

생성형 AI 자가진단의 역할 범위

공간 디자인 실기 교과목 적용 사례에 관한 실행연구

Scope and Role of Generative AI Self-Diagnosis

An Action Research on a Spatial Design Studio Course

주 저 자 : 나상귀 (Na, Sang Kwee) 계원예술대학교 공간연출과 조교수
nask@kaywon.ac.kr

<https://doi.org/10.46248/kidrs.2026.3.707>

접수일 2026. 08. 20. / 심사완료일 2026. 08. 28. / 게재확정일 2026. 09. 07. / 게재일 2026. 09. 30.

Abstract

This action research examines the scope of the role that generative AI self-diagnosis can play in relation to instructor feedback in a spatial design studio course. A six-stage self-diagnostic system, built by a non-developer educator, was applied to 28 students over one semester, and its outputs were compared against expert evaluations and student perceptions. The AI-generated final scores clustered within only the top two intervals on a six-point scale and failed to discriminate differences among students, whereas the two experts assessing the same works showed a strong positive correlation. Because the AI and the experts drew on different evidence and evaluation items, this comparison is exploratory rather than a test of agreement. The diagnostic texts revealed cases in which full marks and defect remarks were output within the same response, and none of the improvement points in the visual stages addressed formal quality—an outcome inseparable from an evaluation design that excluded judgments of taste and style. Correspondingly, 20 of 23 students reported that instructor feedback was more helpful for visual and sensory judgment. The system functioned not as a substitute for instructor feedback but as a complement preceding it.

Keyword

Generative AI(생성형 AI), AI Assessment(AI 판정), Self-Diagnostic System(자가진단 시스템), Spatial Design Education(공간 디자인 교육), Action Research(실행연구)

요약

본 연구는 공간 디자인 실기 교과목에서 생성형 AI 자가진단이 교수자 피드백에 대해 갖는 역할 범위를 검토한 실행 연구이다. 비개발자 교육자가 구축한 6단계 자가진단 시스템을 한 학기 동안 수강생 28명에게 적용하고, AI 진단 로그와 전문가 평가, 학생 인식 설문을 대조하였다. AI가 산출한 최종 점수는 6점 척도에서 상위 두 구간에만 분포 하여 학생 간 차이를 변별하지 못한 반면, 같은 대상을 평가한 두 전문가 사이에서는 높은 정적 상관성이 확인되었다. 다만 AI와 전문가는 판단 근거와 평가 항목이 달라, 이 대조는 일치도 검증이 아닌 탐색적 비교에 해당한다. 진단 서술에서는 만점과 결함 지적이 한 응답에 함께 출력되었고, 시각 단계의 개선 사항에 조형적 완성도 지적은 없었다. 이는 조형·미감 판단을 개입 대상에서 제외한 평가 기준 설계와 분리하기 어렵다. 학생 역시 시각·감각적 판단에서는 23명 중 20명이 지도교수 피드백이 더 도움이 되었다고 응답하였다. 본 시스템은 교수자 피드백의 대체재가 아니라 그 이전 단계의 보완재로 기능하였다.

목차

1. 서론

- 1-1. 연구 배경 및 목적
- 1-2. 연구 방법 및 구성

2. 이론적 배경

- 2-1. 실기 교육에서의 피드백과 크리틱
- 2-2. AI 교육 피드백의 가능성과 한계
- 2-3. 본 연구의 위치

3. 시스템 설계 및 운영

- 3-1. 설계 배경과 진단 구조
- 3-2. 진단 프롬프트와 판정 구조
- 3-3. 운영 중 조정

4. 시스템 적용 결과

- 4-1. 시스템 활용 양상
- 4-2. AI 진단 결과와 전문가 평가의 대조
- 4-3. AI 진단 서술의 질적 분석

1. 서론

1-1. 연구 배경 및 목적

생성형 AI의 발전은 교육 현장에서 학습 지원과 평가 보조의 새로운 가능성을 열어주고 있다. 특히 최근의 멀티모달 모델은 텍스트뿐만 아니라 이미지를 해석하여 피드백을 산출하는 수준으로 발전하였다. Mollick과 Mollick은 생성형 AI가 교육자를 도구 사용자에서 학습 경험의 설계 주체로 이동시키고 있다고 보았다.¹⁾ 그러나 이러한 논의는 대체로 언어·개념 중심의 학습 활동을 전제하며, 시각적·감각적 판단이 핵심이 되는 실기 중심 예술 디자인 교육에서 생성형 AI가 어떤 역할을 수행할 수 있는지는 상대적으로 검증되지 않은 영역으로 남아 있다.

본 연구는 이러한 문제의식 아래, 비개발자 교육자가 생성형 AI와의 협업으로 교육 운영 시스템을 구축한 선행 연구를 계승하는 후속 작업으로 기획되었다.²⁾ 앞선 연구는 예술 디자인 학과의 졸업작품 심사 시스템을 설계하고 운영한 실험연구로서, 비개발자 교육자도 자신의 운영 맥락에 맞는 시스템을 직접 만들 수 있음을 확인하였다. 다만 그 논의는 시스템을 구축하고 운영하는 층위에 머물렀으며, 시스템이 산출하는 AI 진단 판정의 타당성은 검증 대상으로 다루어지지 않았다.

이에 본 연구는 동일한 협업 방식으로 개발된 ‘과제 AI 자가진단 시스템’의 적용 결과를 분석 대상으로 삼는다. 이 시스템은 학생이 과제 진행 단계마다 결과물을 제출하면 생성형 AI가 완료·보완·누락 여부를 진단하고 피드백을 제공하도록 설계되었으며, 과제 초기 분석부터 최종 종합 심사까지 총 6단계로 이어진다. 한 학기 동안 공간 디자인 실기 교과목(이하 A교과목) 수

강생 28명에게 적용되었으며, 그 과정에서 AI 진단 로그, 전문가 평가, 학생 인식 설문이라는 세 층위의 자료가 확보되었다.

생성형 AI를 교육 피드백에 활용하려는 시도는 디자인·미술교육과 일반 교육 영역 모두에서 축적되어 왔으나³⁾⁴⁾, 이들 연구는 AI가 산출한 판정이 인간 평가자의 질적 판단과 얼마나 부합하는지를 직접 대조하지는 않았다. 한편 시각·언어모델의 디자인 미학 판단 능력을 평가한 연구는 정교화된 모델조차 인간의 심미적 판단과 뚜렷한 격차를 보이며, 이 격차가 추론 능력을 강화해도 좁혀지지 않는다고 보고하였고⁵⁾, 디자인 크리틱 연구는 여기에 더해 피드백이 오가는 매체 자체가 크리틱의 질에 영향을 줄 수 있음을 보였다.⁶⁾ 이러한 논의는 생성형 AI가 언어·개념적 판단과 시각·감각적 판단에서 서로 다른 수준의 타당성을 가질 가능성을 시사한다.

따라서 본 연구는 2026년 현재의 생성형 AI가 공간 디자인 실기 교과목에서 어느 범위까지 교수자의 역할을 수행할 수 있는지를 검토한다. 이를 위해 동일한 결과물을 대상으로 지도교수와 외부 전문가가 내린 평가와 대조하여 항목별 완료 여부를 판정하는 AI의 진단이 학생 간 완성도의 차이를 변별해 내는지를 확

1) Mollick, E. & Mollick, L., ‘Instructors as Innovators: A Future-focused Approach to New AI Learning Opportunities, With Prompts’, The Wharton School, The Wharton School Research Paper (SSRN 4802463), 2024. 04. pp.1-3.

