인지적·심동적 영역에 국한된 분석의 한계가 존재하므로 학습 로그(수정 횟수, 시행착오 과정 등) 및 생성형 AI의 자연어 처리를 활용한 정의적 영역까지 진단할 수 있는 종합적 평가 체계 구축을 통해 학습자의 정의적 성장 지원이 필요하다.

참고문헌

[1] Song, M., Kim, D., Park, J., Shin, J., Lee, J., Jung, S., & Lim, Y. (2025). *A study on the application of AI models for automated scoring of subjective and essay-type evaluations (Ⅱ)* (Research Report RRE 2025-5). Korea Institute for Curriculum and Evaluation. https://www.kice.re.kr/resrchBoard/view.do?seq=1320&s=kice&m=030103

[2] Ministry of Education. (2026). 2022 *revised elementary and secondary school curriculum* (National Board of Education Notice No. 2026-1 [Separate Volume 1]). https://ncic.re.kr/inv/org/list.do?step1=_1&step2=_17&step3=_24

[3] Seoul Metropolitan Office of Education. (2025, August 5). *Seoul Metropolitan Office of Education begins full-scale development of AI subjective and essay-type evaluation support system.* https://enews.sen.go.kr/news/view.do?bbsSn=189599&step1=3&step2=1#none

[4] Ministry of Education. (2022). *Informatics curriculum (Ministry of Education Notice No. 2022-33* [Appendix 10]). https://ncic.re.kr/inv/org/list.do?step1=_1&step2=_17&step3=_35

[5] Lee, J., & Kim, S. (2024). A study on generative AI-based feedback techniques for tutoring beginners' error codes on online judge platforms. *Journal of The Korea Society of Computer and Information, 29*(8), 191-200. https://doi.org/10.9708/jksci.2024.29.08.191

[6] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1-38. https://doi.org/10.1145/3571730

[7] Wang, J., Liang, Y., Meng, F., Shi, H., Li, Z., Xu, J., Qu, J., & Zhou, J. (2023). Is ChatGPT a good NLG evaluator? A preliminary study. *arXiv*. https://doi.org/10.48550/arXiv.2303.04048

[8] Mohamed, K., Yousef, M., Medhat, W., Mohamed, E., Khoriba, G., & Arafa, T. (2025). Hands-on analysis of using large language models for the auto evaluation of programming assignments. *Information Systems, 128*, 102473. https://doi.org/10.1016/j.is.2024.102473

[9] Mendonça, P. C., Quintal, F., & Mendonça, F. (2025). Evaluating LLMs for automated scoring in formative assessments. *Applied Sciences, 15*(5), 2787. https://doi.org/10.3390/app15052787

[10] Wasik, S., Antczak, M., Badura, J., Laskowski, A., & Sokal, T. (2018). A survey on online judge systems and their applications. *ACM Computing Surveys, 51*(1), 1-34. https://doi.org/10.1145/3143560

[11] Han, S., Liu, X., & Woo, G. (2022). Evaluation methods and classification of online judge systems based on education level and purpose. *KIISE Transactions on Computing Practices, 28*(10), 487–492. https://doi.org/10.5626/KTCP.2022.28.10.487

[12] Chang, S., & Min, B. (2016). A review of the validity of the CSAT indirect writing assessment. *Research on Writing, 29*, 163–209.

[13] Lister, R., Adams, E. S., Fitzgerald, S., Fone, W., Hamer, J., Lindholm, M., McCartney, R., Moström, J. E., Sanders, K., Seppälä, O., Simon, B., & Thomas, L. (2004). A multi-national study of reading and tracing skills in novice programmers. *SIGCSE Bulletin, 36*(4), 119–150. https://doi.org/10.1145/1041624.1041673

[14] Kim, S., Kim, H., Na, W., Min, H., Baek, S., Sung, K., Lee, M., Lee, M., Lee, Y., & Han, G. (2025). Analysis of issues and needs for essay-type assessment (KICE Issue Paper, Vol. 8). Korea Institute for Curriculum and Evaluation. https://www.kice.re.kr/resrchBoard/view.do?seq=1291&s=kice&m=030109

[15] Shin, B., Lee, J., & Yoo, Y. (2024). Exploring automatic scoring of mathematical descriptive assessment using prompt engineering with the GPT-4 model: Focused on permutations and combinations. *The Mathematical Education, 63*(2), 187-207. https://doi.org/10.7468/mathedu.2024.63.2.187

[16] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

[17] Wang, J., Li, H., Han, X., Liu, Z., Hou, L., & Li, J. (2023). On the robustness of ChatGPT: An adversarial and out-of-distribution perspective. *arXiv*. https://doi.org/10.48550/arXiv.2302.12095

[18] Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2023). Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. *arXiv*. https://doi.org/10.48550/arXiv.2310.11324

[19] Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

[20] Han, Y. (2023). Fairness evaluation within large language models through the lens of depression. *ICMLCA* '23: Proceedings of the 2023 4th International Conference on Machine Learning and Computer Application, 108–112. https://doi.org/10.1145/3650215.3650236

[21] Jeon, W. (2024). Risk classification system of the EU AI Act and its implications for Korea's AI policy. *Journal of Korea Information Law, 28*(2), 169-216.

