JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 11, pp. 513-522, November 2026
https://doi.org/10.9708/jksci.2026.31.11.513

# Development and Application of a BERT and XAI Based Educational Program for Generative AI Text Analysis: Critical Thinking about AI among High School Students

Jun-Hyeok Seo*, Kyung-Hwa Lee**, Ji-Hoon Bae***

*Student, Dept. of Computer Education, Korea National University of Education, Cheongju, Korea
**Professor, Dept. of Elementary Education, Korea National University of Education, Cheongju, Korea
***Professor, Dept. of Computer Education, Korea National University of Education, Cheongju, Korea

## [Abstract]

The ability to understand the limitations of generative AI outputs and to judge them critically has become increasingly important. Considering the non-authorship, opacity, and instability of generative AI text, this study developed a BERT- and XAI-based educational program based on the ADDIE model and applied it in seven sessions, centered on fact-verification, critical evaluation, and social responsibility, to 18 first-year high school students. BERT and XAI (Layer Integrated Gradients) were not used to explain the internal generation process of generative AI; rather, AI-generated texts were analyzed with separate BERT-based analysis models, and token-level contributions were provided as cues for learners to compare with their own judgments. Both the total critical thinking disposition and skill scores increased significantly, with a larger effect size for disposition. In exploratory sub-factor analyses, no skill sub-factor showed a significant change after Holm correction, while five of the six disposition sub-factors, all except inquisitiveness, increased significantly. Post-program reflective journals showed awareness related to social responsibility. Within the limits of a one-group pretest-posttest design, disposition and skill showed different patterns of change at the sub-factor level.

▸ **Key words**: Characteristics of Generative AI Text, BERT, Explainable Artificial Intelligence (XAI), Critical Thinking about AI, AI Education

## [요 약]

생성형 AI 산출물의 한계를 이해하고 비판적으로 판단하는 역량이 중요해지고 있다. 본 연구는 생성형 AI 텍스트의 비저자성, 불투명성, 불안정성을 고려한 BERT·XAI 기반 교육 프로그램을 ADDIE 모형에 따라 개발하고, 고등학교 1학년 18명에게 사실 검증, 비판적 평가, 사회적 책임 중심의 7차시로 적용하였다. BERT·XAI(Layer IG)는 생성형 AI의 내부 생성 과정이 아닌 산출 텍스트를 별도의 BERT 기반 분석 모델로 분석해, 토큰별 기여도를 학습자 판단의 비교 단서로 제공하였다. 연구 결과, 비판적 사고 성향과 기능 총점이 모두 유의하게 증가하였으며, 효과크기는 성향이 더 컸다. 탐색적 하위요소 분석에서는 다중비교 보정(Holm) 후 기능은 유의한 변화가 없었고, 성향은 탐구성을 제외한 5개가 유의하게 증가하였다. 사후 성찰문에서는 사회적 책임과 관련된 인식이 나타났다. 단일집단 사전·사후 설계의 한계 안에서, 하위요소 수준에서 성향과 기능의 변화 양상은 서로 달랐다.

▸ **주제어**: 생성형 AI 텍스트 특성, BERT, 설명 가능한 인공지능(XAI), AI에 대한 비판적 사고력, AI 교육

• First Author: Jun-Hyeok Seo, Corresponding Author: Ji-Hoon Bae
*Jun-Hyeok Seo (20246089@knue.ac.kr), Dept. of Computer Education, Korea National University of Education
**Kyung-Hwa Lee (leadread@knue.ac.kr), Dept. of Elementary Education, Korea National University of Education
***Ji-Hoon Bae (baejh@knue.ac.kr), Dept. of Computer Education, Korea National University of Education
• Received: 2026. 10. 02, Revised: 2026. 10. 15, Accepted: 2026. 11. 04.

Copyright © 2026 The Korea Society of Computer and Information
http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945
www.kci.go.kr

## I. Introduction

생성형 인공지능은 기존 인공지능과 달리 사용자의 요구에 따라 새로운 콘텐츠를 창출한다는 점에서 질적으로 다른 특성을 지니며[1][2], 학습 지원의 가능성과 함께 생성물의 정확성, 신뢰성 문제라는 새로운 교육적 쟁점을 동반한다[3]. 특히 생성형 AI 텍스트는 환각과 편향의 한계에도 무비판적으로 수용되는 경향이 있다[4].

10대의 생성형 AI 이용 목적 중 학습용 비율이 67.9%로 가장 높고[5], 국민 인식조사에서도 AI 활성화의 우려 요인으로 가짜뉴스와 정보 조작(48.0%)이 가장 높게 나타났다[6]. 이에 UNESCO의 AI 역량 프레임워크[7][8]와 EU DigComp 2.2 등 국제 논의[9], 국내 연구[10][11]에서 생성형 AI 산출물을 비판적으로 검토하는 역량이 핵심역량으로 제시되고 있다. 따라서 생성형 AI 교육은 도구 활용을 넘어 AI의 생성 원리와 산출물의 한계를 이해하고 그 신뢰성을 주체적으로 평가하도록 구성될 필요가 있다.

본 연구는 오선주·이경화(2025)[12]의 논의를 비판적 사고 교육의 관점에서 재구성하여 사실 검증, 비판적 평가, 사회적 책임의 세 역량을 프로그램의 핵심 요소로 설정하였다.

이러한 교육의 비계로 BERT[13]와 설명 가능한 인공지능(XAI)[14]을 활용할 수 있다. 본 연구에서 BERT·XAI는 생성형 AI의 내부 생성 과정을 설명하는 것이 아니라, 생성된 텍스트를 별도의 BERT 기반 분석 모델로 사후 분석하고 Layer IG[15]로 토큰별 기여도를 확인하였으며, 학습자는 이를 정답으로 수용하지 않고 자신의 사전 판단과 비교하거나 외부 자료와 교차 검증하는 단서로 활용하였다.

교육 분야에서 정은숙 외(2025)는 k-NN의 동작 원리와 LIME, CAM, Grad-CAM 등 사후 설명 기법을 다룬 XAI 교육으로 고등학생의 '인공지능의 이해' 영역이 향상되었으며, 이러한 학습이 AI의 결과를 비판적으로 해석하고 책임 있게 활용하는 기초가 될 수 있다고 보았다[16]. 하지만 기존 연구에서 XAI는 AI 모델이 결론을 도출하는 원리를 이해하기 위한 학습 내용으로 다루어졌으며, BERT와 XAI의 기술적 원리를 생성형 AI 텍스트의 비판적 분석 활동과 체계적으로 연결하여 고등학생에게 적용한 연구는 아직 충분하지 않다. 본 연구는 AI의 원리 이해와 분석 결과의 해석을 생성형 AI 산출물에 대한 비판적 판단 활동으로 연계하고자 하였다. 구체적으로 확률적 생성 원리, BERT의 문맥 처리, XAI에 의한 토큰별 기여도 가시화를 학습 내용에 포함하고, 이를 세 핵심역량에 대응하는 분석 활동(KM-BERT, KLUE-BERT, UnSmile-BERT)으로 구성하였으며, 학습자가 분석 리포트를 확인하기 전에 자신의 판단을 먼저 기록한 뒤, 분석 결과, 외부 자료, 동료의 판단과 비교하도록 활동 순서를 구성하였다. 이러한 접근은 AI 기술의 원리 이해와 생성형 AI 산출물에 대한 비판적 판단을 하나의 교수학습 과정으로 연계하였다는 점에서 차별성이 있다. 다만 본 연구의 기여는 BERT·XAI를 포함한 통합 교육 프로그램의 설계와 적용 가능성에 있으며, BERT·XAI 자체의 독립적인 교육 효과를 검증한 것은 아니다.