2) 나상귀, ‘예술·디자인 학과 졸업작품 심사를 위한 생성형 AI 협업 기반 시스템 구축에 관한 실험연구 — 비개발자 교육자 사례를 중심으로’, 한국디자인리서치학회, 디자인리서치, Vol.11, No.2, 2026. 06. pp.746-755.

3) 장윤경·윤인화·박지원·제갈유진·이지연, ‘이미지 생성 AI 도구의 이해와 미술교육의 활용 가능성 탐구’, 한국조형교육학회, 조형교육, Vol.88, 2023. 12. pp.277-298.

4) 김동삼·류다현·박규동·강혜진, ‘AI기반 실시간 온라인 교육 맞춤형 피드백 시스템 효과성 평가’, 한국컴퓨터교육학회, 컴퓨터교육학회 논문지, Vol.27, No.7, 2024. 10. pp.11-23.

5) An, A., Sun, S., Huang, D., Cheng, M., Gao, Y., Li, J., Qiao, Y. & Bian, J., ‘Can Vision-Language Models Assess Graphic Design Aesthetics? A Benchmark, Evaluation, and Dataset Perspective’, ICLR 2026 (International Conference on Learning Representations), 2026. 03.

6) 안지산·박주령, ‘디자인 크리틱의 교육심리학적 고찰과 사용자경험 가이드라인’, 한국브랜드디자인학회, 브랜드디자인학연구, Vol.19, No.2, 2021. 06. pp.159-172.

인한다. 여기에 AI가 판정 근거로 제시한 진단 서술과 학습자의 수용 양상을 더하여 언어·개념적 판단과 시각·감각적 판단 사이에서 AI 자가진단 결과의 유효성이 어떻게 갈라지는지를 검토한다.

1-2. 연구 방법 및 구성

본 연구는 연구자가 시스템의 개발자이자 운영자로 현장에 참여하여 문제를 인식하고 개입을 실행한 뒤, 그 결과를 성찰적으로 분석한다는 점에서 선행 연구와 마찬가지로 실행연구(Action Research)의 접근을 취한다.⁷⁾ 선행 연구가 시스템을 구축하고 정비하는 사이클 자체를 분석 대상으로 삼았다면, 본 연구는 그렇게 구축된 시스템이 산출한 판단의 역할 범위를 검토하는 데 초점을 둔다는 점에서 후속 사이클에 해당한다.

연구의 범위는 2026년 A교과목 수강생 28명의 운영 데이터로 한정하며, 연구 자료는 세 가지로 구성된다. 첫째, 시스템에 기록된 학생별 AI 진단 로그로 단계별 시도 횟수와 점수, 완료·보완·누락 판정 결과 및 진단 서술이 포함된다. 둘째, 지도교수(연구자 본인) 1인과 외부 전문가 1인이 최종 발표 현장에서 AI 결과를 열람하지 않은 상태로 실시한 14개 항목의 평가 자료이다. 셋째, 학기 종료 약 1개월 후 시행한 학생 인식 설문(응답 25명)으로 AI 피드백에 대한 신뢰도와 실제 반영 여부, 지도교수 피드백과의 비교 인식 등을 다룬다. 설문은 성적이 확정된 이후 익명으로 수집하였으며, 연구 목적과 자료 활용 범위, 참여 여부가 성적과 무관하고, 미참여에 따른 불이익이 없음을 고지한 뒤, 이에 동의한 응답자만 응답하도록 구성하였다. 자가진단 점수는 성적 산출에 반영하지 않았다. 학생 개인을 식별할 수 있는 정보는 분석에 사용하지 않았다.

분석은 기술통계와 질적 코딩을 병행하여 수행한다. AI 진단 점수의 분포와 전문가 평가 점수를 대조하여 AI 판정의 변별력을 검토하고, AI가 산출한 진단 서술과 학생 설문의 서술형 응답을 범주화하여 이러한 결과가 나타나게 된 기제를 확인한다. 다만 본 연구의 소표본 조건을 고려하여, 상관계수는 참고치로 제시하며 이를 근거로 한 인과적 해석이나 표본을 넘어서는 일반화는 시도하지 않는다.

본 논문은 제2장에서 실기 교육의 크리틱과 AI 피드백에 관한 선행 연구를 검토하고, 제3장에서 자가진단 시스템의 설계와 운영 중 조정을, 제4장에서 적용

결과를 다룬 뒤, 제5장에서 연구 결과를 종합하여 의의와 한계, 향후 과제를 제안하는 순으로 구성된다.

2. 이론적 배경

2-1. 실기 교육에서의 피드백과 크리틱

실기 중심 예술디자인 교육에서 크리틱(critique)은 교수자와 학습자, 혹은 학습자 상호간에 지속적인 피드백이 오가는 교수 방식으로 폭넓게 활용되어 왔다. 크리틱은 참여자 구성과 진행 방식에 따라 여러 유형으로 나뉜다. 국내 디자인전공 학생 437명을 대상으로 네 가지 크리틱 유형을 비교한 연구에서는 비대면·온라인 형태로 진행된 크리틱일수록 학습자가 체감하는 목표의 명확성과 과제 집중도, 학습자 간 상호작용 정도가 유의하게 낮아지는 것으로 나타났다.⁸⁾

이 결과는 크리틱의 질이 피드백 제공자의 전문성뿐 아니라, 피드백이 오가는 매체와 상호작용 방식에 의해서도 좌우될 수 있음을 시사한다. 본 연구가 다루는 AI 자가진단 시스템은 텍스트 기반의 비동기적 온라인 피드백을 제공한다는 점에서 시각·감각적 판단 영역에서 구조적으로 불리한 조건을 안고 있다고 볼 수 있다.

2-2. AI 교육 피드백의 가능성과 한계

생성형 AI가 교육자의 역할을 확장할 수 있다는 논의는 여러 방향에서 전개되어 왔다. Mollick과 Mollick은 시뮬레이션, 멘토링, 코칭 등 프롬프트 기반 활용을 검토하며, 교원의 내용 전문성과 교수법 전문성이 새로운 교육기술 설계의 기반이 될 수 있다고 보았다.⁹⁾ 미술교육 영역에서도 이미지 생성 AI 도구가 학습자의 표현 활동을 보조할 가능성이 탐색된 바 있다.¹⁰⁾

한편 실제 교육 현장에 도입된 AI 피드백 시스템의 작동 방식을 살펴보면, 그 피드백의 성격이 대체로 정량적 지표에 근거하고 있음을 확인할 수 있다. 선행 연구에서는 웹캠 기반 시선추적 데이터로 학습자의 참여 수준을 판정하고 일정 임계값을 기준으로 자동화된 피드백을 제공하는 시스템이 개발되어, 기존 화상회의 플랫폼보다 높은 사용성 평가를 받았다.¹¹⁾ 다만 이 시스

7) McNiff, J., 『Action Research: All You Need to Know』, SAGE Publications, 2017, pp.9-12.

8) 안지산·박주영, Op.cit., 2021, pp.159-172.

9) Mollick, E. & Mollick, L., Op.cit., 2024, pp.1-3.

10) 장윤경 외, Op.cit., 2023, pp.277-298.

11) 김동심 외, Op.cit., 2024, pp.11-23.

템이 다루는 시선 체류 시간과 같은 지표는 측정 단계에서 이미 수치의 형태를 갖추고 있어, 판정 기준을 임계값으로 환원하는 데 무리가 없다. 반면 본 연구가 다루는 기획의 완성도나 시각적 결과물의 설득력은 그러한 환원이 전제되지 않은 대상이다.

