교육과정평가연구
The Journal of Curriculum and Evaluation
2026, Vol. 29, No. 2, pp.275~303
DOI: https://doi.org/10.29221/jce.2026.29.2.275

학생부종합전형 서류평가의 AI 예비평가 모델 설계 및 탐색적 다중 모델 검증

황희돈 (명지대학교 직원)*

요약

학생부종합전형 서류평가는 평가자의 정성적 판단에 기반하므로, 평가자의 경험과 대학별 평가 환경에 따른 판단 편차는 평가의 일관성 및 공정성과 밀접하게 관련된다. 이를 완화하려면 대학 고유의 평가 기준을 일관되게 적용할 보조 도구가 필요하다. 이에 본 연구는 입학사정관의 평가를 보조하는 AI 도구가 대학 고유의 평가 기준을 충실히 구현하여 평가의 구성 타당성을 확보하고, 아울러 평가자 간 편차까지 완화할 수 있는지를 탐색적 다중 모델 비교로 검증하였다. 5개 대학 공동연구(임진택 외, 2022)의 평가항목을 바탕으로 AI 예비평가 모델(A: LLM 자체 채점 기준, B: 사실 추출 레이어, C: 루브릭 없는 통제군, D: 전문가 기준 부호화)을 구성하고, 30건의 모의 학생기록을 네 모델과 서로 다른 대학 전문가 2인이 독립 평가하였다.

분석 결과, 범용 모델은 순위 합의(Kendall W=.859)는 양호하였으나 등급 분류(Fleiss κ=.165)에서는 낮은 일치도를 보였다. 서로 다른 대학 전문가 간 일치도는 κ=.397이었으며, 불일치는 모두 인접 등급 차이에 머물러 평가 기준 차이를 반영하는 교차 기관 평가 기준선임을 보여준다. 모델D는 전문가A의 판단을 κ=.648 수준으로 재현하였고, 전문가B와의 일치도는 κ=.261이었으나, 두 전문가 판정이 일치한 18건에서는 77.8%의 일치율로 네 모델 중 가장 높았다.

평가 기준의 부호화는 대학별 기준 적용을 전제로 유효하며, 사실 추출 레이어는 AI 워딩 과다 기록의 사실성 저하를 효과적으로 감지하였다. AI 기반 평가 보조 도구는 수험생을 평가자·연도 편차로부터 보호하고, 실제 노력과 성장을 보인 학생이 가려지지 않도록 하는 공정성 강화에 기여할 수 있다.

주제어: 학생부종합전형, AI 서류평가, 전문가 기준 부호화, 구현 충실도, 수렴 타당도

* 단독저자, don2020@mju.ac.kr

I. 서 론

1. 연구의 필요성

학생부종합전형(이하 학종)은 2027학년도 기준 전체 대학 모집인원의 약 23.7%를 차지하며, 수시모집 내에서는 학생부교과전형에 이어 두 번째로 큰 비중을 차지한다(한국대학교육협의회, 2025). 학종은 학교생활기록부를 중심으로 지원자의 학업역량, 진로역량, 공동체역량 등을 종합적으로 평가하는 전형으로, 표준화된 점수가 아닌 정성적 기록에 대한 전문가 판단을 요구한다. 이러한 규모의 정성평가가 매년 반복되는 만큼, 평가의 일관성과 신뢰성 확보는 전형의 공정성을 좌우하는 핵심 과제다.

학종 서류평가에는 세 가지 구조적 문제가 존재한다. 첫째, 평가자 간 일관성 문제다. 류영철(2016)은 대입 전형에서 동일 지원자에 대한 평가자 간 점수 편차가 크게 나타남을 보고하였으며, 평가자 내 비일관성과 결합되어 지원자의 합격 여부가 배정된 평가자에 의해 좌우될 수 있다는 공정성 우려가 제기된다(노성, 2021; 양림, 2020). 더욱이 각 대학 입학사정 업무에는 매년 신규·전보·위촉 평가자가 유입되며, 정성평가 기준을 조직 내부에서 상세히 공유하더라도 평가자 개인의 경험·휴리스틱 차이는 훈련만으로 완전 균일화되기 어렵다. 결과적으로 동일 대학 지원 수험생조차 '기관 내·연도 간 평가자 변이(within-institution, across-year evaluator variance)'에 노출되어 평가 품질의 예측 가능성에 일정한 위험을 지닌다. 둘째, AI 활용 워딩(이하 AI 워딩)의 증가다. 본 연구에서 AI 워딩이란 구체적 사실이나 경험의 기술보다 추상적·수식적 표현의 비율이 과도하게 높아진 기술 양태를 의미한다. 예컨대 '탁월한 리더십을 발휘하였음'은 학생부 전반에 구체적 활동이 뒷받침될 때는 정당한 평가이지만, 관련 활동 기록 없이 미사어구로만 기재되면 사실 기반이 결여된 표현에 해당한다. 최근 생성형 AI의 보편화로 이러한 양태의 빈도 증가 가능성이 있다. 셋째, 대학 간 평가 기준의 이질성 문제이다. 각 대학은 고유한 인재상과 학과 특성에 따라 서로 다른 평가 기준을 운용하며, 동일 지원자가 대학에 따라 상이한 평가를 받을 수 있다.

한편, 2022 개정 교육과정 기반 2028 대학입시제도 개편(교육부, 2023)은 내신 9등급제의 5등급 상대평가 전환과 학생부 정성평가 확대를 예고하였다. 고교학점제 시행으로 학생별 교과 선택이 개별화되고 내신 변별력이 축소되는 상황에서 정성적 서류평가의 중요성은 더욱 높아지고 있다. 5개 대학 공동연구(임진택 외, 2022)는 학종 공통 평가요소로 3개 역량 10개 항목 프레임워크를 도출하여 대학 간 표준화 기반을 마련하였으나, 이를 대규모 평가 현장에서 일관되게 적용하기 위한 기술적 지원 방안은 충분히

탐색되지 않았다.

이러한 맥락에서 본 연구는 LLM(Large Language Model)을 활용한 AI 예비평가 모델을 설계하고, 범용 모델과 전문가 기준 부호화 모델의 비교를 포함하여 그 타당성과 신뢰성을 탐색적으로 검증함으로써 학종 서류평가 보조 도구로서의 가능성을 탐색한다.

2. 연구 목적 및 연구 문제

본 연구의 목적은 공통 평가요소 기반 AI 예비평가 모델을 설계하고, 모델 구조의 차이가 평가 결과에 미치는 영향을 평가자 간 일치도·구성 타당도·구현 충실도의 세 측면에서 탐색적으로 검증하는 것이다. 본 연구는 국내 대학의 학생부종합전형 서류평가 맥락에서, 대표적 공통 평가요소 프레임워크인 임진택 외(2022)를 범용 모델(A·B·C)의 참조 기준으로 차용한 탐색적 다중 모델 검증 연구(exploratory multi-model comparative study)의 성격으로 수행되었으며, 특정 AI 모델의 일반화된 타당성 주장보다는 네 가지 설계 구조가 각각 어떤 장·단점을 가지는지를 비교·서술하는 데 초점을 둔다. 이를 위해 다음의 연구 문제를 설정하였다.

연구문제 1 (평가자 간 일치도): 공통 평가요소 프레임워크 기반 범용 AI 모델(A, B, C) 간 평가 일치도는 어느 수준이며, 루브릭 유사 자체 생성 기준의 제공과 사실 추출 레이어의 유무는 일치도에 어떤 영향을 미치는가?

연구문제 2 (구성 타당도): 국내 학종 서류평가에서 범용 AI 모델의 판정은 사전 설계된 역량 수준 및 전문가 평가와 어떻게 수렴·발산하며, 이를 통해 각 모델이 포착하는 평가 구인(construct)의 특성은 무엇인가?

연구문제 3 (사실성 감별): 사실 추출 모델(B)의 fact_density_score는 AI 워딩 과다 학생부의 사실 기반 비율 저조를 정량적으로 감별하는가, 그리고 해당 지표는 최종 등급 판정에 유효하게 반영되는가?

연구문제 4 (부호화 방법론의 특성): 국내 학종 서류평가 맥락에서, 전문가 개인의 암묵적 평가 기준을 규칙 기반으로 부호화한 모델(D)은, 설계자(전문가A)의 판정에 대한 구현 충실도와 독립 평가자(전문가B)에 대한 수렴 타당도의 두 축에서 범용 모델과 어떻게 구별되는가?

각 연구문제에 대응하는 결과는 Ⅳ장 2~3절(연구문제1·2), Ⅳ장 4절(연구문제3), Ⅳ장 5~7절(연구문제4)에서 제시되며, 각 절 첫 문장에 해당 연구문제를 명시한다.

Ⅱ. 이론적 배경

1. 학생부종합전형의 평가 체계

학생부종합전형은 학교생활기록부를 중심으로 지원자의 역량을 정성적으로 평가하는 대입 전형이다. 5개 대학 공동연구(임진택 외, 2022)에서는 공동연구를 통해 학종 평가의 공통 요소를 체계화하였으며, 3개 역량과 10개 세부 평가항목으로 구성된 프레임워크를 제안하였다.

