THE JOURNAL OF KOREAN ASSOCIATION OF COMPUTER EDUCATION Volume 29, Issue 4, 2026

컴퓨터교육학회 논문지 2026년 제29권 제4호
https://doi.org/10.32431/kace.2026.29.4.002

생성형 AI 기반 학생평가 플랫폼 연구

A Generative AI-Based Student Assessment Platform

신상윤† · 강신천††
Sangyun Shin† · Shincheon Kang††

요약

본 연구는 생성형 AI의 광범위한 활용으로 학생 산출물의 외형적 완성도와 실제 이해 수준 간의 괴리가 확대되는 상황에서, 학습자의 실제 이해를 정확하게 평가하는 데서 발생하는 문제를 다룬다. 이러한 변화는 교사의 평가 부담을 증가시키지만, 학생이 무엇을 어느 정도 이해하고 있는지를 진단하는 일은 교육 평가의 본질적인 요소로서 여전히 중요하다. 이에 본 연구에서는 학습자의 이해 수준을 다각적으로 진단하고 교사의 전문적 판단을 보완할 수 있는 참고 자료를 제공함으로써 평가 부담을 경감하는 생성형 AI 기반 학생평가 플랫폼을 개발하였다. 플랫폼의 프로토타입은 국내외 선행연구 분석을 통해 도출한 여섯 가지 설계 원리를 기반으로 설계·구현되었다. 학습자의 이해는 맞춤형 이해 점검 문항 응답과 동료 채점 결과와 AI 채점 결과 간의 일치도 및 유사도 분석을 통해 평가된다. 또한 전문가를 대상으로 한 예비적 사용성 평가 결과, 플랫폼은 직관적인 구조, 교사의 평가 업무 경감, 수업 적용 가능성 측면에서 강점을 보였다. 반면, 플랫폼 내 튜토리얼의 부재, 중요 안전장치의 일부 비일관성, 이해 중심 피드백이 학습자의 행동 변화로 충분히 이어지지 못하는 한계는 개선 과제로 도출되었다. 이러한 결과는 차기 버전 플랫폼의 개선 방향 설정과 향후 후속 연구를 위한 시사점을 제공한다.

주제어 인공지능, 생성형 AI, 학생평가 자동화, 생성형 AI 기반 학생평가, 학생평가 플랫폼

ABSTRACT

This study addresses the challenge of accurately assessing learners' actual understanding as the widespread use of generative AI widens the gap between the polished appearance of student work and genuine comprehension. Although this trend increases teachers' evaluation burden, identifying what students understand and to what extent remains fundamental to educational assessment. To address this issue, we developed a generative AI-based student assessment platform that diagnoses learners' understanding from multiple perspectives and provides supplementary data to support teachers' professional judgment. The platform prototype was developed based on six design principles derived from domestic and international research. Learners' understanding is assessed through customized comprehension questions and by examining the concordance and similarity between peer grading and AI grading results. A preliminary usability study with 10 experts identified strengths in intuitive navigation, reduced teacher workload, and classroom applicability, while also revealing areas for improvement, including the absence of an in-app tutorial, inconsistencies in critical safety mechanisms, and the limited behavioral impact of understanding-focused feedback. These findings inform directions for platform refinement and future classroom-based studies.

Keywords Artificial Intelligence, Generative AI, Automated Student Assessment, Generative AI-Based Student Assessment, Student Assessment Platform

†정회원 국립공주대학교 컴퓨터교육과 교육학 박사
††종신회원 국립공주대학교 사범대학 컴퓨터교육과 교수(교신저자)

논문투고 2025년 09월 30일
심사완료 2026년 02월 12일
게재확정 2026년 02월 23일
발행일자 2026년 04월 13일

1. 서론

최근 몇 년간 생성형 인공지능(Generative Artificial Intelligence, 이하 생성형 AI) 기술의 발전은 사회 전반에 혁신적인 변화를 일으켜 사회 구조의 근본적인 변화를 이끌고 있다. 생성형 AI란 이용자의 특정 요구에 따라 텍스트·이미지·음성·영상 등 다양한 형태의 결과물을 능동적으로 생성(generate)해 내는 인공지능 기술을 의미한다. 기존까지의 딥러닝 기반 AI 기술이 단순히 기존 데이터를 기반으로 예측(prediction)하거나 분류(classification)하는 데 주력했던 것과 달리, 생성형 AI는 질문의 의도와 맥락을 파악하여 이용자가 요구한 질문이나 과제를 해결하기 위해 스스로 학습·추론하고 새로운 결과를 생성한다는 점에서 한 단계 진화한 AI 기술이라 할 수 있다[1].

이러한 기술적 도약은 창의적·복합적 문제 해결과 의사결정과 같이 전통적으로 지식 노동자에게만 가능하다고 여겨졌던 영역으로 AI의 활용 범위를 급격히 확장시키고 있으며, 2030년경에는 생성형 AI가 다수의 업무에서 중간 수준의 인간 노동자가 낼 수 있는 결과물을 대등하게 만들어 낼 것이라는 추측도 나오고 있다[2]. OpenAI의 ChatGPT, Google의 Gemini, xAI의 Grok, DeepSeek 등 다양한 생성형 AI 모델이 잇달아 등장하며 이러한 변화는 더욱 가속화되고 있다.

특히 교육 분야는 생성형 AI의 잠재력이 크게 주목되는 영역이다. 대규모 언어 모델을 기반으로 한 생성형 AI는 학습 자료 제작, 개인화 학습, 자동화 평가, 실시간 피드백 등에서 새로운 가능성을 제시하며, 이는 전통적 교수·학습 및 평가 패러다임을 근본적으로 재구성할 잠재력을 지닌다. 국내·외 다양한 연구에서 생성형 AI 기술의 교육적 활용에 대한 논의가 활발히 이루어지고 있으며[3, 4], 이러한 연구 흐름은 점차 학교 현장으로 확산되며, 실질적인 교육 혁신으로 이어지고 있다. 교사와 학생 또한 생성형 AI를 활용한 혁신적 교육 방식을 경험하고 이를 실질적으로 적용하는 사례가 증가하고 있는데, 실제 2023년 이화여자대학교 미래교육연구소와 교육부가 실시한 생성형 AI 활용 실태조사 결과에 따르면, 학생의 79.2%, 교사의 61%가 생성형 AI 사용 경험이 있는 것으로 나타났다[5]. 이처럼 학교 현장에서 생성형 AI의 활용이 빠르게 확산되면서, 학습 효과에 대한 우려 또한 함께 제기되고 있다. 특히 학생들이 과제를 손쉽게 완성하거나 정보를 즉시 획득하면서 자기 주도적 사고와 깊이 있는 학습 경험이 약화될 수 있다는 지적이다. 실제 최근 연구에 따르면 거대언어모델(LLM)을 활용해 글을 작성할 경우 기억력, 언어 구성력, 창의성이 전반적으로 저하되는 경향이 나타났으며[6], KPMG 캐나다 조사에서도 생성형 AI가 단기적으로는 성과 향상에 도움이 되지만 장기적으로는 학습 효과를 약화시킬 수 있음이 보고되었다[7].

이러한 상황에서 단순히 생성형 AI의 사용을 제한하거나 차단하는 방식은 미래 사회에 필수적인 디지털 리터러시와 비판적 사고 역량을 함양할 기회를 학생들에게 제공하지 못할 뿐 아니라, AI 시대의 학습 주체로 성장할 가능성까지 제약할 수 있다[8, 9]. 따라서 기술 활용을 무조건 억제하기보다는, 부작용을 최소화하면서 교육적 효과를 극대화할 수 있는 평가적 대안을 마련하는 접근이 필요하다. 핵심은 학생이 생성형 AI를 활용하더라도 그 과정에서 실제적인 이해가 형성되고 있는지 확인하고 촉진할 수 있는 평가 체계를 구축하는 것이다. 오늘날 많은 학습자가 생성형 AI를 이용해 과제나 보고서를 손쉽게 작성하고 있는 현실을 고려할 때, 산출물의 표면적 완성도를 평가하는 것만으로는 학생이 실제로 무엇을 알고 이해했는지를 정확히 파악하기 어렵다. 특히 제출된 결과물이 높은 완성도를 보이더라도 그것이 곧바로 학습자의 실제 이해 수준을 의미하지 않을 수 있으며, 이러한 산출물과 실제 이해 간의 간극은 교사에게 평가 부담을 안기는 구조적 한계로 이어진다.

