https://doi.org/10.30775/KMES.55.3.191
음악교육연구 제55권 제3호 pp. 191-215 (2026년 8월) ISSN 1229-4179 (Print)
Korean Journal of Research in Music Education Vol.55, No.3, (August, 2026) ISSN 2713-3788 (Online)

폐쇄 로컬 구조 OSS 기반 음악교육 서술형 LLM AI 평가도구 개발

Development of an Open-Source Based AI for Automated Scoring of Constructed Responses in Music Education Using Large Language Models

신기호*
Giho Shin

초록 본 연구는 학교 현장 적용을 전제로, 외부 전송을 최소화하는 폐쇄 로컬 구조에서 구동되는 OSS 기반 LLM 서술형 AI 평가도구를 개발하고 그 성능과 채점 안정성을 검토하였다. 평가도구는 문항 입력-루브릭 자동 생성/편집-답안 업로드-루브릭 기반 상・중・하 분류-근거 제시-개선 피드백 출력-엑셀 다운로드의 흐름으로 구현되었다. 연구는 서양음악사 서술형 답안 200건(상・중・하 라벨 포함)을 대상으로 (1) 3수준 분류 성능(Accuracy, Macro-F1), (2) 답안 순서 무작위 재배열 후 반복 채점 간 일치도를 분석하였다. 또한 LLM의 변동성과 관대 채점 경향을 통제하기 위해, LLM의 역할을 기준별 판정・근거 추출로 제한하고 RBSS 등 결정론적 신호를 결합한 하이브리드 채점 로직을 적용하였다. 결과적으로 반복 실행에서도 점수・순위 구조가 전반적으로 유지되었고, 판정 확신도(confidence)와 RBSS의 관계는 판정 신뢰도 해석 가능성을 시사하였다.

주제어: 인공지능, 음악교육, AI 평가도구, 폐쇄 로컬 구조, 오픈소스(OSS), 거대언어모델(LLM), 루브릭 기반 평가

Abstract This study developed and evaluated an open-source, locally deployed LLM-based assessment tool for constructed responses in music education, designed to minimize external data transfer and support school-level privacy constraints. The system implements an end-to-end workflow—item input, automatic rubric generation and teacher editing, Excel-based response upload, rubric-based three-level classification, evidence citation, personalized improvement feedback, and Excel export. Using 200 labeled constructed responses in a Western Music History unit, we examined (1) classification performance and (2) scoring consistency across repeated runs after random shuffling of response order under fixed model and inference settings. To mitigate known issues in LLM scoring, the tool restricts the LLM to criterion-level judgments and evidence extraction while computing final scores via deterministic Python logic, combining rule-based scoring signals (RBSS) with model metadata in a hybrid architecture. Results indicated that the overall score and rank structure remained stable across runs, and the observed relationship between confidence and RBSS suggests that confidence can be interpreted as a decision-certainty signal rather than a direct proxy for correctness.

Key words: AI, music education, AI assessment tool, local deployment, open-source software(OSS), large language model(LLM), Rubric-based evaluation

* Corresponding author, E-mail: music.dr.shin@knue.ac.kr
Lecturer, Korea National University of Education, 250 Taeseongtabyeon-ro, Gangnae-myeon, Heungdeok-gu, Cheongju-si, Chungbuk 28173, South Korea

Received: 1 March 2026, Reviewed (Revised): 9 April 2026 (16 April 2026), Accepted: 17 April 2026
© 2026 Korean Music Education Society.

Ⅰ. 서론

1. 연구의 배경 및 필요성

최근 거대언어모델(Large language model, LLM)은 문장 생성 능력을 넘어, 제시된 기준에 따라 글을 분류하고 평가 의견을 생성하는 수준까지 빠르게 확장되고 있다. 특히 서술형 평가 장면에서 LLM은 루브릭을 조건으로 주었을 때 학생 응답을 분석해 점수와 코멘트를 산출할 수 있으며, 실제 학생 글을 대상으로 자동채점・피드백 수행을 통해 가능성과 한계를 점검하려는 연구도 국내에서 보고되고 있다(Im & Im, 2024).

그러나 교육평가에 LLM AI를 곧바로 적용하기에는 몇 가지 문제에 대한 검토가 필요하다. 첫째, 웹 기반 생성형 AI는 프롬프트 조건에 따라 채점과 피드백이 달라질 수 있어, 후속적인 정확성 검증과 변동성 통제의 필요가 지적된다(Choi & Park, 2023). 둘째, LLM 자동채점은 인간 채점과 비교해 평가자로서의 객관성을 유지하지 못하고 주관이 개입되어 관대하게 채점되는 경향이 나타나거나, 인간 채점자료와 통합할 때 절대적 지식에 대한 논증이 이루어지는 탓에 교육과정에서 제시하는 서술적 지식의 학년별 특수성이 지지되지 않는 등 평가학적 쟁점이 보고된 바 있다(Ham et al., 2024). 셋째, 무엇보다 학교는 학생・학부모의 개인정보를 다루는 기관으로서, 교육활동・교육평가・상담 등이 일반적인 웹 클라우드 기반 AI 플랫폼을 통해 이루어질수록 개인정보 수집・이용 약관에 의하여 개인정보의 유출 또는 제3자 제공 위험도 함께 커질 수 있다는 문제의식이 꾸준히 제기되어 왔다(Jung, 2021).

이러한 조건을 고려할 때, 학교 현장에 적합한 LLM AI 평가도구는 ‘성능’뿐 아니라 ‘운영 방식’에서부터 객관성과 학년별, 지역별 유연성을 가진 교육 환경, 그리고 정합성을 가져야 한다. 특히 클라우드 기반 서비스에서는 제공자가 구조적으로 반드시 이용자 데이터를 보관・처리할 수밖에 없어, 개인정보 처리・위탁・국외 이전 등 법・정책적 쟁점이 발생할 수 있다는 논의가 있다(Kim, 2021). 따라서 교육평가 자료(학생 서술 응답, 채점 결과, 피드백)를 다루는 도구는 외부 전송을 최소화하고 기관 내부에서 통제 가능한 내부망을 통한 로컬 폐쇄적 형태로 설계될 필요가 있으며, 본 연구가 지향하는 폐쇄 로컬 구조 OSS 기반 LLM AI 평가도구는 이러한 요구에 대한 현실적인 대안이 될 수 있다.

이를 위해서는 평가도구의 폐쇄적 구조 설계뿐만 아니라, AI를 통한 평가 검증의 과정에서 객관성을 유지할 수 있는 루브릭을 설계하고 이에 따라 LLM AI의 주관적 판단을 지양하며, 주어진 루브릭에 따라 적절한 교과와 학년에 따른 평가를 내릴 수 있는 인공지능 모델의 개발이 필연적이라고 할 수 있다.

한편, 2022 개정 음악과 교육과정은 평가를 학습 과정에 통합하여 지속적으로 실시하

고, 성취기준을 평가 준거로 삼아 공정성・타당성・신뢰성을 확보하도록 요구한다. 또한 학생의 지식・이해, 과정・기능, 가치・태도를 통합적으로 평가하여 음악적 성장과 발달을 촉진하고 학습 향상을 추구하도록 제시하며, 학생을 평가 과정에 포함해 학습 성찰 기회를 제공하고 적절한 피드백으로 학습 의미를 재확인하게 해야 한다고 명시한다 (Ministry of Education, 2022).

이에 따라 본 연구에서 개발하고자 하는 LLM AI 평가도구의 가장 핵심적인 기능 중 하나로서 학습자 응답에 따른 '맞춤형 개선 피드백'이 포함되어 있다. 학생들은 단순히 서술형 평가에 대한 점수를 부여받는 것 뿐만 아니라 LLM AI의 평가를 통해 스스로에게 부족한 점은 무엇인지, 잘 작성한 부분은 무엇인지 객관적인 관점에서 맞춤형 언어 피드백을 부여받음으로써 평가가 갖는 중요한 가치인 '학습자 성장을 지원하는 평가'로서의 순기능을 올바르게 제공할 수 있도록 AI를 구성하는 것이 매우 중요하다 할 수 있다.

LLM 기반 자동채점은 두 가지 측면에서 교육적 의의를 갖는다. 첫째, 루브릭을 조건으로 하여 학생 응답을 빠르게 분석하고 채점・피드백을 산출할 수 있어, 평가가 점수 산출에 머무르지 않고 학습 개선을 지원하는 방향으로 작동하도록 돕는 보조 도구가 될 수 있다. 둘째, 2022 개정 교육과정이 강조하는 평가를 통한 '학생의 성찰'과 '적절한 피드백'은 결국 개별 학습자의 상태를 진단하고 다음 학습 행동을 제안하는 과정으로 볼 수 있는데 이에 AI 평가 모델은 각 학생을 위한 적절한 학습 개선 방안을 생성할 수 있다는 점에서 학습자의 성장을 지원하는 1:1 맞춤형 피드백 형태로 평가의 교육적 가치를 제고할 수 있다.

