

컴퓨터교육학회 논문지 2026년 제29권 제5호
<https://doi.org/10.32431/kace.2026.29.5.003>



성취기준-NCS 연계 기반 직업계고 자동문항 생성 프레임워크 설계 연구: ‘측량’ 교과 사례

Design and Validation of an Achievement Standards-NCS Linked Automatic Item Generation Framework for Vocational High Schools: The Case of Surveying

곽윤미[†] · 이시훈^{††}
 Yunmi Kwak[†] · Sihoon Lee^{††}

요약

본 연구는 2022 개정 교육과정과 고교학점제 환경에서 직업계고 전문교과의 평가 타당도와 현장 적합성을 제고하기 위해 성취기준-NCS(국가직무능력표준) 연계 기반 자동문항생성(AIG) 프레임워크를 설계하였다. 설계과학연구방법론(DSRM)을 적용하여 핵심 요소 도출-데이터셋 구축-프로토타입 시연-전문가 검증을 수행하였다. 또한 Webb의 DOK(Depth of Knowledge) 모형을 적용하여 문항유형과 난이도 기준을 정교화하였다. 핵심 성과는 RAG(Retrieval-Augmented Generation) 기술 기반 근거 제시(외부 루프)와 교사 피드백 기반 품질을 개선(내부 루프)을 결합한 이중 순환(dual-loop) 구조를 구현한 데 있다. 이는 LLM(대규모 언어모델)의 할루시네이션 문제를 보완하고, 설명 가능한 AI(XAI) 관점에서 직업계고 전문교과 평가의 교육적 타당도와 신뢰도를 확보할 수 있는 구체적 설계 대안을 제시했다는 점에서 의의가 있다.

주제어 자동문항생성, 직업계고, 성취기준, NCS, 설명 가능한 AI, HITL

ABSTRACT

This study designed an Automatic Item Generation (AIG) framework based on the linkage between achievement standards and National Competency Standards (NCS) to enhance the evaluation validity and field suitability of specialized subjects in vocational high schools under the 2022 Revised National Curriculum and the High School Credit System. Applying the Design Science Research Methodology (DSRM), the research process encompassed core element derivation, dataset construction, prototype demonstration, and expert validation. Furthermore, Webb's Depth of Knowledge (DOK) model was employed to refine the criteria for item types and difficulty levels. The primary achievement of this study lies in the implementation of a dual-loop structure that integrates Retrieval-Augmented Generation (RAG)-based evidence provision (outer loop) with teacher feedback-driven quality improvement (inner loop). This approach is significant as it mitigates the hallucination issues inherent in Large Language Models (LLM) and provides a concrete design alternative to ensure the educational validity and reliability of vocational evaluations from the perspective of Explainable AI (XAI).

Keywords Automatic Item Generation, Vocational High School, Achievement Standards, NCS, Explainable AI, Human-in-the-loop

†정회원	한국교원대학교 대학원 교육학과 박사과정
††정회원	한국교원대학교 뇌AI기반 교육대학원 강사
논문투고	2025년 11월 10일
심사완료	2026년 01월 29일
게재확정	2026년 02월 25일
발행일자	2026년 05월 11일

1. 서론

1.1 연구의 배경 및 필요성

2022 개정 교육과정과 2025년 고교학점제의 전면 도입은 직업계고 교육 현장에 성취기준 중심의 교육과정 설계, 최소 성취수준 보장지도, 과정 중심 평가의 강화를 요구한다[1, 2]. 이러한 변화는 학생의 과목 선택권 확대와 책임교육 강화라는 긍정적 측면과 함께 다과목 지도와 새로운 교육과정 운영에 따른 교사의 업무 부담 가중의 현실적 과제를 동시에 부여하고 있다[3].

특히 산업 현장과의 연계성이 핵심인 직업계고 전문교과에서 평가 문항은 학생의 실무 역량을 측정하는 중요한 척도이므로 현장성과 타당성을 확보하는 것이 중요하다. 그러나 기존의 평가는 정형화된 지필 평가 중심으로 이루어져 성취기준과의 연계성이 낮고, 급변하는 산업 기술과 실무 상황을 반영하는 데 한계를 보여왔다. 더불어 교사들이 양질의 평가 문항을 지속적으로 개발하고 검증하는 데 투입해야 하는 시간과 노력은 부담으로 작용하고 있다.