이러한 맥락에서 학생 평가 영역에서의 생성형 AI 활용은 주목할 만하다. 생성형 AI 기반 학생평가 플랫폼은 교사의 추가적인 부담을 최소화하면서 형성평가(formative assessment)를 보다 자주·정밀하게 수행할 수 있도록 지원하며[10], 산출물의 외형적 완성도만으로는 포착하기 어려운 학습자의 실제 이해 수준을 진단·파악하여 교사의 전문적 판단을 보완하는 참고 자료를 제공한다는 점에서 교육적 가치가 크다.

이에 본 연구는 변화하는 교육 환경에 부합하고 교사에게 유용한 평가 정보를 제공할 수 있는 새로운 형태의 학생평가 플랫폼을 개발하는 것을 목적으로 한다. 기존의 AI 기반 학생평가가 자동 채점과 같은 개별 기능 중심의 도구 활용에 머무른 데 비해, 본 연구에서 개발하고자 하는 생성형 AI 기반 학생평가 플랫폼은 학생이 평가 활동에 직접 참여하는 평가 과정을 포함함으로써 하나의 통합된 평가 체제를 지향한다. 이를 위해 생성형 AI 기반 학생평가 플랫폼 개발을 위한 연구를 단계적으로 진행해 왔으며, 본 논문은 그 두 번째 단계 연구로서 생성형 AI 기반 학생평가 플랫폼의 설계와 프로토타입 개발을 다룬다. 1차 연구에서는 중등교사를 대상으로 생성형 AI를 활용한 학생평가에 대한 인식 조사를 실시하였으며, 그 결과 중등교사는 학생평가 분야에서 생성형 AI 기술의 교육적 활용 가능성에 대해 대체로 긍정적인 인식을 갖고 있는 것으로 나타났다. 또한, 생성형 AI 기반 학생 평가가 중등학교 현장에 적용되기 위해 필요한 요소로는 간편하고 직관적인 형태의 평가 플랫폼 구축이 가장 중요한 과제로 확인되었다[11].

이에 본 연구는 중등학교에서 적용 가능한 생성형 AI 기반 학생평가 플랫폼을 설계·구현하고, 전문가 사용성 평가를 통해 그 교육적 활용 가능성을 예비적으로 검증함으로써 실질적인 현장 적용의 토대를 마련하고자 한다. 본 연구의 구체적 연구 문제는 다음과 같다.

첫째, 생성형 AI 기반 학생평가 플랫폼의 설계 원리와 핵심 기능은 무엇인가?

둘째, 개발된 프로토타입에 대한 전문가 사용성 평가는 어떠하며, 플랫폼의 강점과 개선 과제는 무엇인가?

2. 이론적 배경

2.1 생성형 AI 시대의 학생평가

학생평가(student assessment)는 교육목표에 비추어 학습자가 수업 내용을 얼마나 이해하고 성취했는지를 확인하고, 이를 토대로 교수·학습의 질을 개선하는 데 목적이 있다[12, 13]. 잘 설계된 평가는 학습 성과를 진단하는 핵심 요소로서, 학습자의 변화와 성장을 지원하기 위해 다양한 자료를 체계적으로 수집·분석함으로써 깊이 있는 학습을 촉진한다[13, 14].

효과적인 평가를 위해서는 교사가 교과의 특성과 수업 목표를 충분히 반영해 문항을 개발하고, 학생 수준에 적합한 과제를 채점하며, 수행 과정과 관찰 결과를 토대로 개별화된 피드백을 제공해야 한다. 이러한 평가는 학습 성과를 정확히 진단하고 학습 성장을 촉진하는 데 필수적이지만, 동시에 상당한 시간과 전문성을 요구한다. 더욱이 교사는 학습 계획 수립, 피드백 제공, 교실 운영 등 다양한 업무를 병행해야 하므로 평가에 필요한 자원을 충분히 투입하기 어려운 실정이다. 이로 인해 평가의 부담이 과중해지고 오류 발생 가능성도 높아질 수 있다[15]. 최근 과정 중심 평가가 강조되면서 교사의 평가 부담은 더욱 심화되고 있으며, 이에 따라 평가 방식의 혁신은 중요한 교육 과제로 부상하고 있다[16].

과정 중심 평가의 대표적 형태인 서·논술형 평가는 학습자의 사고 과정과 이해 수준을 정밀하게 파악할 수 있다는 점에서 교육적으로 큰 의미가 있다. 그러나 학생 개별 응답을 세밀히 분석하고 개별 피드백을 제공해야 하므로 교사에게 막대한 시간과 노력이 요구된다. 단순히 정답 여부를 확인하는 방식이 아니라 학습자의 사고 과정을 함께 평가해야 하기 때문에 교사의 시간적 부담은 더욱 커진다. 그럼에도 불구하고 학생평가는 학습의 질을 제고하고 깊이 있는 이해를 진단하기 위한 핵심적 수단이므로, 이를 소홀히 할 수 없다.

최근에는 대규모 언어모델(LLM)에 기반한 생성형 AI 도구가 빠르게 확산되면서 전통적 평가 방식이 직면한 한계를 지적하며, 기존 학생평가 방식의 지속 가능성에 대한 문제 제기가 늘어나고 있으며[17], 생성형 AI 시대에 적합한 새로운 평가 패러다임을 모색해야 한다는 논의도 활발히 전개되고 있다[12, 18]. 전통적 학생평가가 직면한 한계는 크게 두 가지로 요약할 수 있다.

첫째, 학습자의 실제 이해 수준을 검증하기 어렵다는 점이다. 생성형 AI 서비스의 손쉬운 접근성과 편의성으로 인해 학생들은 과제나 학습 활동에 이를 적극적으로 활용하고 있다. 이때 학생이 제출한 결과물이 실제로 자신의 사고 과정을 반영한 것인지, 아니면 생성형 AI가 제공한 답변을 단순히 변형·재구성한 것인지를 교사가 판별하기란 쉽지 않다. 특히 서·논술형 과제, 탐구 보고서, 발표 자료와 같이 산출물의 완성도가 외형적으로 높게 나타나기 쉬운 과제 유형에서는 이러한 문제가 더욱 두드러진다. 겉으로는 논리적으로 완결성이 높고 오류가 없다 하더라도, 그것이 반드시 학습자의 내적 이해와 비판적 사고에 기반했다고 단정하기 어렵다.

둘째, 그로 인해 교사에게 과도한 평가 부담이 가중된다는 점이다. 교사는 학생 산출물의 표면적 완성도가 아닌 다양한 응답을 다각도로 분석해 학생의 실제 이해 수준을 파악해야 한다. 이러한 심층적 진단은 평가가 단순한 성취 측정을 넘어 학습의 성장을 촉진하는 본래 목적을 달성하는 데 필수적이다. 그러나 현실적으로 모든 학생의 결과물을 대상으로 깊이 있는 분석을 수행하는 것은 막대한 시간과 노력을 요구하며, 추가적 평가 업무 부담으로 이어진다.

이러한 상황에서 단순히 생성형 AI 사용을 금지하거나 표절 여부만을 확인하는 방식은 근본적 해결책이 될 수 없다. 오히려 교사는 학생이 제출한 결과물이 실제로 학습자의 개별적 이해를 기반으로 작성된 것인지 면밀히 분석할 수 있어야 한다. 이를 위해 학습자의 사고 과정과 이해의 깊이를 드러낼 수 있는 체계적이고 정밀한 '실제 이해 수준 진단' 평가 체계가 요구된다.

결국, 생성형 AI 시대의 학생평가는 학습자의 실제 이해와 사고 과정을 정밀하게 파악·평가하는 방향으로 전환되어야 한다. 이러한 변화가 이루어져야만 평가는 학습의 질을 향상시키고 학생의 성장을 지원하는 본래 목적을 실현할 수 있다. 따라서 교사의 평가 부담을 완화하면서도 학생의 실제 이해 수준을 다각도로 진단할 수 있는 새로운 평가 지원 방안을 모색하는 일이 시급하다.

2.2 AI 기반 학생평가의 역할과 기능

최근 교사의 평가 부담을 줄이면서도 평가의 본질적 기능을 유지할 수 있는 대안으로 인공지능(AI) 기술이 주목받고 있다. AI는 기존 평가 절차 에서 반복적·소모적인 업무를 자동화하여, 교사가 채점 업무에서 벗어나 수업 설계나 개별 학생 지도 등 본질적 교육 활동에 더 많은 시간을 할애할 수 있도록 돕는다[19]. 또한 AI는 학생의 내면적 사고 과정이나 학습 진전 정도처럼 교사가 전통적 방식만으로는 지속적·정밀하게 파악하기 어려운 영역을 심층적으로 분석할 수 있게 한다. 이를 통해 학습자의 숙달 수준에 근거한 개별화된 학습 제언을 제공함으로써 수업 개선뿐 아니라 맞춤형 학습 경로 설계에도 활용될 수 있다[20]. 즉, AI는 단순히 채점을 보조하는 수준을 넘어, 교사가 학생의 이해도와 학습 궤적을 보다 정확히 진단하고 그 결과를 교수 전략과 수업 계획에 반영하도록 지원하는 핵심 평가 도구로 자리매김하고 있다.