생성형 모델의 판정 자체를 인간 평가와 대조한 연구는 언어 산출물 영역에서 먼저 이루어졌다. GPT로 대규모 영어 에세이를 채점해 인간 기준 점수와 대조한 연구는 일정 수준의 정확도를 확인하면서도 완전한 일치에는 이르지 못하여, LLM 채점을 인간 평가의 대체가 아닌 보조 도구로 규정하였다.¹²⁾ 실제로 이 연구에서도 모델 점수는 좁은 구간에 겹쳐 분포하여 수준 구분이 어려웠고, 원 등급과의 정확 일치율은 54.33%에 그쳤다. 그러나 이러한 검증은 에세이 등 텍스트 산출물에 한정되어 있으며, 시각 산출물이 중심이 되는 실기 영역에서 이루어진 사례는 찾기 어렵다.

이러한 간극은 시각-언어모델의 성능을 다룬 선행 연구에서 보다 직접적으로 드러난다. 그래픽 디자인의 미학적 품질을 4개 차원, 12개 지표로 세분화하여 다수의 모델을 평가한 연구에 따르면, 가장 우수한 모델조차 인간의 심미적 판단과 뚜렷한 격차를 보였다. 결함의 존재 여부를 판단하는 과제보다 결함의 위치를 정밀하게 특정하는 과제에서 성능이 현저히 낮았으며, 이는 무엇이 잘못되었는지를 가려내는 일과 어디가 어떻게 잘못되었는지를 짚어내는 일이 모델에게 서로 다른 난도의 과제임을 뜻한다. 특히 추론 능력을 강화한 모델도 일반 모델보다 나은 결과를 내지 못하였는데, 이는 이러한 한계가 연산 능력의 문제가 아니라 시각적 미학 판단이라는 과제 자체의 특성에서 비롯됨을 시사한다.¹³⁾ 이 구분은 본 연구의 분석에서도 중요한 참조점이 된다. 진단이 결함의 유무를 확인하는 데 머무는지, 아니면 무엇을 어떻게 고쳐야 하는지까지 짚어내는지는 피드백의 교육적 유용성을 가르는 지점이기 때문이다.

이상의 선행 연구를 종합하면 생성형 AI의 교육적 활용 가능성에 대한 낙관적 전망과 구현된 시스템의 정량 지표 의존성, 텍스트 영역에 한정된 일치도 검증, 그리고 시각적 판단 영역에서 확인된 모델 성능의 한

계가 서로 다른 층위에서 공존하고 있음을 알 수 있다. 이는 AI 자가진단 결과의 유효성이 언어개념적 판단과 시각·감각적 판단 사이에서 비대칭적으로 나타날 수 있다는 본 연구의 문제의식에 이론적 근거를 제공한다. 나아가 결함의 존재 여부를 가리는 과제와 그 위치를 특정하는 과제 사이에 성능 격차가 존재한다는 보고는 판정을 몇 개의 등급으로 환원하는 진단 구조가 질적 차이를 담아내지 못할 가능성을 시사한다. 다만 이들 연구는 모델의 성능을 벤치마크 환경에서 재거나 시스템의 사용성을 보는 데 초점을 두었으며, 실제 교육 현장에서 운영된 시스템의 판정이 같은 대상을 평가한 교수자의 판단과 어떤 관계에 놓이는지를 직접 대조한 사례는 찾기 어렵다. 본 연구는 한 학기 동안 운영된 시스템의 산출물을 대상으로 이 대조를 수행한다는 점에서 선행 연구가 다루지 못한 지점에 접근한다.

2-3. 본 연구의 위치

본 연구는 비개발자 교육자가 생성형 AI와의 협업으로 교육 운영 시스템을 구축한 사례를 보고한 선행 연구의 연장선에 있으나, 질문의 층위를 ‘시스템을 어떻게 만들 것인가’에서 ‘그 시스템이 산출하는 판단이 교수자 판단과 어떤 관계에 놓이는가’로 이동시킨다는 점에서 차별된다. 앞서 검토한 크리티크 이론과 AI 피드백 선행 연구는 본 연구의 분석 축과 작업 가설을 뒷받침한다. 분석 축은 언어개념적 판단과 시각·감각적 판단의 구분이며, 가설은 AI 자가진단 결과의 유효성이 두 영역 사이에서 비대칭적으로 나타날 것이라는 것이다.

다만 본 연구는 AI 진단 판정의 타당성을 검증하는 것이 아니라, 동일 대상에 대한 전문가 판단과 탐색적으로 대조하여 AI 자가진단이 교수자 피드백에 대해 갖는 역할 범위를 가능하는 데 목적을 둔다. AI와 전문가 판정 근거로 삼은 자료와 평가 항목, 척도가 서로 달라 동일한 구인을 측정하지 않으므로, 이하의 대조는 일치도를 검증하는 작업이 아니라 두 판정이 어떤 관계에 놓이는지를 살피는 작업이다. 진단 항목의 구성이 평가하고자 한 바를 제대로 담고 있는지는 검토 대상에 포함하지 않는다.

한편 학생 인식 조사가 다루는 것은 판정과 전문가 판단의 관계가 아니라 피드백의 유용성과 수용 양상이다. 학습자가 도움이 되었다고 응답하는 것과 그 판정이 전문가의 판단과 부합하는 것은 서로 다른 층위의 문제이며, 본 연구는 이 둘을 구분하여 서술한다. 제4장에서 4-2절과 4-3절은 AI 진단 판정과 서술을, 4-4절은 유용성과 수용 양상을 다루며, 4-5절에서 두 층

12) Mizumoto, A. & Eguchi, M., ‘Exploring the Potential of Using an AI Language Model for Automated Essay Scoring’, *Research Methods in Applied Linguistics*, Vol.2, No.2, 2023. 08. pp.1-13.

13) An, A. et al., *Op.cit.*, ICLR 2026.

위를 함께 놓고 판단 영역별 차이를 검토한다. 다음 장에서는 이러한 토대 위에서 실제 시스템이 어떻게 설계되고 운영되었는지를 분석한다.

3. 시스템 설계 및 운영

3-1. 설계 배경과 진단 구조

선행 연구에서 다른 시스템이 교수자의 최종 평가를 기록·집계하는 도구였다면, 본 연구의 시스템은 학생이 과제를 제작하는 과정에서 스스로 점검할 수 있도록 설계된 자가진단 도구라는 점에서 성격이 다르다. 교수자의 평가를 대신하는 것이 아니라, 교수자에게 결과물을 가져가기 이전 단계에서 학생이 진행 상황을 확인하게 하는 데 목적이 있다.

시스템은 A교과목의 과제 진행 순서에 따라 6단계(이하 STAGE 1~5 및 FINAL)로 구성되며, 각 단계는 3~6개의 세부 평가 항목을 둔다. 학생이 해당 단계까지 작성한 과제를 PDF로 제출하면 AI가 각 항목을 완료·보완·누락으로 판정하고 항목당 2점·1점·0점으로 환산하여 단계별 점수를 산출한다. 단계별 평가 항목의 구성은 [표 1]과 같다.