이러한 맥락에서 직업계고 전문교과 교사는 전공 이론 지식과 실무 지식을 갖추어야 하며, 학습자 맞춤 평가 설계와 디지털 기술 활용 역량도 동시에 요구되는 ‘이중 전문성(dual expertise)’이 강조되고 있다[4]. 최근 대규모 언어모델(LLM)의 발전은 자동문항생성(Automatic Item Generation, AIG) 분야의 유망한 가능성을 열었지만, 개별 교사에 의한 범용 AI 활용은 생성 문항의 일관성과 품질을 담보하기 어렵고, 직업계고 전문교과 NCS 연계성을 충분히 반영하지 못할 수 있다.

AIG 관련 선행연구에서는 주로 다지선다형 문항 중심으로 직업 및 대학 수준에서 적용되며, 특징 공학(feature engineering), 프롬프트 공학(prompt engineering) 등을 통해 생성 문항의 효율성과 일관성을 확보하려는 시도가 이루어지고 있다[5, 6]. 그러나 LLM 기반 AIG는 언어적 유연성과 다양한 주제를 다루는 데 강점을 보이지만, 생성 문항의 교육적 타당성 검증과 품질 관리 체계 부재라는 한계를 지닌다[7]. 이에 따라 인간 전문가의 구조적 평가 절차를 포함하는 자동문항생성(AIG) 프레임워크의 개발과 인간-기계 협업(Human-in-the-loop, HITL) 기반 접근에 대한 필요성이 강조되고 있다[8].

따라서 단순한 프롬프트 활용 수준을 넘어 직업계고 전문교과의 성취기준을 체계적으로 분석하고 이를 문항생성 규칙과 연계하여 평가의 정합성과 실무성을 확보하는 구조화된 접근이 필요하다. 이러한 접근은 교사의 문항 개발 부담을 경감시킬 뿐 아니라 교과협의회나 전문적 학습공동체에서 표준화된 평가 도구를 공동 개발·공유하는 협업 기반을 마련하여 직업계고 평가의 질적 혁신을 이끄는 전략적 대안이 될 수 있다.

1.2 연구 목적 및 연구 문제

본 연구의 목적은 2022 개정 교육과정과 고교학점제 환경에서 직업계고 전문교과 평가의 타당도와 현장 적합성을 제고하고 교사의 평가 전문성을 지원하기 위해 성취기준-NCS 매핑 및 전문가 피드백과 연동되는 AI 기반 AIG 프레임워크를 설계하는 데 있다.

본 연구는 프레임워크를 설계하고 파일럿 수준의 시연과 전문가 검토를 통해 구조적 타당성과 적용 가능성을 점검함으로써 향후 실증적 검증의 기초를 마련하고자 한다. 이를 위해 다음과 같은 연구 문제를 설정하였다.

첫째, 직업계고 전문교과 성취기준에 최적화된 AI 기반 AIG 프레임워크의 설계 방안은 무엇인가?

둘째, 개발된 프레임워크의 문항생성 성능은 어떠한가?

셋째, 생성된 문항은 성취기준 정합성, 실무성, 교육적 타당성을 확보하는가?

이러한 탐색을 통해 직업계고 교육 현장에서 평가의 질과 효율성을 동시에 향상시키고 성취기준-NCS 연계 기반 AIG 프레임워크의 구조적 재현성과 교육적 활용 가능성을 제시하고자 한다.

2. 이론적 배경

2.1 직업계고 교육과정과 평가의 특수성

2025년 고교학점제의 도입은 직업계고에도 교육과정의 유연성과 학생 수요 중심 운영을 요구하면서 평가 방식에도 변화를 유발하였다. 김시은 외[9]는 정책 문서 분석을 통해 고교학점제 아래 교사에게 요구되는 전문성이 ‘다과목 지도 전문성’, ‘학업성취수준 관리 전문성’, ‘평가 다각화 전문성’, ‘진로·학업 연계 지도 전문성’과 같이 다면적 차원에서 동시에 요구된다고 지적하였다. 이는 산업 현장 실무와 교육과정 연계성을 중심으로 운영되는 직업계고 전문교과의 특수성과 맞닿는 부분이다.

