컴퓨터교육학회 논문지 2026년 제29권 제8호
https://doi.org/10.32431/kace.2026.29.8.013

생성형 AI 기반 서술형 평가 지원 시스템의 설계·개발 및 성능 평가: 원격대학 사례를 중심으로*

Design, Development, and Performance Evaluation of a Generative AI-Based Constructed-Response Assessment Support System: A Case of a Distance University

안해연† · 양유정††
Haeyeon Ahn† · Youjung Yang††

요약

본 연구는 원격대학의 서술형 평가를 지원하기 위한 생성형 AI 기반 평가 지원 시스템을 설계·개발하고, 실제 과제 데이터를 활용하여 그 성능과 적용 가능성을 검토하였다. 선행연구 분석 및 전문가 검토를 통해 루브릭 기반 평가, 평가 근거와 피드백 제공, 교수자 검토·수정이 가능한 인간 개입형 평가 구조를 시스템의 설계 원리로 도출하였다. 이를 바탕으로 과제 생성과 제출, AI 평가, 교수자 검토 및 결과 제공 과정을 지원하는 웹 기반 시스템을 구현하였다. 두 교과목의 실제 학생 과제물 450건을 분석한 결과, 교수자 점수와 AI 점수 간 정합성은 두 교과목 모두 QWK .800 이상, nMAE .078 이하로 나타나 비교적 높은 수준의 채점 일치도와 낮은 상대적 평균 오차가 확인되었다. 교수자 32명을 대상으로 한 파일럿 사용자 평가에서는 시스템 활용 의도가 5점 척도에서 평균 4.18점(SD = .97)으로 나타나, 향후 활용에 대한 긍정적인 인식이 확인되었다. 본 연구는 생성형 AI를 교수자의 평가 판단을 대체하는 도구가 아니라, 평가와 피드백 업무를 지원하는 인간 개입형 평가 지원 도구로 활용하기 위한 설계 및 적용 방향을 제시하였다.

주제어 생성형 인공지능, 서술형 평가, 평가 지원 시스템, 원격대학

ABSTRACT

This study designed and developed a generative AI-based assessment support system for constructed-response assessment in a distance university and examined its performance and applicability using authentic assignment data. Based on prior research and expert review, three design principles were identified: rubric-based assessment, provision of scoring rationales and feedback, and a human-in-the-loop structure that enables instructors to review, revise, and confirm AI-generated results. A web-based system was then implemented to support assignment creation, submission, AI-based assessment, instructor review, and result delivery. Analysis of 450 assignments from two courses showed relatively high agreement between instructor and AI scores, with QWK values of .800 or higher and nMAE values of .078 or lower. In a pilot evaluation involving 32 instructors, the mean intention to use the system was 4.18 on a five-point scale (SD = .97), indicating positive initial perceptions regarding its future use. The findings suggest that generative AI can be used as a human-in-the-loop assessment support tool rather than as a replacement for instructors' judgment.

Keywords Generative AI, constructed-response assessment, assessment support system, distance university

†정회원 국가교육위원회 사무관
††정회원 한국방송통신대학교 선임연구원 (교신저자)

논문투고 2026년 06월 15일
심사완료 2026년 07월 09일
게재확정 2026년 07월 15일
발행일자 2026년 08월 27일

* 본 논문은 2025년 한국방송통신대학교 원격교육혁신 연구원의 연구보고서의 내용을 기반으로 작성하였음(정책과제 25-11).

1. 서론

원격교육은 시간과 장소의 제약 없이 학습할 수 있는 교육 환경을 제공함으로써 고등교육의 접근성을 확대해 왔다. 그러나 대규모 학습자를 대상으로 운영되는 원격대학에서는 교수자 1인이 많은 학생의 과제 평가와 피드백을 담당해야 한다는 한계가 존재한다. 국내 K 대학의 경우 2025년 기준 약 9만 명의 학부 재학생이 재학하고 있으며, 전임교원 1인당 학생 수는 평균 534명에 이른다[1]. 이러한 환경에서는 서술형 과제를 신속하고 일관되게 평가하고 개별화된 피드백을 제공하는 데 상당한 어려움이 발생할 수 있다.

서술형 평가는 학습자의 개념 이해, 논리적 사고 및 지식 적용 능력을 확인할 수 있다는 점에서 중요하다. 그러나 수기 중심의 평가 방식은 다수의 과제를 처리하는 과정에서 평가 지연, 피드백의 편차 및 교수자의 업무 부담을 초래할 수 있다[2-4]. 특히 학습자와 교수자 간 상호작용이 제한될 수 있는 원격교육에서는 평가 결과와 피드백이 학습자의 성취와 학업 지속을 지원하는 핵심 요소이므로, 평가의 효율성과 일관성을 함께 확보할 수 있는 기술적 지원이 필요하다.

최근 대형언어모델(Large Language Model, 이하 LLM)의 발전은 이러한 문제를 보완할 수 있는 가능성을 제시한다. LLM은 자연어의 의미와 맥락을 분석하고, 사전에 설정된 루브릭에 따라 서술형 응답의 점수와 피드백을 생성할 수 있다. 이에 따라 생성형 AI 기반 서술형 평가 시스템은 교수자의 반복적인 채점 업무를 줄이고 다수의 학습자에게 신속한 피드백을 제공할 수 있는 평가 지원 도구로 주목받고 있다[5-9].

생성형 AI 기반 평가에 관한 선행 연구는 주로 AI 점수와 인간 평가자 점수 간 일치도, 평가 정확도 및 반복 채점의 안정성을 검증해 왔다. 연구 결과에 따르면 적절한 루브릭과 평가 지침이 제공될 경우, AI 평가가 인간 평가와 일정 수준 이상의 정합성을 보일 수 있지만, 과제 유형, 평가 기준 및 모델에 따라 성능 차이가 발생하는 것으로 나타났다[4, 9, 10-12]. 그러나 기존 연구는 특정 모델의 채점 성능을 비교하거나 제한된 실험 자료를 분석한 경우가 많아, 실제 교육 환경의 평가 절차와 교수자의 요구를 반영한 시스템 설계·적용 연구는 상대적으로 부족하다.

생성형 AI 기반 서술형 평가가 실제 교육 환경에서 활용되기 위해서는 높은 채점 성능뿐 아니라 교수자가 평가 과정에 개입할 수 있는 구조가 필요하다. AI가 산출한 점수와 피드백의 근거를 교수자가 확인하고, 필요한 경우 이를 수정하여 최종 평가를 확정할 수 있어야 한다. 따라서 생성형 AI는 교수자의 판단을 대체하는 평가 도구가 아니라, 채점 기준(루브릭)의 일관된 적용과 피드백 작성을 지원하는 협력적 평가 도구로 설계될 필요가 있다[13]. 특히 평가 결과가 실제 학업성적과 연결되는 상황에서는 루브릭 기반 평가, 평가 근거의 제시, 교수자 검토 및 수정 기능을 포함하는 인간 개입형 구조가 중요하다.

이에 본 연구는 원격대학에서 활용 가능한 생성형 AI 기반 서술형 평가 지원 시스템을 설계·개발하고, 실제 학생 과제물을 활용하여 그 적용 가능성을 검토하는 데 목적을 두었다. 특히 본 연구는 생성형 AI를 활용한 자동 채점에 초점을 두기보다, AI 기술이 교수자의 평가 의사결정, 루브릭 기반 채점, 피드백 생성 및 평가 결과 검토 과정을 어떻게 지원할 수 있는지를 살펴보고자 하였다.

이를 위해 Richey와 Klein의 설계·개발 연구 방법[14]에 따라 문헌분석 및 전문가 검토를 수행하여 시스템의 설계 원리를 도출하고, 이를 반영한 웹 기반 시스템을 구현하였다. 개발된 시스템은 과제 생성과 제출, 루브릭 기반 AI 평가와 피드백 생성, 교수자 검토·수정, 최종 평가 결과 제공의 전 과정을 지원하도록 설계되었다. 이후 두 교과목의 실제 학생 과제물을 활용한 성능 평가와 교수자 대상 파일럿 사용자 평가를 통해 시스템의 평가 성능과 교육적 활용 가능성을 종합적으로 검토하였다.

본 연구의 연구 문제는 다음과 같다.

첫째, 교수자의 평가 요구를 반영한 생성형 AI 기반 서술형 평가 지원 시스템의 설계 원리는 무엇인가?