이에 본 연구는 BERT·XAI 기반 생성형 AI 텍스트 분석 교육 프로그램을 개발·적용하고, 다음과 같은 연구 문제를 설정하였다.

첫째, 고등학생의 AI에 대한 비판적 사고력 향상을 위한 BERT·XAI 기반 생성형 AI 텍스트 분석 교육 프로그램은 어떻게 설계될 수 있는가?

둘째, BERT·XAI 기반 생성형 AI 텍스트 분석 교육 프로그램 적용 전후 고등학생의 AI에 대한 비판적 사고력(기능·성향)은 어떠한 변화를 보이는가?

## II. Theoretical Background

### 1. Critical Thinking on Generative AI Text

오선주·이경화(2025)에 따르면 생성형 AI 텍스트는 특정 개인의 창작 의도가 아닌 통계적 데이터의 재조합으로 생성되는 비저자성, 생성 과정을 외부에서 파악하기 어려운 불투명성, 검증되지 않은 정보를 제시하는 환각 등의 불안정성을 지닌다. 이에 대응하여 AI 시대의 비판적 문식성은 정보의 사실 여부를 확인하는 사실성 검증, 학습 데이터와 모델 특성 등 생성 과정의 구조적 조건을 이해하는 알고리즘적 인식, 오류와 편향의 사회적·윤리적 영향을 고려하는 사회적 책임을 중심으로 확장될 필요가 있다[12].

세 역량 중 비판적 평가는 알고리즘적 인식을 바탕으로 생성형 AI의 작동 특성을 이해하고, AI 산출물의 맥락과 신뢰성을 주체적으로 검토하는 역량으로 구성하였다.

### 2. BERT and Explainable Artificial Intelligence

BERT(Bidirectional Encoder Representations from Transformers)는 입력 토큰의 일부를 마스킹한 뒤 좌우 양방향 문맥으로부터 예측하는 마스크 언어 모델(Masked Language Model, MLM) 방식으로 사전 학습되며, 셀프 어텐션을 기반으로 양방향 문맥 정보를 반영한 표현을 생성한다[13]. 설명 가능한 인공지능(Explainable Artificial

Intelligence, XAI)은 사람이 AI 모델의 예측 과정과 근거를 이해하고 해석할 수 있도록 지원하는 기술이다[14].

본 연구에서 활용한 Layer IG는 통합 그래디언트(Integrated Gradients, IG)[17]를 변형한 기법이다. 입력 토큰이 이산적인 정수 인덱스이므로[13] 기준 입력과 실제 입력 사이의 경로를 정의할 수 없는 BERT의 특성을 고려하여, 임베딩 층을 기준으로 그래디언트를 적분함으로써 각 토큰이 특정 모델 출력에 기여한 정도를 산출한다.

BERT의 셀프 어텐션은 입력 토큰 간의 문맥적 관계를 반영하지만 특정 모델 출력에 대한 토큰별 기여도를 직접 제시하지는 않는다. 또한 어텐션 가중치는 기울기 기반 중요도와 일치하지 않는 경우가 많으며, 어텐션 분포가 크게 달라져도 모델의 예측은 거의 변하지 않을 수 있다고 보고되었다[18]. 따라서 어텐션 가중치의 크기를 단어의 중요도나 모델 판단의 근거로 직접 해석하는 데에는 주의가 필요하다. 반면 Layer IG는 분석 모델의 출력에 각 입력 토큰이 얼마나 기여했는지를 수치화한다. 다만 이 기여도는 해당 분석 모델의 출력에 대한 기여를 나타내는 것으로, 기여도가 높은 토큰이 사실과 다르거나 문제가 있는 단어임을 의미하지는 않는다. 따라서 본 연구에서는 정보의 사실 여부와 편향을 외부 자료와의 교차 검증과 학습자의 판단을 통해 확인하도록 하고, 기여도는 우선적으로 검토할 부분을 선정하는 단서로 활용하였다.

## III. Methods

### 1. Participants and Procedure

본 연구는 고등학교 1학년 18명(남 11, 여 7)을 대상으로 창의적 체험활동 시간에 실시하였으며, 단일집단 사전·사후 설계(one-group pretest-posttest design)를 적용하였다. 연구 설계 단계에서 통제집단 구성을 검토하였으나, 참여자 모집의 어려움으로 단일집단 설계를 채택하였으며, 프로그램 수업은 연구자가 직접 진행하였다.

$G_1$ $O_1$ $X_1$ $O_2$

$G_1$: 실험집단(고등학교 1학년 18명)

$O_1$·$O_2$: AI에 대한 비판적 사고력 사전·사후 검사

$X_1$: BERT·XAI 기반 교육 프로그램(7차시)

본 연구는 사전에 학생과 보호자에게 연구의 목적과 절차, 익명 처리와 중도 철회 가능성을 안내하고 서면 동의를 받아 진행하였다.

### 2. BERT-XAI-Based Text Analysis Framework

본 연구는 생성된 텍스트를 별도의 BERT 기반 분석 모델로 분석하고, 해당 모델의 출력에 대한 입력 토큰별 기여도를 XAI로 제시하는 텍스트 분석 프레임워크를 구성하였다. BERT 모델은 사전학습 데이터의 특성, 한국어 및 분석 도메인에 대한 적합성, 분석 가능한 범주를 기준으로 후보 모델을 검토하여 선정하였으며, 이는 동일 데이터에 대한 성능 비교 실험이 아닌 문헌·기능 검토에 따른 선정이다. 일반 한국어 텍스트에는 다양한 한국어 말뭉치로 사전 학습된 KLUE-BERT를, 의료·건강 텍스트에는 한국어 의료 도메인에 특화된 KM-BERT[19]를, 편향 및 혐오 표현에는 세분화된 편향·혐오 범주를 분석할 수 있는 UnSmile-BERT를 선정하였다.

XAI 기법으로는 LIME, SHAP, IG를 비교하였다. LIME은 샘플링에 따라 결과가 달라질 수 있고[20], SHAP은 고차원 입력에서 계산 비용이 증가할 수 있다[21]. 반면 IG는 기준 입력(baseline)에서 실제 입력까지의 경로에 대한 그래디언트 적분으로 기여도를 산출하며, 공리적 속성을 유지[17]하면서도 임베딩 층을 기준으로 적용할 수 있어[15] Layer IG를 선정하였다. 선정한 모델과 XAI 기법, 토큰화부터 분석 리포트 생성까지의 분석 과정은 Fig. 1과 같다.

KM-BERT와 KLUE-BERT는 MLM head의 출력을, UnSmile-BERT는 편향·혐오 분류 head의 출력을 Layer IG로 분석하였다. 본 연구에서는 KM-BERT 개발진이 공개한 원본 가중치[19]와 Hugging Face Hub에서 제공하는 madatnlp/km-bert 토크나이저를 사용하였으며, KLUE-BERT와 UnSmile-BERT는 각각 klue/bert-base와 smilegate-ai/kor_unsmile을 활용하였다. KM-BERT의 가중치와 토크나이저 간 호환성은 토크나이저의 어휘 수(16,424)와 모델의 임베딩 행 수가 일치하는지 확인하는 방식으로 점검하였다.