한편 직업계고 전문교과 교사에게는 교육학적 지식뿐 아니라 실무적 전문성을 함께 갖춘 이중 전문성이 요구된다. 김남희 외[4]는 델파이 연구를 통해 직업계고 교사의 전문성을 ‘직업적’과 ‘교육학적’으로 구분하며, 전자는 전공 분야 이론 및 실무 지식, 산업현장 대응력, 디지털 전환 대응까지 포함하고, 후자는 교수·학습 설계, 피드백 제공, 디지털 기술의 교육적 활용 등을 포함한다고 제시하였다. 그럼에도 한국직업능력연구원의 조사 결과에 따르면 전문교과 교사는 전공 이론 지식에 대한 자신감은 높지만 산업현장과의 협력 역량은 상대적으로 낮게 인식되고 있는 것으로 나타났다[1]. 이는 이중 전문성의 균형적 함양이 아직 충분히 이루어지지 않았음을 보여준다.

추가적으로 NCS 기반 교육과정 운영 역량 중 교사의 평가 설계 능력이 핵심 역량으로 강조된다는 사실은 직업계고 전문교과의 평가 문항이 성취기준과 NCS 능력단위에 기반하여 정교하게 설계되어야 함을 방증한다.

2.2 성취기준 기반 교육평가 이론

성취기준은 국가 교육과정에서 학습자가 도달해야 할 학습 성과를 구체적으로 진술한 준거로, 교수·학습 및 평가의 적합성을 담보하는 핵심 장치이다[2]. 성취평가제는 이러한 성취기준을 근거로 학생의 학업 성취 정도를 절대평가 방식으로 판정하여 교육의 공정성과 책무성을 강화하는 제도적 기반을 제공한다[1]. 특히 직업계고 맥락에서 성취평가제는 산업 현장에서 요구되는 최소한의 직무 수행 역량을 학습자가 확보하도록 보장하는 제도적 장치이며, 학습 결손 예방과 공정한 교육 기회 제공의 근거가 된다. 우원재[3]는 고교 학점제 도입에 따른 교사 인식 조사에서 성취평가제가 교사의 책무성과 학생의 학습권 보장을 동시에 실현하는 수단임을 강조하였다. 따라서 직업계고 전문교과의 성취기준 기반 평가는 단순한 학업 성취 판별을 넘어 직업 역량 인증과 교육적 형평성 실현을 동시에 추구하는 평가 패러다임으로 볼 수 있다.

교육평가 연구에서는 성취기준을 행동동사 수준에서 분석하여 인지적 요구 수준과 문항 난이도의 적합성을 검증하는 접근이 일반적이다. 박고운, 최숙기[10]는 국어과 읽기 영역 AIG 연구에서 성취기준-행동동사 분석을 바탕으로 문항 난이도의 체계적 구성이 가능함을 보였으며, 이를 통해 평가의 교육적 타당성을 강화할 수 있음을 입증하였다. 또한 Bloom의 인지 과정 분류체계가 제시하는 ‘기억-이해-적용-분석-평가-창안’의 위계가 성취기준의 인지적 요구 수준을 구체화하는 데 유효하다는 점을 보여주었다.

한편, Bloom의 분류가 학습 목표의 인지적 위계를 중심으로 평가의 적합성을 확보하는 데 초점을 둔다면 Webb(1997)의 DOK(Depth of Knowledge) 모형은 과제가 요구하는 사고의 깊이와 인지적 복잡성 수준을 구분하는 데 중점을 둔다. 선행연구는 2022 개정 초등 교육과정의 성취기준을 DOK 관점에서 분석하여 다수의 성취기준이 DOK 1-2 수준(기억·기초적 기술)에 편중되어 있음을 보고하며, 고차적 사고(DOK 3 이상)를 요구하는 과제 개발의 필요성을 강조하였다[11]. 이는 평가가 단순한 사실 재생이 아닌 복합적 사고와 문제 해결을 요구하는 방향으로 설계되어야 함을 시사한다.