AI 기반 평가가 생성하는 학습 데이터는 단순 성취 진단을 넘어 학습 분석(Learning Analytics)을 통한 다층적 활용 가능성을 지닌다. 학습 분석은 학습자의 활동 데이터를 수집·분석해 현재 상태를 설명(descriptive)하고, 미래 성취를 예측(predictive)하며, 학습 전략을 처방(prescriptive)하는 일련의 과정을 포함한다. 이러한 분석 결과는 곧바로 적응형 학습(adaptive learning)으로 연결되어, 학생 개인의 수준과 요구에 적합한 학습 경로를 제시한다. 특히 한국교육학술정보원(2023)은 AI와 적응형 학습을 위한 분석 범주와 예시 지표를 제안하였으며, Table 1은 이를 요약·정리한 것이다[21].

Table 1. Example Analyses for AI and Adaptive Learning Applications

Category	Description
Descriptive	· Scores from formative and summative assessments · Number of logins and total learning time · Subject-specific progress levels · Time spent per subject · Number of learning-note entries · Frequency of study-plan creation and reminders · Class-wide and individual average data · Comparative analytics by student
Predictive	· Probability of achieving specific learning goals · Identification and early alerts for at-risk students · Probability of course completion based on grades and performance data
Prescriptive	· Feedback addressing the affective domain · Curriculum or content recommendations · Personalized learning-strategy suggestions · Provision of analytic summary reports

* Summarized from AI Digital Textbook Development Guidelines (Korea Education and Research Information Service, 2023).

이상의 범주와 지표는 평가를 단순 점수화 중심에서 설명-예측-처방의 순환 분석 구조로 확장하며, 수집된 학습 데이터를 수업 설계와 개별화 지원에 직접 연결하는 근거를 제공한다.

이처럼 AI 기반 학생평가는 채점 자동화를 넘어 학습 분석과 적응형 지원을 통합하는 방향으로 진화하고 있다. 또한 최근 생성형 AI의 발전은 문항 자동 생성과 개인화된 피드백 제공을 가능케 하여 평가 정밀도와 개별화를 높였고, 문항 설계-평가 시행-사후 분석에 이르는 전 주기의 혁신 가능성을 열었다. 이에 본 연구 역시 생성형 AI를 학생평가 맥락에 적용하여 그 적용 가능성과 시사점을 탐색하고자 한다.

3. 연구 방법 및 절차

본 연구는 설계·개발연구 방법론 중 유형 1에 해당하는 산출물 및 도구 연구를 적용하였다. 연구 절차는 분석, 설계, 개발 세 단계로 구성하였으며, 실제 수업 적용을 통한 교육적 효과 검증은 후속 연구 과제로 설정하였다. 각 단계별 주요 내용은 Table 2에 제시하였다.

먼저, 분석(Analysis) 단계는 선행 연구에서 이미 수행되었으며, 생성형 AI의 교육적 활용 및 학생평가 관련 선행 연구를 검토하고, 중등교사 대상 설문을 통해 생성형 AI의 교육적 활용에 대한 인식과 경험, 특히 평가 영역의 요구를 규명하였다. 그 결과, 평가 부담 완화와 과정 중심 평가를 지원하는 직관적 평가 플랫폼의 필요성이 도출되었으며, 이러한 결과는 본 연구의 플랫폼 설계·개발을 위한 기초 자료로 활용되었다.

다음으로, 설계(Design) 단계에서는 요구 분석 및 관련 선행 연구 분석 결과 토대로 플랫폼 설계를 위한 핵심 기능 기반 설계 원리를 도출하였다. 도출된 설계 원리는 '실제 이해 수준 진단 중심 평가', '평가 판단 근거의 다층화', '맞춤형 피드백 제공', '교사의 평가 주도권 보장', '직관적인 플랫폼 구조', '윤리성과 신뢰성 확보'의 여섯 가지이다. 이를 바탕으로 과제 생성 및 제출, 맞춤형 문항 응시, 동료평가, 결과 확인으로 이어지는 평가 사이클을 설계하고, 각 단계의 기능이 설계 원리를 반영하도록 구체화하였다. 또한 시스템 아키텍처, 개체-관계 다이어그램(ERD), 메뉴 구조도 및 화면 설계안을 구체화하여 개발의 청사진을 마련하였다.

이후 개발 단계에서는 설계안을 기반으로 생성형 AI 기반 학생평가 플랫폼의 프로토타입을 구현하였다. 개발 환경은 Python 기반 Streamlit 프레임워크를 활용하였으며, 데이터 관리를 위해 SQLite 데이터베이스를 구축하였다. 문항 및 피드백 생성과 채점에는 OpenAI의 GPT-4 모델을 활용하였고, 모든 AI 처리 과정에서 동일한 파라미터를 적용하여 결과의 일관성을 유지하였다. 구현된 프로토타입은 과제 관리, 학습자 산출물 기반 문항 생성, 동료평가 및 AI 채점, 결과 리포트 제공 기능을 포함한다.

플랫폼의 프로토타입 개발 완료 후에는 컴퓨터교육, 교육공학, 교육평가 전공 교수, 중등학교 교사 등 전문가 10인을 대상으로 사용성 평가를 실시였다. 사용성 평가 결과는 각 문항별 평균과 표준편차를 산출하여 플랫폼의 기능적 안정성, UI/UX 적합성, 교육적 활용 가능성을 점검하였으며, 정량적 결과와 개방형 응답 결과를 종합적으로 분석 및 반영하여 시스템을 보완하였다.

Table 2. Research Procedure, Content, and Outputs

Phase	Content	Outputs
Analysis	· Review of prior studies on the educational use of generative AI and student assessment · Examination of teacher needs analysis survey results from a prior study · Identification of key needs in student assessment	· Literature review findings · Teacher survey results
↓		
Design	· Derivation of core platform design principles based on prior studies and needs analysis results · Design of an assessment cycle consisting of task submission, item generation, assessment, and result review · Structural design including system architecture, ERD, and task flow diagrams	· Design principles · Conceptual architecture & menu structure diagram · Screen design specification
↓		
Development	· Implementation of a prototype platform using Python-based Streamlit · Development of item and feedback generation functions using GPT-4 · Conducting expert usability evaluation and refining system functions and UI/UX	· Platform prototype · User manual

4. 생성형 AI 기반 학생평가 플랫폼 설계

4.1 기능 기반 플랫폼 설계 원리 도출

AI 기반 학생평가와 관련한 국내외 선행 연구[12, 16, 22-25]와 중등교사를 대상으로 한 요구 분석 결과를 종합적으로 분석한 결과를 토대로 플랫폼의 설계 원리를 도출하였다. 특히, 학습자의 실제 이해 수준을 진단하여 평가의 본래 목적을 충실히 수행함과 동시에, 교사의 전문적 판단을 보완하는 보조 자료를 제공한다는 연구 목적을 설계 원리에 긴밀히 연계하였다. 아울러 선행 연구에서 제시된 AI 평가의 한계 및 가능성과 현장 교사의 요구를 대조·비교하는 과정을 통해, 이론적 논의와 실제 교육 현장을 연결할 수 있는 기능 중심의 설계 원리를 도출하고자 하였다. 이러한 과정을 통해 핵심 기능에 기반한 설계 원리를 다음과 같이 도출하였다.

첫째, 실제 이해 수준 진단 중심 평가이다. 본 플랫폼은 맞춤형 이해 점검 문항과 동료평가 활동을 통해 평가의 목적이 단순히 점수를 산출하는 것을 넘어, 학습자의 실제 이해 수준과 사고 과정을 진단하는 것을 평가의 핵심 목적으로 설정한다.

둘째, 평가 판단 근거의 다층화이다. 맞춤형 문항 응답 결과와 동료평가 결과 등 다층적 진단 정보를 제공하여 교사가 이를 비교·종합해 학생의 이해 수준을 해석·판단할 수 있도록 지원한다.

셋째, 맞춤형 피드백 제공 원리이다. 이해 점검 문항 응시 결과와 동료평가 데이터를 통합해 학습 주제·평가 준거별 이해도를 진단하고, 그 결과에 기반한 개인화된 피드백을 제공한다. 이는 교사에게는 전문적 판단을 보완하는 보조적 자료로, 학생에게는 학습 지침으로 기능할 수 있다.