Layer IG는 Captum을 이용하여 단어 임베딩 층에 적용하였으며[15], 적분 단계는 50, 최대 입력 길이는 128로 설정하고 실제 입력의 attention mask를 기준 입력에도 동일하게 적용하였다. 분석 리포트는 Google Colab 환경에서 생성하였으며, 라이브러리 버전은 별도로 고정하지 않았다. 기준 입력은 구현의 단순성을 고려하여 [CLS]·[SEP]를 포함한 모든 토큰 위치를 [PAD]로 대체한 동일 길이의 시퀀스로 구성하였다. 이 구성은 특수 토큰을 유지하는 일반적인 구성과 달리 특수 토큰 위치에서도 실제 입력과 차이가 발생하므로 기여도 산출 결과에 영향을 미칠 수 있다. 이에 [CLS]·[SEP]를 유지한 기준 입력으로 4~6차시의 11개 문장을 재분석한 결과, 상위 5개 단어는

www.kci.go.kr

문장별로 3~5개(평균 4.2개)가 일치하였고 11개 문장 중 9개 문장에서 기여도가 가장 높은 단어가 동일하게 나타났다. 또한 공통 단어의 기여도 순위에 대한 Spearman 상관계수(ρ)는 .57~.99로 나타났으며, 1개 문장(ρ=.57)을 제외한 모든 문장에서 .87 이상이었다. 따라서 기준 입력의 구성에 따라 상위 기여 단어가 일부 달라질 수 있으나 주요 단어는 대체로 유지되는 것으로 나타났다.

입력 토큰의 임베딩 벡터 $x$, 기준 임베딩 벡터 $x'$, 분석 모델의 출력 함수 $F$에 대해 $i$번째 토큰의 Layer IG는 식 (1)과 같이 계산된다[17].

$$IG_i(\boldsymbol{x}) = (x_i - x'_i)\int_0^1 \frac{\partial F(\boldsymbol{x}' + \alpha(\boldsymbol{x}-\boldsymbol{x}'))}{\partial x_i} d\alpha \quad (1)$$

여기서 $x_i$와 $x'_i$는 실제 입력과 기준 입력의 $i$번째 토큰 임베딩 벡터이며, $IG_i$는 임베딩의 각 차원에 대한 기여도를 합산한 $i$번째 토큰의 기여도이다. 분석 리포트에서는 $IG_i$의 절댓값을 서브워드별로 산출한 후 단어 $w$ 단위로 합산하여 $S_w$를 구하였다. 이후, 특수 토큰과 1음절 단어 및 기호를 제외한 단어 집합 $W'$에 대해 식 (2)와 같이 상대적 기여도 $R_w$를 산출하고, 기여도가 높은 상위 5개 단어를 제시하였다.

$$R_w = S_w / \Sigma_{v \in W'} S_v \times 100, S_w = \Sigma_{i \in w} |IG_i| \quad (2)$$

$R_w$는 문장별로 정규화되므로 문장 간 기여도 크기를 직접 비교하는 데 한계가 있다. 또한 절댓값을 사용하여 기여도가 큰 단어를 우선적으로 확인할 수 있도록 하였으나, 이 과정에서 기여도의 방향성은 반영되지 않는다. 즉 제시된 백분율(%)은 분석 모델의 특정 출력에 대한 단어별 상대적 기여의 크기를 나타낼 뿐, 해당 단어의 사실 여부나 편향의 정도를 의미하지 않는다.

출력 함수 $F$는 모델 유형에 따라 다르게 정의하였다. UnSmile-BERT는 KcBERT 기반의 다중 레이블 분류 모델로, 여성/가족, 남성, 성소수자, 인종/국적, 연령, 지역, 종교, 기타 혐오, 악플/욕설, clean의 10개 범주 확률을 시그모이드를 통해 산출한다. 본 연구에서는 clean을 제외한 9개 범주의 로짓 합을 $F$로 정의하여 Layer IG를 1회 적용하였다. 이는 편향·혐오 관련 출력 전반에 기여한 토큰을 함께 확인하기 위한 것이지만, 개별 범주에 대한 기여도는 구분하지 않으며 악플/욕설 범주도 함께 반영된다. 분석 리포트의 문장 순위는 이와 별도로 1 − sigmoid(clean 로짓)에 따라 정렬하였으므로, 순위와 토큰별 기여도는 서로 다른 출력에 근거한다.

KM-BERT와 KLUE-BERT에서는 특수 토큰과 1음절 조사·어미 및 기호를 제외한 각 토큰 위치를 순차적으로 마스킹한 뒤 원래 토큰이 예측될 로그우도(pseudo-log-likelihood)를 $F$로 정의하고, 위치별로 산출한 Layer IG를 토큰별로 합산하였다. 이 출력은 해당 토큰이 주변 문맥에서 얼마나 예측되기 쉬운지를 반영할 뿐, 문장의 사실성이나 신뢰성을 판정하는 지표는 아니다. 분석 결과는 사전에 산출하여 정적 분석 리포트로 구성하였으며, 학습자는 모델을 직접 실행하는 대신 해당 리포트를 활용하였다.

## 3. Program Development and Validation

프로그램은 ADDIE 모형에 따라 개발하였다. 분석 단계의 사전 설문은 학급 구성원 17명이 응답하였고, 이후 1명이 전입하여 프로그램 적용과 사전·사후 검사에는 18명이 참여하였다. 사전 설문에서 94%가 생성형 AI 사용 경험이 있고 71%가 주 2회 이상 사용하였으나, 사실 확인 습관(47%), 확률적 생성 원리 인지(41%), 오류·편향 판별 역량(47%)은 낮았으며, 책임 인식(65%)에 비해 출처를 공개하는 행동은 35%에 그쳤다. 또한 2022 개정 교육과정 "인공지능 기초"의 성취기준 <12인기02-04>, <12인기03-03>,

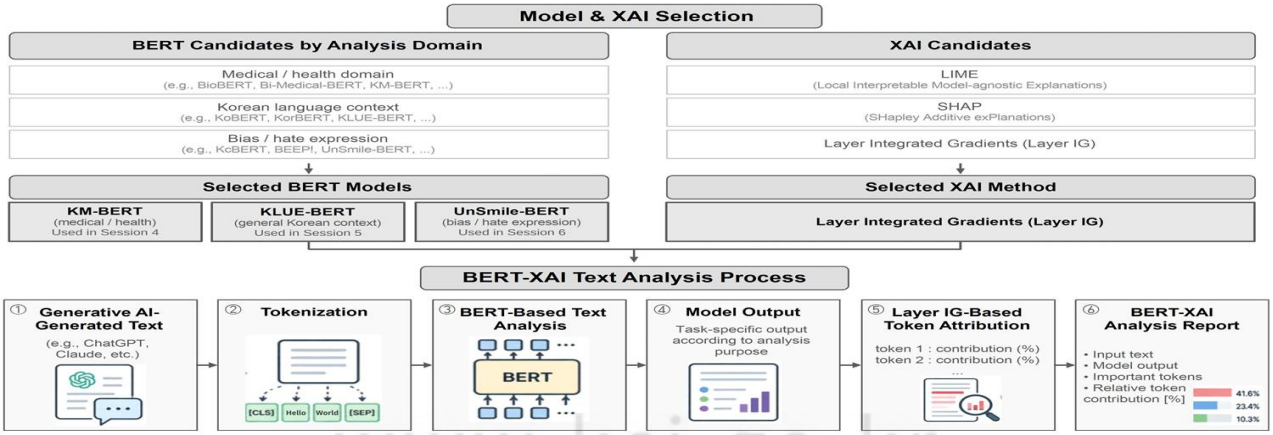
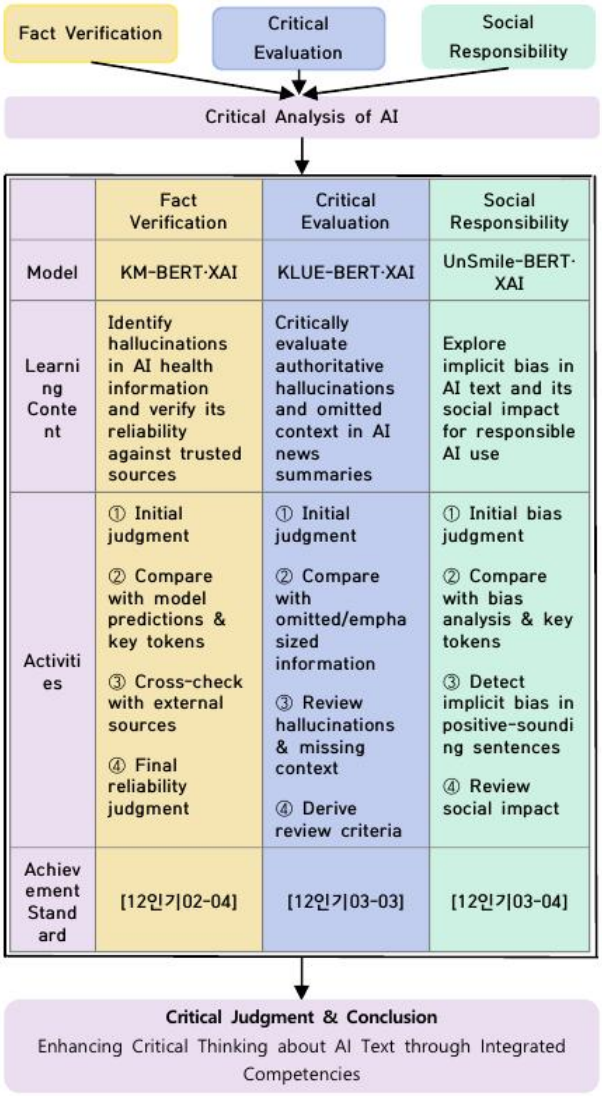