둘째, 개발된 생성형 AI 기반 서술형 평가 지원 시스템의 점수는 교수자 점수와 어느 수준의 정합성과 오차를 보이는가?

셋째, 교수자는 개발된 생성형 AI 기반 서술형 평가 지원 시스템의 유용성, 사용 편의성 및 활용 가능성을 어떻게 평가하는가?

2. 이론적 배경

2.1 원격대학에서의 서술형 평가

원격교육은 대규모 학습자를 대상으로 비대면·비동시적 방식으로 운영되므로, 학습자의 이해 수준과 수행 결과를 지속적으로 파악하기 어렵다는 특성을 지닌다. 이러한 환경에서 평가는 학습 성취를 확인하는 절차일 뿐 아니라, 학습자의 현재 수준을 진단하고 이후 학습 방향을 조정하도록 지원하는 교수·학습 요소로 기능한다[15, 16].

서술형 평가는 선택형 평가에 비해 학습자의 개념 이해, 논리적 사고 및 지식의 적용 과정을 구체적으로 확인할 수 있다는 장점이 있다. 그러나 대규모 원격 강좌에서 교수자 한 명이 모든 응답을 직접 채점하고 개별화된 피드백을 제공하기에는 상당한 시간과 노력이 요구된다[17, 18]. 이로 인해 채점과 피드백이 지연되거나 평가 기준의 일관성이 약화될 수 있으며, 학습자는 자신의 수행을 적시에 점검하고 이후 학습에 반영하기 어려워질 수 있다[15, 19].

따라서 원격대학의 서술형 평가 지원 시스템은 반복적인 채점 업무의 부담을 완화하는 동시에, 교수자의 전문적 판단과 피드백 제공을 효과적으로 지원할 수 있도록 설계될 필요가 있다.

2.2 생성형 AI 기반 서술형 자동 채점과 성능 검증

에세이 자동 채점(Automated Essay Scoring, AES)은 컴퓨터 알고리즘을 활용하여 학습자의 서술형 응답에 점수를 부여하는 기술이다. 초기에는 어휘 빈도, 응답 길이, 문법 오류와 같은 표면적 특성을 중심으로 점수를 예측하였으나, 최근에는 대규모 언어모델(Large Language Model, LLM)을 활용하여 응답의 의미, 논리 구조 및 주제 적합성을 종합적으로 평가하는 방식으로 발전하고 있다[20].

LLM 기반 채점은 과제 설명, 루브릭 및 학습자 응답을 자연어 프롬프트로 제공하여 별도의 과제별 모형 학습 없이 점수와 평가 근거를 생성할 수 있다는 장점이 있다. 그러나 채점 결과는 모델, 프롬프트, 루브릭 및 과제 특성에 따라 달라질 수 있으며, 선행연구에서도 인간 평가자와의 일치도와 반복 안정성에 대해 상반된 결과가 보고되었다[4, 21, 22]. 따라서 생성형 AI 기반 채점의 활용 가능성은 실제 교육 과제에서 인간 채점과의 일치도와 오차를 통해 검증할 필요가 있다.

인간–AI 채점 비교에서는 하나의 지표보다 상호 보완적인 지표를 함께 활용하는 것이 적절하다. QWK는 점수 범주 간 가중 일치도를, PCC는 두 점수 간 선형적 관련성을, MAE는 실제 점수 차이의 평균 크기를 나타낸다[23, 24]. 다만 MAE는 원점수 척도에 의존하는 지표이므로, 점수 범위가 서로 다른 교과목 간 오차 수준을 비교하기 위해 MAE를 해당 교과목의 점수 범위로 나누어 정규화한 nMAE를 함께 산출하였다.

2.3 생성형 AI 평가 지원 시스템의 설계 원칙

생성형 AI 기반 서술형 채점을 교육 현장에 적용하기 위해서는 평가 기준의 명시성, 결과의 설명 가능성 및 인간의 감독 가능성이 확보되어야 한다. 먼저 생성형 AI가 채점 항목과 점수 수준을 일관되게 해석하도록 하기 위해서는 평가 요소와 수준별 수행 특성이 구체적으로 제시된 루브릭이 필요하다. 평가 기준이 모호한 경우 모델이 평가 요소의 중요도를 임의로 해석하거나 동일한 수준의 응답에 서로 다른 점수를 부여할 가능성이 있다[25]. 따라서 과제 설명, 루브릭, 출력 형식 및 채점 규칙을 구조화하고 동일한 조건에서 채점을 수행하는 것이 중요하다.

둘째, 생성형 AI 기반 평가는 최종 점수뿐 아니라 점수 산출의 근거를 사용자가 확인할 수 있도록 설계되어야 한다. 설명 가능성은 모델의 내부 추론 과정을 모두 공개하는 것을 의미하기보다, 응답의 어떤 부분이 어떠한 루브릭 기준에 근거하여 평가되었는지를 교수자가 검토할 수 있도록 정보를 제공하는 것을 의미한다[26, 27]. 이러한 기준 연계형 설명은 오류를 발견하고 평가 결과를 교육적으로 활용하기 위한 기초 자료가 된다.

셋째, 생성형 AI의 채점 결과는 인간의 검토와 조정이 가능한 구조에서 활용되어야 한다. 교육평가는 학습자의 성취와 직접 연결되므로 AI가 최종 평가 권한을 독립적으로 행사하기보다, 교수자의 전문적 판단을 지원하는 보조 평가자로 기능하는 것이 적절하다[28, 29]. 교수자가 AI 점수와 평가 근거를 확인하고 필요한 경우 수정할 수 있는 human-in-the-loop 구조는 채점 오류에 대응하고 최종 평가에 대한 교육적 책임을 유지하기 위한 조건이다.

3. 연구 방법

본 연구는 원격대학 교육 환경에서 활용 가능한 '생성형 AI 기반 서술형 평가 지원 시스템'을 구축하기 위하여 Richey와 Klein이 제안한 설계·개발 연구(Design and Development Research) 방법론을 적용하였다. 설계·개발 연구는 특정 교수설계 산출물이나 모델의 개발 과정을 체계적으로 기록하고, 그 결과물의 타당성을 검증함으로써 실천적 지식을 도출하는 데 목적이 있다[14].

연구 절차는 크게 (1) 설계 원리의 도출, (2) 평가 지원 시스템 설계 및 개발, (3) 성능 평가, (4) 파일럿 사용자 평가의 네 단계로 구성된다. 각 단계는 원격대학의 실제 맥락에서 수행되었으며, 개발된 시스템에 대한 성능평가와 파일럿 사용자 평가를 통해 실제 교육 현장에서의 적용 가능성을 검토하였다. 구체적인 연구 절차는 Figure 1과 같다.

① Derivation of Design Principles	② System Design and Development
▪ Review of prior studies and existing practices ▪ Expert review and requirement analysis ▪ Derivation of design principles	▪ Design of system structure and functions ▪ Prototype development ▪ System implementation and integration
③ Performance Evaluation	④ Pilot User Evaluation
▪ Comparison of AI and instructor scores ▪ Quantitative performance analysis ▪ Visual analysis of score distributions	▪ Operation of a pilot test with instructors and students ▪ Evaluation of usability and usefulness ▪ Analysis of qualitative feedback

Figure 1. Research Procedure

3.1 생성형 AI 기반 서술형 평가 지원 시스템의 설계 원리 도출

본 연구는 원격대학에서 교수자의 서술형 평가를 지원하기 위한 생성형 AI 기반 평가 시스템을 설계·개발하고자 하였다. 이를 위해 생성형 AI 기반 평가, 자동 평가, 설명 가능한 인공지능, 인간 중심 AI, 교육평가의 책임성과 관련된 선행 연구 및 정책 문헌을 분석하였다. 문헌분석의 목적은 생성형 AI를 교육평가에 적용할 때 발생할 수 있는 주요 쟁점을 도출하고, 이를 시스템 설계 요구사항과 구현 원리로 전환하는 데 있었다.

문헌분석 결과, 생성형 AI 기반 서술형 평가 시스템을 교육 현장에 적용하기 위해서는 명시적인 루브릭을 통한 평가 기준의 일관된 적용[30], AI가 산출한 점수와 피드백의 판단 근거를 확인할 수 있는 설명 가능성[26], 그리고 교수자가

Table 1. Final Design Principles and System Implementation Based on Expert Consultation

Design Principle	Expert Feedback	Final Reflection in System Design
Rubric-based assessment framework	Criterion-level feedback should be provided.	Criterion-level feedback function was added.
Explainability of judgment criteria and rationale	Feedback for students should be strengthened.	Corrective feedback for students was added.
Human-in-the-loop assessment structure	Instructor support functions are needed for efficient review.	Bulk download and visualization functions were added.