다만 이러한 가능성이 곧바로 '도입의 정당성'을 의미하지는 않는다. 앞서 언급한 것처럼 LLM 자동채점은 변동성, 관대 채점 경향, 비교가능성 문제 등 평가학적 쟁점이 보고되고 있으며, 교육정보를 다루는 기관 특성상 개인정보 보호와 현장 적용 적합성을 동시에 충족해야 한다. 본 연구는 음악교과의 다양한 수행평가 장면 가운데, 감상 및 비평 활동과 연결된 서술형 평가 상황을 예시 맥락으로 삼아 AI 평가도구의 가능성을 검토하였다. 특히 서양음악사 단원은 학습자의 개념 이해, 시대적 맥락 파악, 음악적 특징에 대한 서술이 요구된다는 점에서 루브릭 기반 서술형 평가의 적용 가능성을 확인하기에 적절한 사례로 보았으나, 본 도구의 구조 자체는 특정 학교급이나 특정 단원에 한정되지 않으며, 교사 또는 평가자가 문항과 루브릭을 구성하는 방식으로 다양한 학교급과 평가 상황에 적용될 수 있도록 설계하였다. 따라서 본 연구는 공개형 서비스 기반의 '범용 LLM 채점'이 아니라, 로컬 환경에서 구동되는 폐쇄 로컬 구조 OSS 기반 LLM AI 평가도구를 개발하고, 루브릭에 기반한 채점 결과가 어느 정도의 재현성과 일관성을 보이는지를 검토함으로써, 학교 현장 적용을 위한 기초 근거를 마련하는 것에 의의가 있다.

2022 개정 음악과 교육과정은 평가를 학습 과정에 통합하여 지속적으로 실시하고, 성

취기준을 준거로 공정성・타당성・신뢰성을 확보하도록 요구한다. 또한 학생을 평가 과정에 참여시키고, 학습 과정과 상황을 면밀히 파악해 적절한 피드백을 제공함으로써 학습 의미를 재확인하도록 제시한다. 이러한 평가 방향은 단순 채점의 자동화만이 아니라, 루브릭 기반의 진단과 피드백이 학습을 지원하도록 설계된 평가 도구의 필요성을 강화한다.

국내에서 LLM 기반 자동채점 연구는 우선 교사 채점과의 비교를 통해 가능성과 한계를 탐색하는 방식으로 전개되어 왔다. 예를 들어, 세계지리 서・논술형 문항을 대상으로 교사와 챗지피티(ChatGPT)의 채점 결과를 비교한 연구는 프롬프트에 평가기준만 제시한 조건보다 평가기준+예시답안 또는 평가기준+예시답안+비판적 채점 지시를 포함한 조건에서 교사 채점과의 상관이 높게 나타날 수 있음을 보고한다. 특히 문항 성격(정답이 비교적 명확한 문항 vs 채점자의 주관 개입이 큰 문항)에 따라 일치 양상이 달라질 수 있음을 함께 제시하여, LLM 채점의 활용은 문항 특성과 루브릭 설계・프롬프트 구성에 의해 좌우됨을 시사한다(Seong & Shin, 2023).

한편, GPT-4 기반 자동채점 시스템으로 초등 과학탐구활동보고서를 평가한 연구는 LLM 채점이 인간 채점에 비해 관대하게 채점되는 경향을 보일 수 있으며, 다국면 채점자 모형 분석에서 인간 채점 자료와 통합할 경우 비교가능성이 지지되지 않는 결과를 보고하였다. 이는 LLM 자동채점을 '인간 채점 대체'로 단정하기보다, 채점 편향 및 비교가능성 문제를 고려한 운영 원리(검수, 보조적 활용, 재현성 관리)가 필요함을 강조하는 근거로 해석될 수 있다. LLM이 제공하는 교육적 가치는 '점수 산출'보다 '피드백 생성'에서 더 직접적으로 드러날 수 있으며, 국내 연구들은 피드백의 즉시성, 확장성, 프롬프트에 따른 변동성을 주요 쟁점으로 보고한다. 고등학교 학생 글을 대상으로 챗지피티의 채점과 피드백을 점검한 연구는 챗지피티가 신속한 채점과 더불어 작문 맥락에 폭넓게 적용 가능한 피드백 제공 가능성을 제시하면서도, 피드백 생성의 변동성을 줄이기 위한 프롬프트 엔지니어링의 중요성을 강조한다(Choi & Park, 2023).

최근에는 모델 간 비교를 통해 자동채점과 피드백의 신뢰도 문제를 정량적으로 다루는 연구도 등장했다. 여러 생성형 AI 모델과 인간 채점자를 대상으로 논술 글의 반복 채점과 피드백 생성을 수행한 연구는 생성형 AI가 일정 수준의 점수・피드백 일관성을 보일 수 있으나, 모델에 따라 점수 편향과 피드백 일관성 부족이 나타날 수 있으며, 단독 활용보다는 교사의 보조 도구로 활용하는 접근이 타당하다는 결론을 제시하였다(Kim, 2025).

이상의 선행연구는 LLM이 루브릭에 기반으로 서술형 응답을 평가하고 피드백을 생성할 수 있음을 보여주면서도, (1) 프롬프트 조건에 따른 변동성, (2) 문항・영역에 따른 일치도 차이, (3) 관대 채점 및 비교가능성 문제 등 평가학적・운영적 과제를 동시에 제기한다.

또한 국내 연구의 다수는 ChatGPT/GPT-4 등 상용 모델 및 API 기반 환경에서 수행되었

으며, 루브릭 기반 채점의 안정성을 높이기 위해 교차 채점 또는 반복 판정 같은 전략이 제안되고 있다. 이는 학교 맥락에서 중요하게 제기되는 개인정보・평가자료 통제 요구와 연결될 때, 외부 전송을 최소화하는 폐쇄적 구조의 로컬 구동형 평가도구 필요를 뒷받침하는 간접적인 근거가 된다. 더 나아가 2022 개정 음악과 교육과정이 강조하는 '학습 과정 통합 평가'와 '적절한 피드백 제공'의 방향은 채점 결과의 재현성을 통제된 환경에서 검증하면서, 루브릭 정합적 피드백을 함께 산출하는 평가도구 개발 연구를 요구한다.

따라서 본 연구는 선행연구의 성과를 바탕으로, (1) 루브릭 기반 3수준(상・중・하) 분류 성능을 정량 평가하고, (2) 동일 데이터에 대해 답안 순서를 무작위로 섞는 반복 평가를 통해 채점 일관성을 점검함으로써, 2022 개정 교육과정의 평가 방향에 부합하는 로컬 기반 LLM 평가도구 개발・검증의 기초 근거를 마련하고자 한다.

2. 연구문제

이에 본 연구는 폐쇄 로컬 구조 OSS 기반 LLM 평가도구를 개발하고, 서양음악사 서술형 응답 데이터를 대상으로 루브릭 기반 분류 성능과 채점 일관성 및 재현성을 중심으로 다음의 연구문제를 설정한다.

RQ1. 루브릭 기반 LLM AI 평가도구는 서양음악사 서술 답안을 얼마나 정확하게 분류하는가?
RQ2. 서술 답안 DB의 무작위 배열 후 반복 채점 간 일치도는 어느 정도 수준인가?
RQ3. 루브릭 기반 LLM AI 평가도구의 판정 확신도(confidence)는 점수 산출 및 오류 사례와 어떤 관계를 보이는가?

3. 연구 범위 및 제한점

1) 연구 범위

본 연구는 2022 개정 음악과 교육과정이 제시하는 평가 방향, 즉 평가를 학습 과정에 통합하여 지속적으로 실시하고 성취기준을 준거로 공정성・타당성・신뢰성을 확보하며, 학습 과정과 상황을 파악해 적절한 피드백을 제공하도록 한 원칙을 출발점으로 한다. 또한 교육과정은 평가 결과를 교수・학습에 피드백으로 적용하여 수업 계획・방법・자료 개선에 활용하고, 학생의 역량 함양에 반영할 것을 제시한다.

이러한 방향에 맞추어 본 연구는 다음 범위에서 폐쇄 로컬 구조 OSS 기반 LLM 서술형

AI 평가도구를 개발・검토한다.

(1) 교과 및 내용 범위

본 연구의 정량 검증은 음악교과 내 서양음악사 관련 서술형 응답을 중심으로 수행하였다. 이는 감상 및 비평 활동과 연계된 서술형 수행평가 장면에서 학습자의 개념 이해와 설명 능력을 비교적 분명하게 확인할 수 있다는 점을 고려한 것이다. 다만 개발된 도구의 구조는 특정 학교급이나 특정 단원에 한정되지 않으며, 교사 또는 평가자가 평가 문항과 루브릭을 직접 구성함으로써 다양한 학교급과 교과 상황에 적용할 수 있도록 설계하였다.

(2) 평가 방식 범위

본 연구의 핵심 문항은 서술형이며, 교육과정에서 제시하는 다양한 평가 방법 가운데서・논술형 평가, 보고서/포트폴리오/프로젝트/토의・토론/온라인 평가 등 “다면적 평가 방법 활용”의 취지를 기술적으로 지원하는 방향으로 설계한다. 특히 “과정 중심 평가”의 실행과 학습 지원을 위해, 채점 결과와 함께 개선 피드백을 출력하는 기능을 포함한다.