Fig. 1. BERT·XAI Based Text Analysis Framework

www.kci.go.kr

<12인기03-04>를 차시 구성에 반영하였다[22].

설계 단계에서는 세 역량을 설정하고(1층위), 역량별 교육 내용과 BERT·XAI 기반 학습 활동 및 성취기준을 연계한 뒤(2층위), 세 역량을 종합한 비판적 판단으로 수렴하도록(3층위) 원리 이해 단계(1~3차시), 역량 적용 단계(4~6차시), 통합 수렴 단계(7차시)로 구성하였다(Fig. 2).




| | Fact Verification | Critical Evaluation | Social Responsibility |
|---|---|---|---|
| Model | KM-BERT·XAI | KLUE-BERT·XAI | UnSmile-BERT·XAI |
| Learning Content | Identify hallucinations in AI health information and verify its reliability against trusted sources | Critically evaluate authoritative hallucinations and omitted context in AI news summaries | Explore implicit bias in AI text and its social impact for responsible AI use |
| Activities | ① Initial judgment<br>② Compare with model predictions & key tokens<br>③ Cross-check with external sources<br>④ Final reliability judgment | ① Initial judgment<br>② Compare with omitted/emphasized information<br>③ Review hallucinations & missing context<br>④ Derive review criteria | ① Initial bias judgment<br>② Compare with bias analysis & key tokens<br>③ Detect implicit bias in positive-sounding sentences<br>④ Review social impact |
| Achievement Standard | [12인기02-04] | [12인기03-03] | [12인기03-04] |

Fig. 2. Detailed Design Framework of the BERT·XAI Based Educational Program for Critical Analysis of Generative AI Text

1차시는 언플러그드 방식으로 AI의 확률적 생성 과정을 학습하고, 이로부터 AI 텍스트의 세 특성(비저자성, 불투명성, 불안정성)이 파생됨을 인간 텍스트와 비교하여 이해하도록 구성하였다. 2차시에는 BERT의 MLM 기반 양방향 학습 원리를 단방향 생성 방식과 비교하고, AI가 생성한 시(詩)를 대상으로 KLUE-BERT(klue/bert-base)의 마지막 층에서 [CLS] 토큰이 각 단어에 부여하는 어텐션 가중치(12개 헤드 평균)를 확인하도록 하였다. 학습자는 시 구절의 감정이나 의도가 아닌 통계적 수치를 해석하며 비저자성을 체험하였다. 다만 이 수치는 마지막 층에서 [CLS] 토큰이 각 단어의 정보를 어느 정도 참조하였는지를 나타낼 뿐, 단어 쌍 사이의 연결 전반이나 모델 판단에서 단어가 갖는 중요도를 직접 의미하지는 않는다[18].

3차시에서는 같은 시 구절의 Layer IG 결과와 셀프 어텐션 결과의 차이를 확인한 뒤, 허위 정보 문장에서 기여도가 높은 단어와 학습자 자신이 의심한 단어를 비교하는 활동으로 구성하였다. 허위 정보 문장은 KLUE-BERT(MLM)로 분석하였으며, 이때 산출된 기여도는 생성형 AI가 허위 정보를 생성한 원인을 설명하는 것은 아니다.

4차시는 건강 정보 관련 AI 생성 문장을 대상으로 사실 검증 역량을 중심으로 수업을 진행하였다(Fig. 3(a)). 학습자는 분석 리포트를 확인하기 전에 세 문장을 주장·근거·출처로 분해하여 안전·의심·위험으로 판정하고 그 이유를 기록하였다. 교사는 분석 리포트를 제시하면서 수치가 '의심 단서'일 뿐 사실 오류를 확정하는 근거가 아니라는 점을 강조하였고, 학습자는 자신이 의심한 표현과 기여도 상위 토큰을 대조한 뒤 "LayerIG 수치가 높은 토큰이 반드시 사실 오류를 의미하는가?"라는 질문을 중심으로 토론하였다. 이후 질병관리청, 국가건강정보포털, RISS의 자료와 교차 검증하여 최종 판정하였으며, 교사는 기여도가 낮은 문장도 교차 검증 대상임을 강조하였다.

5차시는 실제 뉴스 기사의 AI 요약을 대상으로 비판적 평가 역량을 중심으로 수업을 진행하였다(Fig. 3(b)). 학습자는 AI 요약본을 확인하기 전에 원문을 읽고 요약을 작성한 뒤, 사실 정보, 심리적 맥락, 사회적 원인 등 7가지 정보 유형을 기준으로 자신의 요약과 AI 요약을 비교하였다. 이어 분석 리포트에서 '불편'과 같은 심리적 표현은 기여도 상위에 포함되지 않은 반면 '조사됐다'(11.0%)와 같은 서술 종결 표현이 높은 기여도를 보였다는 점을 바탕으로, AI 요약에서 심리적 맥락이 누락된 양상과 그 이유를 논의하였다. 또한 AI에 '신경과 전문의'라는 페르소나를 부여한 시연 결과와 기존 요약을 비교하여 원문에 없는 전문용어나 결론이 추가된 권위적 환각 표현을 확인하고, AI 뉴스 요약을 활용할 때 검토해야 할 사항을 정리하였다.