넷째, 교사의 평가 주도권 보장 원리이다. AI가 생성한 채점과 피드백 등에 대해 최종적인 교육적 판단은 교사에게 하도록 하여 평가의 전문성과 교육적 신뢰성을 유지한다.

다섯째, 간편하고 직관적인 플랫폼 구조 원리이다. 사용자 친화적인 구조를 통해 플랫폼 이용 시 불필요한 탐색 비용과 인지 부하를 줄인다.

여섯째, 윤리성과 신뢰성 확보 원리이다. 개인정보 수집 최소화와 AI 채점 근거를 투명하게 제시하여 신뢰가능한 평가 환경을 보장한다.

본 연구는 이러한 설계 원리를 기반으로 플랫폼의 주요 기능을 구체화했으며, 각 원리에 대한 설명과 플랫폼 구현 기능은 Table 3에 제시하였다.

Table 3. Design Principles Based on the Core Functions of a Generative AI–Based Student Assessment Platform

No	Design Principle	Description	Needs Analysis Basis	Platform Functions	References
1	Assessment centered on diagnosing students' depth of understanding	Diagnoses students' understanding and thinking beyond score calculation.	· Teachers reported that score-oriented assessment is insufficient to capture students' understanding processes. · The need for process-oriented assessment in performance tasks was emphasized.	· Artifact-based item generation · Peer assessment support	Seong et al. (2024); Baidoo-Anu & Ansah (2023)
2	Multi-layered evidence for assessment judgments	Provides multiple sources of evidence to support teachers' judgments.	· The need for an assessment judgment structure that does not rely on a single scoring result was identified.	· Separate presentation of item and peer assessment results	Park(2025); Miao & Holmes(2023)
3	Provision of Personalized Feedback	Provides feedback tailored to students' understanding levels.	· A significant amount of time and effort was reported to be required for providing individualized feedback.	· Personalized feedback based on understanding analysis	Kim et al. (2024); Seong et al.(2024); Baidoo-Anu & Ansah (2023)
4	Preserving teachers' decision-making authority in assessment	Supports teachers' final decision-making authority in assessment.	· Concerns were raised about potential errors and overreliance on generative AI, highlighting the need to preserve teachers' professional judgment.	· Teacher review and revision of AI-generated results	Miao & Holmes(2023); Baidoo-Anu & Ansah (2023)
5	Intuitive platform structure	Enables assessment tasks with minimal cognitive load.	· Teachers perceived intuitive usability as a prerequisite for classroom adoption of AI-based assessment tools.	· Role-based interface · Task-oriented menus	Park(2025); Jakob Nielsen(2024)
6	Securing Ethical and Trustworthy Practices	Ensures a fair and trustworthy assessment environment.	· Common concerns were reported regarding bias, inaccuracy, and data privacy in generative AI-based assessment.	· Minimal data collection · Storage of AI rationales	Park(2025); Miao & Holmes(2023)

4.2 플랫폼 구조 설계

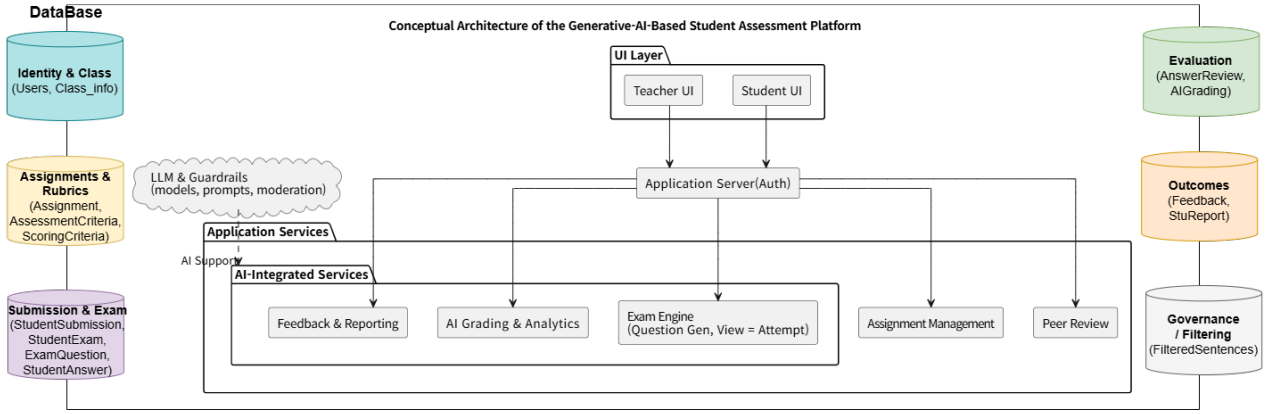


Figure 1. Conceptual Architecture of the Generative AI-Based Student Assessment Platform

Figure 1은 플랫폼의 개념적 아키텍처를 나타내며, 과제 생성부터 피드백에 이르는 전체 평가 사이클이 6개 논리 도메인(총 14개 테이블)으로 조직된 운영 데이터 저장소와 서비스 모듈 간의 일관된 데이터 흐름으로 연결되어 있음을 보여준다.

Figure 2은 플랫폼의 핵심 엔티티와 그 관계를 도식화한 것으로, 관계 설정은 과제 생성, 제출·응시, 문항·답안, 동료 평가·AI 채점, 피드백·결과 확인까지의 전 평가 과정을 포괄한다.

Figure 2. ERD for the Generative AI-Based Student Assessment Platform

본 플랫폼은 사용자 유형별 과업 중심의 2단계 메뉴 구조를 제시하며, Figure 3과 같다.

교사 영역은 사용자 관리, 과제 관리, 활동 관리, 성적 처리로 구성하였고, 학생 영역은 과제 제출, 시험 응시, 동료 평가, 결과 확인의 실제 작업 흐름을 반영하였다. 초기화·삭제·재생성 등 세부 기능은 화면 내부에서 처리하여 메뉴 계층을 단순화하고 탐색 효율을 높여 인지적 부하를 줄였다. 이러한 설계는 얕고 예측 가능한 네비게이션을 권고하는 UX 휴리스틱의 원칙과 부합한다.

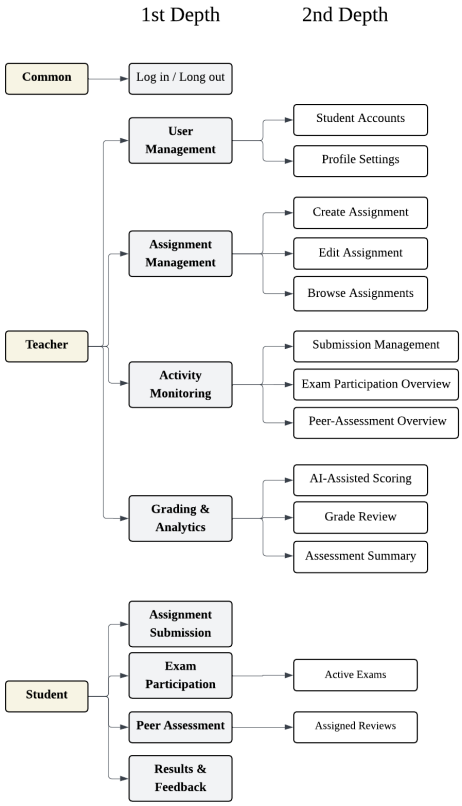


Figure 3. Menu Structure of the Generative AI-Based Student Assessment Platform

4.3 플랫폼 화면 설계

본 연구에서 제시하는 설계안은 플랫폼 개발에 앞서 마련된 개념적 청사진(conceptual blueprint)으로, 시스템의 구조와 각 기능이 어떤 방식으로 유기적으로 연결되어야 하는지를 체계적으로 표현한 것이다. 본 설계안은 앞서 도출한 설계 원리가 실제 플랫폼 환경에서 구현될 수 있는 방식을 구체적으로 보여주어, 단순한 이론적 구상에 머무르지 않고 실제 개발 단계에서 적용 가능한 지침으로 활용될 수 있도록 마련되었다. 또한 참고해야 할 세부 설계 요소(UI 구성, 데이터 흐름, 예외 처리 등)까지 포함하고 있어, 단순한 기능 배치도를 넘어 UI 설계와 구현 지침을 긴밀히 연계하는 핵심 가이드라인으로 기능한다.

본 논문에서는 학생 사용자의 핵심 기능인 '문항 응시'와 '동료 평가'를 중심으로 화면 설계안을 제시한다. 먼저, Figure 4는 학생용 시험 응시 화면을 나타낸다.