(3) 도구(시스템) 범위

외부 API 의존을 최소화하기 위해 로컬 환경에서 구동되는 OSS LLM (20b) + 올라마(Ollama) + 파이썬(Python) 기반으로 평가도구를 구현한다. 구현 기능은 (가) 문항 입력, (나) 루브릭 자동 생성 및 사용자 수정/재생성, (다) 학생 답안 엑셀 업로드, (라) 루브릭 기반 상・중・하 평가, (마) 평가 근거 및 개선 피드백 생성, (바) 결과 엑셀 다운로드로 구성한다.

(4) 자료 및 성능 검증 범위

라벨(상・중・하)이 포함된 서술형 답안 200개 데이터를 대상으로 한다. 연구문제(RQ1)에 따라, 3수준 분류의 정확도(accuracy)와 매크로 평균 F1 점수(macro-averaged F1 score)를 산출한다. 또한 채점 일관성을 위해, 동일 데이터에 대해 답안 순서(및 응답자 순서)를 무작위로 섞어 반복 평가하고, 반복 간 평가 점수의 일치율/변동률을 산출한다.

2) 연구 제한점

본 연구는 “기초 AI 개발 및 성능 검증” 단계에 초점을 두므로, 다음의 제한점을 갖는다.

(1) 모델・버전・로컬 환경에 따른 재현성 한계

본 연구는 폐쇄 로컬 환경에서 모델과 파라미터를 통제하여 동일 입력에 대해 가능한 일관된 출력을 얻고자 한다. 그러나 LLM 기반 시스템은 모델 버전, 올라마 런타임, 하드웨어 및 추론 옵션에 따라 출력 특성이 달라질 수 있으며, 본 연구 결과는 본 연구에서 설정한 로컬 구동 환경과 파라미터 조건에 의존한다. 따라서 동일한 성능 지표가 다른 모델/버전/환경에서 그대로 재현된다고 단정하기 어렵다.

(2) 피드백 적절성 평가의 주관성 및 검증 범위

2022 개정 교육과정은 학습 과정과 상황을 파악해 적절한 피드백을 제공함으로써 학습 의미를 재확인하도록 제시한다. 본 연구는 이에 부합하도록 '채점+개선 피드백' 기능을 도구 설계에 포함하지만, 피드백의 "유용성"은 내용 적절성・구체성・실행 가능성 등 질적 요소가 포함되며, 평가자(교사) 관점의 평정 기준과 훈련, 평정 신뢰도 확보가 필요하다. 본 연구는 개발 단계의 성격상 피드백의 교육적 효과를 실험적으로 검증하기보다, 루브릭 기반 피드백 생성 가능성을 제시하고 향후 검증을 위한 평가틀을 논의하는 수준으로 제한한다.

(3) 검증 범위

본 연구의 정량 검증은 서양음악사 단원의 서술형 답안과 3수준(상・중・하) 및 10점 척도 분류를 중심으로 수행하였다. 교육과정이 제시하는 다양한 평가 방법을 모두 동일 수준으로 실증 검증하기에는 범위가 제한적이며, 다른 단원・학년・과제 유형으로의 일반화는 후속 연구에서 확장 검증이 필요하다. 반면 감상, 비평, 창작 의도 서술과 같이 음악적 가치 판단이나 해석, 표현의 다양성이 크게 개입되는 문항은 본 연구의 검증 범위에 포함되지 않았으며, 이러한 영역에서 동일한 수준의 루브릭 생성 및 채점 안정성이 확보되는지는 후속 연구를 통해 별도로 검토할 필요가 있다. 더불어 본 연구의 평가도구는 루브릭 기반 개선 피드백을 산출할 수 있도록 설계하였다. 다만 본 연구는 기초 개발 및 성능 검증 단계에 초점을 두고 있으므로, 피드백의 실제 교육적 효과나 활용 양상까지를 실증적으로 검토하지는 않았다. 이 부분은 향후 교사 평정, 학습자 반응, 수업 적용 사례를 포함한 후속 연구를 통해 보다 심층적으로 검토될 필요가 있다.

Ⅱ. 이론적 배경

1. 폐쇄 로컬 구조 OSS 기반 평가도구

학교에서 생성되는 평가 자료는 단순한 학습 기록이 아니라 학생의 성취 수준과 학습 이력을 포함하는 교육 정보이며, 실무적으로는 학생 식별 정보와 결합되어 처리되는 경우가 많다. 국내 논의에서는 교육기관이 개인정보처리자로서 개인정보 보호 관련 법령의 적용을 받는다는 점과 교육 영역에서 개인정보 처리의 특수성이 함께 다루어진다(Jung, 2021).

이러한 조건에서 외부 서비스에 의존하는 평가 시스템은 데이터 처리의 위탁, 보관, 이전과 같은 법・정책적 쟁점을 동반하기 쉽다. 클라우드 환경에서의 개인정보 보호에 관한 국내 연구들은 클라우드 이용이 구조적으로 제삼자 처리를 수반할 수 있으며, 그에 따라 보호 조치와 관리 체계가 요구된다는 점을 검토한다. 따라서 학교 단위에서 평가도구를 설계할 때에는 데이터가 외부로 전송되는 구조를 최소화하고 기관 내부에서 통제 가능한 실행 환경을 확보하는 접근이 필요하다(Kim, 2021).

또한 LLM 기반 채점은 동일 문항과 동일 루브릭을 사용하더라도 프롬프트 구성이나 운영 조건에 따라 결과가 달라질 수 있다는 점이 반복적으로 지적되어 왔다. 서술형 채점에 ChatGPT를 적용한 연구는 채점 기준과 예시답안 제시 방식 등 프롬프트 조건에 따라 교사 채점과의 관계가 달라질 수 있음을 보고한다(Seong & Shin, 2023). GPT 기반 자동채점 연구 역시 인간 채점과 비교했을 때 관대한 채점 경향과 비교가능성의 문제가 나타날 수 있음을 보여준다(Ham et al., 2024). 이와 같은 변동성과 편향 가능성은 운영 환경의 통제와 재현성 확보를 평가도구 설계 원리로 포함해야 함을 시사한다.

본 연구가 채택한 폐쇄 로컬 기반 OSS 도구는 개인정보 보호 요구에 부합하는 실행 구조를 제공하는 동시에, 모델 버전과 추론 파라미터를 고정하여 동일 입력에 대한 출력의 일관성을 점검할 수 있다는 점에서 연구 설계상 장점을 가진다. 이 맥락에서 폐쇄 로컬 구조 OSS 기반 AI 평가도구는 단지 기술 선택이 아니라, 학교 맥락의 제약을 고려한 평가 운영 모델로 이해할 수 있다.

2. 음악교육 서술형 LLM AI 평가도구

2022 개정 음악과 교육과정은 음악 교과 역량을 감성 역량, 창의성 역량, 자기주도성 역량, 공동체 역량, 소통 역량으로 설정하고, 삶의 맥락 속에서 연주, 감상, 창작 활동이 이루어지도록 교육과정을 구성한다(Ministry of Education, 2022). 이러한 구성은 학습자가 무엇을 알고 있는지의 확인을 넘어, 음악적 개념을 이해하고 이를 근거로 설명하거나

해석하며 자신의 사고를 구성해 표현하는 성취를 평가해야 한다는 요청으로 연결된다.

음악과 평가는 학습의 과정에 통합되어 지속적으로 이루어져야 하며, 성취기준을 준거로 공정성, 타당성, 신뢰성을 확보하고 수행 주체와 과정을 관찰하고 확인해야 한다는 방향을 제시한다. 또한 지식과 이해, 과정과 기능, 가치와 태도를 통합적으로 평가하여 학생의 음악적 성장과 발달을 촉진하고, 학생을 평가 과정에 참여시켜 학습을 성찰하게 하며 적절한 피드백을 제공해 학습한 내용의 의미를 재확인하도록 안내한다. 이 지점에서 서술형 응답은 학습자의 이해 수준과 사고 전개, 개념 적용의 흔적을 드러낼 수 있는 평가 산출물로 기능한다.

평가 방법 측면에서도 교사 평가, 자기 평가, 동료 평가 등 다양한 방법을 제시하며, 학습의 과정과 결과를 함께 평가하여 과정 중심 평가가 실행되도록 요구한다. 평가 결과는 교수학습으로 순환되어 수업 계획과 방법, 자료를 개선하는 데 활용되고 학생의 역량 함양에 반영되어야 한다는 점도 명시된다. 성적 산출에 머무르기보다 학습 지원 체계로 작동해야 한다는 점을 교육과정 차원에서 분명히 하고 있다.