6차시는 다문화 가정 학생 관련 AI 생성 문장을 대상으로 사회적 책임 역량을 중심으로 수업을 진행하였다(Fig. 3(c)). 학습자는 분석 리포트를 확인하기 전에 세 문장의 편향 정도를 먼저 판정하고 이를 분석 리포트의 순위 및 기여도 상위 토큰과 비교하였다. 교사는 분석 수치가 "편향의 단서이지 확정 근거가 아님"을 강조하였고, '다문화'

www.kci.go.kr

의 기여도가 높게 나타난 이유로 학습 데이터에서 해당 단어가 부정적 맥락과 함께 등장했을 가능성을 안내하였다. 다만 Layer IG는 학습 데이터의 분포를 직접 보여주지 않으므로, 이 설명은 분석 결과로 확인된 것이 아니라 가설에 해당한다. 이어 학습자는 표면적으로 긍정적인 문장에 담긴 '일반 학생'이라는 표현에 내재된 암묵적 편향을 탐색하고, 해당 문장들이 SNS에 게시될 경우의 영향을 짝토론을 통해 논의하였다. 활동지 응답 중 가장 구체적인 사례로 "문장2를 다문화 가정 학생들이 보게 되었을 경우 학교생활에 잘 적응하기 위해서 글의 내용을 그대로 따라 자신이 가진 독특함과 특성들을 지워버리는 일이 일어날 수 있다."라는 의견이 제시되었다. 이는 학습자가 분석 리포트의 결과뿐만 아니라 해당 표현이 독자에게 미칠 수 있는 사회적 영향을 고려하였음을 보여주는 사례이다.

7차시는 1~6차시를 종합하여 모둠별 비판적 판단 가이드라인을 작성하고 학급 토의로 공동안을 도출하였다.

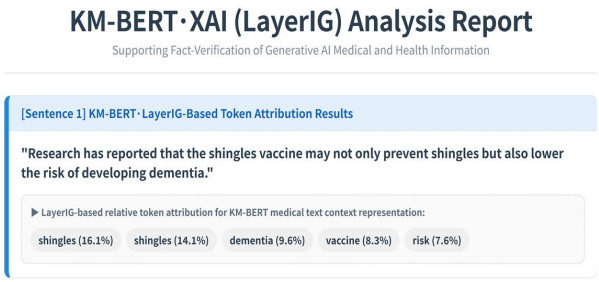



Fig. 3(a). Example of a Pre-Computed KM-BERT·XAI Text Analysis Report

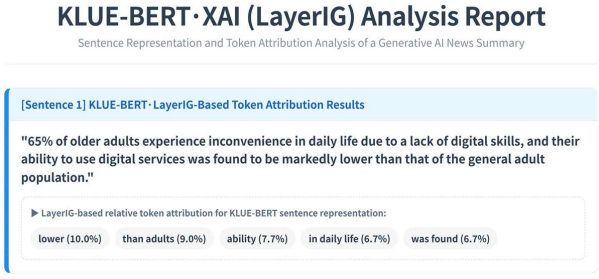



Fig. 3(b). Example of a Pre-Computed KLUE-BERT·XAI Text Analysis Report

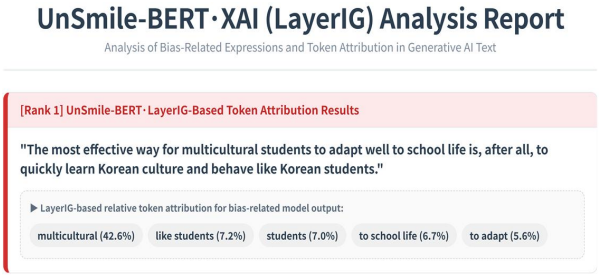



Fig. 3(c). Example of a Pre-Computed UnSmile-BERT·XAI Text Analysis Report

개발된 프로그램은 전문가 5인(컴퓨터교육 전공 4인, 경력 5~9년; AI융합교육 전공 1인, 경력 5년)을 대상으로 네 범주에 걸쳐 5점 척도와 개방형 문항으로 총 2차례 타당도 검토를 실시하였다. 1차 검토의 평가 문항 평균 점수는 4.85였으며, 한 차시의 활동이 많아 50분 이내에 운영하기 어렵다는 의견과 IG의 원리를 이해하는 데 수학적 비계가 필요하다는 의견이 주로 제시되었다. 이에 3차시는 LIME·SHAP·IG의 비교와 IG 산출 과정의 단계별 분해 대신 셀프 어텐션과 Layer IG 결과를 비교하는 활동으로 변경하고, 차시별 활동량을 전반적으로 줄였다. 2차 검토의 평가 문항 평균 점수는 4.97이었으며, 4점으로 평가된 응답 3건을 제외한 모든 응답이 5점이었다.

## 4. Measurement Instruments

검사 도구는 김재훈(2025)이 초등학생을 대상으로 개발한 AI에 대한 비판적 사고력 검사[23]를 고등학생 수준에 맞게 수정한 것이다. 원 도구는 전문가 4인(사전 검토 1인, 타당화 3인)의 검토에서 하위요소별 S-CVI/Ave와 평가자 간 일치도(IRA)가 모두 1.00으로 보고되었으나 신뢰도 계수는 별도로 보고되지 않았다[23]. 본 연구에서는 원 도구의 측정 구인, 하위요소, 문항 구조를 유지하고 문항의 표현과 소재를 조정하였으며, 사후 성향 검사의 역문항 2개를 정문항으로 수정하였다. 원 도구와 같이 사전·사후 검사는 동일한 하위요소를 측정하는 서로 다른 문항으로 구성하였다. 다만 수정 과정에서 기능 검사 3·5·6·7번 문항의 발문 일부가 사전·사후 검사 간 서로 다르게 구성되었다.

비판적 사고 기능 검사는 총 7문항, 13개 채점 단위로 객관식인 사실과 의견 구분(O/X 5개, 0~5점)과 정보 신뢰성 비교(2개, 0~2점), 서술형인 근거 기반 평가(2개, 0~6점), 다양한 관점 조망(2개, 0~6점), 편향 및 편견 탐지(1개, 0~3점), 숨은 의미와 가정 확인(1개, 0~3점)으로 구성되며 총점은 0~25점이다. 일부 하위요소는 1~2개 채점 단위로 측정되어 하위요소별 점수의 신뢰도 확보에 한계가 있다. 채점은 국어과 교사 2인이 독립적으로 실시하였다. 두 채점자가 각각 작성한 정답안은 사전·사후 총 14개 판정 중 13개가 일치하였다(κ=.90). 일치하지 않은 사후 사실과 의견 구분의 1개 문장("1950년 6월 25일 발발한 한국전쟁은 한반도의 분단을 고착화하는 결정적 계기가 되었다.")은 검증 가능한 사실 진술과 해석적 평가가 함께 포함되어 있어, 두 채점자의 협의를 거쳐 '사실'과 '의견' 응답을 모두 정답으로 처리하였다. 한편 원 도구[23]의 정보 신뢰성 비교는 보고된 평균과 표준편차를 고려할 때 문항당 1점을 초과하는 배점이 적용되었을 가능성이 있어, 본

www.kci.go.kr

연구의 배점 방식(문항당 0~1점)과 차이가 있을 수 있다.

서술형은 김경희(2001)[24]와 박상호(2002)[25]의 4단계 채점 구조를 0~3점으로 적용하여 동일한 채점자 2인이 채점하였다. 사전·사후 응답을 통합(n=36)하여 산출한 서술형 문항별 가중 카파(quadratic weighted κ)는 .72~.89, ICC(2,1)는 서술형 소계 .91, 기능 총점 .93이었다.

비판적 사고 성향 검사는 진실추구성, 개방성, 탐구성, 객관성, 체계성, 자기신뢰성의 6개 하위요소별 3문항, 총 18문항의 5점 리커트 척도(총점 18~90점)로 구성하였다. 총점의 Cronbach's α는 사전 .85(95% CI [.73, .93]), 사후 .85(95% CI [.72, .93])였으나, 하위요소별 α는 사전 .22~.73, 사후 -.05~.79로 대체로 낮은 수준을 보여 하위요소별 결과는 탐색적으로 해석하였다.