직업계고 전문교과의 평가는 일반 교과와 달리 산업 현장성과 직무 수행 역량을 함께 측정해야 하는 특수성을 지닌다. NCS는 산업별·직무별로 요구되는 지식, 기술, 태도를 세분화한 기준 체계로, 직업계고 평가에서 성취기준과 긴밀히 연계되어야 한다. 직업계고 교사의 전문성은 교육학적 전문성과 직업적 전문성으로 구성되며, 특히 산업 현장 대응력과 평가 설계 능력이 핵심이라는 지적이 있다[4]. 관련하여, 전문교과 교사들이 산업 현장과의 협력 역량을 상대적으로 낮게 인식하고 있으며, 이에 따라 NCS 기반 교육과정과 평가의 연계 강화가 필요하다는 보고도 제시된 바 있다[1].

2.3 인공지능 기반 AIG 연구 동향

AIG는 인지·심리측정 이론과 계산 모델을 통해 문항을 체계적으로 생성하는 기술로 국내에서도 읽기 평가를 중심으로 개념 정립과 적용 시도가 이루어져 왔다[12]. 초기 AIG 연구는 템플릿이나 모형 기반 접근을 통해 일정 수준의 생산성과 통제 가능성을 확보했으나, 문항 다양성과 실체성 확보, 대규모 운영 시 유지보수의 어려움이라는 구조적 한계를 지속적으로 드러냈다[10].

이러한 한계를 극복하기 위해 최근에는 LLM을 활용한 생성형 접근이 주목받고 있다. 여러 실증 연구들은 제로샷(zero-shot), 퓨샷(few-shot), Chain-of-Thought(CoT) 등 다양한 프롬프트 전략이 문항 품질에 직접적 영향을 미친다는 점을 보여주었으며, 특히 CoT 전략이 형식 적합성, 내용 타당도, 논리적 일관성 측면에서 가장 우수하다는 결과가 보고되었다[10]. 한국어능력시험 읽기 영역을 대상으로 한 연구 또한 문항유형별 프롬프트 구성에 따라 생성 결과의 적합성이 달라졌음을 확인하면서, LLM 기반 AIG의 적용 가능성을 탐색하였다[13]. 더 나아가 예비교사를 대상으로 한 실험에서는 단기간에 맞춤형 문항생성기를 구축·활용할 수 있음을 보여주었으나 실제 수업에서 즉시 활용하기 위해서는 여전히 추가적인 편집과 품질 관리가 필요하다는 한계도 함께 제시되었다[14]. 이러한 연구들을 종합하면, 국내 AIG 연구는 아직 초기 단계에 있으며 실제 수업 맥락에서 다문항 세트 생성, 간접 최소화, 오류 감축 등 해결해야 할 과제가 남아 있음을 알 수 있다.

동시에 교육평가 영역에서는 ‘설명 가능한 AI(Explainable AI, XAI)’에 대한 요구가 강화되고 있다. 법·정책 차원에서는 투명성, 책임성, 공정성, 신뢰성이라는 기본 원칙이 강조되고 있으며[15], LLM의 ‘근거 생성’ 품질이 변동적이라는 평가 결과는 출제 근거, 정답 논리, 참조 맥락을 함께 제공하는 메커니즘의 필요성을 뒷받침한다[16]. 더불어 국내 AI 기반 평가 연구 전반에서도 도구의 타당성·공정성·사용가능성을 동시에 확보해야 한다는 합의가 확산되고 있다[17].

또한 우리나라에서는 자동채점 기술이 AIG보다 앞서 대규모 평가 현장에 적용된 바 있다. 국가수준 학업성취도 평가에서 문장 수준 서답형 자동채점은 인간 채점과 높은 상관을 보여 사람-기계 협업 기반 평가 체계가 신뢰성 있게 운영될 수 있음을 입증하였다[18]. 이러한 경험은 AIG 역시 전문가 피드백과 품질관리 절차를 결합한다면 교육평가 체제에 안전하게 통합될 수 있음을 시사한다.