한편 교육과정은 디지털 플랫폼을 활용한 수업 전략, 온라인 환경에서의 활동과 의견 공유, 인공지능융합 학습 등 디지털 기반 교수학습의 확장을 함께 언급한다. 이러한 환경에서는 학생 산출물의 형태가 다양해지고, 서술형 산출물과 피드백의 양도 증가하기 쉽다. 실제 음악교육 분야에서도 인공지능 기반 음악 창작 도구를 수업에 적용했을 때 학습자의 흥미, 자기효능감, 참여 태도 등이 변화할 수 있음을 보고하며, 교실 맥락에서 AI 기반 창작 수업의 교육적 가능성을 논의한다. 학습자 발달 수준에 적합한 자료를 구성하고, 교사가 주도적으로 학습자 경험을 설계할 필요가 있음을 제시하는 연구도 보고된다(Shin, 2025a). 이와 같은 논의는 음악교육에서 AI가 학습 자료와 학습 경험을 확장하는 도구가 될 수 있음을 보여주며, 평가 영역에서도 학습자 중심의 피드백 제공을 지원하는 도구 개발이 요구된다. LLM 기반 자동채점 연구는 대체로 루브릭을 명시적으로 제시했을 때 채점과 피드백이 구조화될 수 있다는 점을 보여준다. 그 중 임종한과 임종헌(Im & Im, 2024)의 실행연구는 교사가 설정한 루브릭에 따라 논・서술형 답안을 채점하고 평가 점수와 평가 의견을 제공하는 프로그램을 개발하였으며, 다수의 AI 평가교사를 구성해 교차 채점을 수행하는 방식으로 이러한 접근은 LLM을 단일 판정자로 두기보다 보조 평가 체계로 설계하는 방향을 제공한다.

다만 자동채점 결과가 교육평가로서 의미를 가지려면, 점수 산출의 타당성뿐 아니라 학습자의 학업을 지원할 수 있는 성장적 매개로서 개선 방안을 제안할 수 있어야 한다(Lee & Sohn, 2018). 쓰기 교육 맥락에서도 피드백의 효과를 메타분석으로 검토한 연구는 피드백이 의미 있는 효과를 가질 수 있으며 피드백 유형에 따라 효과가 달라질 수 있음을 보고한다(Baek, 2016). 이러한 결과는 LLM 기반 피드백을 설계할 때도 피드백을 단순한

총평으로 제시하기보다, 학습자가 수행할 수 있는 수준의 구체성과 기준과의 정합성을 확보해야 한다는 점을 시사한다(Shin, 2025b).

LLM 기반 피드백은 학습자에게 개별화된 개선 제안을 빠르게 제공할 수 있다는 점에서 과정 중심 평가의 취지와 연결될 수 있다. 2022 개정 음악과 교육과정이 학생을 평가 과정에 참여시키고 학습 과정을 파악해 적절한 피드백을 제공함으로써 학습 의미를 재확인하도록 제시한다는 점은, 평가도구가 피드백 기능을 포함할 필요성을 교육과정 차원에서 뒷받침한다. 다만 앞선 연구들이 보여주듯 LLM 채점은 관대 채점 경향이나 비교가능성의 문제를 가질 수 있으며, 프롬프트 조건에 따라 결과가 달라질 수 있다. 그러므로 교육현장 적용을 전제로 한 평가도구는 점수 계산을 가능한 한 결정론적으로 처리하고, 피드백 생성은 근거 기반으로 제한하며, 동일 조건 교차 검증에서 결과가 유지되는지 점검하는 방식으로 설계될 필요가 있다. 본 연구의 로컬 기반 OSS 평가도구는 이러한 설계 원리를 바탕으로 루브릭 기반 분류와 근거 제시, 개선 피드백 생성을 통합하여 구현한다.

Ⅲ. 연구방법

본 연구는 2022 개정 음악과 교육과정이 제시하는 평가 방향을 실제 학교 환경에서 구현할 수 있는 도구를 개발하고, 그 성능을 정량적으로 검토하는 개발・검증 연구로 설계하였다. 교육과정은 성취기준을 준거로 평가를 설계하고, 학습 과정과 상황을 파악하여 적절한 피드백을 제공함으로써 학습 의미를 재확인하도록 요구한다. 또한 과정 중심 평가가 실행되도록 다양한 평가 방법을 활용하고, 평가 결과가 교수・학습에 활용될 수 있도록 해당 점수의 산출 근거, 개선 사항을 자연어로 제공하여 수업 개선과 학생 역량 함양에 기여하도록 제시한다. 본 연구의 방법은 이러한 요구를 반영하여 루브릭 기반 채점과 개선 피드백 제공 기능을 포함한 평가도구를 로컬 환경에서 구현하고, 서양음악사 서술형 응답 데이터를 대상으로 분류 성능과 채점 일관성을 중심으로 분석하도록 구성하였다.

1. 연구 설계 개요

연구는 두 단계로 진행하였다. 첫째, 폐쇄 로컬 환경에서 구동되는 오픈소스 기반 LLM 평가도구를 구현하였다. 둘째, 도구가 산출한 채점 결과를 라벨이 부여된 데이터와 비교하여 분류 성능을 산출하고, 동일 데이터에 대해 반복적으로 채점을 수행하여 채점 일관성을 점검하였다. 연구문제 RQ1에 해당하는 성능 평가는 정확도와 매크로 평균 F1 점수를

핵심 지표로 삼았다. 채점 일관성 평가는 답안 순서를 무작위로 재배열하여 반복 채점했을 때 동일 답안이 동일 수준으로 분류되는지 확인하는 방식으로 설계하였다.

2. 시스템 개발 환경과 구현 구성

1) 개발 환경

본 연구의 평가도구는 로컬 컴퓨팅 환경에서 오픈소스 기반 LLM을 실행하는 구조로 설계하였다. 모델은 OSS 20b 계열을 사용하였고, 추론 서버는 올라마를 활용하였다. 애플리케이션은 파이썬 기반으로 구현하였으며, 사용자 인터페이스는 간단한 웹 기반 GUI를 제공하여 문항 입력, 루브릭 생성과 수정, 엑셀 업로드, 채점 실행, 결과 다운로드가 가능하도록 구성하였다.

2) 기능 구성

서술형 LLM AI 평가도구는 다음의 기능 흐름을 갖는다. (1) 문항 입력, (2) 루브릭 자동 생성 및 교사 수정・확정, (3) 답안 업로드, (4) 루브릭 기반 채점 및 근거 제시, (5) 개선 피드백 생성, (6) 결과 엑셀 파일 다운로드로 구성하였다.

이를 통해 첫째, 교사가 서술형 문항을 입력한다. 둘째, 교사가 선택한 척도에 맞추어 LLM이 루브릭 초안을 생성한다. 셋째, 교사는 생성된 루브릭을 그대로 사용하거나 수정한 뒤 확정한다. 넷째, 학생 응답이 포함된 엑셀 파일을 업로드하면 도구가 각 응답을 채점하고 개선 피드백을 생성한다. 다섯째, 결과는 엑셀 파일로 내려받을 수 있다.

이러한 구성은 교육과정이 제시한 평가의 다양성과 과정 중심 평가의 실행을 기술적으로 지원하는 방향과 연결된다. 교육과정은 실기 평가와 관찰 평가뿐 아니라 서술형과 논술형 평가, 보고서 평가, 포트폴리오 평가, 프로젝트 평가, 토의・토론 평가, 온라인 평가 등 다양한 방법을 적절히 활용하여 역량 함양 평가가 되도록 안내한다. 또한 학습에 대한 피드백 제공을 통해 음악적 성장이 지속되도록 각 학생별 맞춤형 학습 성취를 위한 개선 방안과 학업 지원을 고려할 수 있다.

3) 채점 로직 및 RBSS 산출 설계

본 연구의 채점 과정은 LLM의 주관적 편향(관대화 및 중간 점수 편향)을 통제하고 재현성을 확보하기 위해, LLM의 의미론적 분석과 결정론적 자연어 처리(Rule-based NLP)

지표를 결합한 하이브리드 아키텍처(Hybrid Architecture)로 설계하였다.

더불어 본 연구에서는 LLM 자동채점에서 보고되는 변동성 및 관대 채점 경향을 고려하여, 채점 과정에서 LLM의 역할을 '기준별 수준 판정과 근거 제시'로 제한하고, 점수 합산과 최종 수준 환산은 파이썬 로직이 결정론적으로 수행하도록 설계하였다. 또한 LLM 출력 형식의 불안정(설명 텍스트 후 JSON 출력, 필드명 변형, 누락)을 고려하여 JSON 추출・복구 및 검수 플래그(needs_review)를 포함하였다.

채점 로직은 크게 의미 추출 단계와 규칙 기반 객관화 단계로 분리된다. 첫째, LLM은 전체 점수를 직접 산출하지 않고, 학생 답안 내에서 루브릭 기준을 충족하는 인용구(Evidence)와 누락 요소(Missing parts)를 추출하는 역할로 제한된다. 둘째, 파이썬 기반의 후처리 시스템이 텍스트의 표면적・구조적 자질을 수치화한 RBSS(Rule-Based Scoring Signal)를 산출한다. RBSS는 서술형 답안의 적절성을 결정론적으로 추정하는 선형 결합(Linear Combination) 모델로, 다음과 같이 정의된다.

$$RBSS = \omega_1 S_{length} + \omega_2 S_{keyword} + \omega_3 S_{spec} - P_{vague}$$

각 지표와 가중치는 선행 테스트를 통해 다음과 같이 설정되었다. 첫째, 답안 길이 점수 ($S_{length}, \omega_1 = 0.35$)는 글자 수를 기준으로 상한과 하한을 설정하여 선형 보간(Linear Interpolation)한 값이다. 둘째, 키워드 일치도($S_{keyword}, \omega_2 = 0.40$)는 루브릭의 상위 기술어에서 추출된 명사 및 전문 용어가 학생 답안에 등장하는 비율이다. 셋째, 구체성 지표($S_{spec}, \omega_3 = 0.15$)는 서양음악사 단원의 특성을 반영하여 연도, 'BWV'와 'Opus' 같은 작품 번호, 고유명사 등의 정규식 패턴 매칭 횟수를 측정한다. 넷째, 불확실성 패널티(P_{vague}, 최대 0.30 감점)는 '잘 모르겠다', '헷갈린다', '아마도', '정확하진 않지만' 등 지식의 불확실성을 나타내는 표현의 빈도에 비례하여 감점하는 요소이다. 이러한 설계는 프롬프트나 출력 형식의 변동 문제로 인해 결과가 흔들리는 가능성을 구조적으로 최소화한다.