<표 1> 공통 평가요소 체계

역량	평가항목	세부 평가 내용
학업역량	학업성취도	대학 수학에 필요한 기본 교과 성적 수준
	학업태도	자기주도적 학습 태도와 의지
	탐구력	지적 호기심과 깊이 있는 탐구 활동
진로역량	전공 관련 교과 이수 노력	전공 관련 교과의 선택과 이수 현황
	전공 관련 교과 성취도	전공 관련 교과의 학업 성취 수준
	진로 탐색 활동	진로 관련 다양한 활동과 경험
공동체역량	협업과 소통	공동체 내에서의 협력적 상호작용
	나눔과 배려	타인에 대한 배려와 봉사 정신
	성실성과 규칙준수	학교 규칙 준수와 성실한 학교생활
	리더십	공동체 발전을 위한 주도적 역할

출처: 임진택 외, 2022

본 연구에서는 3개 역량의 가중치를 학업역량 30%, 진로역량 50%, 공동체역량 20%로 설정하여 AI 평가 모델을 설계하였다. 이 가중치는 임진택 외(2022) 공동연구에서 제시한 것이 아니라 본 연구의 모델 설계를 위해 연구자가 설정한 값이다.

2. 평가자 간 신뢰도와 루브릭

평가자 간 신뢰도(inter-rater reliability)는 정성 평가 질의 핵심 요건으로, 루브릭 제공 여부에 따라 유의미한 차이가 발생한다(류영철, 2016). 교육평가 문헌에서 루브릭(rubric)은 전문가가 이론 기반으로 설계·타당화한 구조화된 채점 기준표로 정의되며(Popham, 1997; Brookhart, 2013; Jonsson & Svingby, 2007의 메타분석), ㉮ 수준별 관찰 가능한 준거 제시, ㉯ 채점자 간 신뢰도 향상, ㉰ 평가 근거의 외부 소명의 세 역할을 담당한다. 이 정의에 따르면 본 연구의 모델A·B가 사용하는 LLM 생성 '자체 채점

기준'은 전통적 루브릭과 형태는 유사하나 설계 주체·타당화 절차가 근본적으로 다르므로 '루브릭'으로 통칭하기보다 '루브릭 유사 LLM 생성물'로 구별 기술하는 것이 타당하다.

최승배, 이영학(2013)은 평가자 간 편차 최소화를 위한 구조적 장치의 필요성을 강조하였고, 최미숙 외(2024)는 신임 입학사정관 공통교육과정 설계에서 평가자 훈련의 중요성을 제시하였다. 그러나 훈련만으로는 앞서 기술한 기관 내·연도 간 평가자 변이를 해소하기 어려우며, 이는 평가 기준을 명시적 규칙 집합으로 부호화하는 접근의 학술적 근거를 제공한다.

3. LLM 기반 자동 채점 연구와 본 연구의 위치

대규모 언어 모델(LLM)을 활용한 자동 채점 문헌은 LLM이 과제 설명·프레임워크로부터 채점 기준을 자율 생성할 수 있음을 보고한다. Pack, Barrett, & Escalante(2024)는 LLM 에세이 자동 채점에서 루브릭 제공이 일관성·타당성을 향상시킴을 보고하였고, Yavuz, Çelik, & Yavaş Çelik(2025)은 EFL 채점에서 루브릭 기반 조건에서 LLM과 사람 채점자 간 높은 일치도를 확인하였다. Ramesh와 Sanampudi(2025)의 체계적 문헌 고찰은 LLM 평가 시스템이 입학사정관과 높은 일치도를 달성할 수 있으나 평가 맥락과 기준의 명확한 제공이 전제 조건임을 보고하였다.

그러나 이 문헌들은 대부분 '사전 주어진 전문가 설계 루브릭을 LLM이 적용'하는 구조이며, 본 연구의 모델A·B와 같이 'LLM이 프레임워크로부터 자율 생성한 기준을 결정론적 알고리즘에 이관'하는 혼합 구조나, 모델D와 같이 '전문가 암묵지를 knowledge engineering 방식으로 부호화하여 결정론적 알고리즘으로 평가'하는 접근은 학종 서류평가 맥락에서 실증된 바가 거의 없다. 학생부의 개인정보 민감성 또한 실증연구 축적을 제약하는 구조적 요인이다.

본 연구는 이 공백에 대응하여 LLM 자율 기준 생성 계보(모델A·B)와 기준 부호화 계보(모델D)를 한 설계 내에서 병행 검토하며, 모의 학생기록 설계로 개인정보 제약을 우회하면서 평가 모델의 구조적 특성을 체계적으로 비교한 국내 초기 수준의 연구라는 점에서 의의를 지닌다.

4. 학생부종합전형의 교육적 효과 및 기본역량의 대학 성과 연계

학생부종합전형으로 입학한 학생들의 대학 성과에 관한 종단 연구들이 축적되고 있다. 황희돈(2017)은 '학생부종합전형 3년의 성과와 고교 교육의 변화' 심포지엄에서 입학전형별 신입생 종단연구를 발표하며 10~16학번 학생부종합전형 입학생과 타 전형 입

학생의 GPA, CLA(교양필수 역량평가), 중도이탈률, 취업률 등을 비교 분석하고 방법론을 발표하였다.

첫째, 학생부종합전형 학생의 중도이탈률은 4개 전형 중 가장 낮았으며, 학과만족도는 가장 높게 나타나 학교생활 몰입도와 진로 부합도가 우수하였다. 둘째, GPA는 타 전형과 유의미한 차이가 없었으나, 취업률은 학생부종합전형이 가장 높았으며 대기업·공공기관 정규직 취업률에서도 우세하였다. 셋째, CLA 분석에서 사회역량(친화력, 영향력, 의사소통, 리더십)이 GPA보다 장기 성과를 더 잘 예측하였으며, 교수들이 대학에서 우수하다고 판단되는 학생들을 추천했을 때 GPA와 CLA가 모두 높은 A그룹에 집중되는 경향이 확인되었다. 이는 학업 성적만으로는 포착되지 않는 진로역량과 공동체역량이 대학 교육 성과와 사회 진출에서 핵심적 역할을 함을 시사한다.

이러한 연구 결과는 학종 서류평가에서 진로역량(50%)과 공동체역량(20%)에 높은 가중치를 부여하는 것의 타당성을 뒷받침하며, AI 보조 평가 도구가 학업 성적의 수치적 측면뿐 아니라 진로 탐색 활동과 공동체 참여의 질적 측면을 적절히 평가해야 함을 시사한다.

백순근(2004)은 평가 도구의 타당도가 평가 맥락에 따라 재구성되어야 하며 다중 증거(multiple sources of evidence)의 통합으로 해석되어야 함을 제시하였는데, 이는 본 연구가 구성 타당도를 설계의도 적중도·전문가 평가 수렴·구현 규칙 재현성의 복수 증거로 통합 해석하는 이론 기반이다. 백순근 외(2017)는 2015 개정 교육과정의 여섯 가지 핵심역량을 측정하기 위한 고등학생용 측정도구를 개발·타당화하여, 학생부종합전형에서 평가하고자 하는 역량이 체계적으로 측정 가능함을 실증하여, 본 연구의 3개 역량 10개 항목이 대학 성과 예측의 선행 지표로서 구성 타당성을 갖는 구인임을 뒷받침한다.

5. 본 연구의 이론적 위치와 비교 축

본 연구의 네 모델은 전통적 루브릭 이론(Popham, 1997; Brookhart, 2013), LLM 자율 기준 생성(Pack, Barrett, & Escalante, 2024 등), 전문가의 암묵적 평가 기준 명시화 전통(Polanyi, 1966; Gaines, 2013)의 세 이론 계보가 교차하는 설계다. 이 교차 지점을 엄밀히 기술하기 위해서는 유사하지만 구분되는 네 층위의 개념을 분리할 필요가 있다. 첫째, 프레임워크는 평가 대상 구인(역량·항목·가중치)을 선언적으로 규정하는 상위 설계 문서로, 본 연구에서는 임진택 외(2022)의 3개 역량 10개 항목 체계(표1)가 이에 해당한다. 둘째, 전통적 루브릭은 전문가가 이론 기반으로 설계·타당화한 구조화된 채점 기준표로 각 평가항목에 수준별 descriptor를 포함하며, 본 연구에서는 외부 참조 개념으로만 등장한다. 셋째, 자체 채점 기준은 LLM이 프레임워크로부터 자율 추론으로

도출한 항목별 판단 기준(모델A·B 공통 전처리 산출물, 평균 15.3개 항목)으로, 전통적 루브릭과 형태는 유사하나 설계 주체·타당화 절차가 다르므로 '루브릭 유사 LLM 생성물'로 구별한다. 넷째, 전문가 부호화 규칙은 전문가의 평가 기준을 자기보고·인터뷰로 명시화한 뒤 조건문·가중치·임계값 등 결정론적 알고리즘으로 번역한 규칙 집합(모델D의 Python 스크립트, 비공개)이다.

이 네 층위 구분에 따라 본 연구의 비교 축은 통상적 '전문가 설계 루브릭 vs AI 생성 루브릭 성능 비교'가 아니라, 'LLM 자율 생성 기준(모델A·B) vs 전문가 부호화 규칙(모델D)'의 방법론 비교로 규정된다. 전통적 전문가 설계 루브릭과의 직접 비교는 본 연구 범위를 넘어서는 후속 과제로 Ⅴ장의 후속 연구 절에서 명시한다.