결국 LLM 기반 AIG는 전통적 템플릿 의존의 한계를 넘어 문항 다양성과 개발 효율성을 높일 수 있는 잠재력을 보여주었고, 동시에 교육책무성 차원에서는 ‘설명 가능한 문항(explainable item)’과 사람-AI 협업 품질관리 체계가 핵심 설계 원리로 자리 잡고 있다. 본 연구는 이러한 동향을 반영하여 성취기준과 NCS를 다차원 기준표로 구조화해 적합성을 확보하고, 정답·오답 해설과 출제 근거·난이도 정보를 포함한 설명 가능한 산출을 제공하며, 전문가 피드백 루프를 통한 반복 개선을 결합함으로써 기존 연구의 한계를 보완

하는 품질보장형 AIG 프레임워크를 제시하고자 한다.

3. 연구방법

3.1 연구설계

본 연구는 직업계고 전문교과 평가 문항 개발 과정에서 드러나는 문항의 신뢰성 저하, 근거 추적의 어려움, 교사의 업무 과중 문제를 해결하기 위해 AIG 프레임워크를 개발하고, 구조적 타당성과 현장 적용 가능성을 탐색적으로 검토하는 것을 목적으로 한다.

이를 위해 본 연구는 설계과학연구방법(Design Science Research Methodology, DSRM)을 적용하였다. DSRM은 실천적 문제를 해결하기 위한 인공물(artifact)을 체계적으로 설계·구축하고, 반복적 검증 과정을 통해 이론적 타당성과 실천적 유용성을 동시에 확보하는 연구 방법이다. 안형준[19]은 LLM을 활용한 미술관 안내 시스템 프레임워크 구축 연구에서 DSRM 절차를 설계 중심 연구(Design-centered Research)로 변형하여 프레임워크의 구조적 타당성을 확보한 바 있다. 본 연구 또한 이러한 접근을 준용하여 직업계고 교육평가의 구조적 문제를 해결하기 위한 성취기준 기반 AIG 프레임워크를 인공물로 설계·개발하였다. 또한 DSRM의 절차를 직업계고 평가 맥락에 맞게 수정하여 문제 인식 및 목표 정의, 프레임워크 설계 및 데이터셋 구축, 시연 및 전문가 검증, 결과 정리 및 개선 제안의 순으로 수행하였다.

Table 1. Research procedures

step	Core Goals	Main activities
①	Identifying field issues and deriving research needs	-Interviews with vocational high school teachers (4 people) -Identifying the limitations of automatic question generation research through literature analysis.
②	Deriving design requirements for an AI-based question generation framework	-Interviews with vocational high school teachers (4 people) -Deriving functional and non-functional requirements for the framework
③	Design and prototype development of an automatic question generation framework	-Building a dataset -Framework structure design -Implementing a prototype based on the framework draft
④	Validation of the framework's validity and usability	-Vocational High School Teacher Survey (4 people): Prototype Usability Review -Deriving improvements to the draft framework -Survey of vocational high school teachers (4 people): Evaluation of the usability of framework improvements and the results of question generation.
⑤	Analysis of research results and presentation of design theory	-Analysis of evaluation results -Derivation of design principles -Suggesting future directions for improvement

3.2 연구절차

본 연구는 DSRM의 절차를 기반으로 ① 문제 인식 및 동기 부여→ ② 설계 목표 및 요구 정의→ ③ 인공물 설계 및 개발→ ④ 시연 및 평가→ ⑤ 결과 해석 및 지식 기여의 다섯 단계로 진행되었다. 각 단계별 핵심 목표와 주요 활동은 Table 1과 같다.

3.3 대상 및 범위

3.3.1 연구 참여자 선정

프레임워크 설계와 결과 평가의 타당성을 확보하기 위해 직업계고 측량 교과를 담당하는 교사 4명을 인터뷰 대상으로 선정하였다. 연구 참여자로 선정된 교사는 모두 측량 교과 지도 경력이 10년 이상으로 전문성과 현장 경험을 고루 갖추고 있다. 연구 참여자의 구체적인 특성은 Table 2와 같다.