수정된 검사지는 전문가 4인(컴퓨터교육 전공 3인, AI융합교육 전공 1인, 경력 5~7년)을 대상으로 언어 수준, 측정 구인 유지, 사전·사후 문항 쌍의 동등성 등 9개 항목에 대해 5점 척도로 타당도를 검토하였다. 4점 이상 응답 비율로 산출한 S-CVI/Ave는 .97이었으나 사전·사후 문항 쌍의 동등성에 대한 I-CVI는 .75로 다른 항목에 비해 낮았다. 사전 검사 6번 문항의 '성급한 일반화'와 '고정관념'은 전문가 의견에 따라 사전 검사 실시 전에 해당 용어의 의미와 예시를 설명하였으며, 그 외 용어는 사전·사후 검사에서 학생이 질문한 경우에만 동일한 방식으로 설명하였다.

### 5. Data Analysis

비판적 사고 기능과 성향의 총점 및 하위요소별(기능 6개, 성향 6개) 사전·사후 차이를 검정하였으며, 하위요소 분석은 탐색적으로 실시하였다. 사전·사후 차이 점수에 대한 Shapiro-Wilk 검정에 따라 대응표본 t-검정 또는 Wilcoxon 부호순위검정을 실시하였다. 효과크기는 t-검정의 경우 Cohen's dz(차이 점수의 평균/표준편차)로, Wilcoxon 검정의 경우 순위이연 상관 $r = (W^+ - W^-)/(W^+ + W^-)$로 산출하였다($W^+$, $W^-$: 사후 증가·감소 쌍의 순위합)[26]. 총점은 두 구성개념 간 효과크기를 비교할 수 있도록 dz와 r을 함께 제시하였다. 총점은 t-검정의 평균 차이에 대한 신뢰구간 또는 Wilcoxon 검정의 Hodges-Lehmann 추정치와 신뢰구간(동점이 있는 경우 정규근사 적용)을 제시하였다. 분석에는 jamovi 2.7(R 4.5)을 사용하였다. 하위요소는 구성개념별 6개 검정을 각각 하나의 묶음으로 설정하고, Holm 방법을 적용하여 유의확률을 보정하였으며(Rj 모듈의 p.adjust 함수 사용), 보정 전·후 유의확률을 함께 제시하였다. 또한 표본 수가 적어 검정 방법에 따라 결과가 달라질 가능성을 고려하여, 모든 변수에 대응표본 t-검정과 Wilcoxon 부호순위검정을 함께 실시하고 결과를 비교하였다. 2문항으로 측정된 정보 신뢰성 비교 하위요소의 문항별 결과도 확인하였다.

양적 결과를 보완하기 위해 프로그램 종료 후 작성된 성찰문(n=18, 8문항)을 활용하였다. 연구자는 전체 응답을 코딩 절차 없이 문항별로 검토하였으며, 세 핵심역량과의 관련성을 중심으로 BERT·XAI 분석 결과에 대한 해석, 자신의 판단과의 비교, 외부 자료를 통한 교차 검증, 편향 및 사회적 영향에 대한 인식이 드러나는지를 살펴보았다. 인용 사례는 각 역량과의 관련성이 명확하고 학습자의 판단 과정이 구체적으로 드러나는 응답을 중심으로 선정하였으며, 맞춤법을 수정하지 않고 제시하였다. 성찰문은 사후에만 수집하였으므로 사전·사후 변화를 직접 확인할 수 없으며, 체계적인 질적 분석 결과가 아닌 양적 결과의 해석을 보완하는 사례 자료로 활용하였다. 특히 검사 도구에서 측정되지 않은 사회적 책임 역량은 관련 인식을 살펴보는 수준으로 제한하였으며, 역량의 향상으로 해석하지 않았다.

## IV. Results

### 1. Quantitative Results

비판적 사고 기능 총점은 사전 16.11점(SD=6.10)에서 사후 18.22점(SD=5.64)으로 통계적으로 유의하게 증가하였다(Wilcoxon, W=98.00, p=.032, r=.63, Hodges-Lehmann 추정치 2.00, 95% CI [0.00, 4.50]; 참고 dz=0.54). 성향 총점은 사전 45.56점(SD=10.29)에서 사후 64.89점(SD=9.13)으로 통계적으로 유의하게 증가하였다(t=5.63, p<.001, dz=1.33, 95% CI [12.08, 26.58]; 참고 r=.96). 모든 변수에서 대응표본 t-검정과 Wilcoxon 부호순위검정의 보정 전 유의성은 일치하였다.

하위요소별 결과는 Table 1과 같다. 기능 영역에서는 Holm 보정 전 다양한 관점 조망 점수가 증가하고 정보 신뢰성 비교 점수가 감소하였으나, 보정 후에는 통계적으로 유의한 하위요소가 없었다. 정보 신뢰성 비교는 두 문항 모두 사후 정답자 수가 감소하였으나(3번 14명→11명, 4번 15명→11명), 문항별 차이는 통계적으로 유의하지 않았다(3번 W=3.00, p=.233; 4번 W=3.50, p=.129). 성향 영역에서는 탐구성을 제외한 5개 하위요소에서 Holm 보정 후에도 통계적으로 유의한 증가를 보여주었다.

### 2. Learner Reflections Related to the Quantitative Results

일부 학생은 XAI의 기여도 수치를 통해 분석 모델의 출력에 상대적으로 크게 기여한 단어나 구절을 확인할 수 있

www.kci.go.kr

Table 1. Pre-Post Comparison by Sub-factor

| Construct | Sub-factor | Pre M(SD) | Post M(SD) | Effect size | p | Holm p |
|---|---|---|---|---|---|---|
| Skill | Fact-opinion discrimination | 3.67(1.19) | 4.00(1.08) | $d_z$=0.24 | .317 | .916 |
| | Evidence-based evaluation | 3.94(2.04) | 4.56(2.04) | $d_z$=0.36 | .150 | .600 |
| | Information reliability comparison | 1.61(0.61) | 1.22(0.73) | r=-.78 | .042* | .209 |
| | Perspective-taking | 2.78(1.63) | 3.94(1.73) | $d_z$=0.60 | .021* | .127 |
| | Bias detection | 2.17(1.15) | 2.28(1.02) | r=.25 | .588 | .916 |
| | Identification of hidden meaning and assumptions | 1.94(1.11) | 2.22(1.00) | r=.42 | .305 | .916 |
| Disposition | Truth-seeking | 7.78(1.93) | 11.67(2.59) | $d_z$=1.05 | <.001*** | .001** |
| | Open-mindedness | 7.83(2.36) | 11.44(2.23) | $d_z$=1.22 | <.001*** | <.001*** |
| | Inquisitiveness | 9.00(2.30) | 9.94(2.21) | $d_z$=0.40 | .112 | .112 |
| | Objectivity | 6.67(2.45) | 11.39(1.72) | $d_z$=1.49 | <.001*** | <.001*** |
| | Systematicity | 7.17(2.31) | 10.17(1.65) | $d_z$=1.06 | <.001*** | .001** |
| | Self-confidence | 7.11(2.30) | 10.28(1.78) | $d_z$=1.00 | <.001*** | .001** |

Note. $d_z$=effect size for the paired-samples t-test; r=rank-biserial correlation for the Wilcoxon signed-rank test, $(W^+ - W^-)/(W^+ + W^-)$, where a negative value indicates a decrease. Holm p = p-values adjusted by the Holm method within each construct (six sub-factors). Sub-factor results are exploratory. *p < .05, **p < .01, ***p < .001.

