컴퓨터교육학회 논문지 2026년 제29권 제6호
https://doi.org/10.32431/kace.2026.29.6.006

AI-인간 협업 기반의 메타분석 1차 문헌 선별 프로세스 설계 및 검증

Designing and Validating an AI-Human Collaborative Process for Primary Literature Screening in Meta-Analysis

양정은† · 김귀훈††
JeongEun Yang† · Kwihoon Kim††

요약

본 연구는 메타분석 1차 문헌 선별의 비효율성을 개선하고자 mBERT 기반의 AI-인간 협업 자동선별 모델을 제안하였다. 제안된 모델을 국내 학술 문헌 1,012편을 대상으로 한 실증 연구에 적용한 결과, 제외 문헌 예측에서 상대적으로 안정적인 예측 경향을 확인하였다. 이를 바탕으로 AI와 인간의 역할 분담을 구조화한 세 가지 협업 시나리오를 설계하고 그 효용성을 검증하였다. 연구 결과, AI를 제2 코더 및 합의규칙, 제3의 코더로 활용 시 인간의 재검토 대상을 전체의 27.7%로 경감하고 76.7%의 문헌을 자동 확정하며, 연구자 간 합의를 지원하는 선별 효율성 및 협업 가능성을 확인하였다.

본 연구는 AI를 단순 대체자가 아닌, 반복적 제외 업무를 자동화하고 판단의 일관성을 보조하는 협업 파트너로 정립하여 새로운 메타분석 방법론의 실증적 토대를 마련하였다.

주제어 인간-AI 협업, 인공지능, 메타분석, 문헌 자동선별, AI기반 스크리닝, mBERT

ABSTRACT

This study proposes an mBERT-based automated AI-human collaborative screening model to address the inefficiency of primary literature screening in meta-analysis. When applied to an empirical study involving 1,012 domestic academic papers, the proposed model demonstrated relatively stronger performance in predicting excluded literature than included literature. Based on these findings, three collaboration scenarios structuring the division of roles between AI and humans were designed, and their effectiveness was verified.

The results indicated that utilizing AI as a second coder, a consensus rule, and a third coder reduced the proportion of literature requiring manual re-review to 27.7% of the total and automatically confirmed 76.7% of the literature, supporting researcher consensus and thereby suggesting the potential for screening efficiency and collaboration.

This study establishes an empirical foundation for a new meta-analysis methodology by positioning AI not as a simple substitute, but as a collaborative partner that automates repetitive exclusion tasks and supports consistency in judgment.

Keywords Human-AI Collaboration, Artificial Intelligence, Meta-analysis, Automated Literature Screening, AI-based Screening, mBERT

†정회원 한국교원대학교 대학원 컴퓨터교육전공 석사
††종신회원 한국교원대학교 인공지능융합 교육전공 부교수(교신저자)

논문투고 2026년 01월 20일
심사완료 2026년 03월 31일
게재확정 2026년 04월 13일
발행일자 2026년 06월 30일

1. 서론

메타분석(Meta-analysis)은 개별 연구들의 결과를 체계적으로 수집, 분석, 평가 및 합성하여 종합적인 결론을 도출하는 과학적 방법론으로 교육, 의학, 사회과학 등 다양한 학문 분야에서 근거기반 연구로 활용되고 있다[1,2]. 메타분석은 체계적 문헌고찰(Systematic Review, SR)의 절차적 틀을 따르며, 특히 문헌의 탐색과 선정과정에서는 PRISMA(Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 가이드라인이 표준으로 준용된다[3]. 분석의 신뢰도를 결정짓는 가장 핵심 단계는 문헌 선별(Screening)인데 연구 목적에 부합하는 문헌을 확보하기 위해 수천 편 이상의 방대한 문헌을 검토해야 하므로 가장 많은 자원이 투입되는 단계이기도 하다. PRISMA 가이드라인은 선정 과정의 투명성과 재현성, 일관성 확보를 위해 최소 두 명이상의 독립된 검토자가 참여할 것을 권고한다. 하지만 실제 연구 현장에서는 가용 자원의 한계로 복수의 전문 코더확보가 쉽지 않거나, 연구자의 수작업에 의존도가 높아 검토 효율성이 저하되고 연구자의 주관적 편향이 개입될 위험이 상존한다[4-8].

최근 발전한 인공지능(Artificial Intelligence, AI), 특히 대규모 언어모델(Large Language Model, LLM) 및 자연어처리(Natural Language Processing, NLP) 기술은 이러한 한계를 완화할 새로운 대안으로 주목받고 있다. 고도화된 텍스트 분석 모델은 문헌의 제목과 초록에서 의미적 맥락 정보를 분석하여 연구 주제의 적합성을 자동선별하거나, 인간 연구자와 협업하여 검토의 정확도를 제고하는 데 기여할 수 있다. 선행연구들은 AI 도입이 문헌 검토 시간을 단축하고 절차적 재현성을 높일 수 있음을 시사한다.

하지만 기존 연구는 주로 의료나 문헌정보학 분야에 집중되어 있고 사회과학 및 교육학 분야의 한국어 학술 문헌을 대상으로 한 실증적 연구는 상대적으로 미비하다. 또한 다수의 연구가 AI모델의 기술적 성능(정확도 등) 평가에 집중하고 있다. 또한 최신 스코핑 리뷰(Scoping review) 연구에 따르면, AI를 활용한 문헌 선별은 환각 및 편향 위험이 있으므로 인간의 감독과 협업이 필수적으로 요구된다[9]. 그러나 실제 선정 절차에서 AI와 인간의 협업이 검토자 간 일치도와 연구의 신뢰성에 미치는 영향에 대한 논의는 부족하다.

이에 본 연구는 한국어 학술 데이터를 기반으로 다국어 BERT(mBERT: Multilingual Bidirectional Encoder Representations from Transformers) 문헌 자동선별 모델을 구축하고, 이를 실증 연구에 적용하여 다음의 세 가지 목적을 달성하고자 한다. 첫째, AI 기반 자동화를 통해 문헌 검토의 부담을 경감하고 효율성을 제고한다. 둘째, 데이터 기반의 일관된 선별 기준 적용함으로써 재현 가능성을 확보한다. 셋째, AI모델과 인간 검토자의 선별 결과를 비교·분석하여 AI가 체계적 문헌 선별의 신뢰성을 보조하는 협업 도구로서 기능할 수 있는지 실증적으로 규명한다.

2. 선행연구 고찰

2.1 문헌선정 자동화의 필요성

연구자 수작업 기반의 문헌 선별은 연구자의 인지적 편향과 장시간 검토에 따른 피로누적으로 인해 검토자 내 신뢰도뿐 아니라 검토자 간 신뢰도에도 불일치 위험을 초래한다. Marshall 외(2019)는 체계적 문헌고찰에서 수천 편의 논문을 대상으로 한 수작업 선별이 수주에서 수개월이 소요되며, 검토자간 불일치가 재현성을 저해하는 주요 요인이라 하였으며, Tran 외(2024)는 인간 검토자만으로 수행한 문헌선정은 평균 소요 기간이 길고 높은 불일치를 보인다며 자동화의 필요성을 보고하였다[5,7]. 또한 일반적인 PRISMA 가이드라인이 인간 연구자 중심으로 문헌선정의 절차 투명성을 강조한다면 최근 의료분야에서 제안된 PRISMA-AI 가이드라인은 AI 중재 연구를 평가 할 때 재현성 확보에 필요한 기술적 세부사항과 편향 통제를 강조한다[10]. 비록 의료분야 AI 연구 보고에 특화되었으나 문헌 선별 도구로 AI를 도입할 때 알고리즘의 편향을 통제하고 기술적 재현성을 확보할 수 있는 투명한 체계가 필수임을 시사한다. 이에 따라 AI기반 문헌선정 자동화는 인간 연구자의 판단을 보조하고 편향을 완화하며, 선정 과정의 투명성과 신뢰성을 제고하는 방법론적 대안으로 논의되고 있다.