[표 1] 단계별 평가 항목 구성

단계	항목 수	평가 항목
STAGE 1 대상 분석	4	대상 선정 이유, 정체성 분석, 선행 사례 분석, 환경 분석
STAGE 2 기획 수립	6	타겟 설정, 목적 설정, 장소 선정, 시간 설정, 내용 설정, 조닝 설정
STAGE 3 컨셉트 구성	4	컨셉트 도출, 레퍼런스 조사, 아이디어 스케치, 무드보드 제작
STAGE 4 3D 렌더링	3	공간 마스터 조감, 아이레벨 투시, 공간 연출 상세
STAGE 5 도면	3	평면도, 입면도, 상세 제작도
FINAL 종합 심사	3	기획-공간 일관성, 물리적·상업적 현실성, 스토리텔링 논리

단계 간 진행은 게이팅(gating) 구조로 통제된다. 특정 단계에서 1회 이상 진단을 받으면 다음 단계가 잠금 해제되며, 완료 판정을 받지 못하더라도 최대 시도 횟수(3회)를 소진하면 다음 단계로 진행할 수 있도록 하였다. 이는 학생이 특정 단계에 머무르며 학기 진도에 뒤처지는 것을 방지하기 위한 설계이다. 이와 달리 FINAL은 초기 설계에서 의도적으로 1회로 제한하였는데, 앞선 단계에서 이미 여러 차례 피드백을 받은 학

생이 마지막으로 제출 가능 여부를 확인하는 자리로 보았기 때문이다. 다만 이 제한은 이후 3-3절에서 서술하는 운영 중 조정을 거쳐 변경되었다.

3-2. 진단 프롬프트와 판정 구조

시스템은 HTML/JavaScript 기반 프론트엔드와 Supabase, Google Gemini API로 구성되었으며 Cloudflare Pages를 통해 배포되었다. 진단에는 Flash-Lite 계열의 경량 모델을 사용하였으며¹⁴⁾, 응답의 일관성을 위해 생성 설정을 temperature 0.1로 고정하였다. 학생이 기획 내용과 시각 산출물이 함께 담긴 PDF를 업로드하면 이 파일이 별도의 텍스트 추출이나 이미지 변환 없이 원본 그대로 모델에 전달되며, 반환된 평가 텍스트는 진단 결과·강점·개선 사항의 세 구간 및 항목별 점수로 구분되어 화면에 표시되는 동시에 데이터베이스에 저장된다.

평가 프롬프트는 서버 측에서만 생성되어 학생에게 노출되지 않는다. 이 프롬프트는 응답 말미에 “완료=2보완=1누락=0”의 규칙에 따른 항목별 점수와 결함 태그를 정해진 데이터 형식으로 출력하도록 하였고, 보완 또는 누락으로 판정한 항목에만 개선 사항을 서술하도록 지시하여 판정과 서술이 서로 대응하도록 구성하였다. 다만 점수는 항목별 3단 판정에서만 산출되며, 서술 구간의 내용이 점수에 직접 반영되지는 않는다.

시각 자료에 대한 판정 기준은 자료의 존재 여부와 판독 가능성, 서술의 논리성으로 규정하였다. 해당 항목의 설명이나 이미지가 없을 때를 누락으로, 자료는 있으나 판독이 어렵거나 구체성이 떨어질 때를 보완으로 판정하도록 지시하였으며, 취향의 차이나 스타일 선택, 감성적 해석의 차이는 개입 대상에서 명시적으로 제외하였다. 교수자가 축적해 온 조정적 평가 기준이나 우수·미흡 사례를 참조 자료로 제공하지 않았고, 프롬프트는 학기 중 수정하지 않았다.

프롬프트는 시도 횟수에 따라 어조를 달리하도록 설계되었다. 1회차는 격려와 최소한의 가이드를, 2회차는 질문 중심의 점검을, 3회차는 제출 가능 여부에 대한 엄격한 심사를 수행하도록 하여, 시도를 거듭할수록 요구 수준이 높아지도록 하였다.

FINAL은 앞선 단계와 다른 방식으로 설계되었다. STAGE 1~5가 각 단계의 산출물을 개별적으로 점검하

14) gemini-3.1-flash-lite-preview. 해당 모델은 연구 종료 시점 전후로 제공사의 지원이 중단되어, 동일 모델을 이용한 재현에는 제약이 있다.

는 자리였다면, FINAL은 학생이 학기 전체에 걸쳐 진행한 작업을 하나의 결과물로 놓고 판단하는 자리였다. 이에 따라 평가 항목을 기획-공간 일관성, 물리적·상업적 현실성, 스토리텔링 논리의 세 가지로 압축하여 항목당 2점씩 총 6점 만점으로 구성하였으며, 앞선 단계보다 항목 수를 줄이는 대신 각 항목이 포괄하는 범위를 넓혔다. FINAL 단계에 사용된 프롬프트의 구성은 [표 2]와 같다.

[표 2] FINAL 단계 평가 프롬프트 구성

구분	내용
판정 원칙	개별 항목의 유무가 아니라, 전체 프로젝트를 관통하는 ‘논리적 일관성’과 ‘설득력’을 냉정하게 평가하라
서술 원칙	긍정적인 평가는 짧게, 치명적 모순이나 현실성 결여는 매우 날카롭게 지적하라
F1	기획-공간 일관성: 기획 목적과 타겟이 공간 디자인과 도면에 논리적으로 반영되었는가
F2	물리적·상업적 현실성: 실제 구현 가능한 스케일과 동선, 상업적 효용성을 갖추었는가
F3	스토리텔링 논리: 프레젠테이션 시 흐름이 끊기지 않고 설득되는가
출력 형식	[종합 평가] / [강점] / [개선 사항] 세 구간 및 항목별 점수
점수 환산	F1·F2·F3 순서대로 우수=2, 보통=1, 미흡=0
결함 태그	LOGIC_GAP, PHYSICAL_CONFLICT, TARGET_MISMATCH, SCALE_ERROR, ETHICS_FLAG

프롬프트는 항목의 유무가 아니라 전체를 관통하는 논리적 일관성과 설득력을 평가하도록 명시적으로 지시하였고, 결함이 있을 경우 날카롭게 지적할 것을 요구하였다.¹⁵⁾ 그럼에도 점수는 세 항목의 3단 판정을 합산하는 방식에서만 산출되었으며, 서술 구간에서 지적된 내용이나 결함 태그는 점수에 반영되지 않았다. 즉 프롬프트가 요구한 판단의 성격과 그 판단이 기록되는 형식 사이에는 간극이 있었다. 어떤 문제를 얼마나 심각하게 보았든 최종적으로 남는 것은 항목별로 부여된 세 등급 중 하나였다.

3-3. 운영 중 조정

시스템은 학기가 진행되는 동안 한 차례 구조 조정을

15) [표 2]에서는 익명화 원칙에 따라 과제의 공간 유형을 지칭하는 표현 1개소를 일반 명사로 대체하였다. 또한 운영 중 FINAL의 시도 횟수를 3회로 확대하였으나 프롬프트 문구는 수정하지 않아, 회차 관련 서술은 초기 설계 시점의 내용을 담고 있다.

을 거쳤다. 조정은 학기 중반, STAGE 3~5의 참여 부재가 확인된 시점에 이루어졌다. 운영 로그를 살펴보면 STAGE 2(기획 수립)까지는 22명이 시도하였으나, 시각적 산출물이 본격적으로 요구되는 STAGE 3(컨셉트 구상)에 진입한 학생은 조정 시점까지 없었다.

이러한 관찰에 따라 두 가지 조정이 이루어졌다. 첫째, STAGE 3~5를 폐지하거나 잠그지 않고 선택 사항으로 남겨둔 채, STAGE 2 완료만으로도 FINAL에 진입할 수 있는 우회 경로를 추가하였다. 둘째, 이로 인해 시각 산출물에 대한 첫 AI 피드백을 FINAL에서 받게 되는 경우가 늘어날 것을 고려하여, 기존 1회로 제한되었던 FINAL의 시도 횟수를 3회로 확대하였다.