Ⅲ. 연구 방법

1. AI 예비평가 모델 설계

본 연구에서는 앞서 기술한 용어 체계에서 정의한 네 층위의 용어 체계(프레임워크·루브릭·자체 채점 기준·전문가 부호화 규칙)에 기반하여 네 가지 구조의 AI 평가 모델을 설계하였다. 모델A~C는 Anthropic사의 Claude 3.5 Sonnet을 활용하되 구현 방식과 처리 과정에서 구조적 차이를 두었으며, 모델D는 전문가의 평가 기준을 규칙 기반으로 부호화한 맞춤형 모델이다. 모델A~C에 동일한 LLM을 사용한 것은 의도적 설계로, 본 연구의 핵심 독립변수가 LLM의 종류가 아닌 모델 구조(루브릭 유무, 사실 추출 유무, 규칙 기반 여부)이므로, LLM 성능 차이라는 교란변수를 통제하여 구조적 차이만의 효과를 관찰하기 위함이다. 다른 LLM을 활용한 비교 검증은 후속 연구 과제로 남긴다.

<표 2> AI 평가 모델 4종의 구조 비교

구분	모델A	모델B	모델C	모델D
명칭	자체 기준 생성 모델	사실 추출 강화 모델	무루브릭 통제군	전문가 기준 부호화 모델
기반	LLM 기준생성+ Python 평가	LLM 기준생성+ Python 평가	LLM 직접평가	규칙 기반(Python)
프레임워크 입력	3개 역량 10개 항목 정의	3개 역량 10개 항목 정의	역량명 + 가중치만	전문가 고유 기준
자체 기준 생성	O	O	X	X (사전 부호화)

구분	모델A	모델B	모델C	모델D
사실 추출	X	O	X	X
핵심 특징	루브릭 기반 자율 평가	사실 기반 검증	직관적 종합 판단	전문가 기준 규칙화
역할	범용 루브릭 모델	사실 검증 모델	통제군	전문가 기준 부호화

모델A(자체 기준 생성 모델, self-generated criteria model)는 앞서 기술한 용어 체계의 'LLM 자율 기준 생성' 계보에 위치한다. 5개 대학 공통 평가요소 프레임워크의 3개 역량 10개 항목 정의를 입력받아 LLM이 항목별 '루브릭 유사 자체 생성 기준'을 생성하고, 이를 결정론적 알고리즘에 이관하여 학생기록을 평가한다. 이 기준은 전통적 전문가 설계 루브릭이 아닌 LLM 생성물이므로 설계 주체·타당화 절차가 본질적으로 구분된다. 루브릭 유사 기준을 도입한 이유는 루브릭 제공이 채점 일관성·타당성을 향상시킨다는 선행 연구 결과(류영철, 2016; Pack, Barrett, & Escalante, 2024)에 근거한다. AI 모델(A~C)은 점수 산출과 함께 평가 근거를 서술하는 내용 요약을 생성한다.

모델B(사실 추출 강화 모델)는 모델A에 사실 추출 레이어를 추가한 것이다. 학생기록의 각 문장을 fact(객관적 사실), experience(구체적 경험), outcome(측정 가능한 결과), evaluative(주관적 평가·수식어)의 네 범주로 분류하여 fact_density_score를 산출하고, 평가 시 evaluative 문장은 근거에서 제외한다. 이는 AI 워딩의 사실성 저하 감별을 위한 설계다.

모델C(무루브릭 통제군)는 역량명과 가중치(30:50:20)만 제공하고 세부 평가항목 없이 종합 판단을 요청하는 통제군으로, 루브릭 제공 여부의 효과 측정을 위한 참조군으로 기능한다.

모델D(전문가 기준 부호화 모델, expert tacit-knowledge encoded model)는 앞서 기술한 용어 체계-(다)의 암묵지 명시화(Polanyi, 1966)·knowledge engineering(Gaines, 2013) 전통에 위치한다. 전문가A의 개인적 평가 휴리스틱을 자기보고식으로 추출하여 결정론적 Python 규칙으로 부호화한 모델(LLM 미사용)이며, 내신 등급 가중치·교과 선택 난이도·진로 일관성 판단 등을 정량 규칙으로 변환하였다. 세부 규칙은 비공개다.

전문가A가 모델D 규칙의 설계자를 겸한 내부 평가자이므로, 모델D와 전문가A의 비교는 본질적으로 설계자-구현물 간 일치를 측정하는 구조에 해당한다. 본 연구는 이를 자가-부호화(self-coding) 구조를 명시적으로 선언하고, 모델D와 전문가A의 비교를 구현 충실도(implementation fidelity, 재현성 점검)로, 모델D와 전문가B의 비교를 독립 수렴 타당도(convergent validity)로 분리하여 해석한다. 이에 따라 κ=.648(전문가A 대비)은 재현성 근거로, κ=.261(전문가B 대비)은 외부 타당도 현 수준으로 각각 해석되

며, 설계자-평가자 겸직에 따른 자가-부호화 편향(self-coding bias)은 연구의 한계 절에 본 연구의 핵심 한계로 명시한다. 이러한 구조에서 모델D는 타당한 범용 AI 평가 도구가 아니라, 각 대학이 자체 입학사정관 기준을 동일 절차로 부호화할 때 구축 가능한 맞춤형 예비평가 도구의 원형(prototype)으로 제시된다.

2. 모의 학생기록 설계

30건의 모의 학생기록을 체계적으로 설계하였다. 지원 학과는 경영학과 15명(S01-S15)과 신소재공학과 15명(S16-S30)으로, 인문·사회 계열과 이공 계열의 학과 특성 차이를 관찰하도록 구성하였다. 각 기록은 학업·진로·공동체역량에 대해 사전 의도된 수준(상/중/하)을 배정하였으며, 특수 사례(성장형, AI 워딩 과다, 특수교육과정)를 포함하였다. 학생부 기재 규정을 준수하여 작성하였고, 블라인드 표현은 익명처리(OO)하였다.

<표 3> 모의 학생기록 설계 배분표

구분	학생ID	설계 특성	교육과정
전형적 사례	S01-S09, S16-S24	학업/진로/공동체 상중하 조합	일반고
성장형	S10, S25	학업역량 저출발-고도달 궤적	일반고
특수교육과정	S11-S15	외국어고 등 교육과정	외국어고 등
특수교육과정	S26-S30	과학고 등 교육과정	과학고 등
AI 워딩 과다	S13, S27	의도적 AI 생성 문체 과다 적용	혼합

성장형 학생(S10, S25)은 AI 모델의 성장 인식 능력 검증, AI 워딩 과다 학생(S13, S27)은 사실 추출 모델의 수식어 과다 감별 능력 검증을 위한 설계다. 기록 텍스트는 교육부 「학교생활기록부 작성 기재요령」에 준거하여 작성함으로써 실제 학생부의 기재 형식·문체 규정을 반영하였으며, 작성 과정에서 LLM(Claude 3.5 Sonnet)을 보조 도구로 활용하였다. 평가 단계에서는 설계의도(역량별 배정 수준)에 접근하지 못하도록 기록 원문·프레임워크만 입력받아 독립 평가하도록 정보 차단하였다. 과목당 세부능력 및 특기사항은 평균 115~175자로 기재규정 상한(500자) 대비 축약되어 있으며, 이는 양이 아닌 질적 차이에 의한 역량 변별을 검증하는 통제 환경 구축을 위한 것이다. 실제 학생부의 풍부한 맥락 정보 영향은 후속 연구 과제다.

3. 전문가 평가

전문가 평가는 학생부종합전형 서류평가 경험이 있는 2인의 평가자가 독립적으로 실시하였다. 전문가A는 모델D 부호화 규칙의 설계자를 겸한 내부 평가자로, 대학 입학 관련 연구·평가 업무에 10년 이상 종사하며 평가기준 설계와 학생부 분석을 주요 업무로 수행해 왔고 전공은 사회과학 분야다. 규칙의 타당화 근거는 10년 이상 실무 반복 적용의 경험적 안정성에 있으며, 모델D는 임진택 외(2022) 프레임워크를 직접 참조하지 않고 전문가A의 실무 경험에 기반해 독립적으로 설계되었다. 단, 공식 타당화 절차를 거친 것은 아니므로 '전문가A ↔ 모델D' 비교는 앞서 명시한 해석대로 구현 충실도로 해석한다. 전문가B는 다른 대학에서 학종 서류평가 실무에 10년 이상 종사하며 평가기준 설계 경험을 보유한 독립 평가자로, 모델D 부호화 과정 및 전문가A 평가 결과에 일절 접근하지 않았다. 각 전문가는 30건에 대해 상/중/하 3등급 범주 평가를 독립 수행하였다. 3등급 범주는 학종 선발 구조(최종 순위 결정)를 반영한 것으로, AI 모델의 세부 점수는 등급 도달 과정을 진단하기 위한 분석 도구다. 전문가A 분포는 상9·중13·하8, 전문가B는 상8·중11·하11로 B가 상대적으로 엄격하였다.

4. 평가 실시 절차

본 연구에서 'LLM'과 'Python 알고리즘'은 서로 다른 층위로, Python은 평가 로직을 구현하는 프로그래밍 환경이고 알고리즘은 LLM이 생성한 자체 채점 기준을 결정 트리 형태로 번역한 결정론적 규칙 집합으로, 외부 기계학습 라이브러리의 범용 모델에 해당하지 않는다. '프레임워크 입력'은 공통 평가요소를 Claude 3.5 Sonnet API의 시스템 프롬프트에 제시하는 in-context 방식을 의미하며, 파인튜닝은 수행하지 않았다. 모델A·B는 Claude 3.5 Sonnet(모델 식별자 claude-3-5-sonnet-20241022, temperature=0)으로 자체 채점 기준 생성 후 Python 규칙 엔진으로 평가하였고, 모델C는 동일 API로 직접 평가, 모델D는 전문가A 규칙을 Python으로 구현한 규칙 기반 시스템이다. 모델A·B·D는 결정론적이므로 재검사 신뢰도 κ=1.0이 수학적으로 보장되어 별도 재검사 실험을 생략하였으며, LLM의 확률적 출력 특성을 지닌 모델C의 재검사 신뢰도 측정은 본 연구에서 수행하지 못하여 연구의 한계 절 제 3 한계로 이관하였다. 30명을 6명씩 5개 배치로 나누어 평가하였으며, 모델C는 프롬프트 캐싱 방지를 위해 독립 세션에서 실행하였다.