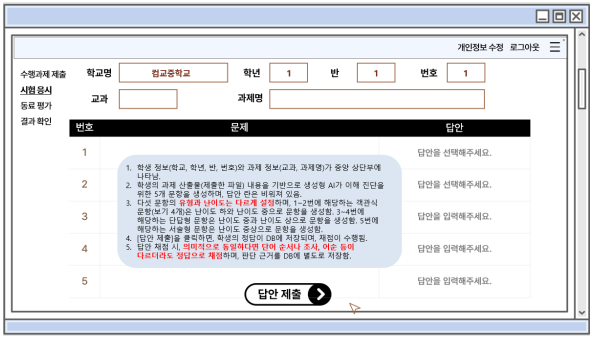


Figure 4. Student Question–Response Interface

문항 응시 화면은 학생이 제출한 개별 산출물을 기반으로 생성형 AI가 학습자의 이해 수준을 진단할 수 있도록 다섯 개의 맞춤형 문항을 자동 생성하도록 설계되었다. 문항은 객관식·단답형·서술형 등 다양한 형태와 난이도로 제시되어 학습자의 이해의 깊이와 전이를 다층적으로 점검한다.

이상적으로는 개별 산출물을 토대로 생성된 문항에 대해 해당 학생이 모두 정답을 도출할 수 있는 것이 바람직하지만, 실제 학교 현장에서는 학생 산출물의 외형적 완성도와 학습자의 실제 이해 간 차이가 발생할 수 있다. 이러한 문제를 보완하기 위해 본 플랫폼은 학생별로 맞춤형 문항을 제시하여 각 학생의 실제 이해를 재확인할 수 있도록 하였다. 또한 문항 생성 과정에 검증 알고리즘을 적용해 오개념·비학문적 진술을 자동 배제함으로써 평가의 타당성을 높이고자 하였다.

향후 프로토타입 구현 단계에서는 의미적 채점(Semantic Grading) 기능을 구현하여, 유사 의미의 다양한 표현도 정답으로 인식하고 AI가 채점 근거를 데이터베이스에 기록해 교사가 사후 검토할 수 있도록 지원하였다.

종합하면, 본 화면은 개별화된 문항을 통해 학습자의 실제 이해 수준을 진단하고, 문항의 타당성 관리를 통해 평가의 질을 제고하고 교사의 평가 부담을 완화하는 핵심 요소로 기능한다.

다음으로, Figure 5는 학생용 동료평가 화면이다.

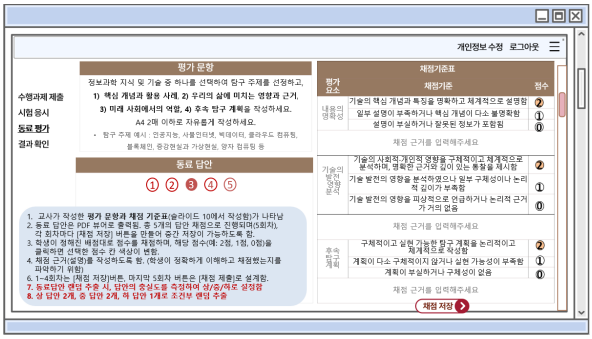


Figure 5. Student Peer–Assessment Interface

이는 학습자의 실제 이해를 다각도로 진단하기 위해 설계되었다. 학생은 교사가 제시한 평가 기준을 적용해 익명·무작위로 배정된 동료의 답안 5건을 평가하며, 단순 정·오답 확인을 넘어 논리성·근거의 타당성·개념 적용 등 여러 차원에서 사고를 확장하도록 유도된다. 각 평가 요소별로 채점 근거를 직접 기재하게 함으로써 학습자는 자신의 판단 과정을 명확히 드러내고 자기 성찰과 메타인지 역량을 강화할 수 있다.

동료평가 결과는 AI 채점과 비교되어 일치도·유사도가 산출되며, 학습자는 자신의 평가가 객관적 기준과 어느 정도 부합하는지를 즉시 확인할 수 있다. 교사는 학생의 채점 근거뿐만 아니라 AI 채점과 학생 채점 결과의 일치도·유사도를 함께 검토함으로써, 학생의 개념 이해 수준과 판단 근거를 다각적으로 파악할 수 있다.

종합하면, 본 화면은 동료평가 과정을 학습 경험으로 전환하여 학습자의 심층적 이해를 진단하고 성장을 지원하는 핵심 기능을 수행한다.

5. 생성형 AI 기반 학생평가 플랫폼 개발

5.1 프로토타입 개발 환경 및 결과

본 연구에서 개발한 생성형 AI 기반 학생평가 플랫폼의 프로토타입은 Python 언어를 기반으로 Visual Studio Code 환경에서 구현하였다. 사용자 인터페이스 및 웹 서비스 환경은 Streamlit 프레임워크를 활용하여 구축하였으며, 이를 통해 별도의 설치 과정 없이 웹 브라우저 환경에서 교사와 학생이 손쉽게 접근하고 활용할 수 있도록 하였다. 이러한 개발 환경 구성은 실제 학교 현장에서의 활용 가능성을 탐색하기 위한 프로토타입 구현이라는 연구 목적에 부합하도록, 접근성과 구현 효율성을 우선적으로 고려한 결과이다. Streamlit의 간결한 구조는 입력–처리–출력 흐름을 직관적으로 설계할 수 있어 교육 현장에서의 접근성과 활용 편의성을 높이는 데 기여한다.

데이터 관리 및 저장은 SQLite 데이터베이스를 기반으로

하였으며, 이는 경량형 구조를 갖춘 관계형 데이터베이스로 프로토타입 단계에서의 신속한 개발과 반복적 수정에 적합하다. 본 연구에서는 사용자 정보, 과제 메타데이터, 학생 제출물, 평가 기준, 동료평가 결과, AI 채점 결과, 개인 성취 리포트 등 총 14개의 테이블로 데이터베이스를 구성하여, 과제 제출부터 평가 결과 제공까지의 전 과정을 체계적으로 기록·관리하도록 설계하였다.

또한 문항 생성, 채점 및 피드백 제공을 위해 OpenAI의 GPT-4 모델을 API 형태로 연동하였으며, 평가 기준과 채점 맥락 등을 명시적으로 구조화한 프롬프트를 통해 평가의 방향성을 통제하였다. 학생 산출물은 교사가 설정한 평가 기준과 함께 프롬프트로 구성되어 AI 모델에 전달되며, 생성된 문항과 채점 결과, 피드백은 구조화된 형태로 처리되어 데이터베이스에 저장되고 즉시 사용자에게 제공된다. 모델 호출 시에는 temperature와 최대 토큰 수를 고정하여 모든 AI 처리 과정에서 동일한 설정을 적용함으로써 결과의 일관성을 확보하였다. 이러한 구조를 통해 본 플랫폼은 사용자 입력, AI 처리, 결과 저장 및 제공으로 이어지는 평가 흐름을 통합적으로 지원하는 프로토타입을 구현하였다.

한편, 본 플랫폼의 채점 과정은 정답 문구의 문자 단위 일치 여부를 넘어, 학습자의 응답이 채점 기준에 제시된 개념 요소와 의미적으로 어떻게 대응하는지를 판단하는 의미적 채점 절차를 포함한다. 이는 대규모 언어모델이 학습자 응답과 채점 기준 간의 개념적 대응 관계를 비교하도록 설계된 프롬프트 기반 방식으로 구현되었다. 해당 채점 프롬프트는 시범 적용 과정에서 연구자의 설계 관점 검토와 교사의 휴먼 피드백을 반복적으로 반영하여 정교화되었으며, AI 채점 결과가 교사의 다른 평가 근거와 함께 비교·종합 가능한 보조적 판단 자료로 활용되도록 설계되었다. 이러한 채점 구조는 정답 재현 중심의 평가를 넘어, 학습자의 이해 수준을 진단하는 이해 중심 평가를 지원하기 위한 시도라 할 수 있다.

Table 4는 프로토타입의 핵심 기능을 역할별로 구분하여 제시한 것이며, Figure 6은 개발한 프로토타입의 실제 구현 화면을 나타낸 것이다.