AI 평가 결과를 검토·수정·확정하고 최종 평가 책임을 유지할 수 있는 인간 개입형 구조[28, 31]가 요구되는 것으로 나타났다. 이에 본 연구에서는 생성형 AI 기반 서술형 평가 지원 시스템의 설계 원리를 ① 루브릭 기반 평가 체계, ② 판단 기준과 근거의 설명 가능성, ③ 교수자 개입형 평가 구조로 설정하였다.

첫째, 루브릭 기반 평가 체계는 생성형 AI 평가의 타당성과 일관성을 확보하기 위한 설계 원리이다. 자동 평가 분야의 선행 연구는 명시적 평가 기준을 구조화하고 이를 일관되게 적용하는 것이 평가 결과의 신뢰성과 타당성을 높이는 데 중요하다고 보았다[30]. 또한 최근 교육 분야의 AI 활용 논의에서도 AI의 판단은 사전에 정의된 교육적 기준에 근거해야 하며, 평가 결과가 임의적 판단으로 산출되어서는 안 된다는 점이 강조되고 있다[28]. 이에 본 연구에서는 교수자가 과제별 평가 기준을 루브릭 형태로 제시하고, 생성형 AI가 해당 루브릭을 기준으로 점수와 피드백을 산출하도록 시스템을 설계하였다. 특히 총점 중심의 단일 평가가 아니라 평가 기준별 점수와 피드백이 함께 제공되도록 하여 교수자가 AI 평가 결과의 적절성을 검토할 수 있도록 하였다.

둘째, 판단 기준과 근거의 설명 가능성은 AI가 산출한 점수와 피드백이 어떠한 평가 기준에 따라 도출되었는지를 교수자와 학습자가 이해할 수 있도록 하기 위한 설계 원리이다. 교육적 맥락에서 AI 기반 평가 결과는 단순히 점수를 제시하는 데 그쳐서는 안 되며, 해당 점수가 산출된 이유와 개선 방향을 함께 제공해야 한다[26]. 설명 가능성은 AI 평가 결과에 대한 사용자의 신뢰를 높이고, 교수자가 평가 결과를 교육적으로 해석·조정할 수 있도록 지원하는 조건이다. 이에 본 연구에서는 시스템이 평가 기준별 피드백, 종합 피드백, 학습자 대상 교정적 피드백을 생성하도록 설계하였다. 이를 통해 AI 평가 결과가 교수자에게는 검토를 위한 자료로, 학습자에게는 학습 개선을 위한 형성적 피드백으로 활용될 수 있도록 하였다.

셋째, 교수자 개입형 평가 구조는 생성형 AI가 평가를 보조하되 최종 판단과 책임은 교수자에게 유지되도록 하기 위한 설계 원리이다. Human-in-the-loop 접근은 AI가 산출한 결과를 인간이 검토·수정·확정할 수 있도록 인간을 의사결정 과정에 포함하는 설계 방식이다[31]. 특히 교육평가에서는 점수 부여가 학습자의 성취 판단과 직결되므로, AI가 산출한 결과를 교수자가 검토하고 수정할 수 있는 절차가 필요하다. 이에 본 연구에서는 AI 자동 평가 결과를 교수자가 확인한 뒤 최종 점수와 피드백을 수정·확정할 수 있도록 하였다.

도출된 설계 원리의 타당성과 현장 적용 가능성을 검토하기 위해 교육공학 전공 연구자 2인, 교육학 전공 연구자 1인, AI 기반 교육혁신 및 원격교육 운영 경험을 보유한 원격대학 교수 2인으로 구성된 전문가 자문을 실시하였다. 자문 결과, 도출된 설계 원리는 전반적으로 타당한 것으로 평가되었으나, 채점 기준별 피드백 제공, 학습자 대상 교정적 피드백 강화, 대규모 평가 결과 관리를 위한 일괄 다운로드 및 시각화 기능 등 교수자 지원 기능의 필요성이 제안되었다. 전문가 자문을 반영하여 확정한 최종 설계 원리는 Table 1과 같다.

3.2 생성형 AI 기반 서술형 평가 지원 시스템의 설계 및 개발

본 연구에서는 도출된 설계 원리를 바탕으로 교수자와 학습자가 모두 활용할 수 있는 웹 기반 생성형 AI 서술형 평가 지원 시스템을 개발하였다. 시스템은 과제 생성, 루브릭 입력, 학습자 과제 제출, AI 자동 평가, 교수자 검토 및 수정, 최종 평가 결과 제공의 절차를 지원하도록 설계하였다. 특히 본 시스템은 AI가 평가를 전적으로 대체하는 방식이 아니라, 교수자의 평가 업무를 보조하고 교수자가 최종 판단을 수행할 수 있도록 지원하는 평가 지원 시스템으로 구현되었다.

시스템은 교수자와 학습자의 역할을 구분하는 역할 기반 접근 구조를 적용하였다. 교수자는 과제와 루브릭을 등록하고, 학습자가 제출한 PDF 파일을 확인하며, AI가 산출한 점수와 피드백을 검토·수정·확정할 수 있다. 학습자는 과제를 제출하고, 교수자가 확정한 평가 결과와 피드백을 확인할 수 있다. 이러한 구조는 생성형 AI 기반 평가에서 요구되는 루브릭 기반 평가, 설명가능성, 교수자 개입형 평가 구조를 시스템 기능 수준에서 구현하기 위한 것이다.

시스템 아키텍처는 Figure 2와 같이 사용자 계층, 프론트엔드, 백엔드 및 데이터 관리, 문서 처리 및 평가 요청, AI 평가 모델, 결과 저장 및 활용 계층으로 구성된다. 교수자와 학생은 React 기반 프론트엔드를 통해 시스템에 접근하며, 프론트엔드는 과제 생성, 제출물 확인, 평가 결과 검토 및 피드백 확인을 위한 사용자 인터페이스를 제공한다.

백엔드 및 데이터 관리 계층에서는 Supabase Edge Function, PostgreSQL Database, Supabase Storage, Authentication 및 Row Level Security를 활용하여 사용자 인증, 데이터 저장, 파일 관리, 접근 제어를 수행한다.

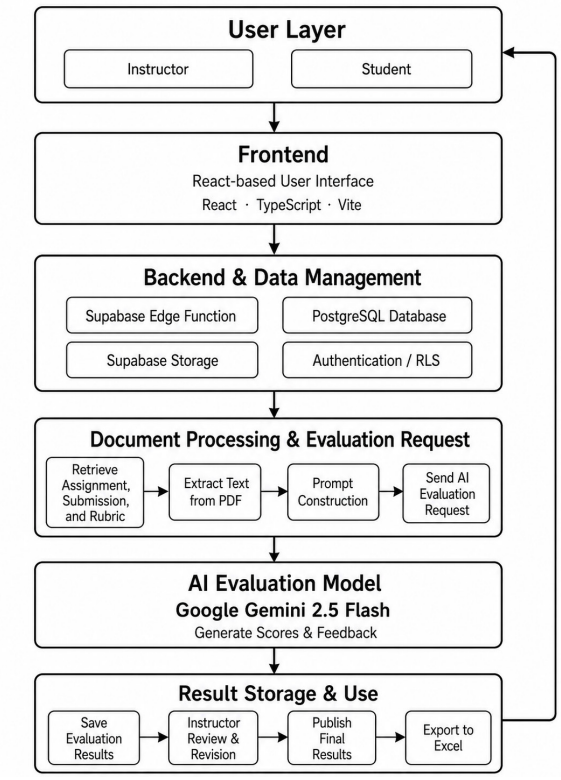

Figure 2. System Architecture of the Generative AI-Based Assessment Support System

문서 처리 및 평가 요청 단계에서는 과제, 제출물, 루브릭 정보를 조회하고, 학생이 제출한 PDF 파일에서 텍스트를 추출한 뒤, 이를 루브릭과 결합하여 AI 평가 프롬프트를 구성한다. 학생 제출물은 문서작성 프로그램으로 작성된 텍스트 기반 PDF였으며, pdfjs-dist 5.4를 이용하여 PDF의 텍스트 레이어를 추출하였다. 손글씨 또는 스캔 이미지 형태의 제출물은 포함되지 않았으므로 별도의 광학문자인식(OCR)은 적용하지 않았다.