[22] Jung, Y. (2025). *Development and validation of a generative AI-based automated grading program for essay-type assessments* (Doctoral dissertation, Seoul National University). https://s-space.snu.ac.kr/handle/10371/221983

[23] Chu, Y., Copur-Gencturk, Y., Li, H., Tang, J., & Yang, K. (2025). LLM-based automated grading with human-in-the-loop. *arXiv*. https://doi.org/10.48550/arXiv.2504.05239

[24] Choi, J., Kim, J., & Kim, H. (2025). Calibrating human and machine: Developing an AI-assisted rater discussion model for automated writing evaluation. *Journal of Educational Information and Media, 31*(4), 1297-1329.

[25] Ministry of Education. (2024). *Achievement standards of*

high school informatics based on the 2022 revised national curriculum. National Curriculum Information Center.
[26] Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213-220. https://doi.org/10.1037/h0026256

정유하

· 2013년 수원대학교 컴퓨터학과(이학사)
· 2025년 성균관대학교 인공지능융합교육전공(석사)
· 2025년~현재 성균관대학교 컴퓨터교육과 박사과정

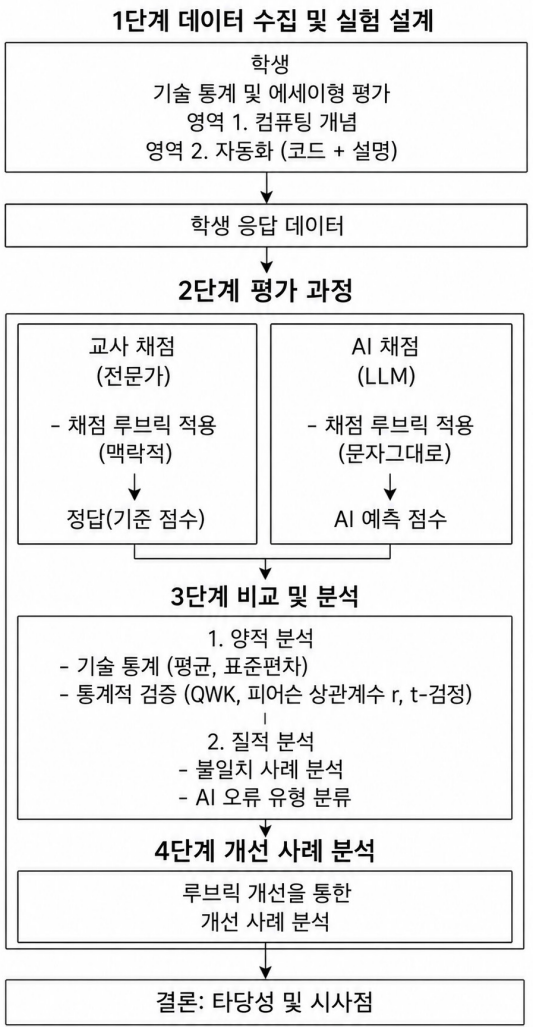
관심분야 : 정보교육, AI 교육, AI 기반 교육평가
jmy8493@g.skku.edu

안성진

· 1998년 성균관대학교 정보공학과(공학박사)
· 1990~1995년 KIST/SERI 연구원
· 1996년 정보통신기술사
· 2022년~현재 기술안보포럼 좌장
· 1999년~현재 성균관대학교 컴퓨터교육과 교수

관심분야 : 네트워크관리, AI 윤리, 기술보호, AI 교육
sjahn@skku.edu

부 록



〈그림 1〉 전체 연구 절차

〈표 1〉 평가 문항

영역	번호	핵심 개념	평가 기준
컴퓨팅 개념	Q1	데이터 압축	데이터 압축의 원리 이해
	Q2	정렬 알고리즘	알고리즘 흐름 설명 및 효율성 비교
	Q3	모델링	모델링이 필요한 이유 설명 및 실생활 문제에 적용
	Q4	인공지능 윤리	인공지능 윤리 문제에 대한 비판적 사고
자동화	Q5	데이터 타입	데이터 타입 오류 식별 및 수정
	Q6	선택 구조	조건문 코드의 출력 예측 및 수정
	Q7	반복 구조	반복문 코드의 실행 흐름 추적 및 수정