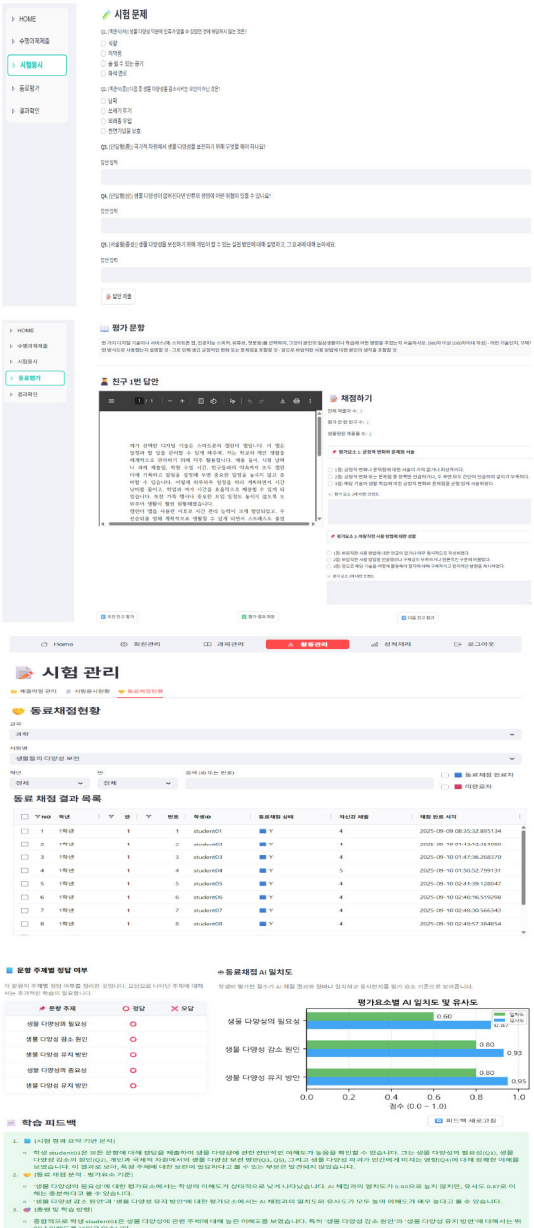


Figure 6. Prototype Development Results

Table 4. Key Platform Functions by User Role

No	Teacher		Student	
	Function	Description	Function	Description
1	Assignment Creation	Enter subject, assignment title, evaluation criteria, scoring scheme, and sample answers to create an assignment	Assignment Submission	Prepare the assignment response and upload as a PDF
2	Activity Monitoring	Track each student's assignment submission, test participation, and peer-review status; view correct/incorrect answers and both student and AI grading rationales	Test Participation	Take comprehension questions generated from the submitted assignment (multiple-choice, short-answer, and essay types)
3	Grade Review	Review test results, AI–student agreement and similarity in peer-review scoring, and AI feedback generation status	Peer Review	Evaluate anonymized peer responses according to the given criteria and record grading rationales
4	Feedback Generation	Generate or regenerate AI-based feedback using test and peer-review results	Result Review	Check AI-generated feedback based on test and peer-review outcomes and use it for self-directed learning

본 연구에서 개발한 생성형 AI 기반 학생평가 플랫폼은 기존 학생평가 방식과 구별되는 구조적 특징을 지닌다. 본 플랫폼은 교사 개인 중심의 평가 수행 구조에서 벗어나 교사와 AI가 협력하는 평가 구조를 지향하며, 교사는 최종 판단의 주체로서 기능하고 AI는 전반적인 평가 과정에서 이를 보조하도록 설계되었다. 특히 본 플랫폼은 개별적인 평가 활동을 병렬적으로 제시하는 데 그치지 않고, 서로 다른 성격의 이해 진단 근거를 하나의 평가 체제 안에서 구조적으로 통합하는 것을 핵심 설계 원리로 삼았다. 이에 따라 이해 진단의 근거를 학생 산출물에 한정하지 않고, 산출물을 기반으로 생성된 맞춤형 이해 점검 문항 응답 결과와 동료평가 결과를 함께 활용함으로써 이해 진단 근거를 다층화하였다. 이러한 다중 근거는 교사가 상호 비교·종합하여 학습자의 이해 수준을 해석할 수 있도록 설계되었다.

5.2 프로토타입 사용성 평가

사용성 평가는 실제 환경에서 프로토타입의 문제점과 개선 필요 요소를 조기에 식별하고 보완점을 도출하는 대표적 내적 타당화 절차로 알려져 있다[26]. 본 연구에서 개발한 생성형 AI 기반 학생평가 플랫폼의 프로토타입은 2025년 8월 25일부터 9월 12일까지 http://aiassess.kr에서 서비스하였으며, 이 기간 동안 컴퓨터 교육·교육 공학·교육 평가 전공 교수 및 연구원, 중등학교 교사 등으로 구성된 전문가 10인을 대상으로 사용성 평가를 실시하였다.

Table 5. Experts Participating in the Usability Evaluation

No	Current position	Area of expertise
1	Professor, K University	Computer Education
2	Professor, K University	Computer Education
3	Professor, S University	Educational Technology
4	Professor, K University	Educational Measurement and Evaluation
5	Researcher, K Foundation	Computer Education
6	Researcher, K Research Institute	Educational Technology
7	Teacher, D Middle School	Science Education
8	Teacher, G Middle School	Science Education
9	Teacher, C High School	Computer Education
10	Teacher, Y High School	Computer Education

본 사용성 평가는 실제 교육 현장에서의 적용 가능성을 탐색하고 프로토타입의 구조적 적절성과 개선 요구 요소를 조기에 도출하기 위한 탐색적 평가의 성격을 지닌다. 따라서 참여자 수를 제한된 전문가 집단으로 구성하고, 플랫폼의 기능 구현, 평가 흐름, 사용자 경험 전반에 대한 심층적 의견을 수집하는 데 중점을 두었다.

사용성 평가 도구는 설문지 방식으로 구성하였으며, 문항은 임철일 외(2009)의 연구, John Brooke (1986)가 제안한 시스템 사용성 척도(System Usability Scale, SUS), 그리고 Jakob Nielsen(2005)의 휴리스틱 평가 항목을 참고하여 본 연구 목적과 맥락에 맞게 재구성하였다.

Table 6. Domains and Indicators of the Usability Evaluation Instrument

Category	Key Indicators	Items (k)
System Reliability	· Frequency of system errors or failures · Loading and data-processing speed · Reproducibility of results under identical conditions	3
Ease of Use	· Intuitiveness and aesthetic quality of menu layout and information architecture · Ability to perform functions with minimal cognitive effort · Suitability for learners' levels and digital literacy	5
Accessibility & Efficiency	· Ease of access to desired information and functions · Consistency of the user interface · Efficiency of navigation · Immediate confirmation of submitted materials	3
Educational Effectiveness	· Contribution to streamlining assessment tasks for teachers, students, and the learning process · Accuracy in measuring learning level and understanding · Support for individualized feedback and promotion of self-directed learning skills	6
User Expectation & Support	· Degree to which expected features are fulfilled · Consistency of screen interface · Availability of step-by-step tutorials · Presence of safeguards against errors or data loss	4
Satisfaction & Intent to Use	· Overall user satisfaction · Intention for continued or future use · Alignment with the practical needs of educational settings	3
Open-ended Feedback	· Strengths, weaknesses, and recommendations for improvement of the platform	1

Table 6은 사용성 평가 도구의 문항 영역과 핵심 지표, 문항 수를 정리한 것이다. 본 평가 도구는 시스템 안정성, 사용 편의성, 정보 접근성과 효율성, 교육적 효과성, 사용자 기대 및 지원, 만족도 및 사용 의향 영역으로 구성하였다.

각 문항의 응답은 5점 리커트 척도(5=매우 그렇다, 4=그렇다, 3=보통이다, 2=그렇지 않다, 1=전혀 그렇지 않다)로 수집하였으며, 정량 결과의 해석을 보완하기 위해 개방형 문항을 병행하였다.

분석은 문항 수준의 기술통계를 중심으로 수행하였다. 각 문항의 평균(M)과 표준편차(SD)를 산출하여 Table 7에 제시하였다. 개방형 응답은 서술적 요약으로 정리하여 정량 결과의 해석을 보완하였다.

Table 7. Item-Level Results of the Usability Evaluation

Category	Item		N	M	SD
System Reliability	1-1	Expectation Alignment	10	4.2	0.75
	1-2	Button/Icon Clarity	10	4.5	0.81
	1-3	Layout Appropriateness	10	4.6	0.66

Category		Item	N	M	SD
Ease of Use	2-1	Layout Intuitiveness	10	4.4	0.66
	2-2	Purpose Clarity	10	4.2	0.98
	2-3	Aesthetics & Simplicity	10	4.4	0.66
	2-4	Low Cognitive Load	10	4.2	0.75
	2-5	Student-Level Suitability	10	4.4	0.80
Accessibility & Efficiency	3-1	Navigation Ease	10	4.7	0.46
	3-2	Interface Consistency	10	4.4	0.80
	3-3	Immediate Submission Access	10	4.4	0.80
Educational Effectiveness	4-1	Assessment Info Clarity	10	4.3	0.78
	4-2	Reduced Teacher Workload	10	4.6	0.66
	4-3	Holistic Achievement View	10	4.4	0.66
	4-4	Process/Outcome Analysis	10	4.2	0.98
	4-5	Personalized Feedback Effect	10	4.2	0.98
	4-6	Self-Directed Learning Support	10	4.0	1.26
User Expectation & Support	5-1	Predictable Execution	10	4.3	0.64
	5-2	Layout Consistency	10	4.5	0.92
	5-3	Tutorial Availability	10	3.9	1.45
	5-4	Critical-Action Safeguards	10	3.6	1.36
Satisfaction & Intent to Use	6-1	Intention to Use	10	4.2	0.87
	6-2	Overall Satisfaction	10	4.3	0.78
	6-3	Classroom Fit	10	4.5	0.67

5.3 결과 해석 및 논의

사용성 평가 결과, 대부분의 문항 평균이 4.0/5.0점 이상으로 나타나 프로토타입의 전반적 사용성은 긍정적으로 평가되었다. 다만 본 결과는 전문가 대상 사용성 평가를 통해 수집된 주관적 인식 자료에 기반한 것이므로, 본 절에서는 이를 중심으로 플랫폼의 교육적 활용 가능성 측면에서의 결과를 해석하고 향후 개선 방향을 논의하고자 한다.