이후 Google Gemini 2.5 Flash 모델이 점수와 피드백을 생성하며, 생성된 평가 결과는 데이터베이스에 저장된다. 교수자는 저장된 AI 평가 결과를 검토·수정한 후 최종 결과를 확정하고, 학습자는 확정된 점수와 피드백을 확인할 수 있다. 시스템의 기술 구성은 Table 2와 같다.

Table 2. Technical Configuration of the Generative AI-Based Essay Assessment Support System

Component	Technology	Main Function
Backend	Supabase	Authentication, data storage, file management, and access control
Frontend	React 18, TypeScript, Vite, Tailwind CSS	Instructor and student interfaces, assessment result display, and user interaction
AI model	Gemini 2.5 Flash	Rubric-based scoring and feedback generation
Database	PostgreSQL	Storage of assignments, submissions, rubrics, assessment results, and user information
Security	Row Level Security, SSL/TLS encryption	Role-based data access control and secure communication
Supported file format	PDF	Submission of student essay assignments
File processing	pdfjs-dist 5.4	PDF text extraction and construction of assessment input data
Assessment method	AI-based automated assessment + instructor review	Instructor review, revision, and confirmation of AI-generated assessment results
Authentication system	Role-based access control	Functional separation between instructors and students

시스템의 주요 기능은 설계 원리와 직접적으로 연결되도록 구현하였다. 첫째, 루브릭 기반 평가 체계를 구현하기 위해 교수자가 과제별 평가 기준을 입력할 수 있도록 하였으며, AI 평가 요청 시 해당 루브릭이 평가 프롬프트에 포함되도록 하였다. AI는 루브릭 기준에 따라 기준별 점수, 기준별 피드백, 종합 피드백을 산출하도록 설정하였다. 이를 통해 평가 기준이 시스템 내부에서 명시적으로 작동하도록 하였다.

둘째, 설명가능성을 확보하기 위해 AI가 단순한 점수만 산출하지 않고, 각 평가 기준에 대한 판단 근거와 개선 방향을 함께 생성하도록 하였다. 교수자는 AI가 제시한 기준별 피드백을 확인함으로써 점수 산출의 근거를 검토할 수 있으며, 학습자는 최종 확정된 피드백을 통해 자신의 과제에서 개선이 필요한 부분을 확인할 수 있다. 따라서 본 시스템의 설명 가능성은 기술적 의미의 모델 내부 해석에 한정되지 않고, 교육적 활용을 위한 평가 근거와 피드백의 명시화로 구현되었다.

셋째, 교수자 개입형 평가 구조를 구현하기 위해 AI 평가 이후 교수자가 평가 결과를 검토·수정·확정하는 절차를 포함하였다. AI가 산출한 점수와 피드백은 최종 평가 결과로 자동 확정되지 않으며, 교수자가 이를 검토한 뒤 필요한 경우 수정할 수 있다. 또한 대규모 평가 상황에서 교수자의 평가 업무 부담을 줄이기 위해 평가 결과 일괄 다운로드 기능과 시각화 기능을 포함하였다. 이를 통해 AI 기반 자동화의 효율성과 교수자 중심 평가 책임성을 함께 확보하고자 하였다.

개발된 시스템의 사용자 인터페이스는 Figure 3과 같다.

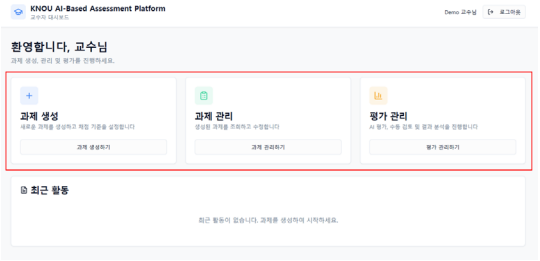

Figure 3. Instructor Dashboard

3.3 AI 평가 모델 및 실행 절차

AI 평가 모델로는 Gemini 2.5 Flash를 활용하였으며, 실제 평가는 2025년 9월부터 10월까지 실시하였다. 별도의 파인튜닝(fine-tuning)을 수행하지 않고, 기본 API 환경에서 시스템 지시문과 생성 매개변수(generation parameters)를 설정하는 방식으로 평가 환경을 구성하였다. API 호출에는 모델 식별자 gemini-2.5-flash를 사용하였으며, 모든 제출물에 동일한 모델과 생성 설정을 적용하였다. 이는 별도의 대규모 학습 데이터 구축이나 모델 재학습 없이도 실제 교육 현장에서 활용 가능한 생성형 AI 기반 평가 지원 시스템을 구현할 수 있는지를 확인하기 위한 것이다.

생성 매개변수는 루브릭 기반 평가 결과의 변동성을 낮추고 충분한 평가 피드백을 생성할 수 있도록 설정하였다. temperature는 0으로 설정하여 응답 생성의 확률적 변동성을 최소화하고자 하였으며, topP는 0.8로 설정하여 응답 생성에 활용되는 후보 토큰의 누적확률 범위를 제한하였다. maxOutputTokens는 800으로 설정하여 평가 기준별 판단 근거와 개선 방향을 포함한 충분한 분량의 피드백이 생성되도록 하였다. presencePenalty와 frequencyPenalty는 모두 0으로 설정하여 신규 표현이나 반복 표현에 대한 별도의 패널티를 부여하지 않았다.

시스템 지시문에는 모델의 역할을 공정한 루브릭 기반 평가 지원자로 규정하고, 제공된 과제 내용, 루브릭 및 학생의 제출물에 근거하여 평가하도록 명시하였다. 또한 루브릭의 평가 항목과 배점 기준에 따라 점수를 부여하고, 평가 기준별 판단 근거와 개선 방향 및 종합 피드백을 생성하도록 지시하였다. 적용한 시스템 지시문은 다음과 같다.

```
You are an impartial assignment evaluator.
Your task is to assess student responses only based on:
· The assignment prompt
· The scoring rubric
· The student response
Do not introduce evaluation criteria that are not specified in the rubric or make unsupported assumptions.

Evaluation Procedure
1. Internalize Criteria
· Understand the assignment's requirements and goals.
· Analyze the rubric, scoring dimensions, and performance levels.
2. Evaluate Student Response
· First Reading: Read once to capture overall impression.
· Second Reading: Re-read using the rubric. Evaluate each dimension, citing specific text as evidence.
3. Generate Feedback
· Provide overall score and dimension-level evaluation.
· Summarize strengths and weaknesses.
· Strengths: Quote effective parts and explain why they meet criteria.
· Improvements: Point out weaknesses and give clear, constructive suggestions.

Rule
All evaluations must rely solely on the rubric and the student's response.
```

3.4 AI 평가 성능 분석

3.4.1 분석 자료 선정

본 연구는 생성형 AI 기반 서술형 평가 지원 시스템의 성능을 실제 교육 맥락에서 검토하기 위해 분석 대상 교과목을 선정하였다. 이를 위해 연구 참여 학과를 공모하였으며, 이 중 학문 분야와 과제 특성의 이질성을 고려하여 2개 교과목을 최종 분석 대상으로 선정하였다. 교과목 선정 시에는 ① 중·대규모 수강생을 보유하고, ② 루브릭 기반 서술형 과제를 운영하며, ③ 학생 과제물과 교수자 평가 기록을 확보할 수 있는지를 주요 기준으로 삼았다. 또한 분석 자료의 다양성을 확보하기 위해 과제 유형, 평가 척도 및 루브릭 구조가 서로 다른 교과목을 포함하였다.

분석 자료는 선정된 2개 교과목의 학생 과제물 원문, 과제별 루브릭, 교수자가 직접 평가한 최종 점수로 구성되었다. 본 연구는 실제 성적 산출이 완료된 이후 사후적으로 수행되었으며, 분석 전 학생 이름, 학번 등 개인식별 정보를 제거하였다. 또한 본 연구는 기관의 연구 자료 활용 승인을 거쳐 수행되었다. 분석 자료의 구성은 Table 3과 같다.

Table 3. Analysis Data Used for Performance Evaluation

No.	Course	Score Range	N
1	Course A	0–15	300
2	Course B	0–30	150

Note. N = number of student responses.