2.2 AI 기반 문헌 자동선별 기술 동향

초기 AI 기반 문헌 자동선별 기술은 텍스트 마이닝(Text Mining)과 머신러닝(Machine Learning, ML)을 활용하여 문헌의 포함/배제 여부를 분류하는 방식으로 시작되었다. O'Mara-Eves 외(2015)는 텍스트 마이닝이 문헌 선별 작업량을 감소시키면서도 높은 재현율을 유지할 수 있음을 입증하였고[4], 이후 시스템화한 ASReview와 같은 오픈소스 도구들이 등장하여[6] 사용자 중심 인터페이스와 능동 학습(Active Learning) 방식으로 불확실성이 높은 문헌을 우선적으로 학습시킴으로써 검토 효율성을 향상시켰다. 최근에는 LLM을 활용한 자동선별 시도가 급증하고 있는데 Tran 외(2024)와 Chibwe 외(2025)는 LLM이 인간과 유사하거나 더 높은 수준의 분류 정확도를 보일 수 있음을 시사하였다[7,11]. 그러나 Lieberum 외(2024)와 Khraisha 외(2024)는 LLM의 높은 운영비용, 환각(Hallucination) 현상, 프롬프트에 따른 결과의 비일관성, 재현성 부족 문제를 지적하며 AI를 완전한 대체재가 아닌 보조 도구로 활용해야 함을 강조하였다[9,12]. 이에 따라 AI를 제2의 검토자로 활용하거나, 인간이 결정 과정에 개입하는 'Human-in-the-Loop (HITL)' 시스템이 대안으로 제안되고 있다[8,13]. 국내의 경우는 한국어 학술 텍스트 처리를 위한 딥러닝 적용 연구가 꾸준히 진행되어 왔다. 김인후 외(2022)와 이용구(2023)는 문헌정보학 분야에서 한국어 학술 텍스트 처리를 위한 BERT 모델 적용 연구를 수행하였으나, 주제 분류나 키워드 추출에 초점을 맞추었고[14,15]. 설아람(2025)은 의료분야 체계적

문헌고찰에서 AI 활용 급증 추세를 분석하며, 국내 연구 환경에서도 효율적인 자동화 도구 도입이 시급함을 시사하였다[10]. 이와 같은 문헌 자동선별 기술의 흐름과 주요 선행연구의 핵심 결과를 종합하여 정리하면 Table 1과 같다.

Table 1. Analysis of Major Prior Research on Automated Literature Screening

Author(year)	Key Technique	Major Outcomes
O'Mara-Eves et al.(2015)[4]	Text Mining, ML(SVM)	30~70% workload reduction, Validated prioritization feasibility
Marshall et al.(2019)[5]	ML(Robot Reviewer) Active Learning	Validated potential for use as a 2nd reviewer
Gartlehner et al.(2019)[13]	Semi-automation (Distiller AI)	Low recall, Insufficient for full automation
van de Schoot et al.(2021)[6]	Active Learning (ASReview), BERT	Maintained 95% recall, 83% workload reduction
Tran et al. (2024) [7]	LLM (GPT-3.5 Turbo)	97% sensitivity, Est. 65% workload reduction
Khraisha et al.(2024)[12]	LLM (GPT-4)	Human-level performance, Noted hallucination risks
Chibwe et al.(2025) [11]	LLM(GPT-3/3.5/4)	(Wastewater epidemiology) Demonstrated cost-efficiency
Kallmes et al.(2025)[8]	LLM, ML (AutoLit/XGBoost)	82~97% recall, 50% time reduction
Lieberum et al.(2025)[9]	LLM Overview (Scoping Review)	Highlighted reproducibility issues, Need for HITL
Kim et al. (2022)[14]	BERT(KoBERT)	(LIS field) F1 0.90 on high-quality data
Lee(2023)[15]	BERT, Transfer Learning	(LIS field) High performance in subject classification
Kang et al. (2024) [16]	BERT, LLM(Llama2)	Proposed optimal model selection guide based on text length
Sul(2025)[10]	ML, LLM (Trend Analysis)	Analyzed increased AI adoption trends in medical reviews

* Note. SVM : Support Vector Machine, Est : Estimated, LIS : Library and Information Science

Table 1에서 나타난 바와 같이 기존 연구들은 주로 개별 도구의 분류 성능 평가나 완전 자동화 가능성을 검증하는 데에 집중해왔으나. 최근에는 LLM의 높은 분류 민감도와 효율성을 입증하면서도 환각 현상과 재현성 문제로 인해 완전 자동화보다는 인간의 참여(HITL)가 필수적임을 역설하고 있다. 그러나 인간과 AI가 1:1로 결합하는 고정된 이중선별 워크플로우 제안에 그치며 실제 연구자의 분석 환경에서 AI 예측 결과를 어느 기준까지 신뢰하여 확정적 판단으로 수용할지, 또 AI와의 역할 분담을 어떻게 설계할 것인지, 구조화된 협업 체계에 대한 논의는 여전히 부족하다. 따라서 본 연구는 불확실성이 있는 생성형 대신 명확한 패턴 학습에 강점이 있는 mBERT 기반 모델을 채택하여 판단의 안정성을 확보하고, 기술적 성능 평가를 넘어 AI의 예측확률 기반의 임계값을 도입하여 다양한 문헌 선별 환경에 유연하게 대응할 수 있는 AI와 인간의 실질적인 협업 프로세스를 설계하고 검증하고자 한다.

3. 연구 방법

3.1 연구 설계

본 연구는 메타분석 과정에서 문헌선정의 효율성과 신뢰성을 제고하기 위해 AI-인간 협업 기반 문헌 자동선별 모델을 설계하고, 이를 국내 학술데이터에 적용하여 실증적으로 검증하고자 한다. 본 연구에서 AI는 인간 연구자의 판단을 대체하는 자동 평가자가 아니라 반복적인 선별과정에서 판단의 일관성을 유지하고 의사결정 효율성을 지원하는 "협업자"로 개념화하였다.

또한 협업 모델을 구현하기 위한 알고리즘 선정 과정에서 LLM과 지도학습 기반 모델의 특성을 검토한 결과, 생성형 모델(Generative Model)은 문맥을 지나치게 포괄적으로 해석하여, 명백한 제외 문헌조차 포함으로 분류하는 과잉 탐지(Over-detection)의 한계가 있었다. 실제 사전분석에서 수집된 전체문헌의 서지 정보를 바탕으로 메타분석 연구자 역할을 부여하고 포함·제외 기준을 지침으로 제공하는 프롬프트를 적용하여 GPT-4 기반 스크리닝을 실시한 결과 대다수 문헌을 포함으로 분류하여 연구자의 재검토 부담을 실질적으로 줄여주지 못하는 문제가 확인되었다.

이에 본 연구는 특정 데이터의 패턴 학습에 최적화되어 명확한 분류 기준을 수립할 수 있는 mBERT 기반의 파인튜닝(Fine-tuning) 모델을 채택하였다. 구체적으로 단일 연구자의 편향 방지를 위해 일정 기간을 두고 반복 재검토로 검토자 내 신뢰도를 검증한 1차선별 결과를 정답 데이터로 하여 mBERT 기반 이진분류 모델을 학습시켰다. 이후 학습된 단일 AI모델을 동일하게 유지한 상태에서 문헌 선별과정 내 AI모델의 활용 방식만을 다르게 하는 세 가지 AI-인간 협업 시나리오를 설계하였다. 각 시나리오는 AI의 개입 수준과 역할에 따라 구분되며, AI는 최종 판단의 주체가 아닌 확률 기반 참조기준 제공자로 작동하도록 구조화하였다 (Table 2).

Table 2. Structure and Roles of AI-Human Collaboration Scenarios

	Scenario 1	Scenario 2	Scenario 3
Application Context	Supplement to multiple-coder system (AI pre-screening)	Single reviewer (Human-AI parallel review)	Cross-review by multiple reviewers (Human-human conflict resolution)
Role of AI	Second coder	Consensus rule provider	Reference indicator for conflict resolution

	Scenario 1	Scenario 2	Scenario 3
Decision-making Structure	AI Sole evaluation (Ternary classification: Include/ Uncertain/ Exclude)	Comparison after independent human-AI evaluations (Agreement: Confirm / Disagreement: Hold)	Third objective reference (Referenced upon human-human disagreement)
Classification & Judgment Rules	Application of conservative threshold, Uncertain assignment	Automatic confirmation only upon identical human-AI judgment	AI predicted probability as auxiliary indicator for disagreed literature
HITL	Intensive manual re-review of AI-classified Uncertain literature	Manual re-review of disagreed literature	Final consensus among reviewers referencing AI results

3.1.1 AI 기반 제2의 코더

첫 번째 시나리오는 AI를 인간 연구자의 판단을 보조·검토하는 제2의 코더로 활용하는 구조이다(Fig. 1).