첫째, 정보 접근성과 효율성 영역의 탐색 용이성(3-1, M=4.70, SD=0.46)은 가장 높은 점수를 보여, 핵심 정보·기능에 대한 접근 경로와 내비게이션 설계가 비교적 안정적으로 구현되었음을 시사한다. 시스템 안정성 영역에서도 레이아웃/배치 적절성(1-3, M=4.6, SD=0.66)과 버튼·아이콘 명확성(1-2, M=4.5, SD=0.81)이 높게 나타나, 시각적 명료성과 조작 용이성 측면에서 설계 완성도가 높은 편으로 해석된다. 다만, 일부 화면에서 교사·학생 메뉴가 동시에 노출된다는 지적이 있어, 사용자 역할별로 필요한 메뉴만 보이도록 하는 역할 기반 메뉴 가시성 제어가 요구된다. 이는 정보 과부하와 탐색 혼선을 줄이기 위한 개선의 여지가 있음을 시사한다. 사용 편의성 영역에서 화면 구성의 직관성(2-1, M=4.4, SD=0.66)과 심미성·간결성(2-3, M=4.4, SD=0.66)은 고르게 우수한 것으로 나타났다.

둘째, 사용자 기대 및 지원 영역의 중요 작업 안전장치(5-4, M=3.6, SD=1.36)와 튜토리얼 제공 여부(5-3, M=3.9, SD=1.45)는 평균이 다소 낮고 분산이 커 개선 여지가 확인되었다. 먼저, 중요 작업 안전장치의 경우 일부 화면에서는 삭제·초기화 등 고위험 작업에 대한 경고·확인이 구현되어 있었으나, 특정 흐름에서는 동일 수준의 보호장치가 일관되게 적용되지 않은 사례가 관찰되었다. 이러한 부분적·비일관적 적용은 해당 문항의 낮은 평균과 높은 분산으로 이어진 것으로 보이며, 보호장치의 전면적·일관적 적용을 위한 개선의 여지가 있음을 나타낸다. 또한, 튜토리얼 제공과 관련하여 평가 참여자에게 사용자 매뉴얼(정적 문서)은 제공되었으나, 페이지·기능 맥락에 맞춘 인앱 안내(예: 툴팁)는 제공되지 않았다. 즉, 문항 5-3이 지시하는 페이지 내 안내와 실제 제공 수단 간 방식의 불일치가 체감 지원 수준을 낮춘 요인으로 해석되며, 맥락형 안내 체계 보강이 필요함을 시사한다.

셋째, 교육적 효과성 영역에서는 학생의 자기주도 학습 역량 향상 지원(4-6, M=4.0, SD=1.26)의 표준편차가 상대적으로 커 응답자 간 평가 차이가 큰 것으로 나타났다. 이는 일부 응답자가 현재 제공되는 이해도 중심 피드백만으로는 목표 설정·실행 등 학습 행동으로의 전환을 촉진하기에 부족하다고 인식했음을 반영한다. 이에 따라 피드백에 다음 단계(활동) 제안과 같은 내용을 포함하여 실제 학습 행동으로의 전환을 지원할 필요가 있다.

넷째, 그럼에도 교사 평가 업무 부담 경감(4-2, M=4.6, SD=0.66)과 학생 성취의 다각적 이해(4-3, M=4.4, SD=0.66)가 높게 나타난 점은, 본 플랫폼이 교사의 평가 업무 지원과 성취 정보 제공 측면에서 활용 가능성이 있다고 인식되었음을 보여준다. 또한 만족도 및 사용 의향 영역의 교육 현장 적합성(6-3, M=4.5, SD=0.67)과 전반적 만족도(6-2, M=4.3, SD=0.78) 역시 긍정적으로 평가되어, 현장 적용 가능성에 대한 우호적 인식을 확인할 수 있었다.

Table 8. Key Issues and Improvement Directions Identified in the Usability Evaluation

Category	Key Issue	Cause	Improvement Direction
Functional Enhancements	Lack of in-app tutorial	User manual provided, but context-specific in-app guidance was insufficient	Strengthen context-aware in-app guidance (e.g., tooltips)
	Partial implementation of critical safety mechanisms	Warnings and confirmations for high-risk tasks were applied only partially and inconsistently	Apply consistent, comprehensive safety mechanisms to all critical tasks
	Absence of AI feedback-editing feature	Currently only an AI feedback regeneration function is available	Add teacher-level editing and refinement functions for AI-generated feedback
Content· Feedback Improvements	Limited support for students' self-directed learning	Understanding-focused feedback alone was insufficient to prompt learning behaviors	Include next-step guidance or activity suggestions to foster learning actions
	Inadequate appropriateness of student-facing feedback	Technical terms and expressions may hinder student comprehension	Provide dual-channel feedback (teacher vs. student) and use plain language and illustrative examples for students

개방형 문항에 대한 응답은 사용성 평가 종료 후 수집된 자유 기술식 의견을 대상으로 연구자가 반복적으로 검토하여 의미 단위로 분절한 후, 유사한 의견을 중심으로 범주화하는 방식으로 수행하였다. 개방형 문항에서 도출된 주요 의견과 시사점은 다음과 같다.

첫째, 학생의 산출물을 기반으로 한 이해 점검 목적의 맞춤형 문항 응시는 의미 있는 활동으로 평가되었다. 다만 과업의 생소함으로 인해 학생과 교사 모두에게 초기 적응 장벽이 존재할 수 있으므로 사전 안내나 간단한 연습 세션 등 초기 적응 지원 장치가 요구된다.

둘째, 학생 평가 활동에 대한 AI 생성 피드백은 교사의 재생성 기능만으로는 충분하지 않으며, 교사가 직접 수정·보완할 수 있는 편집 기능이 필수적이라는 의견이 제기되었다. 이는 교사 개입을 통해 피드백의 정확성과 맥락 적합성을 높일 필요가 있음을 시사한다.

셋째, 동일한 AI 피드백이 교사에게는 학생의 성취를 다각도로 파악하고 전문적 판단을 보완하는보조적 자료로 유용하지만, 학생에게는 전문적 용어와 표현으로 인해 이해가 어려울 수 있음이 지적되었다. 따라서 피드백을 교사용·학생용으로 구분하고, 학생용은 평이한 어휘·예시·다음 단계 제안 중심으로 작성하며, 용어 툴팁을 제공하는 등 수준 적합성을 확보할 필요가 있다. 사용성 평가에서 도출된 주요 개선점과 개선 방향은 Table 8과 같다.

6. 결론 및 제언

오늘날 생성형 AI의 보편화로 학습자의 실제 이해 수준을 정밀히 진단하기 어려워 교사에게는 새로운 평가 부담이 가중되고 있다. 완성도가 높은 산출물이라 하더라도 학생의 개념적 이해를 충분히 반영하지 않을 수 있어, 추가적인 검증과 해석이 요구되는 구조적 한계가 발생하는 것이다.

이러한 맥락에서 본 연구는 교사에게 추가로 요구되는 평가 부담을 실질적으로 경감하면서 학습자의 실제 이해 수준을 다각도로 포착할 수 있는 생성형 AI 기반 학생평가 플랫폼 프로토타입을 설계·구현하고, 전문가 사용성 평가를 통해 전반적 사용성과 현장 적용 가능성을 예비 검증하였다.