〈표 2〉 평가 문항

번호	문항	점수별 학생 응답
Q1	다음 이미지 압축 기법 중 잘못된 설명을 찾고 그 이유를 설명하시오. GIF: ... (생략됨) JPG: jpg를 이용하면 손실 없는 화질을 유지해 사진 촬영 후 친구에게 다양한 색상과 풍부한 디테일이 담겨진 풍경 사진을 보낼 때 사용한다. PNG: ... (생략됨)	[2점] "JPG가 틀린 설명이다. 손실 압축이라 파일 크기가 줄어들고 일부 화질을 버리기 때문에 손실 없이 화질을 유지하는 것이 잘못된 설명이다."
		[1점] "JPG, 화질 손실", "JPG는 손실이 있다.", "JPG, 손실 압축이다."
		[0점] "GIF", "저장 방법이 구려서 화질이 깨지고 픽셀이 보임"
Q6	다음 코드를 보고 물음에 답하시오. score = 85 if score 〉= 90: ...(생략됨)... print("score:", score, "grade:", grade) [하위문항 1] 위 코드의 실행 결과를 작성하시오.	[2점]"score: 85 grade: B"
		[1점] "85 B", "85", "B"
		[0점] "score"

주의. Q6은 하위문항 1만을 의미함 (총 2점).

〈표 3〉 컴퓨팅 개념 문항 루브릭

번호	기준	점수
Q1	이미지 데이터 압축 원리가 틀린 설명 찾기 (1)	8
	이미지 데이터 압축 원리가 틀린 설명에 대해 손실 압축과 품질 저하를 연결하여 명확하게 기술하기 (1)	
Q2	버블 정렬 알고리즘의 개념 설명하기 (3)	9
	버블 정렬 알고리즘의 단계 표현하기 (3)	
	버블 정렬 알고리즘의 효율성 분석하기 (3)	
Q3	모델링의 개념과 필요한 이유 설명하기 (3)	12
	실생활의 문제에서 적합한 모델링 유형 선택하고 이유 설명하기 (3)	
	주어진 자료에서 적합한 모델링 유형을 선택하고 이유 설명하기 (3)	
	일상생활 속에서 모델링의 사례를 찾아 설명하기 (3)	
Q4	VR과 AI에 대한 기술적 특징 설명하기 (3)	15
	VR과 AI에 대한 자신의 입장을 명확히 밝히고 일관성 유지하기 (3)	
	자신의 입장에 대한 윤리적, 사회적 근거 제시하기 (3)	
	반대 의견 제시하고 반박하기 (3)	
	글의 구조 및 표현 (3)	

〈표 4〉 자동화 문항 루브릭

번호	기준	점수
Q5	데이터 유형에 대한 개념 설명하기 (1)	3
	문자열을 정수형으로 변환하여 프로그램 작성하기 (2)	
Q6	정확한 프로그램 실행 결과 제시하기 (2)	7
	'if-elif-else' 구조 설명하기 (3)	
	변경된 프로그램의 결과 예측하기 (2)	
Q7	프로그램의 실행 결과 예측하기 (1)	5
	복합대입연산자의 연산 방법 설명하기 (2)	
	홀수의 합이 나오도록 if문 수정하기 (2)	

〈표 10〉 사례 분석 1

문항	설명
문제	이미지 압축 기법(GIF, JPG, PNG) 중 잘못된 설명을 찾고 그 이유를 설명하시오.
학생 응답	"JPG. JPG는 2565색을 압축할 때 손실이 없지만 256색보다 많은 이미지를 지원되는 색상으로 변환되어 이미지 손실이 발생한다. 따라서 잘못된 설명이다."
교사 채점	1점 (올바른 키워드인 'JPG'를 식별한 부분은 점수를 부여했지만 GIF와 혼동한 명확한 기술적 오개념은 감점했다.)
AI 채점	2점 (만점)
AI 추론 분석	AI는 유창성 편향 때문에 개념적 오류를 탐지하지 못했다.

〈표 14〉 Q6 문항 수정 전 루브릭

개념	점수	루브릭
프로그램 결과 예측	2	프로그램의 실행 결과를 정확하게 작성함.
	1	프로그램의 실행 결과를 부분적으로 작성함.
변화된 프로그램 결과 예측	2	프로그램의 실행 결과를 정확하게 작성함.
	1	프로그램의 실행 결과를 부분적으로 작성함.

주의. 변경되지 않은 항목은 표에서 제외함.