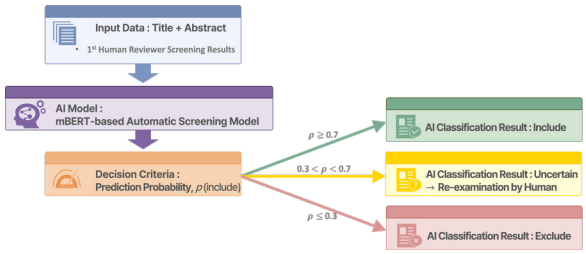


Figure 1. AI-based Second Coder

AI모델은 인간 연구자가 1차로 수행한 문헌 선별 결과를 재분석하며, 단순한 이진 판단 대신 각 문헌에 대해 포함 예측확률을 산출하되 AI의 예측 결과를 그대로 수용하지 않고, 예측확률에 따라 "포함", "제외", "불확실" 세 범주로 구분하는 삼분 의사결정 구조를 적용하였다. 예측확률 임계값은 사전 테스트에 근거하여 재현율 손실을 최소화할 수 있는 보수적 수준인 포함($p \geq 0.7$), 제외($p \leq 0.3$), 불확실($0.3 < p < 0.7$)로 설정하였다. 이는 불확실 영역의 문헌만 인간 연구자의 집중 재검토로 유도하여, AI의 잠재적 오분류 위험을 통제함과 동시에 전체 전수 검토 대비 물리적 효율성을 극대화하기 위함이다. 이를 통해 AI 분류의 불확실성을 명시적으로 드러내고, 최종 판단의 책임은 인간 연구자에게 귀속되도록 설계하였다.

3.1.2 단독 연구자를 위한 AI 기반 합의규칙

두 번째 시나리오는 복수의 검토자를 확보하기 어려운 단독 연구자 환경을 고려하여 설계하였다. AI의 예측확률을 인간 판단의 합의 기준으로 활용하는 방식으로 예측확률이 포함($p \geq 0.6$)과 제외($p \leq 0.4$) 기준을 충족하는 경우에만 인간 연구자의 기존 판단을 자동 확정하고, 기준을 충족하지 못하는 중립구간($0.4 < p < 0.6$)의 문헌은 재검토 대상으로 분류하였다(Fig. 2).

시나리오 1 대비 임계값을 완화한 이유는 재현율 손실을 허용 가능한 범위 내로 통제하면서 자동 확정 비율을 높여 제한된 자원 하에서 물리적 업무 효율성을 극대화하기 위함이다. 이와 같은 합의 규칙을 통해 단독연구자의 검토 부담을 실질적으로 경감하면서, AI와의 교차 검증을 통해 단독 판단에 따른 주관적 자의성을 보완하도록 하였다.

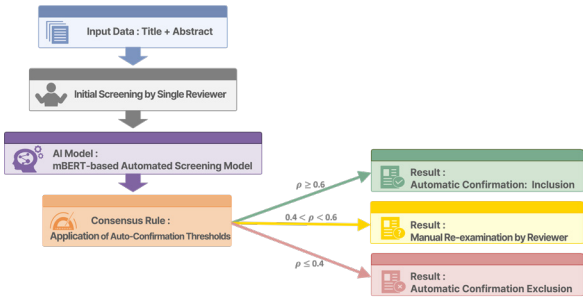


Figure 2. AI-based Consensus Rule for a Single Researcher

3.1.3 AI 기반 제3의 코더(갈등 조정)

세 번째 시나리오는 두 명 이상의 인간 연구자 간 판단 불일치가 발생한 상황을 가정하여 AI를 제3의 코더, 갈등 조정을 위한 확률 기반 참조기준 제공자로 활용하는 구조이다(Fig. 3).

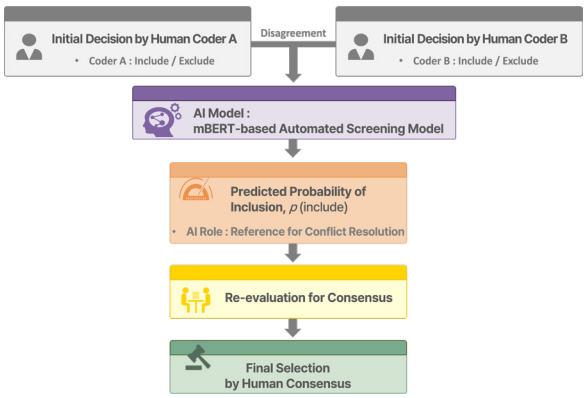


Figure 3. AI-based Third Coder (Conflict Resolution)

AI모델이 제시하는 포함 확률은 특정 판단 방향에 대해 상대적으로 더 높은 근거, 가중치를 의미한다. 단, AI의 제안은 참고 정보로만 활용하며, 최종 결정은 인간 연구자 간의 재논의를 통해 이루어지게 하였다. 이는 AI가 판단의 중재자나 판정자로 기능하는 것을 의도적으로 지양하고 인간 연구자 간의 합의 형성을 지원하는 의사결정 지원 도구로서의 역할로 한정하였다. 특히 연구자 간 개인의 주관적 해석이 충돌할 때, 데이터 패턴에 기반한 객관적 참조 지표로 활

용하여 논의의 방향을 잡고 편향 완화에 기여하도록 설계하였다.

3.2 데이터 수집

본 연구의 데이터는 실증 연구「인공지능 활용 프로그래밍 교육의 효과의 조절변인 분석 및 구조적 관계 탐색」의 문헌 선정 과정에서 확보된 자료로, 2020년부터 2025년까지 국내 학술 데이터베이스를 대상으로 “프로그래밍 교육”, “인공지능” 등의 검색어를 활용하여 수집되었다. 총 1,012편의 학술지 논문 및 학위논문이 확보되었으며, 인간 연구자의 1차 문헌 선별을 통해 123편이 포함, 889편이 제외 논문으로 분류되었다.

3.3 데이터 전처리

수집된 문헌 데이터는 최소 수준의 텍스트 정제를 거친 후, mBERT에 내장된 WordPiece 기반 토크나이저를 활용하여 토큰화하였다. 모델 입력 텍스트는 문헌의 제목과 초록을 결합하여 구성하였으며, 초록이 제공되지 않은 경우는 결측 정보임을 명시적으로 표시하여 입력에 포함하였다. 또한 영어 등 비한국어 문헌에 대해서도 mBERT의 다국어 처리 특성을 활용하여 동일 모델 구조 내에서 처리하였다.

3.4 학습 데이터 구성 및 불균형 처리 전략

AI모델의 학습에는 인간 연구자가 수행한 1차 문헌 선별 정보만을 정답 데이터로 활용하였다. 원문 검토를 통해 통계치 부재 등으로 메타분석 수행요건을 충족하지 못해 제외된 최종 선정 결과는 모델 학습에 포함하지 않고 이후 시나리오 분석 단계에서 AI 판단의 한계를 검증하기 위한 기준값으로만 활용하였다. 이러한 데이터 구성은 AI모델이 접근 가능한 정보의 범위를 인간 연구자의 1차선별 단계와 동일하게 통제함으로써, AI가 인간 연구자의 판단을 보조하는 역할에 충실하도록 설계하기 위함이다.