사용성 평가 결과, 개발된 프로토타입은 전반적으로 긍정적 사용성을 보였으며(다수 문항 M≥4.0), 특히 탐색 용이성, 레이아웃/아이콘 명확성, 화면 구성의 직관성과 심미성이 강점으로 확인되었다. 더불어 교사 평가 업무 부담 경감과 교육 현장 적합성 역시 높은 평가를 받았다. 반면, 교사·학생 역할 기반 메뉴 가시성 미흡, 인앱 튜토리얼의 부재, 중요 작업 안전장치의 비일관성, 피드백의 행동 전환 한계, 학생용 피드백 수준 적합성 보완 등은 개선점으로 도출되었다.

본 연구의 핵심 의의는 생성형 AI의 확산으로 결과물만으로는 학습자의 이해를 충분히 진단하기 어려워진 교육 현실에서, 이를 보완하기 위한 실천적 대안을 구상 수준을 넘어 실제 플랫폼으로 설계·구현하고 그 현장 적용 가능성을 예비 검증했다는 데 있다. 특히 본 플랫폼이 제공하는 학습 과정·결과 데이터와 AI 피드백은 교사의 전문적 판단을 보완하는 보조적 참고 자료로 활용될 수 있으며, 이를 통해 학생 이해 진단을 지원하고 평가 부담을 경감하는 교육적 잠재력을 지닌다.

이상의 결과를 바탕으로 다음과 같이 제언한다.

첫째, 사용성 평가에서 도출된 개선점을 차기 버전에 체계적으로 반영하여 플랫폼의 현장 적합도와 학습 촉진 효과를 단계적으로 높일 필요가 있다. 둘째, 본 연구는 전문가 10인을 대상으로 한 예비적 사용성 평가에 기반하고 있으므로, 향후 연구에서는 생성형 AI 기반 채점 결과의 신뢰성과 타당성을 보다 엄밀하게 검증하기 위해, 인간 채점과의 비교, 오류 사례 유형화, 정량적 성능 지표 산출 등을 포함한 검증 연구를 수행할 필요가 있다. 셋째, 나아가 실제 중등학교 수업 맥락에서 교사와 학생을 대상으로 한 현장 적용 연구를 수행하고, 학급 단위 실험을 포함한 확장된 표본을 통해 플랫폼의 교육적 효과와 활용 가능성을 종합적으로 검증할 필요가 있다.

참고문헌

[1] Yang, J., & Yoon, S. (2023). Beyond ChatGPT: Entering the generative AI era—Cases of media and content generative AI services and strategies for competitiveness. *Media Issue & Trend, 55*(3–4), 62–70. Korea Communications Agency.

[2] Elnaj, S. (2025, March 4). *Generative AI: Ongoing challenges.* Forbes Korea. https://www.forbeskorea.co.kr/news/articleView.html?idxno=340762

[3] Kim, J., Kang, D., & Ko, Y. (2023). A study on educational applications of generative AI: Focusing on the use of ChatGPT [in Korean]. *Journal of the Korean Association of Information Education, 27*(6), 691–703. https://doi.org/10.14352/jkaie.2023.27.6.691

[4] Lee, S. (2024). An Analysis of Research Trends about the Educational Use of Generative Artificial Intelligence : Focusing on Korean Journal. *The Korean Society of Christian Religious Education,* (79), 121-145. https://doi.org/10.17968/jcek.2024..79.006

[5] Jeong, J. (2023, September 7). *Using generative AI appropriately in education.* etnews. https://www.etnews.com/20230907000016

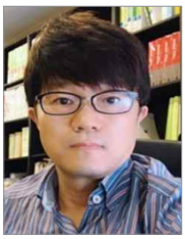
[6] Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X., Beresnitzky, A., Braunstein, I., & Maes, P. (2025). Your brain on ChatGPT: *Accumulation of cognitive debt when using an AI assistant for essay writing task [Preprint].* arXiv. https://doi.org/10.48550/arXiv.2506.08872

[7] KPMG. (2025). *Generative AI and the future of education.* KPMG. https://kpmg.com/ca/AI

[8] Lipman, J., & Distler, R. (2023, January 11). *Schools shouldn't ban access to ChatGPT*. Time. https://time.

com/6246574/schools-shouldnt-ban-access-to-chatgpt/
[9] Ta, R., & West, D. M. (2023, August 7). *Should schools ban or integrate generative AI in the classroom?*. Brookings. https://www.brookings.edu/articles/should-schools-ban-or-integrate-generative-ai-in-the-classroom/
[10] Baker, T., Smith, L., & Anissa, N. (2019). *Educ-AI-tion rebooted? Exploring the future of artificial intelligence in schools and colleges*. NESTA. https://www.nesta.org.uk/report/education-rebooted/
[11] Shin, S., & Kang, S. (2024). Foundational research on utilizing generative AI as a student assessment tool in secondary schools: Based on the analysis of secondary teachers' perceptions and experiences with generative AI. *The Journal of Korean Association of Computer Education, 27*(9), 1–14. https://doi.org/10.32431/kace.2024.27.9.001
[12] Seong, T., Si, K., & Choi, Y. (2024). The era of generative AI: Educational change and the direction of educational evaluation. *Journal of Educational Evaluation Research, 37*(1), 1–28. https://doi.org/10.31158/JEEV.2024.37.1.1
[13] Ministry of Education, & Korea Institute for Curriculum and Evaluation. (2024). *Looking into student assessment in middle school based on the 2022 revised curriculum*(ORM 2024-158-2). Ministry of Education & Korea Institute for Curriculum and Evaluation. https://stas.moe.fo.kr
[14] Mislevy, R., Steinberg, L., & Almond, R. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3-62. https://doi.org/10.1207/S15366359MEA0101_02
[15] Suto, I., Nadas, R., & Bell, J. (2011). Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education, 26*(1), 21-51. https://doi.org/10.1080/02671520902721837
[16] Kim, D., Kim, D., & Kim, K. (2024). Exploring the direction of artificial intelligence to support teachers' student assessment: Focusing on analysis of overseas cases. *Journal of the Korea Contents Association, 24*(5), 10–21. https://doi.org/10.5392/JKCA.2024.24.05.010
[17] Kizilcec, R., Huber, E., Papanastasiou, E., Cramb, A., Makridis, C., Smolansky, A., Zeivots, S., & Raduescu, C. (2024). Perceived impact of generative AI on assessments: Comparing educator and student perspectives in Australia, Cyprus, and the United States. *Computers and Education: Artificial Intelligence, 7*, 100269. https://doi.org/10.1016/j.caeai.2024.100269
[18] Xia, Q., Weng, X., Ouyang, F., Lin, T., & Chiu, T. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education, 21*(1), Article 38. https://doi.org/10.1186/s41239-024-00468-z
[19] Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, 3, 100075. https://doi.org/10.1016/j.caeai.2022.100075
[20] U.S. Department of Education, Office of Educational Technology. (2023). *Artificial intelligence and the future of teaching and learning: Insights and recommendations*. U.S. Department of Education. https://tech.ed.gov
[21] Korea Education and Research Information Service. (2023). *AI digital textbook development guidelines*. Korea Education and Research Information Service. https://www.keris.or.kr/main/ad/pblcte/selectPblcteETCInfo.do?mi=1142&pblcteSeq=13722
[22] Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. https://doi.org/10.54675/EWZM9535
[23] Baidoo-Anu, D., & Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI, 7*(1), 52–62. https://doi.org/10.2139/ssrn.4337484
[24] Park, H. (2025). AI based assessment: Theoretical foundations, specific models, and the directions of classroom application. *Journal of Educational Research, 23*(1), 81–111. https://doi.org/10.31352/JER.23.1.81
[25] Nielsen, J. (2024, January 30). *10 usability heuristics for user interface design*: Nielsen Norman Group. https://www.nngroup.com/articles/ten-usability-heuristics/
[26] Sauer, J., & Sonderegger, A. (2009). The influence of prototype fidelity and aesthetics of design in usability tests: Effects on user behavior, subjective evaluation, and emotion. *Applied Ergonomics, 40*(4), 670-677. https://doi.org/10.1016/j.apergo.2008.06.006

신상윤
· 2017년 국립공주대학교 컴퓨터교육과(교육학사)
· 2021년 국립공주대학교 컴퓨터교육전공(교육학석사)
· 2026년 국립공주대학교 컴퓨터교육전공(교육학박사)
· 2018년 3월~현재 대전광역시교육청 교사
관심분야 : 컴퓨터교육, 인공지능융합교육, 플랫폼 개발
ssyun_@naver.com



강신천
· 1993년 부산교육대학교 (교육학사)
· 1999년 한국교원대학교 교육과정전공(교육학석사)
· 2003년 한국교원대학교 교육공학전공(교육공학박사)
· 2005년 3월~현재 국립공주대학교 사범대학 컴퓨터교육과 교수
관심분야 : 컴퓨터교육, 교육공학, 인공지능융합교육, 플랫폼 개발
godsky@naver.com