시스템 개편은 실행연구가 취하는 예비 운영과 반성적 평가, 재설계와 재실행의 순환적 절차와 궤를 같이 한다.¹⁶⁾ 다만 이 순환은 AI 진단 로직의 성능이나 타당성 문제에서 촉발된 것이 아니라, 학생들의 참여 양상과 학기 진도를 고려한 연구자의 상황적 판단에서 비롯되었다. 이에 본 연구는 운영상의 조정과 AI 진단 판정의 타당성 검토를 서로 다른 층위로 분리하여 다루며, 전자의 결과를 후자의 직접적 원인으로 서술하지 않는다. 다음 장에서는 이와 같이 설계·운영된 시스템이 실제로 어떤 결과를 산출하였는지를 분석한다.

4. 시스템 적용 결과

4-1. 시스템 활용 양상

학기 중 이루어진 구조 조정을 기점으로 학생들의 이용 경로는 두 갈래로 나뉘었다. 단계별 이용 현황은 [표 3]과 같다. 등록 28명 가운데 1회 이상 진단을 시도한 학생은 STAGE 1이 25명, STAGE 2가 22명이었다. 이후 선택 사항으로 전환된 STAGE 3~5를 개별적으로 완료한 학생은 4명, FINAL을 완료한 학생은 16명이었다. 앞의 두 단계는 시도를, 뒤의 세 단계와 FINAL은 완료를 기준으로 집계한 것이므로 이를 단일한 증감 추이로 읽을 수는 없다. STAGE 3~5에서 끊긴 참여가 FINAL에서 다시 이어진 것은 이탈했던 학생이 복귀했기 때문이 아니라, 우회 경로의 추가로 상당수가 이 구간을 건너뛰었기 때문이다. 다만 조정 시점까지 STAGE 3에 진입한 학생이 없었다는 점은, 우회 경로가 이탈의 원인이 아니라 이미 발생한 이탈에 대한 사후 대응이었음을 보여 준다. 시각 산출물이 요구

16) McNiff, J., Op.cit., 2017, pp.9-12.

되는 지점에서 참여가 끊긴 것은 이후 4-3절에서 확인 되는 시각 단계의 진단 특성과 함께 볼 필요가 있다.

[표 3] 시스템 활용 양상

단계	경로	집계 기준	인원	등록 28명 대비
등록	—	—	28명	100%
STAGE 1	공통	시도	25명	89%
STAGE 2	공통	시도	22명	79%
STAGE 3~5	선택	완료	4명	14%
FINAL	최종	완료	16명	57%

4-2. AI 진단 결과와 전문가 평가의 대조

지도교수와 외부 전문가는 최종 발표 현장에서 각자 독립적으로 20명을 평가하였다. 평가 항목은 모두 14개이며, 본 연구의 분석 축에 따라 언어·개념 영역 5항목, 시각·감각 영역 5항목, 그리고 물리적·상업적 현실성과 발표 전달력 각 2항목으로 분류하였다.¹⁷⁾

이 대조에는 몇 가지 조건상의 제약이 있다. 평가는 AI 진단 결과를 참조하지 않고 이루어졌으나, 지도교수는 시스템 운영자로서 학기 중 학생별 진단 기록을 열람한 바 있어 사전 인지의 영향을 완전히 배제할 수는 없다. 또한 AI는 학생이 제출한 PDF만을 근거로 삼은 반면, 전문가 평가자는 최종 발표 현장에서 발표와 질의응답을 함께 접하였다. 판정 구조 역시 AI FINAL은 세 항목 6점 만점, 전문가 평가는 14항목 5점 척도로 서로 다르다. 더욱이 AI가 판단한 세 항목은 기획·공간 일관성, 물리적·상업적 현실성, 스토리텔링 논리로 구성되어 시각·감각 영역을 직접 평가하는 항목을 포함하지 않았으나, 전문가 평가 14항목에는 해당 영역 5항목이 포함되었다. 두 판정은 측정 대상의 구성부터 다르다. 따라서 이하의 대조는 양자가 동일한 구인을 측정한다고 전제하지 않으며, 동일한 결과물에 대해 서로 다른 방식으로 내려진 판단이 어떤 관계에 놓이는지를 확인하는 작업이다.

17) 언어·개념 영역 5항목은 ①대상 정체성 정의의 명확성, ②타겟·목적과 공간 콘셉트의 정합성, ③공간 경험 시나리오의 흐름, ④기획-디자인 논리 연결의 설득력, ⑤핵심 메시지의 명료성이며, 시각·감각 영역 5항목은 ⑥그래픽·키비주얼 완성도, ⑦3D 공간 연출의 표현력, ⑧도면·디테일 완성도, ⑨콘셉트의 신선함, ⑩공간 아이디어 연출의 차별성이다. 나머지 4항목은 ⑪공간 스케일·동선의 현실성, ⑫상업적 효용, ⑬발표 구성·논리 전개, ⑭전달력·태도로, 두 영역 어디에도 속하지 않아 영역별 상관 분석에서는 제외하고 전체 평균에만 포함하였다.

FINAL을 완료한 16명 가운데 최종 발표에도 참여하여 전문가 평가와 대조가 가능한 학생은 15명이다. 대조에 사용한 점수는 각 학생의 마지막 회차 점수이다. 이들의 AI FINAL 점수는 6점 만점이 6명, 5점이 9명으로 상위 두 구간에만 분포하였다. 16명 중 한 명을 제외한 전원이 1회차 진단에서 이미 5점 또는 6점을 받았으며, FINAL을 2회 이상 시도한 학생도 2명에 그쳤다. 상위 구간 집중은 첫 진단에서부터 나타났다. 앞서 서술하였듯 FINAL 프롬프트는 항목의 유무가 아니라 전체를 관통하는 논리적 일관성과 설득력을 평가하도록 지시하였음에도 산출된 점수는 최상위 두 구간에 몰려 학생 간 완성도의 차이를 변별해 내지 못하였다.

같은 대상을 평가한 지도교수와 외부 전문가의 결과는 이와 달랐다. 평가자 간 상관은 $r=.732(n=20)$ 로 높게 나타났으며, 언어·개념 영역($r=.655$)과 시각·감각 영역($r=.681$) 모두에서 뚜렷한 정적 상관이 관찰되었다. 다만 상관은 두 평가가 같은 방향으로 움직임을 보여 줄 뿐 점수의 일치를 뜻하지 않는다. 시각·감각 영역에서도 두 평가자가 학생 간 차이를 유사한 방향으로 변별하였다는 것은, 이들 결과물 사이에 변별 가능한 차이가 실재하였음을 뜻한다. 이 영역별 상관은 발표 관련 2항목을 제외한 값이다. 다만 두 평가자가 같은 발표 현장을 접한 상태에서 여타 항목을 판정하였으므로, 발표에서 받은 인상이 항목 전반의 판단에 영향을 미쳤을 가능성까지 배제되는 것은 아니다.

3자 대조가 가능한 15명을 대상으로 AI FINAL 점수와 전문가 평가 점수의 상관을 확인한 결과는 [표 4]와 같다.