4) 하이브리드 종합 점수 산출 및 정규화 알고리즘

LLM이 산출한 메타데이터와 결정론적 RBSS를 종합(Ensemble)하여, 개별 학생 답안의 최종 종합 점수(Q)를 산출한다. 먼저, LLM의 평가를 수치화한 레거시 점수(S_{legacy})는 누락 요소 수(N_{miss}), 모델의 확신도(C_{LLM}), 근거 문장 수(N_{evid}), 그리고 답안의 이산적 길이 점수(S_{len_discr})를 종합하여 산출된다.

$$S_{legacy} = 0.50\left(1 - \frac{N_{miss}}{3}\right) + 0.25C_{LLM} + 0.20\min\left(1.0, \frac{N_{evid}}{2}\right) + 0.05S_{len_discr}$$

이후, LLM의 편향을 억제하기 위해 결정론적 신호인 RBSS에 70%의 가중치를 부여하여 최종 종합 점수 Q를 도출한다.

$$Q = 0.70 \times RBSS + 0.30 \times S_{legacy}$$

마지막으로, 모델이 모든 학생에게 유사한 점수를 부여하는 현상을 방지하고 교육학적 타당성을 확보하기 위해 순위 기반 분포 강제 정규화(Distribution-Forced Normalization)를 적용한다. 전체 학생을 종합 점수 Q의 내림차순으로 정렬한 뒤, 사전에 설정된 비율(상 30%, 중 40%, 하 30%)에 따라 등급 구간(Band)에 배정한다. 특정 구간에 배정된 학생 수를 N_{band}, 해당 구간의 최고점과 최저점을 각각S_{hi}, S_{lo}라고 할 때, 구간 내에서 r번째 순위($r \in \{0,1,...,N_{band}-1\}$)를 기록한 학생의 최종 정규화 점수($Score_{final}$)는 다음 수식에 따라 백분위수 매핑(Percentile Mapping) 방식으로 산출된다.

$$Score_{final} = S_{hi} - r \times \frac{S_{hi} - S_{lo}}{N_{band}}$$

이를 통해 동일한 수준으로 분류된 학생들 사이에서도 답안의 루브릭 준거 충족 정도의 차이에 따라 점수가 자연스럽게 분산되는 효과를 거둘 수 있으며, 채점의 신뢰도와 변별력을 동시에 확보할 수 있다.

채점 과정은 출력 안정성과 재현성을 높이기 위해 두 단계로 분리하였다. 첫 번째 단계는 기준별 수준 판정 단계이다. LLM은 루브릭의 각 기준에 대해 상・중・하 가운데 하나의 수준을 판정하고, 학생 답안에서 근거가 되는 문장 일부를 직접 인용 형태로 제시하도록 제한하였다. 두 번째 단계는 피드백 생성 단계이다. LLM은 확정된 수준 판정과 근거를 바탕으로 강점, 보완점, 다음 학습 행동을 자연어로 제시하도록 하였다. 점수 산출은 LLM이 임의로 계산하지 않도록 설계하였다. 각 수준에 매핑된 점수는 루브릭에 포함시켜 두고, 최종 점수 합산과 수준 환산은 파이썬 로직이 결정론적으로 수행하도록 구성하였다. 이를 통해 동일 입력에 대한 결과가 프롬프트나 출력 형식 문제로 흔들리는 가능성을 줄이고자 하였다.

3. 루브릭 설계와 적용

LLM AI가 생성하는 루브릭은 세 수준 체계를 기본으로 하였다. 각 기준은 학생 답안에서 관찰 가능한 수행을 중심으로 기술되도록 설계하였고, 학생의 강점과 보완점을 구체적으로 안내할 수 있도록 수준별 서술을 포함하였다. 교육과정은 학생의 학습 과정과 상황을 면밀히 파악하고 적절한 피드백을 제공하여 학습한 내용의 의미를 재확인하도록 요구하며, 배움이 느린 학습자 등을 고려한 개별 맞춤형 피드백 강화를 함께 제시한다. 따라서 본 연구의 루브릭과 피드백 생성 설계는 결과 제시를 넘어 학습 개선으로 연결되는 안내를 포함하도록 구성하였다.

루브릭 생성은 LLM AI를 통해 표 형태로 제안하도록 하였고, 교사(사용자)가 내용이 적합하지 않다고 판단할 경우 문항과 동일 척도 조건에서 루브릭을 재생성하거나 직접 텍스트를 수정하여 확정할 수 있도록 활용도를 높였다. 이는 학교 현장의 다양한 문항과 평가 의도를 반영하기 위한 설계이며, 이 AI 도구를 음악 교과에 한정하는 것이 아닌 타 단원과 타 교과에서도 사용할 수 있도록 하는 보편화 전략으로서 운용할 수 있다.

4. 데이터 구성

성능 검증에는 서양음악사 단원의 서술형 응답 데이터 200개를 사용하였다. 데이터는 상・중・하 라벨이 포함되어 있으며, 각 라벨은 문항의 요구 요소와 루브릭 수준 서술에 부합하도록 구성된 서술 응답으로 이루어져 있다. 데이터는 엑셀 기반으로 관리하였고, 한 행에 한 학생 응답이 입력되며 라벨은 별도 열로 저장하였다.

5. 평가 절차

1) 단일 실행 성능 평가

우선 200개 전체 데이터에 대해 루브릭 기반 채점을 수행하고, 도구의 예측 수준과 데이터 라벨을 비교하여 정확도와 매크로 평균 F1 점수를 산출하였다. 또한 수준 간 혼동 양상을 확인하기 위해 혼동행렬을 함께 산출하였다.

2) 무작위 재배열 반복 평가를 통한 채점 일관성 점검

채점 일관성 평가는 동일한 200개 답안을 대상으로, 응답 순서를 무작위로 재배열한 뒤

동일 조건에서 채점을 반복 수행하여 결과의 일치도를 산출하는 방식으로 진행하였다. 각 반복에서 모델과 추론 파라미터는 고정하였다. 교차 검증은 연구 수행 환경에서 안정적으로 운영 가능한 범위 내에서 설정하였으며, 교차 검증 시 같은 응답이 동일 수준으로 분류되는 비율을 일치율로 산출하였다. 더불어 교차 검증을 통한 정확도와 매크로 평균 F1 점수의 변동 폭을 함께 산출하여 결과 안정성을 점검하였다.

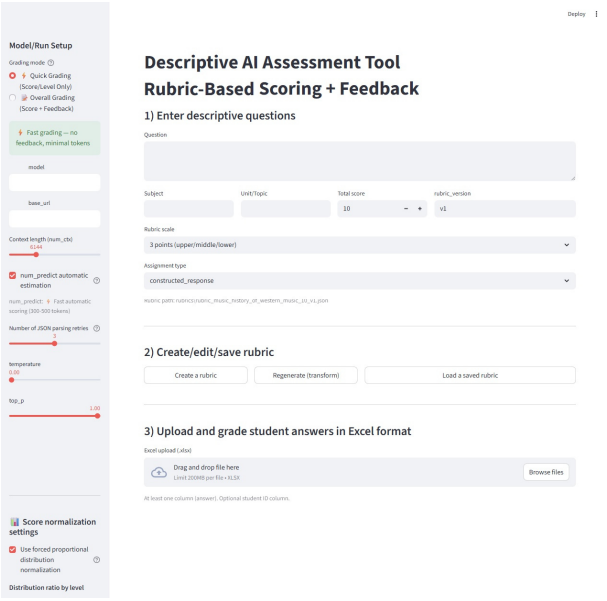
6. 피드백 산출과 탐색적 검토 범위

본 연구는 루브릭 기반 채점 결과를 바탕으로 개선 피드백을 생성하도록 도구를 설계하였다. 교육과정은 학생이 잘하는 점과 보완해야 할 점을 세심하게 파악하고, 목표 도달 여부와 유의미한 향상을 구체적으로 안내하여 학습 개선과 성장의 기회로 삼도록 제시한다. 이에 따라 본 도구는 강점, 보완점, 다음 학습 행동을 포함하는 피드백 텍스트를 출력한다.