3.4.2 분석 절차 및 지표

생성형 AI 기반 서술형 평가 지원 시스템이 교수자 평가 점수와 유사한 평가 결과를 산출하는지를 확인하기 위해 AI 평가 성능을 분석하였다. 실제 제출 과제에 대해 교수자가 부여한 점수를 준거 점수로 설정하고, 동일 과제에 대해 시스템이 산출한 AI 평가 점수와 비교하였다. 해당 교과목은 전문성이 요구되는 전공 교과목으로, 동일 전공 분야의 복수 평가자를 확보하기 어려운 현실적 제약으로 인해 담당 교수 1인의 평가 점수를 준거로 활용하였다.

성능 분석 지표로는 이차 가중 카파 계수(Quadratic Weighted Kappa, 이하 QWK), 피어슨 상관 계수(Pearson correlation coefficient, 이하 PCC), 평균 절대오차(Mean Absolute Error, 이하 MAE)를 활용하였다. QWK는 교수자 평가와 AI 평가 간의 점수 일치도를 확인하기 위한 지표로 사용하였다. PCC는 두 평가 점수 간의 선형적 관련성을 확인하기 위해 산출하였으며, MAE는 AI 평가 점수가 교수자 평가 점수와 평균적으로 얼마나 차이가 나는지를 확인하기 위해 활용하였다. 또한 교과목별 점수 범위의 차이를 고려하여 MAE를 각 교과목의 점수 범위로 나눈 nMAE를 산출함으로써 교과목 간 상대적인 오차 수준을 비교하였다. 이를 통해 AI 평가 결과의 일치도, 점수 경향성, 절대적·상대적 오차 수준을 종합적으로 검토하였다.

3.5 교수자 대상 파일럿 사용자 평가

본 연구에서는 개발된 시스템의 현장 적용 가능성과 사용자 수용 가능성을 탐색하기 위해 교수자 대상 파일럿 사용자 평가를 실시하였다. 이를 통해 개발된 시스템이 실제 교육 현장에서 평가 지원 도구로 활용될 수 있는지와 이에 대한 교수자의 초기 인식을 탐색적으로 확인하였다.

파일럿 사용자 평가는 2025년 11월 12일부터 25일까지 2주간 실시하였다. K 원격대학 소속 교수에게 연구 목적, 시스템 사용 방법 및 온라인 설문 링크를 전자우편으로 안내하여 자발적 참여자를 모집하였다. 참여자는 설문에 응답하기 전에 시스템을 직접 사용하고, 그 경험을 바탕으로 문항에 응답하도록 하였다. 별도의 회원가입 없이 교수자용 및 학생용 데모 계정으로 시스템에 접속할 수 있도록 하였으며, 시스템 사용이 익숙하지 않은 참여자를 위해 사용 절차를 안내하는 온라인 튜토리얼 링크를 함께 제공하였다.

설문 문항은 기술수용모형(TAM)의 지각된 유용성과 지각된 사용편의성[32], 통합기술수용이론(UTAUT)의 행동의도[33]에 관한 기존 측정문항을 바탕으로 생성형 AI 기반 평가 지원 시스템의 사용 맥락에 적합하도록 재구성하였다. 행동의도는 본 연구의 맥락을 고려하여 '활용 의도'로 명명하였다. 설문은 지각된 유용성, 지각된 사용편의성 및 활용의도의 세 구성요인으로 구성하였으며, 각 구성요인은 3개 문항씩 총 9개 문항으로 측정하였다. 이와 함께 시스템에 대한 의견과 개선 요구를 수집하기 위한 개방형 문항 1개를 제시하였다. 구체적인 문항은 Table 4에 제시하였다.

설문 문항의 응답 자료는 요인별 평균과 표준편차를 산출하는 기술통계 방법으로 분석하였다. 개방형 문항의 응답은 전체 응답을 반복적으로 검토하여 응답별 핵심 의견을 요약하고, 유사한 내용을 통합하여 잠정 범주를 구성하였다. 이후 원자료와 범주 간의 부합 여부 및 범주 간 중복을 재검토하여 최종 범주를 확정하였다.

Table 4. Pilot Evaluation Items

Construct	Item	Measurement item
Perceived usefulness	PU1	Using the AI assessment system enables me to complete assessment tasks more quickly and efficiently.
	PU2	Using the AI assessment system improves the quality of assessment.
	PU3	Overall, I find the AI assessment system useful.
Perceived ease of use	PEOU1	I find the AI assessment system easy to learn.
	PEOU2	The functions of the AI assessment system are intuitive and easy to use.
	PEOU3	Learning and using the AI assessment system requires considerable effort.
Intention to use	IU1	I am willing to pilot the AI assessment system.
	IU2	I intend to use the AI assessment system if policy, technical, and educational support is provided.
	IU3	I intend to continue using the AI assessment system once it has been stabilized and validated.

Note. All items were measured on a five-point Likert scale. The survey was administered in Korean, and the items shown are English translations by the authors.

4. 연구 결과

4.1 생성형 AI 기반 평가 지원 시스템 개발

4.1.1 루브릭 기반 평가 체계 반영

첫 번째 설계 원리인 '루브릭 기반 평가 체계'는 '과제 생성' 단계에서 구현되었다. 교수자는 과제 생성 시 평가 항목과 채점 기준을 입력하도록 설계되었으며, 입력 방식은 수동 입력, 텍스트 입력, 파일 업로드 방식 중 선택할 수 있다. 입력된 루브릭은 AI 평가의 기준으로 활용되며, 모든 제출물이 교수자가 설정한 동일한 기준에 따라 평가되도록 한다. 이와 같은 구조는 생성형 AI가 임의적 기준이 아닌, 교수자가 설정한 평가 항목과 기준에 근거하여 평가를 수행하도록 함으로써, 인간 평가자의 판단 논리를 시스템 차원에서 재현하기 위한 장치이다. Figure 4는 과제 생성 단계에서 평가 항목과 기준을 텍스트 형태로 입력하는 화면을 보여준다.

Figure 4. Text Input Function for Scoring Criteria

4.1.2 평가 결과 설명 가능성 구현

두 번째 설계 원리인 '평가 결과 설명 가능성 확보'는 '자동 평가' 단계에서 구현되었다. 교수자는 AI 평가 기능을 활용하여 학생의 제출물을 개별 평가 또는 일괄 평가를 수행할

수 있으며, 모든 제출물은 동일한 평가 기준과 모델과 프롬프트에 따라 평가된다.

그러나 단순히 점수만을 제시하는 방식은 사용자에게 평가 결과를 '불투명한 판단'으로 인식하게 할 수 있으며, 이는 평가 결과에 대한 불신으로 이어질 수 있다. 이에 본 시스템은 AI가 산출한 점수와 함께, 각 점수가 루브릭의 어떤 평가 항목과 기준에 근거하여 도출되었는지를 설명하는 서술형 피드백을 동시에 생성하도록 설계되었다. 생성된 피드백은 평가 항목별 강점과 개선이 필요한 요소를 명시적으로 제시하며, 점수 산출의 판단 근거를 설명할 수 있도록 하였다.

Figure 5는 AI가 산출한 점수와 함께 루브릭의 평가 항목별 성취 수준, 점수 부여 근거 및 개선 사항을 서술형 피드백으로 제공하는 화면을 보여준다.

Figure 5. Provision of Explanation-Based Feedback

이러한 설명 기반 피드백은 교수자가 AI 평가 결과를 검토할 수 있는 근거 자료로 제공되어, 교수자의 평가를 보조하는 역할을 수행할 수 있다. 또한 학습자에게 자신의 성취 수준을 이해시키고, 향후 학습 전략을 조정하도록 지원할 수 있다.

4.1.3 교수자 개입형 평가 구조 구현

세 번째 설계 원리인 '교수자 개입형 평가 구조'는 '교수자 검토' 단계에 반영되었다. 본 시스템은 생성형 AI가 산출한 평가 결과를 교수자가 최종적으로 검토·수정할 수 있도록 설계하였다. 교수자는 학생 제출물을 확인한 후 AI가 산출한 점수를 직접 수정하여 저장할 수 있으며, 평가 결과에 대한 최종 판단 권한을 갖는다. 이러한 구조는 AI가 평가를 독립적으로 확정하는 것이 아니라, 교수자의 전문적 판단을 지원하는 보조 도구로 활용되도록 하기 위한 것이다. Figure 6은 교수자가 학생 제출물과 AI 평가 결과를 확인한 후 점수와 피드백을 검토하고 수정하여 최종 평가 결과를 확정하는 화면을 보여준다.