Table 2. Study participants

Participants	Field of study	Experience in major field
Teacher 1	survey	22 years
Teacher 2	survey	18 years
Teacher 3	survey	14 years
Teacher 4	survey	12 years

3.3.2 분석 대상 자료 및 범위

연구에 활용된 핵심 자료는 문항생성의 기반이 되는 2022 개정 교육과정 측량 교과 성취기준, 실무적 맥락을 제공하는 NCS 능력단위(수준점 측량), 문항의 변별력과 인지적 수준을 통제하기 위한 난이도 기준과 문항유형 지침 그리고 내용의 근거가 되는 교과서와 NCS 학습모듈을 분석 대상으로 포함하였다.

생성된 문항의 품질을 심층적으로 분석하고 프레임워크의 효과성을 명료하게 검증하기 위해 분석 대상 성취기준의 범위를 한정하였다.

분석 대상 성취기준은 '[측기 03-04-01] 수준 측량 장비'를 세울 수 있다.'로 선정하였다. 이 성취기준은 '레벨기 설치'와 '삼각대 세팅'이라는 핵심 실무 기능을 다루면서 NCS 능력단위 '관측하기(1402030202_20v3.4)'의 수행준거인 '레벨, 표적, 삼각대 점검 및 설치'와 직접적으로 연계되어 교과서의 이론적 지식과 현장 실무 표준을 통합하는 프레임워크의 성능을 검증하기에 적합한 요건을 갖추었다.

4. 연구결과

4.1 프레임워크의 핵심 구조와 기능적 요구사항

직업계고 전문교과의 성취기준 기반 AIG 프레임워크를 설계하기 위해 전문가 인터뷰를 통해 프레임워크 설계 시 반

영해야 할 핵심 요소를 선정하고, 이를 기반으로 핵심 구조를 설계하였다. ChatGPT Canvas 기반 프로토타입을 시연하여 프레임워크의 구조적 타당성과 기능적 요구사항을 구체적으로 도출하였다. 또한 핵심 요소 도출 과정에서 전문가 간 판단의 일관성을 점검하기 위해, 이후 4.3절에서 제시하는 정량적 의견 일치도(agreement) 분석을 병행하였다.

AIG 프레임워크의 핵심 반영요소는 이하 3단계의 체계적 과정을 통해 선정되었다.

1) 초기 요소 탐색

선행연구, 2022 개정 교육과정, NCS 학습모듈 분석을 통해 직업계고 성취기준 기반 AIG 프레임워크에 필요한 요소를 다음과 같이 도출하였다.

- (1) 성취기준 코드 입력
- (2) NCS 능력단위 연계 지정
- (3) 문항유형(선다형/단답형 등) 지정
- (4) 난이도(상/중/하) 지정
- (5) 생성 문항의 성취기준 정합성 검토
- (6) 생성 문항의 NCS 실무맥락 반영도 검토
- (7) 문항 내용의 사실적 정확성 검토
- (8) 지정 난이도와와의 적절성(변별력) 검토
- (9) 정답 및 오답의 타당성(매력도) 검토
- (10) 생성 문항의 출제 근거(페이지) 제시
- (11) 정답·오답에 대한 상세 해설 제공
- (12) 교사 피드백(HITL) 반영 기능
- (13) 시나리오 및 상황 제시형 문항 지원
- (14) 문항생성의 재현성
- (15) 시스템 사용 편의성(UI/UX)

2) 1차 전문가 인터뷰 및 요소 정제

1차 전문가 인터뷰를 통해 초기 요소들에 대한 중복 검토와 의미 조정을 실시하여 입력요소, 평가준거, 부가 정보의 3가지 기능적 영역으로 통합 및 재구성하였다.

3) 2차 전문가 인터뷰 및 최종 핵심 요소 확정

프레임워크의 핵심요소를 확정하기 위해 정제된 3개의 기능적 영역을 고려하여 전문가 논의를 거쳐 최종 11개의 핵심 요소를 선정하였다.