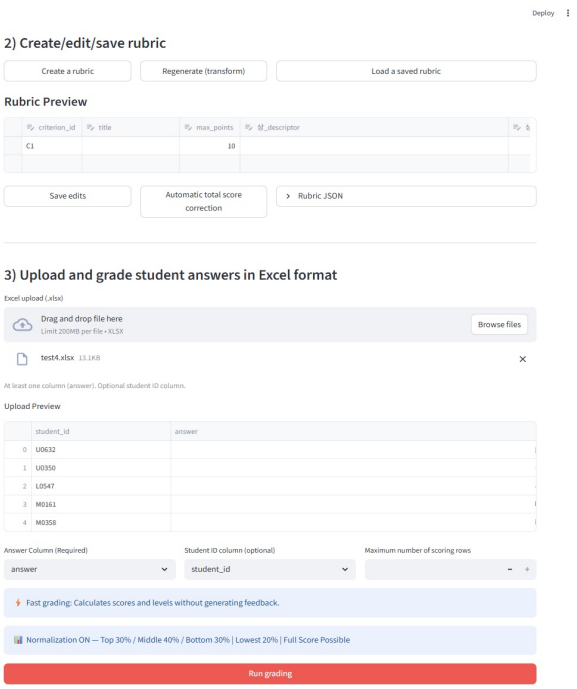
다만 연구문제 RQ3은 탐색적 논의로 두었다. 따라서 피드백의 교육적 효과를 실험적으로 검증하기보다는, 도구가 루브릭과의 정합성을 유지한 채 피드백을 산출할 수 있는 구조를 구현하고, 향후 교사 평정이나 내용 분석을 통해 검토할 수 있는 분석 틀을 논의하는 범위로 제한하였다.

7. 폐쇄 로컬 OSS 기반 LLM AI 개발의 실제

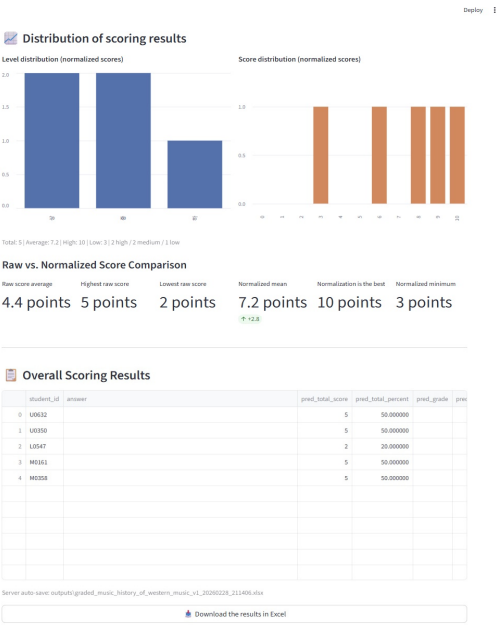
데이터 전처리와 분석은 파이썬 기반으로 수행하였다. 정량 지표 산출에는 분류 성능 평가가 가능한 분석 라이브러리를 사용하였고, 결과는 엑셀 표와 그래프 형태로 정리하였다.



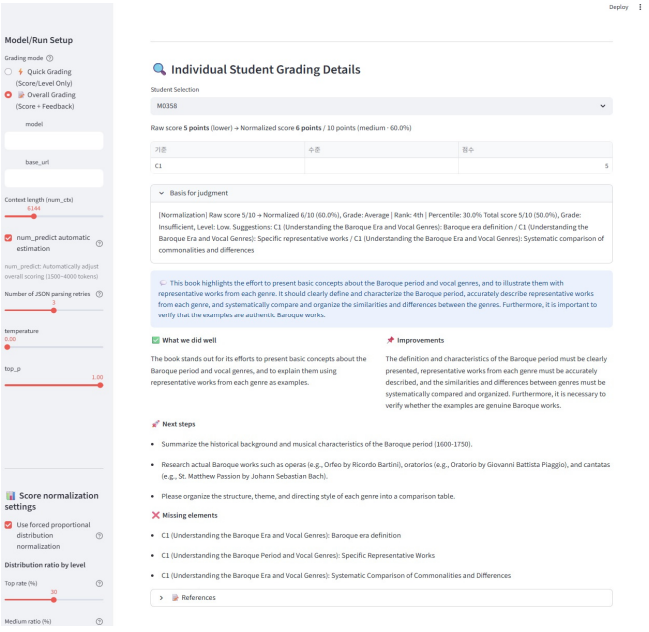
[Figure 1] Graphical User Interface (GUI) of the developed LLM-based AI evaluation tool



[Figure 2] AI-driven automatic rubric generation based on the selected subject and inputted descriptive question



[Figure 3] Chart visualization of normalized score distribution and Excel export functions following the evaluation of descriptive response data



[Figure 4] Personalized feedback for improvement based on learners' descriptive responses

8. 윤리와 데이터 처리

본 연구의 시스템은 로컬 환경에서 구동되며 학생 응답과 산출물은 외부 서버로 전송되지 않도록 설계하였다. AI 도구의 모든 개발 과정에서 연구자가 직접 작성한 파이썬 코드를 중심으로 구성되었으며, 웹 클라우드 기반 LLM AI의 경우 개인정보 유출 등의 통제 불가한 환경을 고려하여, 본 연구는 평가 자료의 폐쇄적 로컬 구조의 안전한 처리와 보관을 전제로 AI 평가도구를 설계하였다.

Ⅳ. 연구결과

1. 서술형 응답 1회차 및 2회차 교차 검증 AI 평가 결과 요약

본 연구는 서양음악사 서술형 응답 200건을 대상으로, 동일한 평가도구를 첫 번째 원 평가(1회차)와 두 번째 응답 순서를 무작위로 재배열한 조건(2회차 교차검증)에서 각각 1회씩 적용하였다. 두 실행 결과는 동일한 응답(학생 ID 기준)을 기준으로 매칭하여 비교 분석하였다. 분석에는 정규화 점수 및 관련 지표(norm_score, norm_percent, score_rank, score_percentile, quality_score)와 판정 확신도와 RBSS를 중심으로 사용하였다.

이때 교차 실행은 “동일한 답안이라도 처리 순서 변화만으로 결과가 달라지는지”를 점검하기 위한 절차로서, 평가도구의 채점 안정성 및 재현성을 검토하기 위한 실험 조건에 해당한다.

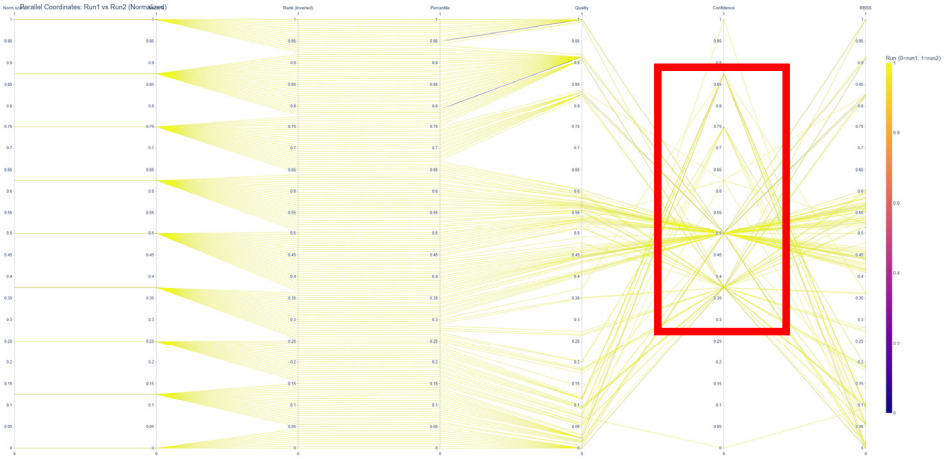
2. 병렬좌표도(parallel coordinates plot) 기반 다변량 패턴 분석

1) Run1-Run2 전체 패턴

[Figure 5]은 Run1과 Run2의 결과를 동일한 축 집합에 정규화하여 병렬좌표(Parallel Coordinates)로 시각화한 것이다. 각 선은 개별 응답(학생) 하나를 의미하며, 축은 정규화 점수(norm_score), 정규화 백분율(norm_percent), 점수 순위(score_rank), 점수 백분위(score_percentile), 품질 점수(quality_score), 판정 확신도(confidence), RBSS로 구성된다.

시각화 결과, 대부분의 선들이 Run1과 Run2에서 유사한 위치를 통과하는 경향을 보였으며, 특히 점수 체계의 핵심인 정규화 점수 순위(norm score-rank)와 백분위 품질 점수(percentile-quality score) 축에서 큰 형태 변화 없이 동일한 흐름이 반복되는 양상이 확인되었다. 이는 응답

순서를 무작위로 바꾸더라도 시스템이 산출하는 정규화 점수와 순위 구조가 전반적으로 유지되었음을 시사한다.



[Figure 5] Parallel coordinates plot: Run1 - Run2 (Normalized)

2) 판정 확신도 축에서 관찰되는 '0.5 수렴' 현상과 해석

병렬좌표도(Parallel plot)에서 특히 주목되는 지점은 '판정 확신도(confidence)' 축의 분포 형태이다. [Figure 5]에서 판정 확신도 값이 0.75-0.35 구간에 주로 분포하면서, 다수의 선이 0.5 주변으로 모이는 수렴 패턴을 확인할 수 있다. 즉, 도구가 산출한 확신도는 과도하게 높게 치우치거나 '항상 확신' 또는 과도하게 낮게 쏠리는 '항상 불확실(uncertainty)'의 형태가 아니라, 중간 수준의 확신도를 중심으로 분포하는 양상을 보였다.