한편 본 연구의 데이터는 포함과 제외 문헌 간 표본 수 불균형을 보인다. 일반적 방식으로 학습을 진행하면 모델이 다수 범주인 제외로 편향되어 학습될 위험이 있다[17]. 층화 추출(Stratified Sampling)은 데이터 불균형 상황에서 클래스 비율을 보전하여 학습 편향을 완화하는 데 효과적인 기법으로 알려져 있어[18], 이를 근거로 본 연구는 학습·검증 데이터 분할 시 층화 추출을 적용하여 클래스 비율을 유지하였고, 모델 학습 단계에서는 비용 민감 학습(Cost-sensitive Learning)을 적용하여 손실 함수(Loss Function)에 클래스 가중치를 부여하였다. 이는 메타분석 1차 선별과정의 핵심 원칙으로 요구되는 누락 최소화(Recall-oriented Screening)를 모델 학습 단계에 기술적으로 구현하고자 한 것이다. 선행 연구들 또한 자동화된 문헌 선별 시스템은 잠재적 포함 문헌의 누락을 방지하기 위해 재현율 확보 우선 원칙을 준수해야 한다는 제언을 하고 있다[4,11,13]. 이에 따라 소수 범주인 ‘포함 문헌’에 대한 오분류(Misclassification) 비용을 높게 설정하여 모델의 학습 편향을 완화하고 높은 재현율을 확보하고자 하였다.

3.5 AI 자동선별 모델 구축

AI 자동선별 모델은 mBERT 사전학습 모델을 기반으로 구축되었으며, 앞서 수립된 전략에 따라 전처리한 데이터를 입력하여 이진 분류 미세조정(Fine-tuning)을 수행하였다. 모델학습에는 선행연구[19]에서 mBERT 계열 모델의 텍스트 분류 과업(Task)에 공식적으로 권장하는 표준 하이퍼파라미터 기본값(학습률 2e-5, 배치크기 16, epoch 3)을 그대로 적용하여 모델의 안정성을 확보하고 과적합을 방지하였다. 데이터 불균형 환경에서 모델이 보이는 범주별 예측 특성과 편향성을 다각도 파악하기 위해 모델의 성능 평가는 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 Score를 종합적으로 산출하였다. 추가적으로 AI모델과 인간 연구자 간의 판단 일치를 검증하기 위해 Cohen‘s Kappa 계수를 산출하여 분석하였다.

4. 연구 결과

4.1 AI 자동선별 모델 학습 및 성능 평가

4.1.1 데이터 구성 및 학습 안정성

본 연구는 총 1,012편의 문헌을 분석 대상으로 하였으며, 이 중 포함 문헌은 123편(12.2%), 제외 문헌 889편(87.8%)으로 클래스 불균형이 뚜렷하였다. 모델 학습 및 검증을 위한 데이터 셋 구성은 Table 3과 같다.

Table 3. Dataset Composition and Label Distribution

Category			Number of Papers		Ratio	
Total			1012		100%	
Human Researcher	Included		123		12.2%	
	Excluded		889		87.8%	
AI Model	Training Set	Include	809	98	80%	12.1%
		Exclude		711		87.9%
	Test Set	Include	203	25	20%	12.3%
		Exclude		178		87.7%

학습 과정에서 손실값(Loss)의 변화는 Table 4, Fig. 4와 같다.

Table 4. Training and Validation Loss by Epoch

Step	50	100	150
Training Loss	0.7087	0.6840	0.5581
Validation Loss	0.6783	0.6333	0.6090

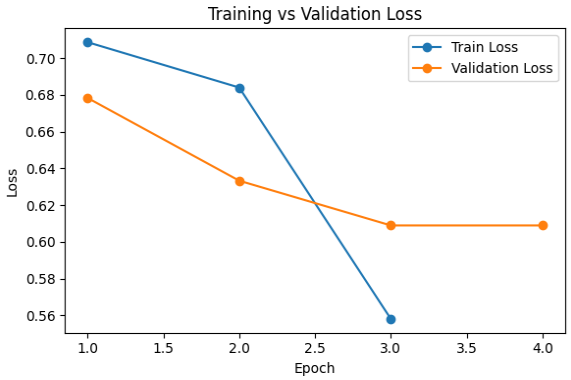


Figure 4. Training and Validation Loss Curves

학습이 진행됨에 따라 학습 손실은 0.7087에서 0.5581로 검증 손실은 0.6783에서 0.6090으로 점진적인 감소 추세를 보였다. 학습 종료 시점까지 검증 손실이 반등하는 전형적인 과적합 패턴이 관찰되지 않았고, 학습 손실과 검증 손실 간의 격차가 크지 않은 상태를 유지하였다. 이러한 손실값의 동반 하락 추세와 좁은 격차를 고려할 때, 본 모델은 학습 데이터에 과도하게 편향되지 않고 비교적 안정적인 수준으로 수렴한 것으로 판단된다.

4.1.2 AI 기반 자동선별 모델의 분류 성능

본 연구에서 구축한 AI 자동선별 모델을 검증 데이터를 바탕으로 성능을 평가한 결과는 Table 5와 같다.

Table 5. Classification Performance Metrics of AI Automated Screening Model

	Accuracy	Precision	Recall	F1 Score	Cohen's Kappa
AI Model	0.7882	0.2955	0.5200	0.3768	0.2607

주요 평가지표인 정확도(0.7882), 정밀도(0.2955), 재현율(0.5200), F1-Score(0.3768)와 함께 Cohen's Kappa계수는 0.2607로 산출되었다. 이는 AI모델과 인간 연구자 간 판단이 약간의 일치(Fair Agreement)를 보이는 수준임을 의미한다[20]. 비록 주관적 해석이 개입될 수 있는 과업의 특성을 고려하면 수용 가능한 유의미한 수치이나, 학습 데이터 불균형에 대비해 비용 민감 학습 등의 처리 전략을 적용하여 소수 클래스에 대한 예측력을 확보하고자 하였음에도 불구하고 정밀도와 재현율이 상대적으로 낮게 산출된 것은 메타분석 선별 데이터가 갖는 압도적인 클래스 불균형(제외 87.8%)과 문헌 정보(제목과 초록)의 모호성에 기인한 것으로 분석되며, AI의 단독 활용에는 분명한 한계가 있음을 시사한다.

4.1.3 전체 문헌 대상 예측양상 분석

본 절에서는 협업 시나리오 적용에 앞서, 학습 데이터를 포함한 전체 문헌을 대상으로 AI모델이 어떠한 예측 분포를 보이는지 분석하였다. 분류 기준으로는 '표준 이진 분류 임계값(0.5)'을 적용하였다. 이는 모델의 일반적인 분류 경향성을 파악하기 위한 목적으로, 이후 협업 시나리오에서 적용되는 확률 기반 의사결정 규칙과는 구별된다.

혼동행렬 분석 결과(Fig. 5) 인간 연구자가 제외로 판단한 문헌 889편 가운데 764편은 AI모델 역시 동일하게 제외로 예측 하였으나, 125편은 포함으로 과대 예측하였다. 반면 인간 연구자가 포함으로 판단한 문헌 123편 중 AI모델도 정확히 포함으로 예측한 문헌은 84편이나, 제외로 예측하여 누락 시킨 문헌은 39편으로 나타났다.

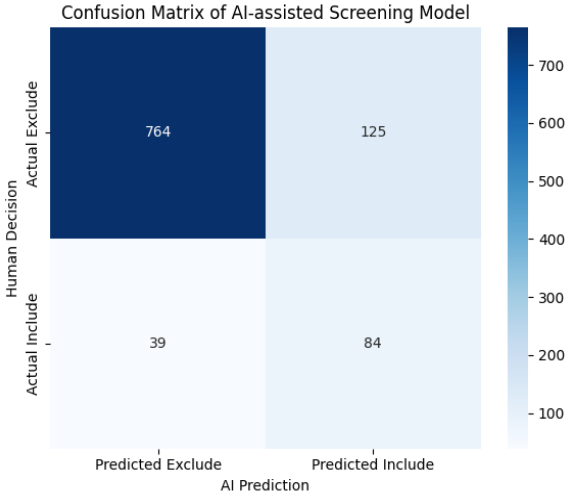


Figure 5. Confusion Matrix of AI-assisted Screening Model

이러한 결과는 AI모델이 제외 문헌에 대해서는 상대적으로 높은 예측 일치를 보이는 반면 포함 문헌에 대해서는 누락의 위험이 존재함을 보여준다. 따라서 실제 선별과정에서는 이러한 누락 위험을 구조적으로 제어할 수 있는 인간 연구자의 검토가 필수적으로 병행되어야 함을 시사한다.

4.2 AI-인간 협업 시나리오별 적용 결과

앞서 분석한 예측 값을 토대로 AI의 역할과 개입 수준을 달리한 세 가지 AI-인간 협업 시나리오의 효과성을 분석하였으며 각 시나리오의 AI모델의 예측확률은 사전 설정한 확률 기반 구간을 적용하였다.