[표 4] AI FINAL 점수와 전문가 평가 점수 간 상관($n=15$)

대조	r
AI ↔ 외부 전문가	-.077
AI ↔ 지도교수	-.078
AI ↔ 전문가 평균	-.082

세 대조 모두 상관계수가 0에 가까웠다. AI 점수 구간별로 전문가 평가를 비교하면 만점을 받은 6명의 평균은 5점 척도 기준 3.83점으로, 5점을 받은 9명의 평균 3.95점보다 오히려 낮았다. 다만 두 집단의 차이는 크지 않으며, 만점 집단에 전문가 평가 최하위권 사례가 포함된 영향이 작용하였다. 그럼에도 AI로부터 만점을 받았으나 전문가 평가에서는 최하위권(2.61점)에 속한 학생이 있었던 반면, 5점을 받고도 최상위권(4.75점)에 오른 학생도 있었다는 점은 두 판정이 서

로 다른 순서를 가리키고 있음을 보여 준다.

한편 AI FINAL 점수가 상위 두 구간에만 분포한다는 것은 상관 분석 자체의 전제를 제약한다. 범위 제한이 매우 큰 변수에서는 상관계수가 안정적인 값을 갖기 어려우며, 이 점에서 위 상관계수는 관계의 부재를 입증한다기보다 앞서 확인한 변별력 부재를 다른 방식으로 확인해 주는 지표에 가깝다. 두 결과가 가리키는 바는 일관된다. 본 시스템이 산출한 AI FINAL 점수는 전문가가 14개 항목을 종합하여 내린 판단과 같은 서열을 그리지 않았으며, 이를 전문가의 종합적 판단을 대신하는 지표로 보기는 어렵다.

4-3. AI 진단 서술의 질적 분석

앞 절이 AI가 매긴 점수를 다루었다면, 본 절에서는 AI가 실제로 어떤 내용의 서술을 산출하였는지를 살펴본다. 분석 대상은 STAGE 3~5를 개별적으로 완료한 4명의 진단 기록 18건이다. 이 세 단계는 뒤로 갈수록 시각 자료의 비중이 커지므로, 판단 영역에 따른 차이를 관찰하기에 적합하다. 분석은 연구자 1인이 수행하였으며, 도출된 범주가 통계적 대표성을 갖지 않는 탐색적 성격의 작업이다. 먼저 18건의 진단 서술에서 개선 사항 구간만을 분리한 뒤, 각 지적이 무엇을 문제 삼고 있는지에 따라 사전 범주를 두지 않고 귀납적으로 묶어 나갔다. 그 결과 항목의 존재 여부를 묻는 지적, 동선·안전·시공 가능성 등 공간 운영에 관한 지적, 조형적 완성도에 관한 지적의 세 범주가 도출되었다. 이후 전체 자료를 다시 읽으며 도출된 범주로 재코딩하고, 한 지적이 둘 이상의 범주에 걸치는 경우에는 그 지적이 개선을 요구하는 대상을 기준으로 하나에 배정하였다. 다만 단일 코더에 의한 분류이므로 코더 간 신뢰도는 산출하지 못하였다.

먼저 단계별 득점률에 뚜렷한 경사가 나타났다. 단계별 진단 결과는 [표 5]와 같다. 언어와 개념이 중심인 STAGE 3은 6건 모두 만점이었으나, STAGE 4는 91.7%, STAGE 5는 77.8%에 그쳤다. 특히 STAGE 3에서는 24개 항목 전부가 완료로 판정되어 세 등급 가운데 하나만 사용되었으며, STAGE 4와 STAGE 5에서는 판정이 세 등급에 분산되었다. 같은 학생이 같은 프로젝트를 이어가며 받은 진단임에도 단계에 따라 득점률이 달라졌다는 것은 AI의 판정이 대상 자료의 성격에 영향을 받았음을 시사한다. 물론 이 경사는 시각 단계에서 AI가 문제를 더 많이 탐지하였다는 뜻으로도 읽힐 수 있으므로, 점수만으로 판단 영역별 비대칭의 방향을 확정하기는 어렵다. 비대칭의 성격은 다음에서

살펴볼 개선 사항의 내용 분석에서 보다 분명하게 드러난다.

[표 5] 심층 완료자 4명의 단계별 진단 결과(18건)

단계	중심 자료	진단 건수	득점/만점	득점률	판정 분포 (완료·보완·누락)
STAGE 3 개념 구상	언어·개념	6건	48/48	100%	24 · 0 · 0
STAGE 4 3D 렌더링	시각	6건	33/36	91.7%	16 · 1 · 1
STAGE 5 도면	시각	6건	28/36	77.8%	12 · 4 · 2

눈여겨볼 것은 AI가 만점을 부여하면서 동시에 결합 태그를 출력한 사례가 4건 있었다는 점이다. 한 학생의 STAGE 3에서 AI는 네 항목을 모두 완료로 판정해 만점을 부여하고도 같은 응답의 개선 사항 구간에서 기획한 앞뒤의 운영 연도가 서로 어긋난다고 지적하였다. 다른 학생의 STAGE 4에서도 세 항목 모두 완료 판정을 받았으나, AI는 같은 구간에서 해당 공간의 안전과 위생에 우려가 있다고 서술하였다. 프롬프트는 보완 또는 누락으로 판정한 항목에 한하여 개선 사항을 서술하도록 지시하였으므로, 전 항목 완료 판정과 결합 지적이 함께 출력된 것은 판정과 서술이 서로 다른 결론에 이르렀음을 뜻한다. 만점 기록 11건 모두에서 개선 사항이 출력되었고, 그중 5건은 완료로 판정한 항목을 다시 ‘보완’으로 표기하였다.

STAGE 4와 STAGE 5의 진단 12건에서 확인된 개선 사항을 내용별로 분류하면, 빠진 항목을 지적한 것이 3건, 동선이나 안전·시공 가능성 등 공간 운영에 관한 것이 9건이었으며, 조형적 완성도 자체를 지적한 사례는 하나도 없었다. AI는 강점 서술에서는 렌더링의 완성도나 색채 대비를 언급하였으나, 개선 사항에서는 항목의 유무나 서술로 따져볼 수 있는 사항만을 지적하였다. 다만 3-2절에서 서술하였듯 조형·미감 판단이 개입 대상에서 제외되었으므로, 조형적 지적의 부재를 모델의 판단 능력만으로 설명할 수는 없다.

4-4. 학생 인식 조사

학기 종료 후 설문에 수강생 28명 중 25명이 응답하였다. 응답자 25명은 앞서 STAGE 1을 시도한 25명과는 별개의 집합이며, 설문이 익명으로 수집되어 두 집단의 개인 단위 대조는 이루어지지 않았다. 설문은 익명 온라인 양식으로 배포하였으며, 사용 경험 여부와

리커트 7문항, 판단 영역별 비교 2문항, 대체 가능성과 적절한 역할에 관한 선택형 문항, 그리고 자유 서술 문항으로 구성하였다. 이 가운데 사용 경험이 없다고 답한 2명을 제외한 실제 사용자 23명을 분석 대상으로 삼았다.

리커트 7문항에 대한 응답 결과는 [표 6]과 같다. 가장 눈에 띄는 것은 실제로 수정·보완했다는 행동 지표(4.39)와 신뢰할 수 있었다는 인식 지표(3.70) 사이의 간극이다. 학생들은 AI 피드백을 전적으로 신뢰하지는 않으면서도 실제로는 그에 따라 결과물을 수정하였다고 응답하였다. 다만 이는 학기 종료 후의 회고적 자기보고이므로, 시스템 로그로 확인되는 실제 수정 행동과는 구분된다.