5. 분석 방법

연구 문제별로 상이한 통계 지표를 조합하여 활용하였다. 모델·전문가 간 일치도는 Cohen κ·Fleiss κ·Kendall τ-b·W·ICC(2,1) 등 다중 지표로 측정하여 순위·범주·연속 점수 차원을 모두 포괄하였고, 구성 타당도 점검은 설계의도 적중률(Hit Rate)과 전문가 평가와의 수렴·변별 지표를 결합하였다. 이러한 지표 구성은 조작적 정의 절에서 정의한 조작적 정의(특히 평정일치도·설계의도·구현 충실도 vs 수렴 타당도)에 기반한다. 각 연구 문제에 대응하는 구체적 분석 지표와 내용은 다음과 같다.

<표 4> 연구 문제별 분석 방법

연구 문제	분석 지표	분석 내용
연구문제1 평가자 간 일치도	Kendall τ-b, Cohen κ(Cohen, 1960), Fleiss κ, ICC(2, 1), Kendall W	모델의 순위·범주 일치 정도(루브릭 효과, 자체 기준 품질 포함)
연구문제2 구성 타당도	Hit Rate	설계의도 대비 AI 판정 정확도
연구문제3 사실성 감별	fact_density_score, B-C 점수 차이	사실 추출 모델의 워딩 감별 효과
연구문제4 부호화 방법론의 특성	전문가-D κ vs 전문가-A/B/C κ	맞춤형 기준의 일치도 향상 효과

주요 지표의 의미는 다음과 같다. Kendall τ-b는 두 평가자의 순위 상관을 측정하며, 값이 1에 가까울수록 순위 배치가 일치함을 의미한다. Cohen κ는 우연 일치를 보정한 범주 일치도로, 두 평가자가 동일한 등급(상/중/하)을 부여한 정도를 나타낸다. Landis & Koch(1977)의 해석 기준에 따르면 κ 〈 .20은 미약(poor), .21-.40은 약간(fair), .41-.60은 보통(moderate), .61-.80은 상당(substantial)한 일치로 분류된다. Fleiss κ는 3인 이상의 다수 평가자 간 범주 일치도, ICC(2, 1)는 절대적 점수 일치도, Kendall W는 다수 평가자의 순위 합의 수준을 각각 측정한다.

범주화 기준으로 모델A~C의 경우 총점 85점 이상 '상', 70점 이상 '중', 70점 미만 '하'로 분류하였다. 모델D는 전문가A의 평가 기준을 부호화하는 과정에서, 해당 전문가의 등급 판정 경향을 사전 설계 단계에 반영하여 87.5점 이상 '상', 77.5점 이상 '중', 77.5점 미만 '하'로 설정하였다. 이 임계값은 사후 교정이 아닌 규칙 설계 시점의 고정 파라미터로, 분석 단계에서 별도의 재조정은 이루어지지 않았다. 역량별 점수는 가중 환산 후 해당 역량 만점 대비 80% 이상 '상', 55% 이상 '중', 55% 미만 '하'로 분류하였다.

6. 주요 지표의 조작적 정의

본 연구의 핵심 지표 정의는 다음과 같다. 설계의도 적중률(Hit Rate)은 모의 학생기록 설계 시 각 학생의 역량별로 배정한 수준(상/중/하)과 AI 판정이 일치하는 비율로 산출된다. 전체 Hit Rate는 30명×3개 역량=90개 셀 기준, 역량별 Hit Rate는 30개 셀 기준의 일치 비율이며, 단위는 퍼센트(%)이다. 본 지표는 절대적 정답과의 일치가 아닌 '기록 설계의 체계성 참조 지표'로, 타당도 최종 판단은 전문가 평가와의 수렴·변별 분석과 결합한 다중 증거 관점(백순근, 2004)에서 이루어진다. 사실성 밀도 점수(fact_density_score)는 내용 분석(Krippendorff, 2019) 관점에서 텍스트의 '정보 신호대 수식적 표현'의 비율을 정량화한 지표다. 학생기록의 문장을 fact(객관적 사실)·experience(구체적 경험)·outcome(측정 가능한 결과)·evaluative(주관적 평가·수식어)의 네 범주로 분류한 뒤, 전체 문장 수 대비 사실·경험·결과로 분류된 문장 수의 비율로 산출한다. 값의 범위는 0~1이며, 값이 낮을수록 미사어구 비율이 높음을 의미한다. 문장 분류는 Claude 3.5 Sonnet이 프롬프트에 제시된 네 범주 정의를 기반으로 수행하였다. 모델D의 미사어구 감점 규칙은 evaluative 비율이 일정 임계값(상한) 이상일 때 고감점, 그보다 낮은 임계값(하한) 이상일 때 중감점을 적용하는 조건 분기 구조이다(구체 임계값·가중치는 비공개). 평정일치도는 Cohen κ(2인 범주; Cohen, 1960), Fleiss κ(3인 이상; Fleiss, 1971), ICC(2,1)(연속 점수; Shrout & Fleiss, 1979), Kendall τ-b·Kendall W(순위)의 다층 지표로 측정하며, 해석 기준은 Landis & Koch(1977)의 5단계 구간(.00-.20 미약/ .21-.40 약간/ .41-.60 보통/ .61-.80 상당/ .81-1.00 거의 완전)을 따른다.

모델D와 전문가의 일치도는 두 층위로 구별 해석한다. 규칙 설계자인 전문가A와 모델D의 일치는 규칙이 설계 의도대로 구현되었는지를 확인하는 엔지니어링 지표로서 구현 충실도(implementation fidelity)로 해석하고, 부호화에 관여하지 않은 전문가B와 모델D의 일치는 독립 수렴 타당도(convergent validity)지표로 해석한다. 이 구별은 이후 통합 해석의 중심축이다.

Ⅳ. 연구 결과

1. 평가 결과 개요

본 절은 이후 Ⅳ장 2~7절의 공통 참조 자료로서, 30명 모의 학생에 대한 4개 모델 및 2인 전문가 평가 결과 전체를 제시한다.

<표 5> 30명 모의 학생 × 4개 모델 총점 및 전문가 평가

ID	모델A	모델B	모델C	모델D	전문가A	전문가B	A등급	B등급	C등급	D등급
S01	94.5	100.0	93.0	98.0	상	상	상	상	상	상
S02	89.3	89.0	84.0	91.8	중	중	상	상	중	상
S03	86.5	89.5	80.0	96.8	상	상	상	상	중	상
S04	79.5	92.0	82.0	93.9	상	상	중	상	중	상
S05	81.0	88.5	73.0	90.0	중	하	중	상	중	상
S06	72.8	81.8	60.0	81.0	중	하	중	중	하	중
S07	66.5	75.0	56.0	75.2	하	하	하	중	하	하
S08	73.0	85.2	69.0	84.9	하	하	중	상	하	중
S09	81.7	91.0	55.0	87.2	중	상	중	상	하	중
S10	79.2	86.5	70.0	82.3	중	중	중	상	중	중
S11	73.2	75.3	75.0	84.8	중	중	중	중	중	중
S12	77.0	80.8	68.0	91.4	중	하	중	중	하	상
S13	75.5	77.7	51.0	54.8	하	중	중	중	하	하
S14	77.7	85.8	75.0	83.1	하	중	중	상	중	중
S15	76.7	86.7	83.0	85.4	중	하	중	상	중	중
S16	99.0	99.0	98.0	96.0	상	상	상	상	상	상
S17	87.2	87.0	84.0	89.4	상	상	상	상	중	상
S18	76.3	94.0	80.0	97.4	상	상	중	상	중	상
S19	82.2	89.2	82.0	93.9	상	상	중	상	중	상
S20	82.2	88.8	78.0	90.9	중	중	중	상	중	상
S21	51.0	79.0	60.0	78.2	중	하	하	중	하	중
S22	41.7	74.3	48.0	77.2	하	하	하	중	하	하
S23	71.7	77.7	72.0	80.1	중	하	중	중	중	중
S24	46.7	80.5	54.0	74.7	중	중	하	중	하	하
S25	74.8	78.7	68.0	84.4	중	중	중	중	하	중
S26	87.3	77.2	75.0	91.7	상	중	상	중	중	상
S27	41.5	68.2	48.0	51.7	하	중	하	하	하	하
S28	93.2	94.0	95.0	97.4	상	중	상	상	상	상
S29	43.7	67.8	47.0	67.3	하	하	하	하	하	하
S30	33.8	80.3	37.0	74.5	하	하	하	중	하	하

범용 모델(A~C)은 상위권과 하위권의 점수 격차가 크고(모델A: 33.8~99.0, 범위

65.2), 모델D는 상대적으로 압축된 분포(51.7~98.0, 범위 46.3)를 보인다. 이는 모델D의 규칙 기반 산출이 극단적 저평가를 억제하는 경향을 시사한다.

2. 범용 AI 모델 간 평가 일치도 및 레이어 효과

본 절은 연구문제 1에 대응한다. 범용 3개 모델(A~C) 간 평가 일치도를 다중 지표로 분석한 결과를 〈표6〉~〈표7〉에 제시한다.