4.2.1 시나리오 1: AI 기반 제2의 코더

시나리오 1은 AI를 제2의 코더로 활용하여 문헌 누락을 방지하고 인간 연구자의 초기 검토량을 경감하는 것을 목적으로 하였다. 전체 1,012편의 문헌을 대상으로 AI의 삼분 분류를 수행하였으며 결과는 Fig. 6과 같다.

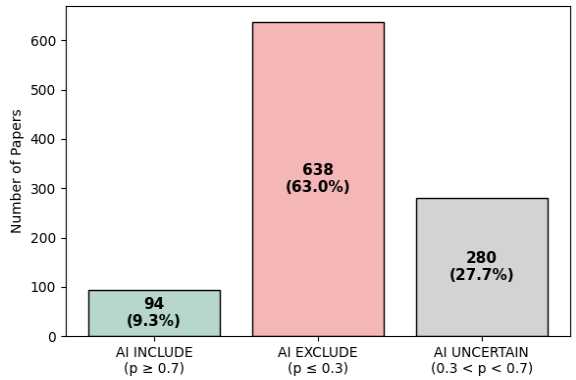


Figure 6. AI-based Second Coder: Three-way Decision Results

AI모델은 전체 문헌의 63%인 638편을 제외로, 9.3%인 94편은 포함으로 판단하였다. 판단이 불확실하여 인간 연구자의 재검토가 필요한 문헌은 27.7%인 280편이다. 이는 인간 연구자가 AI모델을 제2의 코더로 활용 시 약 72.3%에 해당하는 732편에 대해 AI 사전 선별 결과를 확정적 판단(포함/제외)으로 활용할 수 있음을 시사한다.

4.2.2 시나리오 2: AI 기반 합의규칙

시나리오 2는 단독 연구자의 검토에서 인간과 AI의 판단 결과를 대조하여 반복적인 제외 판단을 자동화하여 효율성을 극대화하는 것을 목적으로 하였다. 전체 문헌을 대상으로 인간 연구자와 AI 모델의 예측 결과를 교차 검토하였으며 결과는 Fig. 7과 같다.

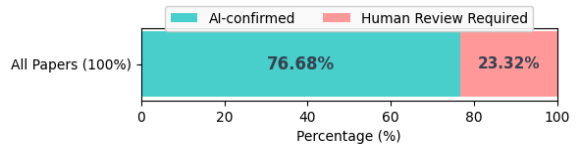


Figure 7. Application Results of AI-based Consensus Rule for a Single Researcher

AI모델의 예측확률이 인간 연구자의 판단과 일치하여 자동 확정 처리된 문헌은 총 776편으로 전체 문헌의 76.7%가 별도의 논의 없이 AI-인간 간 합의 규칙에 의해 처리 될 수 있음을 의미한다. 반면 합의가 성립되지 않은 문헌 125편(12.4%)과 불확실한 문헌 111편(10.9%)은 인간 연구자의 수동 재검토 대상으로 분류되었다.

4.2.3 시나리오 3: AI 기반 제3의 코더(갈등 조정)

시나리오 3은 복수 연구자 간 교차 검토 시에 발생할 수 있는 주관적 의견 불일치를 완화하고 객관적 합의를 지원하기 위해 AI를 참조지표로 활용하는 것을 목적으로 하였다. 인간 연구자의 1차선별 결과를 기준으로 가상의 제2 코더를 구성하고 두 코더 간 불일치 문헌 151편을 추출하여 AI모델의 확률 기반 조정 기준을 적용하였으며 결과는 Fig. 8과 같다. 109편(72.2%)에 대해 AI모델은 제외판단을 상대적으로 지지하는 것으로 나타났으며, 19편(12.6%)은 포함 판단을 지지하는 방향으로 분류되었다. 나머지 23편(15.2%)은 불확실로 분류하였다.

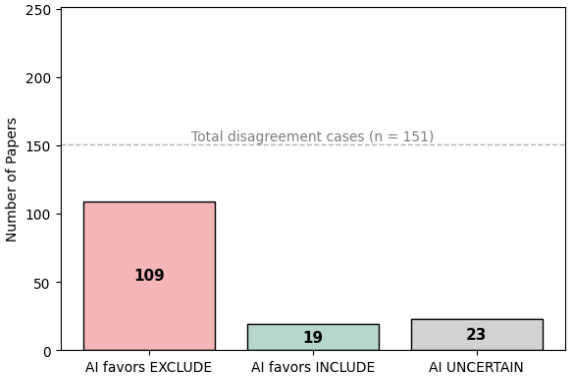


Figure 8. Application Results of AI-based Third Coder

AI모델이 제외를 지지한 문헌을 대상으로 인간 연구자의 1차선별 판단과 비교한 결과, 127편(84.1%)은 인간 연구자 역시 제외로 판단한 문헌이었으며, 24편(15.9%)은 인간 연구자가 포함으로 판단한 문헌으로 확인되었다(Fig. 9).

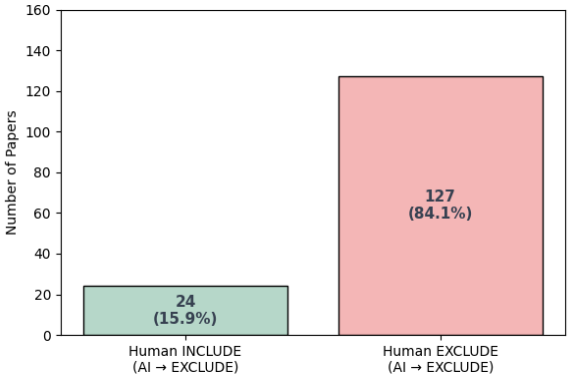


Figure 9. Distribution of Human Screening Decisions for Papers Supported for Exclusion by AI

이는 두 코더 간 판단이 불일치한 상황에서도 AI가 제외를 지지한 문헌의 다수가 인간 연구자의 선별 판단과 일치하는 방향에 있음을 시사한다.

4.3 시나리오의 검토 효율성 및 누락 위험 검증

4.3.1 시나리오 1과 2의 비교

시나리오 1과 시나리오 2의 적용 결과를 보면 인간 연구자의 재검토 부담 수준에는 차이가 나타났으며, Fig. 10과 같다. 시나리오 1에서는 전체 문헌 중 27.7%가 인간연구자의 재검토 대상으로 분류된 반면, 시나리오 2에서는 12.4%로 크게 감소하였다.

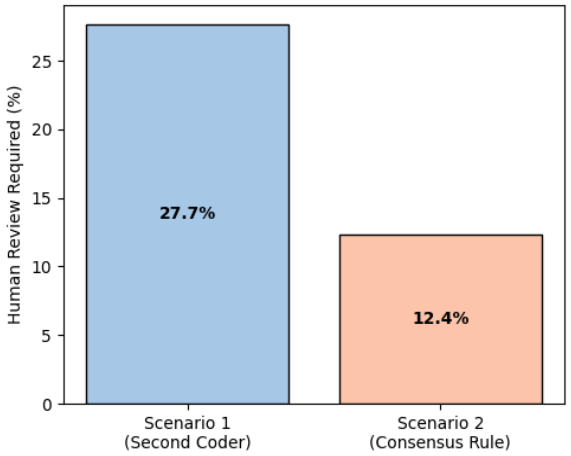


Figure 10. Comparison of Manual Re-review Burden between Scenarios 1 and 2

4.3.2 최종 포함 문헌의 누락 위험 검증

메타분석에서 1차 문헌 선별의 핵심은 "포함 가능성이 있는 문헌의 누락 최소화"이다. 즉, 최종적으로 포함되어야 하는 문헌을 누락하지 않았는지 검증되어야 한다. 이에 따라 실증 연구의 메타분석에 포함된 최종 선정 문헌 15편을 기준으로 AI 모델의 단독 예측 결과(이진 분류, 임계값 0.5 적용)를 추적한 결과, 6편의 문헌에서 누락 위험이 확인되었다. 이러한 6편의 위험 문헌에 시나리오별 규칙을 적용한 결과, 시나리오 1에서는 4편이 불확실 범주로 분류되었으나, 예측확률이 0.3 이하로 도출된 2편은 자동 제외되어 누락이 발생하였다. 반면 시나리오 2에서는 AI가 제외로 판정하더라도 인간 연구자가 포함으로 선별한 경우 재검토로 분류되는 구조적 특성에 따라 6편 전부 자동 제외되지 않고 수동 재검토 목록으로 확인되었다. 시나리오 3에서도 AI는 연구자 간 불일치 상황에서 최종 판정 권한이 없는 참조지표로만 기능하므로 AI가 제외를 지지하더라도 자동 누락이 아닌, 연구자 간 최종 합의 재논의 대상으로 분류됨을 확인하였다.