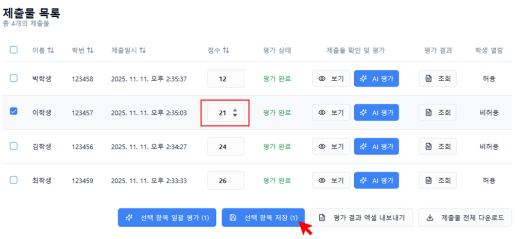

Figure 6. Review and Revision Function for Scoring Results

또한 평가 결과는 엑셀 파일로 일괄 내보내기가 가능하며, 교수자의 수정 여부가 함께 기록되도록 하여 평가 과정의 투명성과 책임성을 확보하였다. 이러한 구조는 생성형 AI 평가를 인간 판단의 대체물이 아닌 전문적 판단을 보조하는 도구로 활용되도록 하기 위한 것이다. 특히 평가 결과가 학점 부여와 직결되는 고등교육 맥락을 고려하여, 평가에 대한 교수자의 최종 결정권을 보장하고 평가 과정의 책임성과 교육적 정당성을 확보하고자 하였다.

4.2 평가 지표 분석

생성형 AI 평가 점수가 담당 교수의 평가 점수와 어느 정도 일치하는지를 검토하기 위해, 교수자가 부여한 점수와 시스템이 산출한 점수 간의 일치도와 상관관계를 분석하였다. 이를 위해 실제 채점 일치 정도를 반영하는 Quadratic Weighted Kappa(QWK), 점수 간의 선형적 관련성을 나타내는 Pearson 상관계수(PCC), 평균적인 점수 차이를 보여주는 평균절대오차(MAE), 그리고 점수 범위 차이를 고려해 오차를 표준화한 nMAE를 산출하였다. 분석 결과, 교과목 A의 과제물 300건에 대해 QWK는 .863, PCC는 .863, MAE는 1.167, nMAE는 .078로 나타났으며, 교과목 B의 경우 QWK는 .800, PCC는 .803, MAE는 .567, nMAE는 .019로 나타났다.

Table 5. Summary of Assessment Performance Metrics

Metric	Course A	Course B
QWK	.863	.800
PCC	.863	.803
MAE	1.167	.567
nMAE	.078	.019

AES 성능 평가에서 일반적으로 QWK .70 이상을 수용 기준으로 삼는 점을 고려할 때[34], 시스템 평가 점수와 담당 교수 점수 간에 비교적 높은 수준의 일치도가 나타났음을 보여준다. PCC 역시 .80 이상이면 높은 수준의 선형 상관 관계로 해석되므로[35, 36], 본 시스템의 평가 점수와 담당 교수 점수의 전반적인 분포도 대체로 유사한 경향을 보였다.

개별 학습자 수준의 점수 차이를 확인하기 위해 절대오차를 추가로 분석한 결과, 교과목 A에서는 전체 사례의 86.7%, 교과목 B에서는 95.3%가 2점 이내의 차이를 보였다.

Table 6. Individual-Level Score Difference Analysis

Metric	Course A	Course B
Number of Cases	300	150
Exact Score Agreement	113 (37.7%)	97 (64.7%)
Absolute Error Within 1 Point	212 (70.7%)	127 (84.7%)
Absolute Error Within 2 Points	260 (86.7%)	143 (95.3%)
Absolute Error Greater Than 2 Points	40 (13.3%)	7 (4.7%)
Mean Score Difference	+0.14	−0.13

정량 지표 분석과 함께 교수자 평가와 AI 평가 간의 점수 관계를 시각적으로 확인하기 위해 산점도, 점수 분포 그래프를 추가로 검토하였다. 산점도 분석 결과, 교수자 점수와 AI 점수는 전반적으로 강한 정적 선형 관계를 보였으며, 이는 앞서 제시한 QWK와 PCC 결과와 일관된다. 또한 점수 분포 분석 결과, AI 평가 점수의 전체 분포는 교수자 평가 점수의 분포와 대체로 유사한 형태를 보였다. 이상의 결과는 개발된 시스템의 평가 점수가 전체적으로 교수자 점수와 유사한 경향을 보이고 있음을 나타낸다.

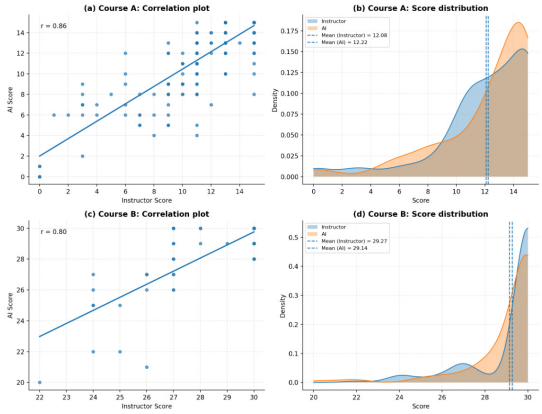

Figure 7. Scatter and Score Distribution Plots

4.3. 파일럿 사용자 평가 결과

4.3.1 교수자 설문 응답 결과

설문 결과, 활용 의도(M = 4.18, SD = .97), 지각된 유용성(M = 3.96, SD = .79)은 비교적 높게 나타났으며, 사용 편의성(M = 3.78, SD = 0.85) 역시 중간 이상 수준으로 확인되었다. 이는 교수자들이 본 시스템의 활용 가능성을 전반적으로 긍정적으로 인식했음을 시사한다.

Table 7. Descriptive Statistics by Construct

Construct	M	SD	Min	Max	α
Perceived usefulness	3.96	.79	1.67	5.00	.87
Ease of use	3.78	.85	1.67	5.00	.87
Intention to use	4.18	.97	1.00	5.00	.96

Note. α = Cronbach's alpha.

다만 본 파일럿 사용자 평가는 표본 수가 크지 않아, 기술통계 중심의 탐색적 분석에 머물러 있으므로, 결과를 일반화하기엔 한계가 있다. 따라서 본 연구의 설문 결과는 생성형 AI 평가 시스템에 대한 초기 수용 가능성과 기대 수준을 보여주는 자료로 해석하는 것이 타당하다.

4.3.2 교수자 자유 응답 결과

파일럿 사용자 평가의 자유 응답 내용을 검토한 결과, 교수자들의 의견은 대체로 전반적인 긍정 인식과 함께, 실제 활용을 위한 기능 개선 요구가 병존하는 양상으로 나타났다. 이러한 결과는 앞서 제시한 양적 설문 결과와 마찬가지로, 교수자들이 생성형 AI 평가 시스템의 활용 가능성을 긍정적으로 인식하면서도, 실제 교육 현장에 적용하기 위해서는 보다 개인화된 활용 가능성과 구체적인 설계 보완이 필요하다고 인식하고 있음을 보여준다.

첫째, 교수자들은 AI 평가 시스템 도입의 필요성과 향후 활용 가능성에 대해 대체로 긍정적인 의견을 나타냈다. 예를 들어, 일부 응답자는 "AI 사용이 점점 일상화되고 있어 평가에도 반영하는 것은 불가피하다", "향후 시스템 활용에 대하여 긍정적이다"와 같은 의견을 제시하였다. 이러한 응답은 생성형 AI 기반 평가 시스템이 미래의 교수·학습 환경에서 일정한 역할을 수행할 수 있다는 기대가 존재함을 보여준다.

둘째, 교수자들은 입력 및 출력 방식의 조정 필요성을 제기하였다. 일부 응답자는 AI가 생성하는 피드백이 지나치게 상세하게 제시될 경우 오히려 실제 수업 맥락에서 활용하기에 부담이 될 수 있다는 점을 지적하였다. 또한 프롬프트의 자율성 필요, 피드백 출력 방식의 조정 가능성 등이 요구되었다. 이러한 의견은 생성형 AI 평가 시스템이 단순히 자동화된 결과를 산출하는 도구에 그치기보다, 교수자가 자신의 교과목 특성과 평가 의도에 맞게 입력과 출력 방식을 유연하게 조정할 수 있는 사용자 중심 시스템으로 발전할 필요가 있음을 시사한다.

셋째, 시스템 확장성과 기능 고도화에 대한 요구도 확인되었다. 응답자들은 생성형 AI 평가 시스템이 특정 기능에 한정된 도구로 머무르기보다, 다양한 사용자 요구와 수업 상황에 맞추어 지속적으로 확장 가능한 구조를 갖출 필요가 있다고 보았다. 구체적으로는 채점 결과를 바탕으로 평가 기준을 수정·보완한 후 다시 적용할 수 있는 기능, 이미지 기반 과제물에 대한 평가 지원, 그리고 다양한 제출 형식에 대한 지원 요구가 제시되었다. 이러한 의견은 향후 생성형 AI 평가 시스템이 실제 원격대학 수업 환경에서 활용되기 위해서는 다양한 과제 유형과 교수자의 운영 방식에 대응할 수 있는 기능적 유연성이 확보되어야 함을 시사한다.