이러한 분포는 운영적 관점에서 다음의 의미를 갖는다. 판정 확신도는 단순한 정확도라기보다 모델의 자기보고형 불확실성(self-reported uncertainty)에 해당한다. 판정 확신도가 지나치게 1.0에 치우치면 경계 사례에서 평가 준거인 '루브릭을 무시'하고 AI 스스로 평가적 주체가 되어 과신이 발생할 수 있고, 지나치게 낮으면 마찬가지로 특정 응답자의 서술 답안이 루브릭을 기반한 판정에 어려움이 발생한 상황으로 볼 수 있다. 본 데이터에서 판정 확신도가 0.5 중심으로 형성된 것은, 모델이 전반적으로 루브릭 기반 판정을 수행하면서도 '모든 사례를 극단적으로 단정'하는 방향으로 치우치지 않았음을 시사한다. 동시에 판정 확신도의 수준은 '요구의 명료성'과 같은 서술 문항의 구조와 루브릭 수준 기술의 구체성/배타성에 의해 영향을 받으므로, 후속 연구에서는 문항・루브릭 설계 변인을 달리하여 판정 확신도 분포 변화도 함께 비교할 필요가 있다.

이를 연구 맥락에서 해석하면, 판정 확신도는 단순한 '정답률'이라기보다 LLM이 자신의 판단을 얼마나 강하게 주장하는지에 가까운 판정의 주관적 확신 수준을 반영하는 지표로 볼 수 있다. 서술형 평가에서 이상적인 평가자는 (1) 과도한 확신으로 루브릭의 경계 사례를 단정하기보다, (2) 루브릭에 근거해 판단하되 애매함을 적절히 인정하며, (3) 필요한 경우 검토 대상으로 남겨 두는 균형적 태도가 요구된다. 이러한 관점에서 판정 확신도 값이 전반적으로 0.5 근방에 형성되는 경향은, 모델이 루브릭 기반 판정을 수행하면서도 극단적 확신으로 치우치지 않는 방향으로 작동했음을 시사한다.

또한 판정 확신도는 '서술 문항의 표현 방식'과 '루브릭 설계의 명료성'에 크게 영향을 받는다. 예컨대 루브릭의 수준 기술이 상호 배타적이고 관찰 가능한 수행으로 구성되면, 경계가 분명해져 판정 확신도가 높아질 가능성이 있다. 반대로 수준 기술이 포괄적이거나 응답의 다양성이 큰 경우에는 판정 확신도가 0.5 주변에 머무르며 경계 판단을 유보하는 양상이 나타날 수 있다. 본 연구 결과에서 확인된 0.5 중심 패턴은, 본 연구의 서술 문항-루브릭 조합이 무리한 확신을 유도하지 않고 극단적인 확신값에 편중되지 않는 분포 양상이 관찰되었다.

3) RBSS 축의 분할 구조와 '실질 판단 신호'로서의 의미

[Figure 5]의 RBSS 축은 판정 확신도와 달리, 구간이 비교적 명확하게 분할되는 양상을 보인다. RBSS는 LLM 호출 없이 답안 텍스트로부터 길이, 핵심 용어 포함, 구체성 패턴, 불확실 표현 패널티 등을 종합하여 계산되는 결정론적 신호이며, 본 도구에서 품질 점수(quality_score)의 핵심 구성요소로 설계되었다.

병렬좌표도에서 RBSS가 적절히 분산・분할된 형태를 보인다는 것은, 답안 텍스트의 '관찰 가능한 특징'에 따라 적절성 신호가 변별력을 갖고 산출되었음을 의미한다. 이는 서술형 응답의 평가에서 흔히 문제되는 '모두가 중간으로 몰리는 현상'을 완화하는 장치로 작동하며, 결과적으로 정규화 점수(norm_score)와 순위(score_rank, score_percentile)가 RBSS 및 품질 점수와 일관된 방향으로와 일관된 방향으로 연결되는 것을 가능하게 한다.

4) '판정 확신도가 높을 때 낮은 점수' 사례의 해석

본 연구 결과에서 인상적인 패턴 중 하나는, 일부 응답에서 판정 확신도가 상대적으로 높음에도 낮은 norm_score(낮은 RBSS/quality_score)로 판단된 선들이 존재한다는 점이다. 이는 표면적으로 '확신이 높으면 점수도 높아야 한다'는 직관과 다르게 보일 수 있으나, 서술형 평가에서는 오히려 의미 있는 현상으로 볼 수 있다.

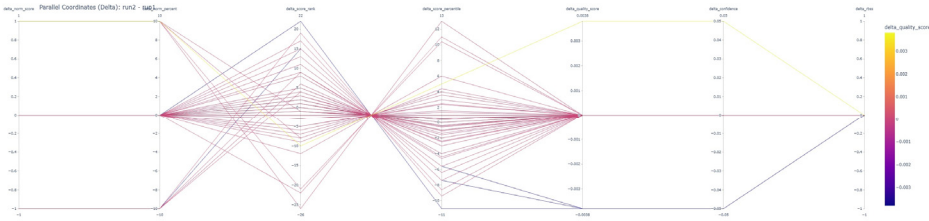
즉, 모델이 '잘 작성한 답안'에 대해서만 확신하는 것이 아니라, 문항 요구와 루브릭 기준을 명확히 충족하지 못한 응답에 대해서도 '결함이 분명하다'고 정확히 판단할 때 확신도가 높아질 수 있다. 이 경우 판정 확신도는 긍정적 판단의 강도가 아니라, 판정의 단정성(명확성)을 나타내는 신호가 된다. 따라서 판정 확신도가 높은데 낮은 점수로 연결되는 사례는, 시스템이 답안의 문제점을 발견하고 이를 점수에 반영했다는 해석이 가능하며, 이는 루브릭 기반 준거 평가가 의도하는 '기준 충족 여부의 명료한 반영'과 부합한다.

다만 이 해석이 타당하려면, 해당 사례들이 실제로 (1) 루브릭 핵심 요소 누락, (2) 개념 오류, (3) 근거 부재, (4) 모호한 표현 과다 등으로 인해 낮게 채점될 만한 이유를 갖는지 질적 점검이 필요하다. 본 연구에서는 검토 필요 플래그(needs_review)와 함께, 후속 연구에서 해당 사례군에 대한 교사 검토를 결합하는 방식으로 타당도 검토를 확장할 수 있다.

3. Delta Parallel Coordinates Plot을 통한 변동성 분석

[Figure 6]는 Run1-Run2의 차이(⊿)를 병렬좌표로 표현한 결과이다. 차이(⊿) 병렬좌표도는 '절대 점수 수준'이 아니라, 지표가 교차 실행에서 '얼마나 변화했는지'를 시각적으로 보여준다.

[Figure 6]에서 대다수 선이 0 근방에 밀집하는 양상이 나타났으며, 이는 1회차와 2회차 사이에서 정규화 점수(norm_score), 점수 백분위(score_percentile), 품질 점수(quality_score), RBSS 등 핵심 지표가 대체로 큰 변동 없이 안정적으로 유지되었음을 시사한다. 일부 응답은 순위(rank) 및 백분위(percentile) 축에서 상대적으로 큰 변화가 관찰되었는데, 이는 정규화 방식이 본질적으로 '순위 기반 배분'을 포함하기 때문에, 품질 점수 변화의 폭은 대체로 0에 수렴하여 나타났으나, 순위 기반 정규화 구조에서는 상당히 적은 차이에도 일부 구간에서 순위 교환(swap)으로 증폭되어 나타날 수 있었다. 따라서 ⊿ plot은 시스템이 불안정하다는 증거라기보다, 순위 기반 정규화가 갖는 민감도(특히 동점 또는 근접 점수 구간에서)를 보여주는 결과로 해석할 수 있다.

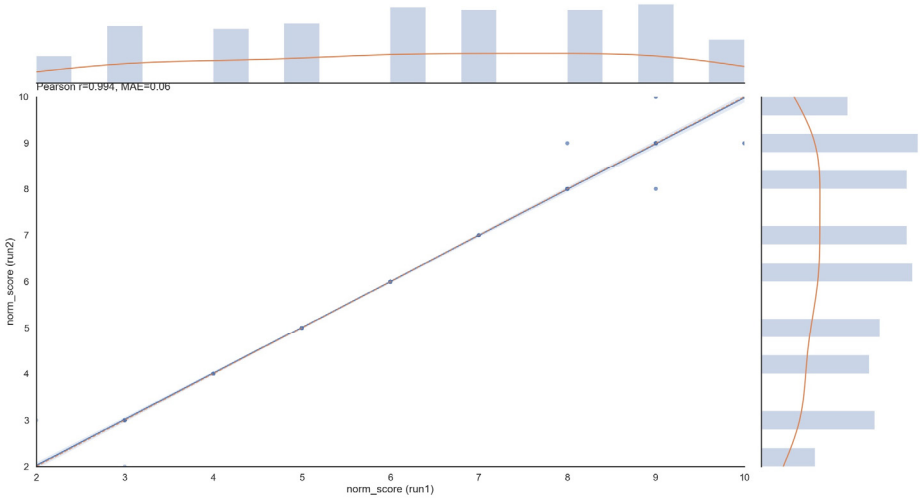


[Figure 6] Delta parallel coordinates plot: Run1 - Run2

4. Jointplot 기반 점수 일치도 분석

[Figure 7]은 Run1의 정규화 점수와 Run2의 정규화 점수 간 관계를 산점도와 주변 분포로 나타낸 것이다. 산점도가 대각선(y = x) 주변에 밀집하고, 상관계수(Pearson *r*)가 매우 높게 나타난 것은 동일 응답이 교차 실행에서도 유사 점수를 부여받았음을 의미한다. 주변 분포 또한 유사한 형태로 나타나 Run1과 Run2의 점수 분포가 크게 달라지지 않았음을 확인할 수 있다. 이는 Run1의 DB와 Run2의 DB의 나열 방식을 서로 다르게 무작위로 섞고 다시 평가했음에도 AI는 동일한 서술 답안에 같은 점수를 부여했음을 시사한다.