5. 결론 및 제언

본 연구는 메타분석 과정에서 1차 문헌 선별에 AI가 인간 연구자의 판단을 어떻게 보조할 수 있는지를 시나리오 기반으로 실증적으로 탐색하였다. 이를 위해 본 분석에서 AI를 대체자가 아닌 협업자로 개념화 하고, mBERT 기반 자동선별 모델을 구축하여 세 가지 협업 시나리오에서 유용성을 검증하였다.

연구 결과, 인간 연구자의 1차선별 결과를 학습한 AI 자동선별 모델은 전반적으로 인간 연구자의 판단을 일정 수준 재현하였으나 메타분석 선별 데이터 특성인 클래스 불균형과 초록 정보의 내재적 모호성에 따른 한계로 인해 포함 문헌보다는 제외 문헌에 대해 상대적으로 높은 예측 일치를 보였다. 이러한 데이터의 한계는 상대적으로 낮은 정밀도 및 재현율과 6편의 문헌 누락 위험으로 나타났다. 이는 단순한 정확도 지표만으로는 AI를 독립적 선별 도구로 활용하기에는 한계가 있음을 시사한다. 또한 본 연구에서 확인된 AI모델의 특성은 문헌 누락을 방지해야하는 메타분석의 엄격성을 고려할 때 AI에게 선정의 권한보다는 명확한 배제의 역할을 부여하는 것이 타당할 수 있음을 시사한다. 즉, AI의 한계를 인간의 고차원적 검토를 보완하는 상호보완적 프로세스로서 연구자의 인지적 자원을 불확실한 영역에 집중시키는 전략적 유용성을 가진다. 이러한 설계 원칙에 따라 분석된 AI-인간 협업 시나리오별 주요 결론은 다음과 같다.

첫째, AI를 인간 연구자의 판단을 재검토하는 제2의 코더로 활용할 때 삼분 의사결정(포함/불확실/제외) 구조는 AI 판단 불확실성을 명시적으로 드러내면서 인간 연구자의 재검토 부담을 전체 문헌의 약 27.7% 수준으로 줄이는 효과를 보였다. AI 자동선별 모델로 다수 코더체계를 유지하면서 실질적인 검토 부담을 완화할 수 있음을 의미한다.

둘째, 단독 연구자를 위한 합의 규칙을 제공하는 판단 지원 도구로 활용 할 때, 전체 문헌의 76.7%가 AI-인간 간 판단 일치에 따라 자동 확정 처리되었다. 비록 모델의 보수적 경향으로 포함 자동 확정 비율은 제한적이었으나 반복적이고 소모적인 제외 판단 과정을 자동화함으로써 연구 효율성을 실질적으로 향상시키는데 기여 할 수 있음을 확인하였다.

한편 시나리오 1에서는 시나리오 2에 비해 불확실 문헌의 비율이 상대적으로 높게 나타났는데 이는 AI모델이 포함과 제외 판단에 대해 높은 확신을 가지는 경우에만 자동 분류를 수행하도록 임계값(0.7)을 보수적으로 설정한 결과이다. 임계값을 0.6으로 완화하면 누락 오류(30편)가 증가하여 선별 안전성이 저하되고, 임계값을 0.8로 상향하면 인간 연구자의 재검토 비율(53.7%)이 상향되어 협업 효율성 저하가 우려된다. 따라서 문헌 누락을 방지하는 안전성과 검토량 경감이라는 효율성 간의 최적 타협점으로 0.7을 채택하였다. 이는 시나리오 1에서의 AI모델은 판단 오류로 인한 누락 위험을 최소화 하는 대신 인간 연구자의 개입 범위를 일정 수준 유지하는 협업 구조임을 의미한다.

셋째, 연구자 간 불일치 상황에서 갈등 조정자로 활용할 때, AI가 최종 판정자가 아닌 합의 형성을 돕는 객관적 참조 기준으로 기능할 수 있는 가능성을 보였다.

결과적으로 본 연구가 제안한 시나리오는 효율성 확보와 동시에 AI의 단독 오분류로 인한 누락 위험을 인간 연구자의 교차 검증과 단계적 협업으로 보완할 수 있음을 보여준다. 이를 바탕으로 연구자 환경에 따른 실제 분석 현장에 적용해본다면, 복수의 코더 확보가 어려운 단독 연구자 환경에서는 시나리오 2를 적용하여 반복적인 제외 업무를 자동화하고 동시에 AI와의 교차 검증을 통해 단독 판단의 편향과 누락 위험을 통제할 수 있다. 반면 복수의 코더가 확보된 팀 단위 연구 환경에서는 시나리오 1과 같이 AI가 사전 선별을 수행하여 전체 검토량을 줄인 후, 인간 연구자 간의 교차 검토 중 발생하는 의견 불일치에 대해서 시나리오 3을 적용

하여 객관적 합의를 도출하는 단계적 접근이 효과적일 수 있다.

종합하면 본 연구는 메타분석 1차 문헌 선별과정에서 AI 자동선별 모델의 가치가 단순한 분류 성능의 우수성에 있는 것이 아니라 협업 설계 방식에 따라 연구의 투명성, 일관성, 효율성을 향상시키는 도구로 확장될 수 있음을 실증적 사례로 탐색하였다는 점에서 의의를 가진다. 특히 인공지능 교육, SW교육 등 관련 연구물이 급증하고 있는 컴퓨터 교육 분야에서 본 연구가 제안한 협업 모델은 방대한 연구물을 체계적으로 정리하고 근거 기반의 연구방법론의 기틀이 될 것으로 기대한다.

그러나 본 연구가 제목과 초록 정보에 한정된 데이터를 활용하여 전체 문맥 파악에는 한계가 있다는 점, 단일 학문 분야의 국내 문헌 중심 분석이라는 점에서 일반화 한계가 있다. 향후 연구에서는 AI-인간 협업 모델의 실효성을 높이기 위해 다음과 같은 후속연구를 제언한다.

첫째, AI 자동선별 모델 구조의 고도화가 필요하다.

명확한 (포함·제외)기준은 규칙으로, 모호한 문헌은 AI가 판단하는 Rule-based+mBERT 하이브리드 모델이나, LLM에 퓨샷러닝(Few-shot Learning)을 적용하여 인간의 판단 논리를 학습시키는 전략이 요구된다[21].

둘째, 데이터 활용 범위의 확장이다. 제목과 초록을 넘어 전문(Full-text)의 데이터를 활용[8,12]하거나 다국어 데이터베이스로 분석 대상을 확장하여 판단의 정확도를 높이는 방법이 필요하다.

셋째, 설명 가능한 AI(Explainable Artificial Intelligence, XAI) 기능의 도입이다. 단순 분류 결과뿐 아니라 근거 기반 판단, 포함·제외의 사유를 자동으로 생성하고 제공함으로써 AI 판단의 투명성과 신뢰성을 강화하는 방향으로 연구가 진행되어야 할 것이다[22].

참고문헌

[1] Hwang, S. (2020). Meta-Analysis using R (2nd ed.). Seoul: Hakjisa.

[2] Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2021). *Introduction to Meta-Analysis*. NJ: Wiley.