<표 6> 범용 3개 모델 간 쌍별 일치도(Kendall τ-b, Cohen κ)

쌍	Kendall τ-b	p-value	Cohen κ	해석
A-B	.600	<.001	.175	순위 보통/ 범주 미약
A-C	.695	<.001	.468	순위 보통/ 범주 보통
B-C	.591	<.001	-.003	순위 보통/ 범주 일치없음

<표 7> 전체 및 역량별 일치도 지표(범용 3개 모델)

지표	전체	학업역량	진로역량	공동체역량
Fleiss κ	.165	.288	.237	.231
ICC(2, 1)	.569	-	-	-
Kendall W	.859	-	-	-

범용 3개 모델은 Kendall W=.859로 순위 수준에서 양호한 합의를 보이나 Fleiss κ =.165로 범주 일치도는 낮다. 역량별 Fleiss κ(학업 .288, 진로 .237, 공동체 .231) 모두 낮아 세부 역량에서도 모델 간 등급 분류 기준에 체계적 차이가 존재한다. 쌍별 비교에서는 A-C(τ=.695, κ=.468)가 가장 일치도가 높고, A-B(τ=.600, κ=.175)·B-C(τ=.591, κ=-.003)에서 모델B의 점수 상향 편향으로 범주 불일치가 현저하다. 모델A·B가 독립 생성한 채점 기준 수가 모두 15.3개로 동일하게 나타나 공통 평가요소 프레임워크가 AI 자체 기준 생성을 안내하는 역할을 함이 확인되었다.

범용 모델은 순위 서열 배치에는 유효하되 등급 기반 판정에는 한계가 있다. 루브릭 제공이 항상 일치도를 높이는 것은 아니며, 사실 추출 레이어와 결합될 때 점수 상향으로 범주 일치가 오히려 저하된다(A-B κ=.175 〈 비루브릭 쌍 평균 .233). 이는 '루브릭의 존재 여부'가 아닌 '설계의 질과 후속 처리와의 상호작용'이 AI 평가 일관성을 결정함을 시사한다.

3. 범용 AI 모델의 설계의도 수렴·발산 및 구성 타당도 해석

본 절은 연구문제 2에 대응한다. 설계의도와 AI 판정 간 Hit Rate를 분석한 결과를 〈표8〉에 제시한다.

<표 8> 설계의도 vs AI 판정 Hit Rate

모델	학업역량	진로역량	공동체역량	전체
모델A	36.7%	60.0%	80.0%	58.9%
모델B	56.7%	56.7%	50.0%	54.4%
모델C	83.3%	86.7%	86.7%	85.6%

모델C의 Hit Rate(85.6%)가 모델A(58.9%)·모델B(54.4%)보다 현저히 높다. 학생부는 교사가 공적 문서로서 긍정적 측면 중심으로 기재하는 구조적 특성을 지니므로, 루브릭 없이 전반적 인상에 기반하는 모델C는 기록의 긍정적 표현 밀도만으로도 설계의도의 단순 상/중/하 분류에 도달한다. 반면 루브릭 기반 모델(A·B)은 10개 항목별로 탐구 깊이, 교과 선택 도전성 등을 세분화하여 판단하므로 단순 수준 분류와 다른 변별을 수행한다. 특히 학업역량에서 모델A(36.7%)와 C(83.3%)의 격차가 가장 컸다. 모델A·B의 Hit Rate가 60% 미만에 그친 점은 모의 기록 생성에 동일 LLM이 활용되었음에도 평가 모델이 설계의도에 종속되지 않고 독립 판단을 수행하였음을 보여준다(모의 기록 설계 절의 정보 차단 설계 확인).

모델C의 높은 Hit Rate는 '깊이 있는 평가' 증거가 아니라 '기록 긍정성 밀도의 표면적 일치도'에 불과하며, 이는 조작적 정의 절의 '설계 일치도(표면적 일치도)' 개념이 설명하는 현상이다. 구성 타당도는 Hit Rate 단독이 아닌 전문가 평가와의 수렴·변별 분석(이후 결과 절)과 결합한 다중 증거(백순근, 2004) 관점에서 판단되어야 한다.

4. 사실 추출 모델의 AI 워딩 감별력 및 등급 반영 효과

본 절은 연구문제 3에 대응한다. 사실 추출 모델(B)의 AI 워딩 감별 결과를〈표9〉·〈표10〉에 제시한다.

<표 9> AI 워딩 학생의 모델별 총점 비교

ID	모델A	모델B	모델C	모델D	전문가A	판정
S13	75.5	77.7	51.0	54.8	하	C, D가 감별
S27	41.5	68.2	48.0	51.7	하	A, C, D가 감별

<표 10> Model B fact_density_score 분포

구분	평균	표준편차	범위
전체(n=30)	.299	.089	[.15, .52]
AI 워딩 학생(S13, S27)	.150	-	[.15, .15]
일반 학생(n=28)	.309	-	[.19, .52]

AI 워딩 학생의 fact_density_score 평균(.150)은 일반 학생 평균(.309)의 약 48%로, 사실 추출 레이어가 AI 워딩의 특성(사실 기반 비율 저조)을 효과적으로 탐지함을 확인하였다. 그러나 모델B는 S13에 77.7점·S27에 68.2점을 부여하여 전문가 판정('하')과 불일치하였으며, 오히려 규칙 기반 모델D가 미사어구 비율 91%를 감지하여 54.8점(하)을 부여함으로써 전문가와 일치하였다. fact_density_score와 B-C 점수 차이 간 Kendall τ-b=-.231(p=.073)로 부적 경향은 관찰되었으나 유의수준에 도달하지 않았다.

사실 추출 레이어는 AI 워딩의 '탐지 지표'로는 유효하나 현재 설계에서는 '등급 반영'까지 이어지지 못한다. 규칙 기반 모델D의 미사어구 감지 사례는 명시적 감점 규칙이 암묵적 LLM 판단보다 AI 워딩 대응에 효과적일 수 있음을 시사한다. 본 결과는 모의 기록 기반의 개념 증명 수준으로, evaluative 표현을 '근거 있는 평가'와 '근거 없는 수식어'로 세분화하는 후속 연구가 요구된다.

5. 전문가 간 평가 신뢰도

본 절은 연구문제 4(전문가 기준 부호화의 구현 충실도·수렴 타당도)에 대한 해석 기준선으로서, 2인 전문가 간 평가 일치도를 분석한다. 모델-전문가 일치도 분석에 앞서, 2인 전문가 간 평가 일치도를 교차 기관 평가 기준선(cross-institutional baseline)으로 제시한다.

<표 11> 전문가 간 평가 일치도

지표	값	해석
Cohen κ	.397	보통 일치
가중κ (linear)	.526	보통 일치
Kendall τ-b	.610	보통~양호 상관
단순 일치율	60.0% (18/30)	-

두 전문가는 서로 다른 대학에서 학생부종합전형 평가를 수행한 경험을 지닌 독립적 평가자로서, 상/중/하의 등급 경계 설정에서 차이를 보였다. 12건의 불일치는 모두 인접 등급 간 차이(상↔중 또는 중↔하)로, 2단계 차이(상↔하)는 한 건도 없었다.

이 패턴은 두 전문가가 상·중·하의 순서 구조를 공유하고 있음을 보여주며, 가중 κ =.526(Cohen, 1968)이 포착하는 방향성 공유와 임계점 차이의 전형에 해당한다. 만약 두 전문가의 차이가 평가 대상에 대한 이해도 격차였다면 2단계 차이 사례가 관찰되어야 하나 본 자료에서는 한 건도 나타나지 않았으므로, 차이의 원천은 이해도가 아닌 등급 경계 임계치의 상이함으로 해석된다. 이러한 해석은 두 전문가가 서로 다른 대학의 평가 문화에서 경험을 축적해 왔다는 점과 정합하며, 전문가별 맞춤형 기준 부호화의 필요성을 논리적으로 뒷받침한다.

6. 전문가 기준 부호화 모델의 구현 충실도 및 독립 수렴 타당도

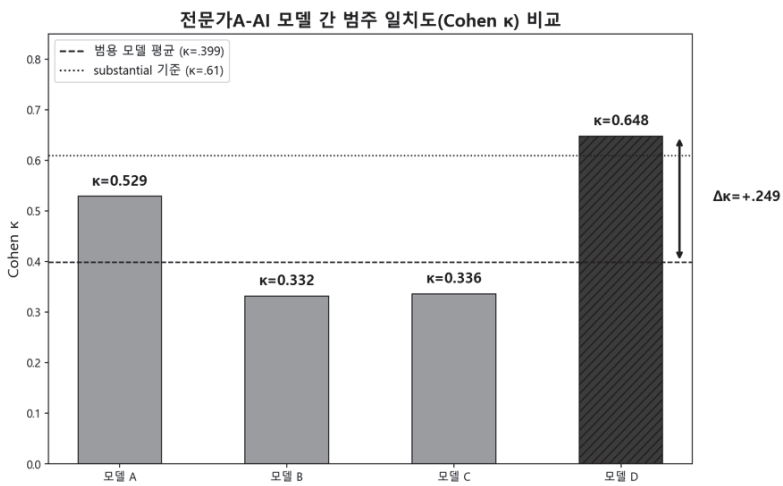
본 절은 연구문제 4(모델D가 설계자 전문가A에 대한 구현 충실도와 독립 평가자 전문가B에 대한 수렴 타당도의 두 축에서 범용 모델과 어떻게 구별되는지)에 대응하며, 해석은 연구 방법 1절의 이중 프레임(‘모델D–A = 구현 충실도’, ‘모델D–B = 수렴 타당도’)을 따른다. 전문가–AI 일치도 비교를 〈표12〉에 통합 제시한다.