5. 결론 및 시사점

5.1 결론

본 연구는 원격대학 환경에서 활용할 수 있는 생성형 AI 기반 서술형 평가 지원 시스템을 설계·구현하고, 실제 과제 데이터를 활용하여 시스템의 평가 성능과 교수자 수용 가능성을 검토하는 데 목적이 있었다. 이를 위해 선행연구 분석과 전문가 자문을 바탕으로 시스템의 설계 원리를 도출하고, 웹 기반 평가 지원 시스템을 개발한 후, 실제 원격대학 운영자료를 활용한 후향적 성능평가와 교수자 대상 파일럿 사용자 평가를 실시하였다.

선행연구 분석과 전문가 자문을 통해 루브릭 기반 평가 체계, 평가 결과의 설명 가능성 및 교수자 개입형 평가 구조를 핵심 설계 원리로 도출하였다. 이를 바탕으로 교수자가 과제와 루브릭을 등록하고, 학습자가 과제를 제출하면 AI가 점수와 피드백을 생성하며, 교수자가 AI 평가 결과를 검토·수정·확정할 수 있는 웹 기반 시스템을 구현하였다. 특히 AI가 산출한 결과를 자동으로 최종 평가 결과로 확정하지 않고, 교수자가 최종 판단과 평가 책임을 유지하도록 설계하였다.

성능평가 결과, 교과목 A에서는 QWK=.863, PCC=.863, MAE=1.167, nMAE=.078로 나타났으며, 교과목 B에서는 QWK=.800, PCC=.803, MAE=.567, nMAE=.019로 나타났다. 두 교과목 모두 QWK와 PCC가 .80 이상으로 나타나, 개발된 시스템의 점수가 담당 교수의 점수와 비교적 높은 수준의 일치도와 선형적 관련성을 보이는 것으로 확인되었다.

K 원격대학 소속 교수 32명을 대상으로 실시한 파일럿 사용자 평가에서는 활용 의도(M=4.18), 지각된 유용성(M=3.96), 지각된 사용편의성(M=3.78)이 전반적으로 긍정적인 수준으로 나타났다. 이는 개발된 시스템이 교수자의 평가 업무와 피드백 제공을 지원하는 도구로 수용될 가능성이 있음을 보여준다. 그러나 파일럿 평가는 제한된 수의 자발적 참여자를 대상으로 시스템 사용 직후의 초기 인식을 조사한 것이므로, 이를 시스템의 장기적 효과나 실제 활용행동으로 일반화하는 데에는 주의가 필요하다.

본 연구의 기여는 특정 생성형 AI 모델의 채점 성능을 확인한 데 한정되지 않는다. 첫째, 기존의 모델 또는 프롬프트 중심 자동채점 연구와 달리, 생성형 AI 평가에 요구되는 교육적·책임적 조건을 루브릭 기반 평가, 판단 근거의 설명 가능성 및 교수자 개입형 평가 구조라는 설계 원리로 구체화하고 이를 실제 시스템 기능으로 구현하였다. 둘째, 교수자 점수를 AI 성능 검증을 위한 준거로만 활용하는 데 그치지 않고, 교수자가 AI가 생성한 점수와 피드백을 검토·수정·확정하여 최종 판단과 책임을 유지하는 human-in-the-loop 평가 절차를 구현하였다. 셋째, 공개 데이터가 아닌 원격대학 두 교과목의 실제 과제물, 루브릭 및 교수자 점수를 활용한 성능평가와 교수자 대상 파일럿 사용자 평가를 병행함으로써, 생성형 AI의 채점 가능성을 넘어 실제 평가 지원 시스템으로서의 적용 조건을 검토하였다. 이러한 점에서 본 연구는 생성형 AI 기반 서술형 평가 연구의 범위를 모델 수준의 성능 검증에서 평가 시스템의 설계와 운영으로 확장하였다는 의의를 갖는다.

5.2 시사점

본 연구의 결과는 생성형 AI 기반 서술형 평가 시스템의 설계와 활용에 다음과 같은 시사점을 제공한다.

첫째, 생성형 AI 기반 평가의 성능은 모델 자체의 능력뿐 아니라 평가 기준의 구조화 방식에 영향을 받는다. 본 연구에서 적용한 루브릭 기반 평가 구조는 AI가 교수자가 설정한 평가 기준을 바탕으로 점수와 피드백을 산출할 수 있도록 하였다. 이는 생성형 AI를 교육평가에 활용하기 위해서는 프롬프트 수준의 단순 지시를 넘어, 평가 기준을 명시적으로 구조화하고 시스템 기능과 연계하는 설계가 필요함을 시사한다.

둘째, 생성형 AI 기반 평가 시스템은 교수자의 최종 판단을 대체하기보다 지원하는 방향으로 설계될 필요가 있다. 본 연구의 시스템은 AI가 산출한 평가 결과를 교수자가 검토·수정·확정할 수 있도록 구현하였다. 이는 고등교육 평가에서 책임성과 공정성을 확보하기 위해 교수자 개입형 평가 구조가 중요함을 보여준다. 따라서 생성형 AI 평가 시스템은 자동화 수준을 높이는 것뿐 아니라, 교수자가 AI 평가 결과를 해석하고 통제할 수 있는 기능을 함께 제공해야 한다.

셋째, 생성형 AI 기반 평가 시스템의 현장 활용을 위해서는 사용자 중심의 기능 개선이 필요하다. 파일럿 사용자 평가 결과는 교수자들이 시스템의 활용 가능성을 긍정적으로 인식하고 있음을 보여주었으나, 실제 수업 맥락에서 안정적으로 활용되기 위해서는 교과목별 루브릭 조정, 평가 결과 검토 편의성, 피드백 수정 기능, LMS 연계 등 추가적인 개선이 요구된다. 이는 생성형 AI 기반 평가 시스템의 확산이 기술 개발만으로 이루어지기 어렵고, 교수자의 평가 실천을 지원하는 운영 환경과 함께 설계되어야 함을 시사한다.

5.3 연구의 한계 및 제언

본 연구는 생성형 AI 기반 서술형 평가 지원 시스템의 설계와 적용 가능성을 실제 교육 맥락에서 검토하였으나, 다음과 같은 한계를 지닌다.

첫째, 본 연구는 단일 기관의 두 개 교과목을 대상으로 시스템을 적용하였기 때문에 연구 결과를 다양한 고등교육 맥락으로 일반화하는 데에는 제한이 있다. 후속 연구에서는 서로 다른 학문 분야, 과제 유형, 평가 기준을 포함한 다기관 연구를 통해 시스템의 적용 가능성과 성능의 안정성을 추가로 검토할 필요가 있다.

둘째, 본 연구는 특정 시점의 Gemini 2.5 Flash 모델을 기반으로 시스템을 구현하였다. 생성형 AI 모델은 지속적으로 업데이트되므로, 모델 버전 변화에 따라 평가 결과가 달라질 가능성이 있다. 후속 연구에서는 다양한 생성형 AI 모델과 버전에서 동일한 설계 원리가 안정적으로 작동하는지

비교 분석할 필요가 있다.

셋째, 분석 대상 교과목의 전문성과 전공 교수 인력의 제한으로 인해 담당 교수 1인이 부여한 점수를 준거 점수로 활용하였다. 이에 따라 교수자 평가에 포함될 수 있는 개인적 판단의 영향을 충분히 통제하지 못하였으며, 평가자 간 신뢰도를 검증하지 못했다는 한계가 있다. 후속 연구에서는 해당 전공에 대한 전문성을 갖춘 복수의 평가자를 확보하고, 평가자 간 일치도와 합의 과정을 바탕으로 준거 점수를 설정함으로써 AI 평가 성능을 더욱 엄밀하게 검증할 필요가 있다.

넷째, 본 연구는 AI 평가 성능과 교수자 인식을 중심으로 분석하였으며, 학습자 관점에서의 피드백 활용 효과는 직접 검증하지 못하였다. 향후 연구에서는 AI 기반 피드백이 학습자의 과제 수정, 자기조절학습, 학업 성취에 미치는 영향을 분석함으로써 생성형 AI 평가 시스템의 교육적 효과를 보다 종합적으로 검토할 필요가 있다.