[3] Page, M., McKenzie, J., Bossuyt, P., Boutron, I., Hoffmann, T., Mulrow, C., Shamseer, L., Tetzlaff, J., Akl, E., Brennan, S., Chou, R., Glanville, J., Grimshaw, J., Hróbjartsson, A., Lalu, M., Li, T., Loder, E., Mayo-Wilson, E., McDonald, S., McGuinness, L., Stewart, L., Thomas, J., Tricco, A., Welch, V., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. https://doi.org/10.1136/bmj.n71

[4] O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Systematic Reviews*, 4, 5. https://doi.org/10.1186/2046-4053-4-5

[5] Marshall, I., & Wallace, B. (2019). Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, 8, 163. https://doi.org/10.1186/s13643-019-1074-9

[6] Van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., & Oberski, D. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3, 125-133. https://doi.org/10.1038/s42256-020-00287-7

[7] Tran, V., Gartlehner, G., & Yaacoub, S. (2024). Sensitivity and specificity of using GPT-3.5 Turbo models for title and abstract screening in systematic reviews and meta-analyses. *Annals of Internal Medicine, 177*(6), 791-799. https://doi.org/10.7326/M23-3389

[8] Kallmes, K., Thurnham, J., Sauca, M., Tarchand, R., Kallmes, K., & Holub, K. (2025). Human-in-the-loop artificial intelligence system for systematic literature review: Methods and validations for the AutoLit review software. *Cochrane Evidence Synthesis and Methods, 3*(6), e70059. https://doi.org/10.1002/cesm.70059

[9] Lieberum, J., Töws, M., Metzendorf, M., Heilmeyer, F., Siemens, W., Haverkamp, C., Böhringer, D., Meerpohl, J., & Eisele-Metzger, A. (2025). Large language models for conducting systematic reviews: On the rise, but not yet ready for use – a scoping review. *Journal of Clinical Epidemiology, 181*, 111746. https://doi.org/10.1016/j.jclinepi.2025.111746

[10] Sul, A. (2025). Utilization of AI in conducting medical systematic reviews. *Journal of Health Technology Assessment, 13*(1), 1-7.

[11] Chibwe, K., Mantilla-Calderon, D., & Ling, F. (2025). Evaluating GPT models for automated literature screening in wastewater-based epidemiology. *ACS Environmental Au, 5*(1), 61-68. https://doi.org/10.1021/acsenvironau.4c00042

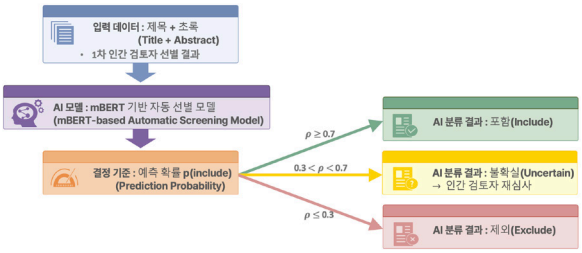
[12] Khraisha, Q., Put, S., Kappenberg, J., Warraitch, A., & Hadfield, K. (2024). Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Research Synthesis Methods, 15*(4), 616-626. https://doi.org/10.1002/jrsm.1715
[13] Gartlehner, G., Wagner, G., Lux, L., Affengruber, L., Dobrescu, A., Kaminski-Hartenthaler, A., & Viswanathan, M. (2019). Assessing the accuracy of machine-assisted abstract screening with DistillerAI: A user study. *Systematic Reviews, 8*(1), 277. https://doi.org/10.1186/s13643-019-1221-3
[14] Kim, I., & Kim, S. (2022). Automatic classification of academic articles using BERT model based on deep learning. *Journal of the Korean Society for Information Management, 39*(3), 293-310. https://doi.org/10.3743/KOSIM.2022.39.3.293
[15] Lee, Y. (2023). A study on automatic classification of subject headings using BERT model. *Journal of the Korean Society for Library and Information Science, 57*(2), 435-452. https://doi.org/10.4275/KSLIS.2023.57.2.435
[16] Kang, K., Li, Y., & Hong, A. (2024). Construction and Performance Analysis of an Automatic Classification Model for Domestic Academic Papers Using BERT and Llama2. *Korean Review of Corporation Management, 15*(4), 103-128.
[17] He, H., & Garcia, E. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239
[18] Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks, 106*, 249-259. https://doi.org/10.1016/j.neunet.2018.07.011
[19] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://aclanthology.org/N19-1423/
[20] Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310
[21] Akinseloyin, O., Jiang, X., & Palade, V. (2024). A question-answering framework for automated abstract screening using large language models. *Journal of the American Medical Informatics Association, 31*(9), 1939-1952. https://doi.org/10.1093/jamia/ocae166
[22] Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765-4774). Curran Associates, Inc. https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions

양정은
· 2005년 경상대학교 경제학과(학사)
· 2026년 한국교원대학교 컴퓨터교육전공(석사)
관심분야 : 프로그래밍 교육, AI 교육, AI활용 및 교육데이터 기반 학습자 맞춤형 교육설계 등
jisyang7@gmail.com

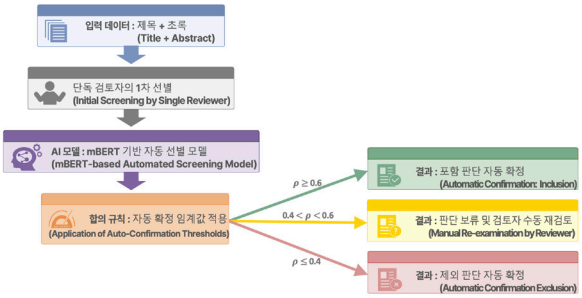
김귀훈
· 1998년 KAIST 전기및전자공학과(공학사)
· 2000년 KAIST 전기및전산학과(공학석사)
· 2019년 KAIST 전기및전자공학부(Ph.D.)
· 2000년~2005년 LG 데이콤 주임연구원
· 2005년~2020년 ETRI 실장, 책임연구원
· 2020년~현재 한국교원대학교 인공지능융합교육전공 교수
· 2006년~현재 ITU-T SG11 Rapporteur, Editor
· 2022년~현재 한국AI융합교육연구소 소장
· 2024년~현재 교육정보원 원장
관심분야 : 인공지능융합교육, AI디지털교과서, 지능형에지컴퓨팅, 강화학습
kimkh@kace.ac.kr

부 록

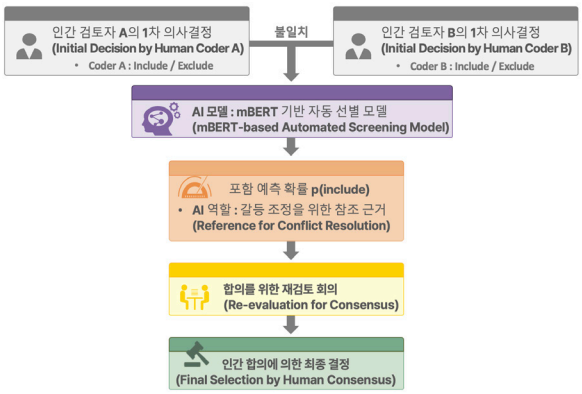
본 논문의 영문(초록과 표) 작성과 문장 개선을 위해 ChatGPT 5.3과 Gemini 3.1을 사용하였으며, 저자가 최종 내용을 검토하고 편집하였습니다.



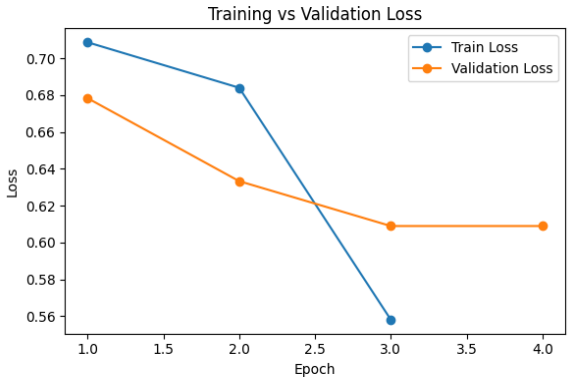
[그림 1] AI 기반의 제2의 코더



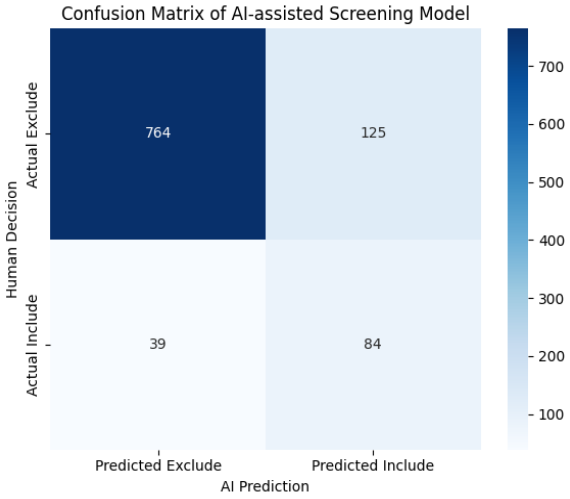
[그림 2] 단독연구자를 위한 AI 기반 합의규칙



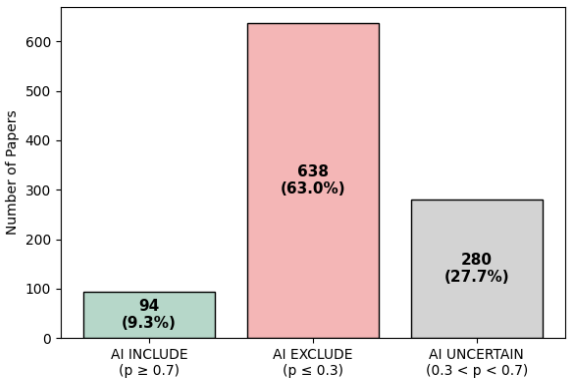
[그림 3] AI 기반의 제3의 코더(갈등 조정)