〈표 11〉 AI 평가 특성의 질적 분석: 한계와 강점

주요 원인	근본 원인(편향)	오류 유형	하위 요인	구체적으로 관찰된 사례
1. 한계				
모델 처리	유창성 및 완전성 편향	과대평가	유창성 편향	내용이 틀려도 문장 구조가 완벽하면 높은 점수를 부여함.
			논리적 오류 간과	길고 유창한 문장 안의 논리적 모순을 모델이 탐지하지 못함.
		과소평가	단편적 답변 간과	모델이 완전한 문장을 선호하여 독립적으로 맞는 핵심 키워드도 감점함.
			학습자 언어에 대한 비유연성	학술적 용어가 부족한 학생의 단순한 설명을 감점함.
	단순 키워드 매칭	과대평가	맥락 없는 키워드 수용	논리적·인과적 연결이 없어도 단순히 같은 키워드가 존재하면 점수를 부여함.
	형식적 경직성	과소평가	입력 형식	줄바꿈 같은 사소한 형식 차이에도 감점함.
루브릭 설계	정의되지 않은 추상적 지시어	과소평가	정의되지 않은 추상적 용어	구체적 예시 없이 "부분적으로"라는 표현을 해석하지 못해 부분 점수를 부여하지 않음.
2. 강점				
모델 역량	문법 독립적 논리 평가	–	알고리즘 중심 평가	과정 중심 평가 도구로서의 가능성을 검증함.
			기존 시스템과의 차별성	문법 오류가 있어도 올바른 논리 흐름에는 부분 점수를 부여함.

〈표 12〉 사례 분석 2

문항	설명
문항	사망한 가족 구성원을 위한 VR과 AI 기술을 활용한 복원 기술 도입에 대한 자신의 생각을 서술하시오. (조건: 기술적 특성을 포함할 것)
학생 응답 (일부)	"... VR 기술은 요즘 떠오르는 기술이다. (중략) 장점도 있지만 단점도 있다. ... 따라서 나는 사회적 도입에 반대한다."
교사 채점	9/15점 (필수 하위 기준인 '기술적 특성 설명'이 누락되어 감점함.)
AI 채점	11/15점 (과대평가됨)
AI 추론 분석	AI가 추상적인 평가 기준을 느슨하게 해석하여 유려한 문장에 대해 기술적으로는 불완전한 응답을 과대평가함.

〈표 13〉 Q1 문항 수정 전 루브릭

개념	점수	루브릭
이미지 데이터 압축 기법 이해하기	2	이미지 데이터 압축 원리가 틀린 설명을 찾아 틀린 이유를 명확하게 기술할 수 있다.
	1	이미지 데이터 압축 원리가 틀린 설명을 찾을 수 있으나, 틀린 이유에 대해 부정확하게 기술할 수 있다.
	0	응답이 모두 부적절하다.

〈표 15〉 Q1 문항 수정 후 루브릭

Concept	점수	루브릭
이미지 데이터 압축 기법 중 틀린 설명 찾기	1	'JPG', 'JPEG', 'jpg', 'jpeg'라는 단어 언급이 있다.
	0	'JPG', 'JPEG', 'jpg', 'jpeg'라는 단어 언급이 없다.
이미지 데이터 압축 기법의 설명에서 틀린 이유 작성하기	1	답안에 다음 두 가지 개념(A, B)이 모두 명시적으로 드러나야 한다. 개념 A (원인): '손실', '압축', '데이터 삭제', '정보를 버림'과 같은 단어 언급 개념 B (결과): '화질 저하', '깨짐', '복구 불가', '흐려짐'과 같은 단어 언급
	0	jpg의 특징에 대해 설명하지 못했다. [오답 처리 예시 – 절대 1점 주지 말 것] "JPG는 화질이 안좋다." (결과만 있고 원인인 '손실, 압축'과 같은 단어 언급이 없으므로 0점) "JPG는 손실 압축이다." (원인만 있고 결과인 '화질, 복구'와 같은 단어 언급이 없으면 0점) "JPG가 용량이 작아서." (오개념이므로 0점)

〈표 16〉 Q6 문항 수정 후 루브릭

개념	Pts	루브릭
프로그램 결과 예측	2	프로그램 실행 결과인 'Score: 85 Grade: B'.를 정확하게 작성함.
	1	프로그램 실행 결과의 일부 (예를 들면 'B' 또는 '85 B')를 작성함.
변화된 프로그램 결과 예측	2	프로그램 실행 결과인 'Score: 75 Grade: C'.를 정확하게 작성함.
	1	프로그램 실행 결과의 일부 (예를 들면 'C' or '75 C')를 작성함.

주의. 변경되지 않은 항목은 표에서 제외함.