게 되었다고 응답하였으나("결과에 가장 큰 영향을 미친 특정 단어들과 구절을 찾아낼 수 있다"), 정보의 사실 여부를 판별하는 데 어려움을 경험했다는 응답도 나타났다("가짜를 찾는것이 힘들었던 것 같고 독해 능력의 한계를 느꼈던 것 같다"). 이러한 응답은 근거 기반 평가의 사후 평균이 증가하였으나 통계적으로 유의하지 않았던 결과(dz=0.36, p=.150)와 함께 살펴볼 수 있다.

정보 신뢰성 비교와 관련해서는 정보의 중요도에 따라 검증 강도를 달리해야 한다는 응답과 함께, AI 전반에 대한 경계심을 드러낸 응답("역시 AI는 무조건 믿으면 절대 안된다")이나 근거가 제시된 문장이 거짓으로 판명되어 기존 신뢰 판단 기준에 혼란을 경험한 응답("그 문장이 거짓이라고 나왔을 때 충격을 당했습니다")이 있었다. 이러한 사례는 AI에 대한 경계심이 개별 정보의 신뢰성을 변별하는 판단으로 반드시 이어지는 것은 아닐 가능성을 보여준다. 다만 일부 학생의 응답에 근거한 해석이므로, 이를 정보 신뢰성 비교 점수 감소의 원인으로 단정할 수는 없다.

다양한 관점 조망과 관련하여 6차시에서 자신의 판단, 분석 결과, 동료의 판단이 서로 다를 수 있음을 확인한 응답("여러 관점을 알게 되어서 흥미로웠다")이 있었다. 성향과 관련해서는 무비판적 수용에 대한 경계, AI의 분석 결과를 하나의 관점으로 인식하는 태도, 출처 확인과 교차검증에 관한 응답이 나타났다. 반면, 실생활에서의 구체적인 적용 방안을 제시하지 못한 응답도 있어 이를 탐구성의 양적 결과와 함께 살펴볼 수 있다. 검사에서 측정되지 않은 사회적 책임과 관련해서는 민감한 주제를 다룰 때 신중한 접근이 필요하다는 인식이 나타났다.

## V. Conclusion and Discussion

BERT·XAI 기반 교육 프로그램을 고등학교 1학년 18명에게 적용한 결과, 성향과 기능 총점이 모두 유의하게 증가하였으나, 두 영역의 효과크기를 각각 동일한 지표로 비교한 결과, 성향(dz=1.33, r=.96)이 기능(dz=0.54, r=.63)보다 크게 나타났으며, 하위요소 수준에서는 성향의 5개 하위요소에서만 보정 후에도 유의한 증가가 확인되었다. 이는 성향과 기능이 동일한 양상으로 변화하지 않았다는 신경림 외(2005)의 종단연구[27]와 유사한 맥락에서 살펴볼 수 있다. 또한, 태도와 인식에 해당하는 성향에 비해 실제 판단 수행에 해당하는 기능의 하위 능력별 변화에는 보다 반복적인 분석·검증 경험이 필요할 가능성을 시사한다.

정보 신뢰성 비교의 감소는 Holm 보정 후 통계적으로 유의하지 않았으나, 두 문항 모두 사후 정답자 수가 감소하여 그 원인을 탐색적으로 검토할 필요가 있다. 이에 대해서는 측정과 학습의 두 측면에서 여러 설명이 가능하다. 측정 측면에서 이 하위요소는 객관식 2문항(0~2점)으로 구성되어 소수의 응답 변화에도 평균 점수가 크게 달라질 수 있으며, 사전 3번 문항에만 '(주관적인)'이라는 단서가 포함되어 있었고, 정답 외 선택지에도 규범적 주장이 포함되어 문항의 해석과 정답 판단에 이견이 발생할 가능성이 있다. 학습 측면에서는 근거나 출처가 제시된 정보도 검증 결과와 다를 수 있다는 경험이 '근거 제시 여부' 중심의 기존 판단 기준에 영향을 미쳤을 가능성이 있다. 따라서 후속 적용에서는 근거와 출처 자체의 타당성을 교차 검증하는 활동을 보강하고, 신뢰성 비교 문항 수를 늘리는 한편 선택 이유를 함께 서술하도록 할 필요가 있다.

학술적으로는 생성형 AI 텍스트의 특성과 비판적 사고의 핵심역량을 연계하여 BERT·XAI 기반 교수·학습 활동

www.kci.go.kr

으로 구체화하는 설계 방식을 제시하였고, 교육적으로는 학습자가 기여도를 무비판적으로 수용하지 않고 자신의 판단과 외부 근거와 비교·검토하도록 프로그램을 구성하였다. 또한 분석 리포트를 사전에 산출하여 별도의 코드 실행 필요 없이 활동을 진행할 수 있었다. 따라서 향후 문장 입력을 통해 분석 리포트를 생성하는 웹 기반 도구와 기여도 해석 시 유의점을 포함한 교사용 안내 자료를 개발·제공한다면, 교육 현장의 활용 가능성을 높일 수 있다.

본 연구의 제한점은 다음과 같다. 첫째, 통제집단이 없는 단일집단 사전·사후 설계로 관찰된 변화를 프로그램의 효과로 직접 귀인하거나 연구 결과를 일반화하는 데 한계가 있다. 검사 형식에 대한 익숙함, 성숙, 기대와 요구 특성, 연구자가 수업을 진행한 영향, 자기보고식 성향 검사의 사회적 바람직성, 응답 기준의 변화(response shift) 등이 함께 작용했을 가능성이 있다. 따라서 성향 하위요소의 큰 효과크기(dz=1.00~1.49)는 신중하게 해석할 필요가 있다. 또한 프로그램에 교차 검증, 토의 등 여러 교수·학습 요소가 포함되어 있어 BERT·XAI 활동만의 독립적인 효과를 구분하기 어렵다. 둘째, 7차시의 비교적 짧은 적용으로 기능의 하위요소별 변화를 충분히 확인하기 어려웠을 가능성이 있다. 셋째, 사전·사후 문항의 동등성을 통계적으로 검증하지 않았으며(동등성 I-CVI .75), 일부 기능 검사 문항은 사전·사후의 발문 조건이 달라 하위요소 점수의 변화에는 프로그램 적용에 따른 변화뿐 아니라 문항 난이도의 차이의 영향이 포함되었을 가능성이 있다. 또한 정보 신뢰성 비교는 2문항으로 측정되어 측정의 안정성에 한계가 있으며, 성향 하위요소의 내적 일관성도 낮게 나타났다. 넷째, 2차시에서 [CLS] 토큰 하나의 어텐션 분포를 단어 간 연결 강도 전반으로 해석하거나, 6차시에서 기여도가 높게 나타난 이유를 학습 데이터의 특성과 연결하여 설명하는 등 일부 교수 안내에는 분석 모델의 출력만으로 직접 확인할 수 없는 해석이 포함되었다. 따라서 이러한 해석은 생성형 AI의 실제 생성 과정이나 학습 데이터 분포를 설명하는 확정적 근거가 아니라, 학습자와 함께 검토할 수 있는 가능한 설명으로 제시할 필요가 있다. 다섯째, 사회적 책임 역량은 검사 도구에서 직접 측정되지 않았으며, 성찰문은 프로그램 종료 후에만 수집되어 코딩 절차 없이 보완적 사례 자료로 검토되었다.