[표 6] AI 피드백 사용에 대한 인식(5점 리커트, n=23)

문항	평균
피드백을 받고 실제로 결과물을 수정·보완했다	4.39
지도교수 컨펌 전 사전 점검이 가능했다	4.26
후배에게 권유하고 싶다	4.04
피드백이 구체적이어서 무엇을 고칠지 알 수 있었다	4.00
AI가 내 작업을 정확히 이해한 것으로 느껴졌다	3.83
AI 피드백의 내용을 신뢰할 수 있었다	3.70
PDF 변환이 번거로웠다	2.26

판단 영역별로 AI와 지도교수 피드백 가운데 어느 쪽이 더 도움이 되었는지 묻는 결과는 [표 7]과 같다.

[표 7] 판단 영역별 AI-지도교수 피드백 비교(n=23)

판단 영역	AI 우위	비슷	지도교수 우위
언어·개념적 판단	3	10	10
시각·감각적 판단	0	3	20

언어·개념적 판단에서는 평가가 나뉘면, 시각·감각적 판단에서는 23명 중 20명이 지도교수 피드백이 더 도움이 되었다고 응답하였으며 AI가 더 도움이 되었다는 응답은 한 건도 없었다. AI 피드백의 대체 가능성에 대해서는 보조 역할에 적합하다는 응답이 13명, 일부 영역에서는 대신할 수 있다는 응답이 10명이었으며 양 극단의 응답은 없었다. 시스템의 가장 적절한 역할을 묻는 문항에서도 지도교수의 확인을 받기 전 스스로 점검하는 예비 단계라는 응답이 15명으로 가장 많았다.

서술형 응답에서도 이러한 인식이 확인되었다. 한 학생은 AI가 요소의 내용보다 요소의 존재를 중요하게

판단하는 것 같다고 서술하였으며, 다른 학생은 초반에 단어만 들어가면 통과가 가능하여 통과 자체를 목표로 삼게 되었다고 밝히면서도 최종 제출 이후 교수자의 피드백과 AI의 피드백이 문제점 부분에서 유사하게 나왔다고 덧붙였다. 또 다른 학생 역시 특정 단어만 입력하면 통과되는 경우가 있어 완결된 평가 체계로 보기는 어렵다고 응답하였다. 한편 정지된 문서로만 평가가 가능하고 실시간 피드백을 받을 수 없어 후반 작업에서는 사용하지 않았다는 응답도 있었다. 반면 시간과 공간의 제약 없이 피드백을 받을 수 있다는 점을 가장 큰 장점으로 꼽거나, 교수자에게 연락하기를 어려워하는 학생에게 유용하다고 본 응답도 확인되었다.

4-5. 판단 영역별 비대칭 종합

이상의 결과는 서로 다른 층위에서 동일한 방향으로 수렴한다. 전문가 평가와의 대조에서는 AI 점수가 두 구간에만 분포하여 학생 간 차이를 변별하지 못하였고, 진단 로그에서는 만점과 결함 지적이 동시에 출력되었으며 시각 단계의 개선 사항 가운데 조형적 완성도에 대한 지적은 전무하였다. 학생 인식에서도 시각·감각 영역에서 23명 중 20명이 지도교수 피드백이 더 도움이 되었다고 응답하였다. 여기에 학생 스스로의 설명이 더해져, 형식적 요소의 존재 여부에 치중한 판정이라는 같은 방향을 가리킨다. 다만 이 치중은 평가 기준 설계와 무관하지 않으므로, 모델 특성만으로 귀속시킬 수는 없다. 다음 장에서는 이러한 결과를 종합하여 그 함의를 정리한다.

5. 결론

본 연구는 공간 디자인 실기 교과목에서 한 학기 동안 운영된 AI 자가진단 시스템의 데이터를 통해, 본 시스템이 산출한 AI 자가진단 결과가 어느 범위까지 교수자의 역할을 수행할 수 있는지를 검토한 것이다. AI 진단 로그와 전문가 평가, 학생 인식 설문을 대조한 결과는 다음과 같이 정리된다.

첫째, 본 시스템에서 항목별 완료 여부를 판정하는 체크리스트형 구조는 질적 변별을 대리하지 못하였다. AI FINAL 점수는 5점과 6점 두 구간에만 몰려 학생 간 차이를 변별해 내지 못하였으며, 같은 대상을 평가한 지도교수와 외부 전문가 사이에서는 시각·감각 영역을 포함해 높은 정적 상관($r=.732$)이 확인되었다.

FINAL 프롬프트가 논리적 일관성과 설득력의 평가를 명시적으로 지시하였음에도 결과가 달라지지 않았다는 것은, 변별의 실패가 지시의 부재에서 비롯된 것이 아님을 뜻한다. 판정을 세 등급으로 환산하는 출력 구조, 서로 다른 결과물에도 유사한 판정이 산출되는 모델 응답의 특성¹⁸⁾, 조형-미감 판단을 개입 대상에서 제외한 평가 기준 설계가 원인으로 놓일 수 있으나, 본 연구는 판정 구조가 다른 시스템과의 비교나 동일 조건에서의 반복 산출 확인을 수행하지 않았으므로 이 가운데 무엇이 작용하였는지는 가릴 수 없다. 진단 텍스트에서도 같은 양상이 확인된다. 만점 판정과 결함 지적이 같은 응답 안에서 함께 출력되어 판정과 서술이 어긋난 사례가 있었으며, 시각 단계의 개선 사항 가운데 조형적 완성도 자체에 대한 지적은 나타나지 않았다. 이는 조형-미감 판단을 평가 기준에서 제외한 설계와 분리하기 어려우며, 시각-언어모델이 결함의 존재 여부를 판단하는 과제보다 그 위치를 특정하는 과제에서 낮은 성능을 보인다는 선행 연구와도 부합한다.¹⁹⁾

둘째, 이러한 비대칭은 판단 대상의 성격뿐 아니라 피드백이 오가는 매체의 성격과도 맞물려 있다. 비대면-온라인 형태의 크리틱일수록 학습자가 체감하는 목표의 명확성과 상호작용 정도가 낮아진다는 선행 연구는, 텍스트 기반의 비동기적 피드백이라는 본 시스템의 조건이 시각-감각 영역에서 구조적으로 불리하게 작용했을 가능성을 시사한다. 한 학생은 정지된 문서로만 평가가 가능하고 실시간 피드백을 받을 수 없어 후반 작업에서는 시스템을 사용하지 않았다고 응답하였다.

셋째, 학생들은 AI 피드백을 전면적으로 신뢰하지 않으면서도 실제로는 그에 따라 결과물을 수정하였다고 보고하였다. 수정·보완했다는 응답(4.39)이 신뢰할 수 있었다는 응답(3.70)보다 높게 나타났다는 것은 AI의 교육적 유용성이 반드시 학습자의 전면적 신뢰를 전제로 하지는 않는다는 뜻이다. 시간과 공간의 제약 없이 피드백을 받을 수 있다는 점을 장점으로 꼽은 응답에서 드러나듯, 학생들은 AI를 옳은 답을 주는 존재가 아니라 부담 없이 물어볼 수 있는 존재로 취급하였으며 이러한 낮은 진입 장벽이 신뢰 수준과 무관하게 실제 사용으로 이어진 것으로 해석된다.

넷째, 자가진단 시스템은 판정 기준이 명시적일수록 학습자가 그 기준을 실질적 개선이 아닌 통과와 대상으로 인식하게 될 위험을 안고 있다. 초반에 단어만 넣

으면 통과가 가능하여 통과 자체를 목표로 삼게 되었다는 응답이 그 사례에 해당한다. 이 학생이 최종 제출 이후 교수자와 AI의 피드백이 유사하게 나왔다고 덧붙인 점은 AI의 진단 항목이 방향성 측면에서는 유효했음에도 그 시점에는 학습자가 이를 개선 신호로 받아들이지 않았음을 의미한다. 이는 AI가 지적인 내용이 통과·미통과라는 이분법적 판정으로 환산되는 과정에서 학습자에게 전달된 결과로 볼 수 있다.