이러한 한계에도 불구하고 본 연구는 생성형 AI 기반 서술형 평가 지원 시스템을 실제 원격 고등교육 맥락에서 설계·구현하고, 실제 학생 과제물을 활용하여 AI 평가 성능과 교수자 수용 가능성을 함께 검토하였다는 점에서 의의가 있다. 특히 본 연구는 생성형 AI 기술을 단순한 자동 채점 도구로 활용하는 데 그치지 않고, 루브릭 기반 평가, 설명 가능한 피드백, 교수자 검토·수정 절차를 포함한 인간 개입형 평가 지원 시스템으로 구현하였다는 점에서 AI 융합 교육 분야에 시사점을 제공한다. 본 연구의 결과는 향후 생성형 AI 기반 평가 지원 시스템을 설계·개발하고, 교수자의 전문적 판단과 평가 책임을 유지하면서 이를 단계적으로 도입하기 위한 기초자료로 활용될 수 있을 것이다.

참고문헌

[1] Academyinfo. (2025, August 7). *Disclosure information of Korea National Open University.* Retrieved October 31, 2025, from https://www.academyinfo.go.kr/pubinfo/pubinfo0020/list.do?schlId=0000273

[2] Golchin, S., Garuda, N., Impey, C., & Wenger, M. (2025, January). Grading massive open online courses using large language models. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. Di Eugenio, & S. Schockaert (Eds.), *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 3899–3912). Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.263/

[3] Najafi, H., Rolheiser, C., Harrison, L., & Håklev, S. (2015). University of Toronto instructors' experiences with developing MOOCs. *International Review of Research in Open and Distributed Learning, 16*(3), 233–255. https://doi.org/10.19173/irrodl.v16i3.2073

[4] Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence, 6,* Article 100234. https://doi.org/10.1016/j.caeai.2024.100234

[5] Choi, I., Park, S., & Lee, Y. (2025). Development of an essay-type assessment for teaching aptitude and personality using generative AI-based automated scoring. *Korean Journal of Educational Research, 63*(6), 215–250. https://doi.org/10.30916/KERA.63.6.215

[6] Emirtekin, E. (2025). Large language model-powered automated assessment: A systematic review. *Applied Sciences, 15*(10), Article 5683. https://doi.org/10.3390/app15105683

[7] Meyer, J., Jansen, T., Schiller, R., Liebenow, L., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence, 6,* Article 100199. https://doi.org/10.1016/j.caeai.2023.100199

[8] Stahl, M., Biermann, L., Nehring, A., & Wachsmuth, H. (2024). *Exploring LLM prompting strategies for joint essay scoring and feedback generation.* arXiv. https://doi.org/10.48550/arXiv.2404.15845

[9] Tang, X., Chen, H., Lin, D., & Li, K. (2024). Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments. *Heliyon, 10*(14), Article e34262. https://doi.org/10.1016/j.heliyon.2024.e34262

[10] Ahn, H. (2025). Exploring the potential of generative AI in essay-based assessments: Evidence from the Regents Exam data. *Journal of Educational Evaluation, 38*(3), 823–846. https://doi.org/10.31158/JEEV.2025.38.3.823

[11] Li, H., Chen, C., Fan, K., Young-Johnson, C., Lim, S., & Feng, Y. (2025). *Agreement between large language models and human raters in essay scoring: A research synthesis.* arXiv. https://doi.org/10.48550/arXiv.2512.14561

[12] 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), Article 100050. https://doi.org/10.1016/j.rmal.2023.100050

[13] Kim, J., & Jo, I. (2025). A qualitative comparison of human and ChatGPT assessments in undergraduate argumentative writing. *Journal of Educational Technology, 41*(3), 921–958. https://doi.org/10.17232/KSET.41.3.921

[14] Richey, R., & Klein, J. (2007). Design and development research: *Methods, strategies, and issues.* Routledge. https://doi.org/10.4324/9780203826034

[15] Gaytan, J., & McEwen, B. (2007). Effective online instructional and assessment strategies. *American Journal of Distance Education, 21*(3), 117–132. https://doi.org/10.1080/08923640701341653

[16] Swan, K., Shen, J., & Hiltz, S. (2006). Assessment and collaboration in online learning. *Journal of Asynchronous Learning Networks, 10*(1), 45–62. https://doi.org/10.24059/olj.v10i1.1770

[17] Balfour, S. (2013). Assessing writing in MOOCs:

Automated essay scoring and calibrated peer review. *Research & Practice in Assessment, 8*, 40–48. https://eric.ed.gov/?id=EJ1062843

[18] Sorensen, C. (2015). An examination of the relationship between online class size and instructor performance. *Journal of Educators Online, 12*(1), 140–159. https://doi.org/10.9743/JEO.2015.1.3

[19] Kulkarni, C., Bernstein, M., & Klemmer, S. (2015). PeerStudio: Rapid peer feedback emphasizes revision and improves performance. *In Proceedings of the Second (2015) ACM Conference on Learning @ Scale* (pp. 75–84). Association for Computing Machinery. https://doi.org/10.1145/2724660.2724670

[20] Shermis, M., & Burstein, J. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410606860

[21] Bui, N., & Barrot, J. (2025). ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. *Education and Information Technologies, 30*, 2041–2058. https://doi.org/10.1007/s10639-024-12891-w

[22] Zubaer, A., Granitzer, M., Geschwind, S., Graf Lambsdorff, J., & Voss, D. (2025). GPT-4 shows comparable performance to human examiners in ranking open-text answers. *Scientific Reports, 15*, Article 35045. https://doi.org/10.1038/s41598-025-21572-8

[23] Ramineni, C., Trapani, C., & Williamson, D. (2015). *Evaluation of e-rater® for the Praxis I® Writing Test* (ETS Research Report No. RR-15-03). Educational Testing Service. https://doi.org/10.1002/ets2.12047

[24] Yang, K., Raković, M., Li, Y., Guan, Q., Gašević, D., & Chen, G. (2024). Unveiling the tapestry of automated essay scoring: A comprehensive investigation of accuracy, fairness, and generalizability. *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(20), 22466–22474. https://doi.org/10.1609/aaai.v38i20.30254

[25] Yancey, K., Laflair, G., Verardi, A., & Burstein, J. (2023). Rating short L2 essays on the CEFR scale with GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 576–584). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.49

[26] Khosravi, H., Buckingham Shum, S., Chen, G., Conati, C., Tsai, Y., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence, 3*, Article 100074. https://doi.org/10.1016/j.caeai.2022.100074

[27] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[28] 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://www.ed.gov/sites/ed/files/documents/ai-report/ai-report.pdf

[29] UNESCO. (2023). *Guidance for generative AI in education and research*. UNESCO. https://doi.org/10.54675/EWZM9535

[30] Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *The Journal of Technology, Learning and Assessment, 4*(3). https://ejournals.bc.edu/index.php/jtla/article/view/1650

[31] Memarian, B., & Doleck, T. (2024). Human-in-the-loop in artificial intelligence in education: A review and entity-relationship (ER) analysis. *Computers in Human Behavior: Artificial Humans, 2*(1), Article 100053. https://doi.org/10.1016/j.chbah.2024.100053

[32] Davis, F. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340. https://doi.org/10.2307/249008

[33] Venkatesh, V., Morris, M., Davis, G., & Davis, F. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425–478. https://doi.org/10.2307/30036540

[34] Williamson, D., Xi, X., & Breyer, F. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

[35] Mukaka, M. (2012). Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal, 24*(3), 69–71. https://doi.org/10.4314/mmj.v24i3

[36] Schober, P., Boer, C., & Schwarte, L. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia, 126*(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

안해연

· 2021년 서울특별시교육연구정보원 교육연구사
· 2024년 이화여자대학교 교육학과(문학박사)
· 2026년 경희대학교 학술연구교수
· 2025년~현재 이화여자대학교 초빙교수
· 2026년~현재 국가교육위원회 사무관

관심분야 : AIED, 빅데이터 분석, 교육정책
haeyeonahn@korea.kr

양유정

· 1997년 연세대학교 교육학과(문학사)
· 2000년 연세대학교 대학원 교육공학전공(교육학석사)
· 2013년 연세대학교 대학원 교육공학전공(교육학박사)
· 2016년~현재 한국방송통신대학교 원격교육혁신연구원 선임연구위원

관심분야 : AIED, 원격고등교육, 성인학습자
yjyang2016@mail.knou.ac.kr