<표 12> 전문가-AI 일치도 비교(전문가A·B 기준 통합)

비교 쌍	τ	κ	일치율	κ 해석
전문가A vs 모델A	.672	.529	21/30 (70%)	보통
전문가A vs 모델B	.497	.332	17/30 (57%)	약간~보통
전문가A vs 모델C	.663	.336	17/30 (57%)	약간~보통
전문가A vs 모델D(구현 충실도)	.791	.648	23/30 (77%)	상당
전문가B vs 모델A	.502	.239	15/30 (50%)	미약
전문가B vs 모델B	.486	.223	14/30 (47%)	미약
전문가B vs 모델C	.479	.275	16/30 (53%)	미약
전문가B vs 모델D(수렴 타당도)	.463	.261	15/30 (50%)	미약
전문가A 대비 범용 모델 평균(A-C)	.611	.399	18.3/30 (61%)	-
모델D – 범용 평균 차	+.180	+.249	+16%p	-

전문가A vs 모델D의 κ =.648은 모델D 규칙이 전문가A 휴리스틱에서 유도된 구조이므로 상단 범용 모델 κ들과 동일 척도로 병렬 비교될 수 없으며, 구현 충실도(implementation fidelity)지표로 해석한다. κ가 .700대에 도달하지 못한 주된 요인은 ㉮ 규칙으로 언어화되지 않는 전문가A 휴리스틱의 잔여 영역(Gaines, 2013), ㉯ 경계선 학생에 대한 전문가의 시점별 미세 판단 변화를 모델 D의 고정 임계값이 반영하지 못하는 점, ㉰ 임계값이 설계 단계에서 확정되어 사후 조정이 불가능한 점이다. 반면 전문가B vs 모델D의 κ =.261은 독립 수렴 타당도의 현 수준으로, 두 전문가 간 κ =.397(표

11)을 상한선이 아닌 교차 기관 평가 기준선(cross-institutional baseline)으로 재정의하는 통합 해석 절 논리에 따라, 모델D의 결함이 아니라 '전문가A의 평가 기준을 내재화한 모델이 다른 대학 전문가B의 기준으로 평가된 결과의 구조적 필연'으로 해석된다. 세부 해석은 통합 해석 절에 통합 제시한다.



[그림 1] 전문가-AI 모델 간 범주 일치도(Cohen κ) 비교

두 전문가가 동일 등급을 부여한 18건(60%)에 대한 각 모델 일치율을 분석하였다.

<표 13> 전문가 동일 판정 기준 모델별 일치율

모델	전문가 동일 판정 건 일치	일치율
모델D	14/18	77.8%
모델A	12/18	66.7%
모델B	11/18	61.1%
모델C	11/18	61.1%

두 전문가가 판정이 일치한 명확한 사례에서 모델D의 일치율(77.8%)이 가장 높았다. 일치 18건 중 D 불일치 4건(S08: A·B '하'→D '중' 임계값 초과 / S24: A·B '중'→D '하' 규칙 감점 누적 / 나머지 2건: 경계선 인접 등급)의 공통 원인은 고정 임계값이 사람 평가자의 맥락적 가변성을 흡수하지 못함과 결정 트리 변수 합산이 학생부 전체 맥락을 근사하기 어려움이다. 4건 모두 2단계 차이가 아닌 인접 등급 차이로, 77.8%는 전문가 기준 부호화 접근이 현실적으로 도달 가능한 최대 일치율로 해석된다.

<표 14> 전문가-모델 불일치 사례 분석

SID	전문가A	전문가B	A	B	C	D	비고
S08	하	하	중	상	하	중	B 상향 편향
S13	하	중	중	중	하	하	D가 미사어구 감별
S05	중	하	중	상	중	상	전문가 간 인접 불일치
S12	중	하	중	중	하	상	D 상향

<표 15> 통합 분석

지표	3모델(A-B-C)	5인 평가자(4모델+EA)	6인 평가자(4모델+EA+EB)
Fleiss κ	.165	.358	.330
Kendall W	.859	.817	.752

모델D + 전문가 2명의 조합에서 Fleiss κ=.430으로 가장 높은 범주 일치를 보였다. 6인 평가자 체계에서 Fleiss κ(.330)와 Kendall W(.752)가 5인 평가자 대비 소폭 하락한 것은 전문가B의 상대적으로 엄격한 평가 기준이 반영된 결과이다.

7. 성장형 학생 감지

본 절은 연구문제 1(루브릭 유사 자체 생성 기준과 사실 추출 레이어의 효과)에 대한 부가 분석으로서, 1학년 성취 대비 3학년까지의 향상이 뚜렷하도록 설계된 성장형 사례에서 각 모델이 성장 궤적을 어떻게 포착하는지를 검토한다. 성장형으로 설계된 학생에 대한 모델별 평가 결과를 〈표16〉에 제시한다.

<표 16> 성장형 학생의 모델별 총점 비교

ID	모델A	모델B	모델C	모델D	전문가A	전문가B	A-C 차이
S10	79.2	86.5	70.0	82.3	중	중	+9.2
S25	74.8	78.7	68.0	84.4	중	중	+6.8

성장형 학생은 1학년 학업 성취가 낮으나 3학년까지 여러 역량에서 뚜렷한 향상을 보이도록 설계된 사례로, AI 모델이 시계열적 변화를 포착하는 능력을 검증하기 위한 것이다. S10에서 루브릭 모델(A: 79.2, B: 86.5)이 통제군(C: 70.0)보다 9.2~16.5점 높은 점수를 부여하였고, S25에서도 유사한 패턴(A: 74.8 vs C: 68.0, 차이+6.8)이 관찰되었다. 이는 루브릭 기반 모델이 학업태도와 탐구력 향상 등 성장 관련 항목을 체계적으로 평가함으로써, 1학년의 낮은 내신에도 불구하고 이후의 성장 궤적을 점수에 반영한 것으로 해석된다.

모델C가 상대적으로 낮은 점수를 부여한 것은, 루브릭 없이 전반적 인상에 기반할 경우 초기의 낮은 성취가 전체 평가를 끌어내리는 경향을 시사한다. 반면 모델D(S10: 82.3, S25: 84.4)는 규칙 기반으로 학년별 성적 변화를 반영하여 모델C보다 12~16점 높은 점수를 산출하였다. 두 성장형 학생 모두 전문가A와 전문가B가 '중'으로 평가하여 전문가 간 완전 일치를 보였으며, 루브릭 모델(A)과 맞춤형 모델(D)이 이와 동일한 등급을 부여하였다.

학종에서 성장 가능성은 주요 평가 요소 중 하나다. 본 결과는 루브릭이나 명시적 규칙이 AI 모델의 성장 궤적 인식을 유도하며, 이를 통해 초기 성취가 낮더라도 지속적 성장을 보인 학생이 적절히 평가받을 수 있는 구조를 확보할 수 있음을 시사한다.

본 절의 결과는 연구문제 1이 포괄하는 '루브릭 효과'의 구체적 사례 증거다. 즉 이후 결과 절에서 확인된 루브릭 유사 자체 생성 기준의 일반적 효과(순위 합의 W=.859)가, 성장형이라는 특수 사례에서는 '학년별 성취 변화의 체계적 포착'이라는 구체적 기제로 발현됨을 보여준다. 루브릭 모델(A·B)과 규칙 모델(D)이 통제군(C) 대비 성장형 학생에게 일관되게 높은 점수를 부여하였고, 이 판정이 두 전문가 동일 판정('중')과도 일치한다는 점은, 루브릭·규칙의 세분화된 항목 구조가 시계열 성장 정보를 평가에 반영하는 통로로 기능함을 시사한다. 이는 이후 통합 해석 절에서 논의하는 '대학별로 상이한 평가 기준을 수용하는 모델 설계 접근의 기술적 실현 가능성' 논지를 뒷받침하는 구체적 사례 증거로 기능한다.

V. 논의 및 결론

1. 연구 요약

본 연구는 국내 대학의 학생부종합전형 서류평가 맥락에서 4종의 AI 예비평가 모델을 설계·검증한 탐색적 다중 모델 비교 연구다. 범용 모델(A·B·C)은 순위 수준(Kendall W=.859)에서 양호한 합의를 보였으나 범주 일치도(Fleiss κ=.165)는 낮았으며, 루브릭 유사 자체 생성 기준과 사실 추출 레이어의 결합은 점수 상향 편향으로 범주 일치도를 저하시켰다.

사실 추출 모델B의 fact_density_score는 AI 워딩 학생의 사실성 비율을 일반 학생의 약 48%로 감지하는 탐지 지표로 유효하였으나, 현재 설계에서는 등급 반영 효과가 제한적이었다.

전문가 기준 부호화 모델D는 설계자 전문가A에 대해 κ=.648의 구현 충실도를 보였으나, 독립 평가자 전문가B에 대해서는 κ=.261로 범용 모델 수준에 머물렀다. 두 전문가 간 일치도 κ=.397을 '사람 평가자 간 일치도 상한선'이 아닌 '교차 기관 평가 기준선'으로 재정의하는 통합 해석 절의 해석에 따르면, 모델D의 전문가B 대비 낮은 수렴도는 모델 결함이 아닌 '전문가A의 평가 기준을 내재화한 모델이 다른 대학 전문가B의 기준으로 평가된 결과의 구조적 필연'이다.