우선, 문항생성 입력요소(4개 항목)는 1단계의 성취기준 코드, NCS 능력단위, 문항유형, 난이도 지정 요소가 AI에게 전달하는 핵심적인 입력조건으로 합의되었다. 다음으로 문항 품질 평가준거(5개 항목)는 1단계의 성취기준 정합성, NCS 반영도, 내용 정확성, 난이도 적절성, 정답/오답 타당성 요소가 생성된 문항을 평가하는 5대 준거(루브릭)로 확정되었다. 마지막으로 제공 부가 정보(2개 항목)는 1단계의 출제 근거, 상세 해설 요소가 교사의 교육적 판단을 돕기 위해 AI가 제공해야 할 2가지 핵심 부가 정보로 확정되었다.

Table 3. Key Elements for Designing the AIG Framework

area	Key elements
Question generation input elements	1. Achievement standards 2. Practical Context (NCS Learning Module) 3. Question type 4. Difficulty
Item Quality Evaluation Criteria	1. Consistency of achievement standards 2. Reflection of practical context 3. Accuracy of content and terminology 4. Difficulty Appropriateness 5. Validity of correct and incorrect answers
Additional information provided	1. Detailed explanations of correct and incorrect answers 2. Basis for the question

교사 피드백 기능과 재현성 요소는 프레임워크의 특정 항목이 아닌 6단계 프레임워크(Fig. 1 참고)의 작동 방식(HITL)과 품질 목표에 반영하는 것으로 최종 합의되었다. 이 과정을 통해 3개의 기능적 영역 핵심요소는 최종적으로 Table 3에 제시된 11개의 핵심 반영요소로 수렴 및 확정되었다.

전문가들은 “성취기준과 NCS 간 매핑 구조의 명확화가 학습의 연계성과 평가의 실효성을 높인다”고 평가하였으며, “LLM은 반드시 교사가 지정한 근거 자료에 기반하여 문항을 생성해야 한다”고 강조하였다.

이러한 결과는 프레임워크 설계의 중심 원리로 ‘교육과정 정합성’과 ‘출처 기반 신뢰성’을 설정하게 된 근거가 되었다. 프레임워크 설계의 핵심 요소를 반영하여 데이터셋을 Table 4와 같이 설계하였다.

Table 4. Dataset composition

division	structure	Key Contents	Design Purpose
Dataset A	Achievement Standards – Action Verbs – Bloom Cognitive Domain Mapping	Standardization of behavioral areas by subject achievement standards	Determining the cognitive level of the items
Dataset B	Difficulty criteria and adjustment rules	Setting specific criteria for distinguishing between high, medium, and low difficulty levels	Automatically adjust question difficulty
Dataset C	Question type rules	Defining stem – option – answer patterns by question type	Ensuring consistency in question structure by type
Dataset D	Achievement Standards – NCS Competency Unit Expansion Mapping	Structuring the link between school education and NCS field work	Ensuring consistency between practical skills and curriculum

프레임워크의 핵심 구성 요소인 데이터베이스를 구축하기 위해 연구 자료를 구조화하였다. 프레임워크의 규칙 기반 제어를 위해 성취기준, NCS 매핑표, 난이도 및 문항유형 기준표 등 4종의 규칙 데이터를 표준화된 CSV(comma-separated values) 형식으로 구조화하였다. 각 파일의 구

조와 내용은 Table 5와 같다. 이 과정을 통해 LLM의 확률적 해석에 의존하지 않고, 시스템이 코드 기반으로 정확하게 규칙을 조회할 수 있는 기반이 마련되었다.

Table 5. Dataset structure and details

division	column	Content (example)
DatasetA	Achievement Standard Code	측기 03-01-01
	Achievement Standard Contents	Can explain the principles of leveling
	action verbs	explain
	Content elements	Leveling principles, field book entry method
	Bloom's Cognitive Domain	understanding
DatasetB	Difficulty	middle
	Bloom's Cognitive Domain	Application, Analysis
	Key Features	The ability to apply acquired knowledge or concepts to new situations, or to break down information into its components and understand their relationships.
DatasetC	Question type	Best answer
	Production principle	-Select the most appropriate answer from the answer sheet. -Answers to social phenomena or application cases -Used primarily to assess the level of application, as well as retention and understanding of knowledge and concepts.
	Example sentence	Which of the following is the most appropriate example of the principle being utilized in the following passage?