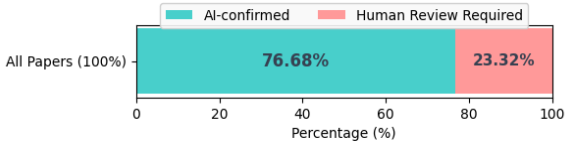
[그림 4] 학습 손실과 검증 손실 곡선



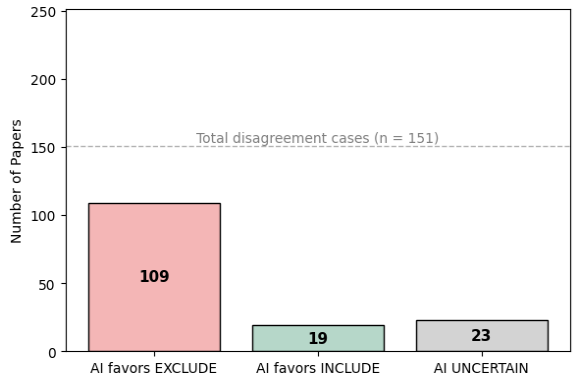
[그림 5] AI 자동선별 모델의 혼동행렬



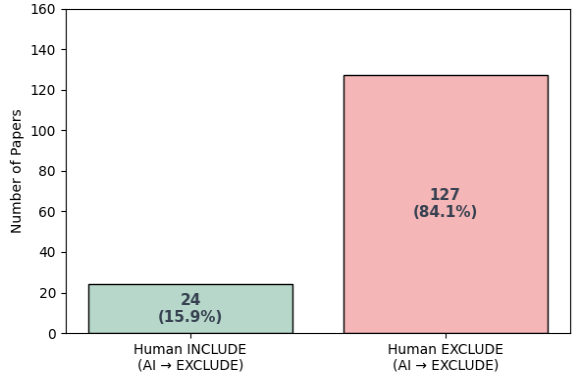
[그림 6] AI 기반 제2의 코더: 삼분 구조 결과



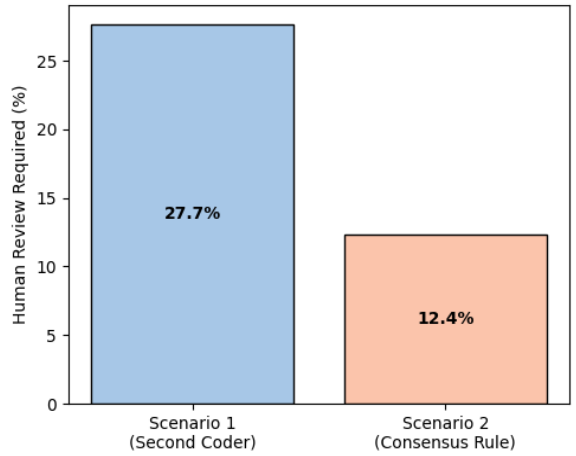
[그림 7] 단독연구자를 위한 AI 기반 합의규칙 적용 결과



[그림 8] AI 기반 제3의 코더- 적용 결과



[그림 9] AI모델의 제외 지지 문헌에 대한 인간 연구자 선별 판단 분포



[그림 10] 시나리오 1과 2의 재검토 부담 수준 비교

〈표 1〉 문헌선정 자동화 관련 주요 선행연구 분석

연구자(연도)	주요 기술	주요 연구 결과
O'Mara-Eves et al.(2015)[4]	Text Mining, ML(SVM)	업무량 30~70% 절감, 우선순위화의 실무적용성 확인
Marshall et al.(2019)[5]	ML(Robot Reviewer) Active Learning	제2 검토자로서의 활용 잠재력 검증
Gartlehner et al.(2019)[13]	Semi-automation (Distiller AI)	낮은 재현율, 완전 자동화에는 불충분
van de Schoot et al.(2021)	Active Learning (ASReview), BERT	재현율 95% 유지, 업무량 83% 절감
Tran et al. (2024)	LLM (GPT-3.5 Turbo)	민감도 97%, 업무량 약 65% 절감 추정
Khraisha et al.(2024)	LLM (GPT-4)	인간 수준의 성능 도달 단, 환각 위험 존재
Chibwe et al.(2025)	LLM(GPT-3/3.5/4)	하수 역학 분야, 비용 효율성 입증
Kallmes et al.(2025)	LLM, ML (AutoLit/XGBoost)	재현율 82~97% 기록, 시간 50% 단축
Lieberum et al.(2025)	LLM Overview (Scoping Review)	재현성 문제 제기, 인간참여 필요성 강조
김인후· 김성희(2022)	BERT(KoBERT)	문헌정보학분야, 고품질 데이터에서 F1 0.90 달성
이용구(2023)	BERT, Transfer Learning	문헌정보학분야, 주제명 분류, 특정범주 고성능
강광선 외(2024)	BERT, LLM(Llama2)	텍스트 길이에 따른 최적 모델 선택 가이드 제시
설아람(2025)	ML, LLM (Trend Analysis)	의료분야 체계적 문헌고찰에서 AI도입 증가

〈표 2〉 AI-인간 협업 시나리오의 구조와 역할

	Scenario 1	Scenario 2	Scenario 3
적용 상황	다수 코더 체계 보완 (AI 사전 선별)	단독 검토 연구자 (인간-AI 병렬 검토)	복수 연구자 간 교차 검토 (인간 간 갈등 조정)
AI 역할	제2 코더	합의규칙 제공	갈등 조정 참조 지표
의사 결정 구조	AI 단독 평가 (삼분 분류: 포함/ 불확실/ 제외)	인간-AI 독립 평가 후 대조 (일치: 확정/ 불일치: 보류)	제3의 객관적 지표 (인간 간 불일치 시 참조)
분류 및 판단 규칙	보수적 임계값 ($p \geq 0.7$) 적용, '불확실' 할당	AI와 인간 동일 판단 시에만 결과 자동 확정	불일치 문헌 대상 AI 예측확률 보조 지표화
인간 연구자 개입	AI 분류 불확실 문헌 집중 수동 재검토	불일치 문헌 인간 수동 재검토	AI 결과 참조 연구자간 최종 합의

〈표 3〉 데이터셋(Dataset) 구성 및 라벨 분포

구분			문헌 수		비율	
전체 수집 문헌			1012		100%	
인간 연구자	포함 문헌		123		12.2%	
	제외 문헌		889		87.8%	
AI 모델	학습용 (train)	포함	809	98	80%	12.1%
		제외		711		87.9%
	검증용 (test)	포함	203	25	20%	12.3%
		제외		178		87.7%

〈표 4〉 Epoch별 학습 손실 및 검증 손실

Step	50	100	150
학습 손실	0.7087	0.6840	0.5581
검증 손실	0.6783	0.6333	0.6090

〈표 5〉 AI 자동선별 모델의 분류 성능 지표

구분	Accuracy	Precision	Recall	F1 Score	Cohen's Kappa
AI 모델	0.7882	0.2955	0.5200	0.3768	0.2607