하지만 본 연구의 논의 목적상, 단일 지표의 상관보다도 [Figure 5]과 [Figure 6]에서 확인되는 다변량 관계(quality_score-confidence-RBSS-norm_score의 연결 구조)가 평가도구 설계의 핵심 작동 원리를 더 직접적으로 보여주므로, 본 장에서는 parallel plot 기반 결과를 중심으로 해석하였다.



[Figure 7] Jointplot with Marginal Distributions: norm_score: Run1 - Run2

연구 결과를 정리하면 다음과 같다. 첫째, 가장 눈여겨볼 점으로 [Figure 5]에서 개발된 LLM AI 평가도구의 판정 확신도는 0.5를 중심으로 분포하여, 과도한 확신 또는 과도한 불확실성으로 치우치지 않는 균형적 판단 양상을 보였으며, 이는 서술 문항과 루브릭 설계가 모델의 주관성 표출을 어떻게 제어하는지 보여주는 결과로 해석된다. 둘째, RBSS는 텍스트 기반 결정론적 신호로서 변별력을 갖게 분산되었고, 정규화 점수와 순위 구조가 RBSS · quality_score 축과 일관되게 연결되면서 집단 분포 통제와 변별을 동시에 달성하

는 경향이 확인되었다. 셋째, 판정 확신도가 높은데 낮은 점수로 판정된 사례는, 모델이 결함이 분명한 응답에 대해서도 단정적으로 판단할 수 있음을 시사하며, 루브릭 준거에 기반한 결함 탐지 신호로서 판정 확신도의 역할을 제안함으로써 이를 해석할 수 있다.

V. 결론

이 연구는 학교 현장 적용을 전제로 폐쇄 로컬 구조 OSS 기반 LLM 서술형 평가도구를 개발하였다. 2022 개정 음악과 교육과정의 '과정 중심 평가'와 '학습 지원을 위한 피드백' 요구를 구현하기 위해, 문항 입력-루브릭 편집-채점-근거 제시-개선 피드백 산출을 하나의 수행 흐름으로 통합하였다. 이 도구는 음악교육 현장에서 교사가 다수의 서술형 응답을 일정한 준거에 따라 검토해야 하는 부담을 보조할 수 있다는 점에서 실천적 의미를 갖는다. 특히 감상 및 비평 활동과 연결된 서술형 수행평가 장면에서, 학습자의 이해 수준을 일정한 기준 아래 구조화하여 파악하고, 그 결과를 바탕으로 성장지원적 보완 방향을 제시할 수 있다는 점에서 교육적 가치를 갖는 LLM AI 평가 모델의 의의를 갖는다.

2022 개정 교육과정에서의 수행평가 또는 서술형 문항의 평가 방향은 평가를 단순 점수 산출로서 바라보는 것이 아니라 학습자의 학업을 지원하는 피드백 체계로 확장하여 이해할 수 있다. 이상적인 평가는 학생의 음악적 성장과 발달을 지원할 수 있도록 평가 내용과 단계를 계획하고, 학습 과정과 상황을 면밀히 파악하여 적절한 피드백을 제공함으로써 학습한 내용의 의미를 재확인하도록 안내한다. 본 연구가 설계한 AI 평가도구는 이러한 교육평가의 교육 지원적 지침을 구현하기 위해 서술형 문항 입력과 루브릭 생성 및 편집 기능을 제공하고, 학생 응답 엑셀 업로드 이후 루브릭 기반 수준 판정과 근거 제시, 맞춤형 개선 피드백 산출까지 하나의 흐름으로 연결하였다. 특히, 점수 합산과 최종 점수 산출은 결정론적 파이썬 로직으로 처리하고 LLM AI의 역할을 '기준별 판정・근거 추출・피드백 생성'으로 제한함으로써, 변동성과 관대 채점 경향을 운영 차원에서 통제하는 설계를 제안하였다.

학교 현장에서 개인정보 보호를 위해 교내 데이터의 외부 전송을 최소화해야 한다는 요구는 AI 평가도구 설계의 기술적 선택을 넘어 운영 원리로 다루어질 필요가 있다. 웹 클라우드 컴퓨팅 환경에서는 서비스 제공자가 이용자 개인정보를 보관하고 처리하는 구조가 발생할 수 있다. 본 연구가 제안하는 폐쇄 로컬 구조 AI 평가도구는 평가 자료가 외부로 유출되는 경로를 구조적으로 줄이고, 교육 기관의 주체인 학교가 통제 가능한 환경에서 평가와 피드백을 수행할 수 있도록 한다는 점에서 현장 적합성을 강화하는 방향으로 AI 평가도구를 위치시킬 수 있다.

이를 통해 본 연구는 2022 개정 음악과 교육과정이 요구하는 과정 중심 평가와 학습자 맞춤형 피드백의 방향을 기술적으로 지원하는 한 가지 AI 구현 모델을 제시하였다. 다만 본 연구의 핵심 성과는 LLM AI 평가도구의 학습자 응답 분류 성능 수치 자체에만 있는 것이 아니라, 폐쇄 로컬 구조 OSS 기반 운영, 루브릭 기반 판정, 근거 기반 피드백이라는 객관적인 서술형 평가 설계 원리를 통해, 학교 현장에서 요구되는 교사들의 어려움과 교육 피로도를 낮출 방안을 모색하고 교육 현장의 제약을 개선할 수 있는 하나의 대안을 구체화하여 제안하고, 교차 검증 절차를 통해 타당도 높은 AI 평가도구를 개발하는 일련의 과정을 통해 교육 현장의 어려움 개선에 해당 연구가 기여할 수 있기를 바란다.

References

Baek, J. P. (2016). Meta-analysis of feedback effects in Korean writing education. *Urimal Yeongu, 44*, 251-281. https://doi.org/10.18110/urimal.2016..44.251

Choi, S. K., & Park, J. I. (2023). Exploring the utilization of ChatGPT for composition assessment in the era of artificial intelligence. *Journal of CheongRam Korean Language Education, 95,* 65-109. https://doi.org/10.26589/jockle..95.202309.65

Ham, E. H., Park, S. Y., Lee, B. Y., Lee, S. H., Lee, Y. G., & Hong, Y. J. (2024). Characteristics of GPT-4 automated scoring of scientific inquiry competency. *Journal of Educational Information and Media, 30*(3), 713-742. https://doi.org/10.15833/KAFEIAM.30.3.713

Im, J. H., & Im, J. H. (2024). Development and implementation of an AI-based automatic scoring program for constructed-response items using the ChatGPT API: An action research study. *Journal of Educational Innovation Research, 34*(1), 349-370. https://doi.org/10.21024/pnuedi.34.1.202403.349

Jung, S. W. (2021). Problems and improvement plans of personal information protection laws and regulations in the education sector. *Journal of Law and Human Rights Education, 14*(1), 25-54. https://doi.org/10.35881/HLER.2021.14.1.2

Kim, G. J. (2025). ChatGPT vs. Gemini: The potential and limitations of generative AI in automated scoring and feedback. *Journal of CheongRam Korean Language Education, 103*, 195-225. https://doi.org/10.26589/jockle..103.202501.195

Kim, H. S. (2021). A legal and policy review of personal information protection in cloud computing. *Journal of Law and Policy Research, 21*(4), 93-122. https://doi.org/10.1792

6/kaolp.2021.21.4.93

Lee, B. N., & Sohn, W. S. (2018). Meta-analysis of feedback effects: Differences by feedback, learning tasks and learner characteristics. *Journal of Educational Evaluation, 31*(3), 501-529. https://doi.org/10.31158/JEEV.2018.31.3.501

Ministry of Education (2022). *2022 Revised National Curriculum*. Ministry of Education.

Seong, J. W., & Shin, B. C. (2023). Exploring the feasibility of automatic scoring of written test using ChatGPT: Focusing on the world geography written test. *Journal of the Association of Korean Geographers, 12*(3), 415-432. https://doi.org/10.25202/JAKG.12.3.3

Shin, G. H. (2025a). The effects of AI-based create music education on elementary students' perceptions. *Korean Journal of Teacher Education, 41*(4), 233-249. https://doi.org/10.14333/KJTE.2025.41.4.11

Shin, G. H. (2025b). Exploring teaching-learning strategies for early childhood musical development through AI-generated children's songs: Focused on the 2019 Revised Nuri Curriculum. *Korean Journal of Arts Education, 23*(4), 475-492. https://doi.org/10.23317/kaae.2025.23.4.025