종합하면 본 연구는 AI가 사람 평가자를 대체함을 주장하는 연구가 아니라, 서로 다른 대학 전문가 간 구조적 평가 기준 차이(κ=.397)의 실증과, 이를 범용 단일 모델이 아닌 대학별 기준 부호화 방식으로 수용하는 설계 접근을 탐색적으로 제시한 사례 연구다. 모델D의 가치는 대학 기준의 충실한 재현(정확성, 구현 충실도 κ=.648)과 이를 수험생에 일관 적용하는 공정성의 두 축에 있으며, '대학 기준 부합→ 일관 적용→ 수험생 공정성'의 인과로 연결된다.

2. 통합 해석

본 절에서는 Ⅳ장의 결과를 관통하는 세 가지 논지를 집약한다.

가. 교차 기관 평가 기준선과 제도적 이질성 가설

전문가A-전문가B 일치도(κ=.397)는 '사람 평가자 간 일치도의 상한선'이 아니라, 서로 다른 대학 평가 문화에서 훈련된 두 평가자 간의 교차 기관 평가 기준선(cross-institutional baseline)으로 해석된다. 12건 불일치가 모두 인접 등급에 국한되고(2단계 차이 0건) 가중 κ=.526이 방향성 일치를 포착한다는 점은, 두 전문가의 차이가 '이해도 격차'가 아닌 '등급 경계 임계치의 대학별 상이함'에 기인함을 시사한다. 이는 n=2 표본 기반 '제도적 이질성 가설'과 일치하는 탐색적 관찰이며, 전면 규명은 다기관 후속 연구가 필요하다.

나. 모델D의 방법론적 기여와 이중 해석

모델D는 전문가A의 판단 규칙을 결정론적 알고리즘으로 재현하여 κ=.648 수준의 '상당한' 구현 충실도(implementation fidelity)를 보였다. 이는 전문가의 평가 휴리스틱이 명시적 알고리즘 형태로 구현 가능함을 보여주는 결과로서, 개인화된 평가 모델 설계의 방법론적 가능성을 시사하는 독립적 기여를 갖는다. 다만 이 지표는 설계자와 구현물 간 재현성 점검에 해당하므로, κ=.648을 수렴 타당도의 근거로 확장 해석하는 데에는

제약이 있다. 독립 평가자인 전문가B와의 비교는 외부 타당도의 참조 지표로 활용될 수 있으나, 전문가B가 다른 기관의 독립 기준으로 평가하였으므로 이 비교는 '절대적 수렴 타당도'가 아닌 앞서 논의한 교차 기관 평가 기준선(κ=.397)맥락에서 해석되어야 한다.

모델D-전문가B 간 κ=.261은 범용 모델(κ=.223~.275)과 유사한 수준으로 나타났다. 이는 모델D가 내재화한 전문가A의 기준과 전문가B가 독립적으로 적용한 기준 사이의 구조적 차이가 일치도 지표에 반영된 결과로 해석될 수 있다. 이러한 결과는 모델의 결함이라기보다 대학 간 평가 기준 차이에 따른 구조적 맥락을 반영한 것으로 볼 수 있다.

한편 두 전문가가 일치한 18건을 기준선으로 할 때 모델D의 일치율은 77.8%로 네 모델 중 가장 높게 나타났다. 이는 기관 간 평가 기준 차이가 영향을 미치지 않는 명확한 사례에서 모델D가 두 전문가의 독립 판정과 가장 높은 비율로 부합함을 보여주는 결과로 해석된다.

이러한 맥락에서 모델D는 보편적으로 타당한 AI 평가 도구라기보다, 각 대학이 자체 평가 기준을 동일 절차로 부호화할 때 유효해지는 맞춤형 설계 방법론의 원형(prototype)으로 이해될 수 있다. 본 방법론의 실무적 가능성은 다수 기관에서 동일한 부호화 절차를 적용했을 때 유사한 구현 충실도가 재현되는지, 즉 '기관 간 이식 가능성'의 확보 여부에 달려 있다.

또한 전문가B가 대학A의 평가 기준을 사전 학습한 경우 유사한 판정을 내릴 가능성이 있다는 점(최미숙 외, 2024의 공통교육과정 설계 논리에 부합하는 추론)은, 모델D가 기관별 평가 기준을 체계화한 표준(codified institutional standard)으로서 신규 평가자 교육에 활용될 수 있음을 시사한다. 다만 이러한 활용 가능성은 본 연구에서 실험적으로 검증되지 않았으므로 가설적 함의로 제한된다.

구현 충실도의 재현성 확보 방안으로, 자가-부호화 편향 완화를 위한 기관 내부 독립 부호화 필요성은 Ⅴ장 4절(한계)에서, 다기관 이식성 검증과 모범 평가 사례 벤치마크 등 재현성 검증의 구체 절차는 Ⅴ장 5절(후속 연구)에서 각각 논의한다.

다. 대학별 평가 모델 설계 접근의 학술적·실무적 의의

본 연구 기여는 개별 모델의 절대적 성능이 아니라 학종 서류평가 맥락에서 대학별로 상이한 평가 기준을 수용하는 모델 설계 접근의 기술적 실현 가능성을 실증한 데 있다. 교차 기관 기준선이 보통 수준에 머무는 조건에서 단일 범용 모델 타당도 추구는 본질적 한계에 직면하므로, 본 연구는 이를 '대학별 기준 부호화 후 조합 운용' 원리로 우회한다. 이는 전통적 전문가 설계 루브릭 계보(Popham, 1997; Brookhart, 2013)와 LLM 자율 기준 생성 계보(Pack, Barrett, & Escalante, 2024; Yavuz, Çelik, & Yavaş Çelik,

2025; Ramesh와 Sanampudi, 2025)에 더하여, 학종 서류평가 맥락에서 이 두 계보와 구분되는 세 번째 설계 흐름으로서 기준 부호화(Polanyi, 1966; Gaines, 2013) 기반 결정론적 모델을 국내 초기 수준에서 실증 검토한 점에서 학술적 의의를 갖는다. AI 도구의 결정론적 특성은 재검사 신뢰도 κ=1.0(평가자 내 편차 완전 제거)을, 공통 프레임워크의 일관 적용은 평가자 간 기준 공유를 각각 보장하며, 이러한 구조적 특성이 학종 서류평가의 편차 완화 메커니즘을 구성한다.

실무적으로는 기관 내·연도 간 평가자 변이 완화 장치로 가치를 지닌다. 부호화된 규칙 집합은 ㉮ AI 보조 평가 기준선, ㉯ 신임 입학사정관 교육 명시 교재(최미숙 외, 2024), ㉰ 연도 간 기준 표류 점검 참조 문서로 복합 활용되어, 수험생 관점에서 평가 품질의 예측 가능성과 절차적 공정성을 제고하는 장치로 작동한다. 규칙 기반 모델의 평가 근거 투명성은 설명 가능한 AI(Explainable AI)로서 대입 전형의 공정성·책무성(노성, 2021) 요구에도 부합한다.

3. 제언

본 연구의 결과를 기반으로 다음을 제언한다. 첫째, AI 예비평가 도구는 최종 의사결정이 아닌 평가 보조 도구로 활용되어야 하며, 최종 판단은 전문 평가자의 정성적 검토를 거쳐야 한다. 둘째, AI 보조 도구 도입 시 대학별 전문가 평가 기준 부호화 과정이 포함되어야 한다. 범용 프레임워크만으로는 전문가의 암묵적 판단 기준을 충분히 반영하기 어려우며, 경험 있는 입학사정관의 기준을 인터뷰로 추출·부호화하고 복수 내부 평가자 교차 검증으로 임계값을 교정하는 절차가 실효성을 좌우한다(통합 해석 절 참조). 이 과정에 LLM·파인튜닝은 필요하지 않아 기술적 진입장벽이 낮다. 셋째, 사실 추출 기반 평가는 AI 워딩 대응을 넘어 고교 교사의 학생부 기재 품질 피드백 도구로 활용될 수 있으며, 명시적 감점 규칙과 LLM 사실 추출을 결합한 하이브리드 접근이 검토될 수 있다. 넷째, 평가자 훈련 프로그램의 체계화(최미숙 외, 2024)가 병행되어야 하며, 특히 부호화된 규칙 집합은 신임 입학사정관 교육의 명시적 교재로 기능하여 기관 내 평가 변이 완화의 실질 경로가 된다. 아울러 2028 대입제도 개편에 따른 내신 5등급 전환과 전 전형 정성평가 확대, 고교학점제 시행으로 서류 기반 정성평가 비중이 커지는 환경에서 AI 예비평가 도구의 필요성과 다양한 교육과정 맥락 대응력은 더욱 중요해질 것이다.