이상을 종합하면, 본 연구가 운영한 생성형 AI 자가진단 시스템은 교수자 피드백을 대신하는 대체제가 아니라 교수자에게 결과물을 가져가기 이전 단계에서 요소의 구비 여부를 점검하는 보완재로 자리매김하는 것이 타당하다. 설문에서 가장 많은 응답을 얻은 역할 인식, 즉 지도교수의 확인을 받기 전 스스로 점검하는 예비 단계라는 표현이 이 위치를 정확히 요약한다.

본 연구의 의의는 다음 두 가지로 정리된다. 첫째, AI 피드백의 교육적 역할을 논의할 때 자주 간과되는 판단 영역이라는 변수를 명시적으로 도입하고, 비개발자 교육자가 구축한 시스템이 산출하는 판단을 전문가 판단과 탐색적으로 대조하였다.²⁰⁾ 이는 시스템을 만드는 층위에 머물렀던 선행 연구를 그 시스템이 내리는 판단의 층위로 확장한 것이다. 둘째, 자가진단 시스템이 지닐 수 있는 구조적 역설, 즉 판정 기준이 명시적일수록 학습자가 그 기준을 통과의 대상으로 인식할 수 있다는 위험을 실제 학습자의 진술을 통해 구체적으로 드러내었다.

본 연구는 다음과 같은 한계를 갖는다. 첫째, 연구자가 시스템의 개발자이자 운영자일 뿐 아니라 전문가 평가자 2인 중 1인을 겸하였다. 이를 보완하기 위해 외부 전문가 1인의 독립적 평가를 함께 확보하였으며, 두 평가 결과 사이에는 높은 정적 상관($r=.732$)이 관찰되었다. 다만 상관은 두 평가가 같은 방향으로 움직였음을 보여 줄 뿐 점수의 절대적 일치를 뜻하지는 않으며, 이러한 수치가 연구자 개인의 평가 편향이 없었음을 입증하는 것도 아니다. 연구자는 시스템 운영자로서 학기 중 AI 진단 기록을 열람한 상태였으므로 평가 시점의 독립성에도 제약이 있다.

둘째, 본 연구는 공간 디자인 실기 교과목의 한 학기 운영 사례이며, 회화·시각디자인·영상 등 산출물과 평가 기준이 다른 실기 영역에는 그대로 적용되지 않는다. 표본 규모와 자료의 성격도 해석의 범위를 제한한다. 3자 대조가 가능한 사례는 15명, 학생 설문 응

18) Mizumoto, A. & Eguchi, M., Op.cit., 2023, pp.1-13.

19) An, A. et al., Op.cit., ICLR 2026.

20) 나상귀, Op.cit., 2026, pp.746-755.

답은 25명에 그쳤으며, 설문 또한 회고적 자기보고 방식이어서 기억 감쇠의 영향을 배제할 수 없다. 이에 본 연구는 제시된 상관계수를 참고치로만 해석하였고 인과적 해석, 표본을 넘어서는 일반화는 시도하지 않았다.

셋째, 학생 설문이 익명으로 수집되어 개별 학생의 로그 기록과 연결할 수 없었고, 이에 따라 설문 결과와 로그 데이터의 대조는 집단 수준에서만 이루어졌다. 또한 진단 서술의 질적 분석은 심층 단계를 개별적으로 완료한 4명의 기록 18건에 한정되어, 이들이 시스템 사용에 상대적으로 적극적이었던 학생이라는 점에서 전체 수강생을 대표한다고 보기 어렵다.

넷째, 연구 기간 중 시스템 구조가 동일하게 유지되지 않았다. 학기 중반에 STAGE 3~5가 선택 사항으로 전환되고 FINAL의 시도 횟수가 확대되었으므로, 학생마다 경험한 진단 경로와 피드백 횟수가 달랐다. 이는 실험연구의 순환 절차로서는 자연스러운 조정이나, 산출된 판정을 단일한 조건에서 얻어진 자료로 취급하기 어렵게 만드는 요인이기도 하다.

이러한 한계를 바탕으로 향후 과제를 제안하면 다음과 같다. 첫째, 심층 단계를 완료한 학생과 사용하지 않은 학생을 대상으로 개별 면담을 실시하여, 익명 설문으로는 확인할 수 없었던 개인 단위의 경험과 미사용의 이유를 보완할 필요가 있다. 둘째, 외부 전문가를 추가로 확보하여 주 분석을 외부 전문가 중심 체제로 전환하고 연구자의 평가를 보조 자료로 재배치함으로써, 연구자 겸임에서 비롯되는 객관성의 제약을 구조적으로 완화할 필요가 있다. 셋째, 체크리스트형 판정이 아닌 서술형 질적 피드백을 산출하도록 프롬프트를 재설계한 버전과 현행 시스템을 비교함으로써, 본 연구가 확인한 변별의 실패가 다른 판정 구조에서도 재현되는지를 검증할 필요가 있다. 아울러 조형적 평가 기준과 우수-미흡 사례를 참조 자료로 제공한 조건에서 AI가 조형적 완성도를 판별할 수 있는지를 확인함으로써, 판단 능력의 한계와 기준 설계의 한계를 구분할 필요가 있다. 동일 조건에서 산출되는 판정의 분산을 확인하는 작업도 변별의 실패가 판정 형식과 모델 응답 가운데 어느 쪽에서 비롯되는지를 가리는 데 필요하다.

참고문헌

1. 김동삼·류다현·박규동·강혜신, 'AI기반 실시간 온라인 교육 맞춤형 피드백 시스템 효과성 평가', 한국컴퓨터교육학회, 컴퓨터교육학회 논문지, Vol.27, No.7, 2024
2. 나상귀, '예술·디자인 학과 졸업작품 심사를 위한 생성형 AI 협업 기반 시스템 구축에 관한 실험연구 — 비개발자 교육자 사례를 중심으로', 한국디자인리서치학회, 디자인리서치, Vol.11, No.2, 2026
3. 안지선·박주령, '디자인 크리틱의 교육심리학적 고찰과 사용자경험 가이드라인', 한국브랜드디자인학회, 브랜드디자인학연구, Vol.19, No.2, 2021
4. 장윤경·윤인화·박지원·제갈유진·이지연, '이미지 생성 AI 도구의 이해와 미술교육의 활용 가능성 탐구', 한국조형교육학회, 조형교육, Vol.88, 2023
5. An, A., Sun, S., Huang, D., Cheng, M., Gao, Y., Li, J., Qiao, Y. & Bian, J., 'Can Vision-Language Models Assess Graphic Design Aesthetics? A Benchmark, Evaluation, and Dataset Perspective', ICLR 2026, 2026
6. McNiff, J., 『Action Research: All You Need to Know』, SAGE Publications, 2017
7. Mizumoto, A. & Eguchi, M., 'Exploring the Potential of Using an AI Language Model for Automated Essay Scoring', Research Methods in Applied Linguistics, Vol.2, No.2, 2023
8. Mollick, E. & Mollick, L., 'Instructors as Innovators: A Future-focused Approach to New AI Learning Opportunities, With Prompts', The Wharton School, The Wharton School Research Paper (SSRN 4802463), 2024