후속 연구에서는 다양한 학교와 학습자를 대상으로 통제집단을 두어 프로그램의 효과를 보다 엄밀하게 검증하고, 프로그램을 장기간 적용하고 추적 검사를 실시하여 변화의 지속 여부를 확인할 필요가 있다. 또한 BERT·XAI 활동 포함 여부를 달리한 집단을 비교하여 해당 활동의 고유한 기여를 검토하고, 측정도구의 신뢰도와 타당도를 추가로 검토할 필요가 있다. 아울러 하위요소별 문항 수를 확대한 검사 도구와 사회적 책임 역량을 측정하는 도구를 활용하여 핵심역량을 보다 포괄적으로 평가할 필요가 있다. 이러한 제한점에도 불구하고 본 연구는 BERT·XAI를 학습자의 비판적 판단을 지원하는 교육적 비계로 활용하는 교수·학습 방안을 구체화하고, 생성형 AI 시대의 비판적 사고 교육에 적용할 가능성을 제시하였다는 데 의의가 있다.

## REFERENCES

[1] G. Rani, J. Singh, and A. Khanna, "Comparative Analysis of Generative AI Models," 2023 ICAICCIT, pp. 760-765, Faridabad, India, Nov. 2023. DOI: 10.1109/ICAICCIT60255.2023.10465941

[2] L. Banh and G. Strobel, "Generative artificial intelligence," Electronic Markets, vol. 33, Art. no. 63, Dec. 2023. DOI: 10.1007/s12525-023-00680-1

[3] R. Ganjoo, J. Rankin, B. Lee, and L. Schwartz, "Beyond boundaries: exploring a generative artificial intelligence assignment in graduate, online science courses," Journal of Microbiology & Biology Education, vol. 25, no. 3, Art. no. 00127-24, Dec. 2024. DOI: 10.1128/jmbe.00127-24

[4] E. S. Lee and E. J. Jin, "A case study on ChatGPT 'hallucination' education in college writing classes," Journal of Donam Language and Literature, vol. 45, pp. 211-243, Jun. 2024. DOI: 10.17056/donam.2024.45..211

[5] H. H. Shin, "Analysis of generative AI service usage by age group," KISDI STAT Report, no. 25-07, Korea Information Society Development Institute, Jul. 2025.

[6] S. E. Myeong, "2025 National AI Perception Report: An Analysis of Perception Characteristics Based on AI Understanding and Usage Experience," The AI Report, no. 2025-19, National Information Society Agency, Dec. 2025.

[7] UNESCO, "AI Competency Framework for Students," UNESCO, 2024. DOI: 10.54675/JKJB9835

[8] UNESCO, "AI Competency Framework for Teachers," UNESCO, 2024. DOI: 10.54675/ZJTE2084

[9] National Information Society Agency, "Development of a Digital (AI) Competency Framework for the AI Era: Final Report," Dec. 2024.

[10] S. Y. Choi, "Core competency framework and education plan for future talent in the era of generative AI," Journal of The Korean Association of Computer Education, vol. 27, no. 9, pp. 23-33, Dec. 2024. DOI: 10.32431/kace.2024.27.9.003

[11] U. Y. Kim, "An author without sentience: literacy and critical thinking in the era of generative AI," The Korean Journal of Literacy Research, vol. 16, no. 5, pp. 389-416, Oct. 2025. DOI:

www.kci.go.kr

10.37736/KJLR.2025.10.16.5.12

[12] S. J. Oh and K. H. Lee, "Reconceptualizing critical literacy in the age of AI," Cheongram Language Education, vol. 108, pp. 67-95, Nov. 2025.

[13] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," Proceedings of NAACL-HLT 2019, pp. 4171-4186, Minneapolis, USA, Jun. 2019.

[14] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," IEEE Access, vol. 6, pp. 52138-52160, Sep. 2018.

[15] Captum, "Layer Integrated Gradients," Captum Documentation. [Online]. Available: https://captum.ai/api/layer.html.

[16] E. S. Jung, H. B. Lee, and K. H. Kim, "A Study on ways to strengthen AI literacy capabilities for high school students through explainable AI education," Journal of The Korea Society of Computer and Information, vol. 30, no. 8, pp. 211-220, Aug. 2025. DOI: 10.9708/jksci.2025.30.08.211

[17] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," Proceedings of the 34th ICML, pp. 3319-3328, Sydney, Australia, Aug. 2017.

[18] S. Jain and B. C. Wallace, "Attention is not explanation," Proceedings of NAACL-HLT 2019, pp. 3543-3556, Minneapolis, USA, Jun. 2019.

[19] Y. Kim, J.-H. Kim, J. M. Lee, M. J. Jang, Y. J. Yum, S. Kim, U. Shin, Y.-M. Kim, H. J. Joo, and S. Song, "A pre-trained BERT for Korean medical natural language processing," Scientific Reports, vol. 12, Art. no. 13847, 2022. DOI: 10.1038/s41598-022-17806-8

[20] M. R. Zafar and N. Khan, "Deterministic local interpretable model-agnostic explanations for stable explainability," Machine Learning and Knowledge Extraction, vol. 3, no. 3, pp. 525-541, Jun. 2021. DOI: 10.3390/make3030027

[21] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems, vol. 30, pp. 4765-4774, Long Beach, USA, Dec. 2017.

[22] Ministry of Education, "Ministry of Education Notice No. 2022-33 (Dec. 2022), General Curriculum and Subject Curricula for Elementary, Middle, and High Schools."

[23] J. H. Kim, "Development of a Collaborative Observation Instructional Strategy for Enhancing Critical Thinking about AI," Master's thesis, Seoul National University Graduate School, Aug. 2025.

[24] K. H. Kim, "The effects of problem-based learning on children's critical thinking ability," Master's thesis, Graduate School of Education, Busan National University of Education, Aug. 2001.

[25] S. H. Park, "The effects of self-assessment on enhancing critical thinking," Master's thesis, Graduate School of Education, Incheon National University of Education, Feb. 2002.

[26] D. S. Kerby, "The simple difference formula: An approach to teaching nonparametric correlation," Comprehensive Psychology, vol. 3, Art. no. 1, 2014. DOI: 10.2466/11.IT.3.1

[27] K. R. Shin, J. Y. Ha, and K. H. Kim, "A longitudinal study of critical thinking dispositions & critical thinking skills in baccalaureate nursing students," Journal of Korean Academy of Nursing, vol. 35, no. 2, pp. 382-389, Apr. 2005.

## Authors



Jun-Hyeok Seo received the B.S. degree in Computer Science from Kyonggi University, Korea, in 2018. He is currently pursuing the M.S. degree in the Department of Computer Education at Korea National University of Education, Korea Research interests include computer education, computational thinking and artificial intelligence.



Kyung-Hwa Lee received the M.A. and Ph.D. degrees in Korean Language Education from Korea National University of Education, Cheongju, Republic of Korea, in 1996 and 1999, respectively.

Since 2002, she has been with the Department of Elementary Education, Korea National University of Education, where she is currently a Professor. Research interests include critical literacy, multimodal literacy, digital text reading, and AI-assisted reading and writing.



Ji-Hoon Bae received the B.S. degree in electronic engineering from Kyungpook National University, Daegu, Republic of Korea, in 2000, and the M.S. and Ph.D. degrees in electrical engineering from Pohang University of Science and Technology, Pohang, Gyeongbuk, Republic of Korea, in 2002 and 2016, respectively. From 2002 to 2019, he was a Principal Researcher with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, Republic of Korea. From 2019 to 2024, he was a Faculty Member with the Department of Artificial Intelligence and Big Data Engineering, Daegu Catholic University. In 2024, he joined the Department of Computer Education, Korea National University of Education, Cheongju, Chungbuk, where he is currently an Associate Professor. His research interests include embedded AI, transfer learning, knowledge tracing, deep learning, and computer education. He is a member of Korea Institute of Information Technology (KIIT). He has been serving as the Editor-in-Chief of KIIT domestic journal, since 2025.

www.kci.go.kr