4. 연구의 한계

본 연구는 탐색적 다중 모델 비교 연구로서 다음의 한계를 지닌다. 첫째, 30건의 모의

학생기록 표본은 제한적이며 실제 학생부 데이터 결과와 상이할 수 있다. 학생부 개인정보 제약 하의 현실적 대안으로서 일반고·특목고(외고)·과학고 등 교육과정 유형, 성장형·AI 워딩 과다 등 특수 사례, 경영학과·신소재공학과 두 학과를 포함하여 질적 다양성을 확보하였으나 실제 기록의 복잡성을 완전 반영하지는 못한다. 둘째, 모델A~C는 동일 LLM(Claude 3.5 Sonnet)을 사용하되 구현 방식에 차이를 두었으므로 다른 LLM·구현 방식 비교 검증이 필요하다. 셋째, 모델A·B·D는 결정론적으로 재검사 신뢰도 κ=1.0이 수학적으로 보장되나, 모델C(LLM 직접평가)의 재검사 신뢰도는 미실시되어 후속 확인이 필요하다. 넷째, 전문가 평가 2인·3단계 척도의 한계로, κ=.397의 보통 수준 일치에 그쳤다(12건 불일치 모두 인접 등급, 가중 κ=.526). 다섯째, 본 연구의 가장 본질적 한계는 자가-부호화 편향(self-coding bias)이다. 모델D 규칙을 설계한 연구자가 전문가A를 겸하는 구조로, κ=.648은 타당도 증거가 아닌 '구현 충실도'에 그친다. 이를 완화하기 위해 연구 방법 장에서 해당 구조를 사전 명시하고, 독립 평가자 전문가B를 투입하여 수렴 타당도 기준점을 확보하였으며, 전문가 동일 판정 18건에 대한 D 일치율(77.8%) 분석으로 부분 교차 검증을 수행하였다. 아울러 모델 D의 등급 분류 임계값(87.5/77.5)을 전문가 평가 결과를 확인하기 이전 설계 단계에 사전 고정함으로써, 일치도를 인위적으로 높이기 위한 사후 임계값 조정(post-hoc tuning) 가능성을 절차적으로 배제하였다. 이러한 완화 조치들은 자가-부호화 구조에서 발생할 수 있는 편향 가능성을 절차적으로 차단하는 방법론적 안전장치로 기능한다. 한편 동일 인물이 모델 D 규칙 설계자와 전문가A 평가자를 겸하는 구조는 '한 전문가의 암묵적 평가 기준을 명시적 알고리즘으로 부호화할 수 있는가'라는 본 연구의 핵심 질문에 답하기 위한 의도된 일대일 검증 설계이며, 다기관 외적 일반화는 V장 5절에서 제시한 이식성 검증과 성과 기반 환류 절차를 통해 후속 단계에서 확장 검토된다. 한편 각 대학의 평가 기준은 입학 전형의 기관 비밀 자산에 해당하여 외부 연구자에 의한 동일 규칙 재부호화 검증은 현실적으로 수행하기 어렵다. 따라서 자가-부호화 편향의 완화는 각 기관 내부에서 복수 평가자가 동일 기준을 독립 부호화·교차 검증하는 방식으로 접근할 필요가 있다. 여섯째, 표 5에서 관찰된 바와 같이 범용 AI 모델은 모델별로 상이한 등급 분포 특성을 보였다. 모델 A는 상 7·중 16·하 7로 중등급에 집중되어 양극단 비율이 제한적이며, 모델 B는 상 16·중 12·하 2로 '상' 등급에 강하게 편중된 분포를 보였다. 두 양상 모두 본 연구의 데이터셋 내에서 확인된 객관적 결과이며, 모델 A의 변별 폭 축소와 모델 B의 상향 편중은 각각 다른 형태로 등급 변별력을 제약할 가능성을 시사한다. 다만 이러한 분포 특성이 실제 전형 운영 시 변별력 저하로 직결될지 여부는 대규모 실무 데이터를 통한 추가 검증이 필요한 탐색적 가설 수준의 해석으로 제시한다. 한편 본 연구 사용 30건은 전원 연구자 설계 모의 기록으로 실제 학생 개인정보를 포함하지 않는다.

5. 후속 연구 방향

본 연구에 이은 후속 연구 방향은 다음과 같다. 첫째, 실제 학생부 데이터의 현장 적용과 복수 대학 전문가를 대상으로 한 다기관 교차 검증, 그리고 다수 대학이 각자의 평가 기준을 동일 부호화 절차로 구현했을 때 각 기관 내에서 유사한 구현 충실도가 재현되는지 확인하는 방법론의 '기관 간 이식성 검증'이 필요하다. 이는 본 방법론의 상용화 가능성 판단의 주요 단계에 해당한다. 둘째, 각 대학의 평가 항목별로 설계의도에 부합하는 '모범 평가 사례(exemplar case)'를 구축하고, 모델이 이를 기대 판정과 일치하게 재평가하는지 확인하는 벤치마크 검증이 필요하다. 아울러 맞춤형 규칙 모델(D)과 LLM 활용 모델(A·B)의 하이브리드 접근도 후속 과제로 요구된다. 셋째, AI 평가 결과와 대학 입학 후 GPA·중도이탈률·학과 만족도·비교과 활동 및 역량 발달·취업 성과(진학 포함) 등 다면적 장기 성과 간의 예측 타당도를 추적하는 종단 연구가 수행될 필요가 있다. 더불어 입학 후 우수한 성과를 보인 학생들의 학생부를 역방향으로 사후 분석(retrospective analysis)하여, 그 공통 특성을 평가 규칙에 추가 부호화하는 성과 기반 환류 절차(outcome-based feedback encoding)의 방법론적 정립이 요구된다. 이는 Ⅴ장 4절 다섯째에서 지적한 자가-부호화 구조의 외적 일반화 검증과 평가 모델의 예측 타당도 향상이 동시에 진전될 수 있다.

참고문헌

교육부(2023). 미래 사회를 대비하는 2028 대학입시제도 개편 확정안. 교육부 보도자료 (2023.12.27.).

노성(2021). 공정성 확보를 위한 입학사정관의 사회적 책무와 직무윤리. **한국부패학회보,** 26(2), 105-125.

류영철(2016). 대입 전형 평가자간 신뢰도 분석 및 평가영역별 학업성취도 영향 연구. 융합사회와 공공정책, 10(2), 161-198.

백순근(2004). **학위논문 작성을 위한 교육연구 및 통계분석**. 서울: 교육과학사.

백순근, 신안나, 김연경, 손주영, 윤승혜(2017). 고등학생용 여섯 가지 핵심역량 측정도구 개발 및 타당화 연구. **교육평가연구, 30**(3), 363-395.

양림(2020). 대학입학전형제도의 공정성에 관한 연구. **한국과 국제사회, 4**(2), 113-138.

임진택, 이정림, 방유리나, 이승아, 조민경, 박정선, 안미연, 차정민, 장준호, 김창민, 이수영(2022). **학생부종합전형 공통 평가요소 및 항목 개선 연구**. 2021 고교교육 기여대학 지원사업 공동연구.

최미숙, 서화정, 최상은, 김병주(2024). 신임 입학사정관 공통교육과정 설계 연구. **고등교육, 7**(1), 95-129.

최승배, 이영학(2013). 입학사정관제 평가점수에 대한 객관적인 평가 알고리즘 구현 연구. **한국데이터정보과학회지, 24**(6), 1359-1368.

한국대학교육협의회(2025). **2027학년도 대학입학전형 시행계획**. 한국대학교육협의회.

황희돈(2017). 입학전형별 신입생 종단연구. '학생부종합전형 3년의 성과와 고교 교육의 변화' 심포지엄(경희대 평화의 전당) 발표자료.

Brookhart, S. M. (2013). How to create and use rubrics for formative assessment and grading. Alexandria, VA: ASCD.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213-220.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378-382.

Gaines, B. R. (2013). Knowledge acquisition: Past, present and future. *International*

Journal of Human-Computer Studies, 71(2), 135-156.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130-144.

Krippendorff, K. (2019). *Content analysis: An introduction to its methodology* (4th ed.). Thousand Oaks, CA: SAGE.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

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, 100234.

Polanyi, M. (1966). *The tacit dimension*. Chicago: University of Chicago Press.

Popham, W. J. (1997). What's wrong — and what's right — with rubrics. *Educational Leadership, 55*(2), 72-75.

Ramesh, D., & Sanampudi, S. K. (2025). Large Language Model-Powered Automated Assessment: A Systematic Review. *Applied Sciences, 15*(10), 5683.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology, 56*(1), 150-166.

논문접수 : 2026.4.2. / 수정본 접수 : 2026.4.28. / 게재승인 : 2026.5.11.

ABSTRACT

An Exploratory Multi-Model Comparative Design and Validation of AI Preliminary Evaluation Models for Holistic Admissions Document Review

Heedon Hwang
Staff, Myongji University

Document review in holistic admissions in Korean universities is based on evaluators' qualitative judgment; therefore, differences in evaluators' experience and institutional evaluation environments may directly affect fairness among applicants. To mitigate such variation, a support tool is needed to apply each institution's evaluation criteria consistently. This study examined, through an exploratory multi-model comparison, whether an AI-based evaluation support tool can implement institution-specific evaluation criteria, secure construct validity, and contribute to reducing inter-evaluator variance. Based on the evaluation items from a joint study by five Korean universities (Lim et al., 2022), four AI preliminary evaluation models were constructed: Model A (LLM self-generated criteria), Model B (fact extraction layer), Model C (no-rubric control), and Model D (expert criteria encoding). Thirty mock student records were independently evaluated by the four models and by two expert evaluators from different universities.

The results showed that generic models achieved good rank agreement (Kendall W=.859) but low categorical agreement (Fleiss κ=.165). Agreement between the two cross-institutional experts was κ=.397, and all disagreements were confined to adjacent grades, indicating differences in evaluation criteria across institutions. Model D reproduced Expert A's judgment at κ=.648, while its agreement with Expert B was κ=.261. However, in the 18 cases where the two experts agreed, Model D showed the highest hit rate among the four models at 77.8%.

The findings suggest that encoding evaluation criteria is effective when applied to institution-specific criteria, and that the fact extraction layer effectively detects factuality deficits in records with excessive AI-generated wording. AI-based evaluation support tools may contribute to fairness by protecting applicants from evaluator- and year-based variance and by preventing students' genuine effort and growth from being obscured.

Keywords: holistic admissions, AI document evaluation, expert-criteria encoding, implementation fidelity, convergent validity