4.2 프레임워크 초안 설계

프레임워크의 설계 과정은 4.1.1절의 핵심 요소 도출(Table 3)과 4.1.2절의 데이터셋 구성(Table 4)이 체계적인 논리 구조를 갖도록 설계되었다. 전문가 합의를 통해 도출된 Table 3의 AIG 프레임워크 설계를 위한 핵심 요소는 프레임워크가 준수해야 할 교육학적 설계 원칙에 해당한다. Table 4의 데이터셋 구성은 이러한 설계 원칙을 시스템이 이해하고 실행할 수 있도록 정형화한 것이다. Table 3의 문항생성 입력요소(성취기준, 실무맥락, 문항유형, 난이도)는 Table 4의 데이터셋 A-D로 각각 구현되었으며, 이는 프레임워크 작동의 기반이 되는 설계 규칙이다.

이 데이터셋(규칙)을 활용하여 실제 문항을 생성하고 평가하는 6단계의 실행 절차를 Fig. 1과 같이 프레임워크 초안으로 제시하였다. 각 단계는 성취기준과 NCS 연계 규칙을 적용해 문항의 정합성과 설명가능성을 동시에 확보하도록 설계되었다.

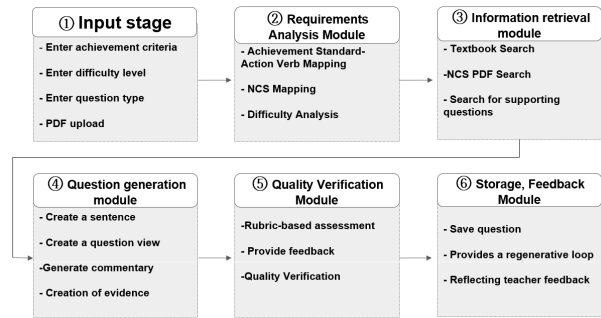


Figure 1. Framework draft

프레임워크 설계 핵심요소와 데이터셋의 논리적 연계성을 강조하기 위해 Fig. 1의 ② 요구분석 모듈이 Table 4의 Dataset A, B, C, D를 모두 참조하여 문항생성 계획을 수립하는 핵심 역할을 수행함을 명시하였다. 또한 Fig. 1의 ⑤ 품질검증 모듈은 단순한 검토가 아니라 Table 3에서 도출된 문항 품질 평가준거 5개 항목을 루브릭으로 직접 적용하여 문항을 평가하고 피드백하는 핵심 단계임을 명확히 기술함으로써 프레임워크의 학술적 타당성과 구조적 완성도를 확보하고자 하였다. 이 구조는 AI의 자동화 능력과 교사의 전문적 판단이 결합된 반자동(Human-in-the-loop) 형태로 설계되었다. 즉 데이터 기반의 문항생성과 교사의 교육적 판단이 상호보완적으로 작동하는 것을 목표로 하였다.

4.3 프레임워크 초안 시연 및 개선사항 도출

프레임워크 초안의 논리적 타당성을 검토하고 실제 현장에서 필요한 기능적 요구사항을 정교화하기 위해 개념적 모델을 활용해 프레임워크 초안의 시연 및 전문가 피드백을 수행하였다. 이때 전문가 의견 일치도 분석은 프레임워크 자체에 대한 총괄적 평가가 아니라, 프레임워크를 통해 생성된 문항의 품질을 평가하기 위한 준거별 판단의 합의 수준을 확인하는 데 목적이 있다. 구체적으로, 전문가들은 생성 문항에 대해 성취기준 정합성, NCS 실무 맥락 반영도, 내용의 사실적 정확성, 난이도 적절성, 정답·오답의 타당성의 다섯 가지 평가준거를 기준으로 5점 리커트 척도로 평정하였으며, 본 연구는 이러한 평정 점수들 간의 일치도를 Cronbach's α , ICC, Kendall's W 지표를 통해 분석하였다.

시연은 ChatGPT Canvas 기반으로 Fig. 1의 6단계 워크플로우를 가시화한 개념적 모델을 구현하여 인터뷰 대상과 동일집단인 교사(전문가) 4명이 직접 시연해 보고 피드백하는 방식으로 진행되었다. 본 시연은 시스템의 성능을 평가하는 것이 아니라 설계의 논리적 흐름과 현장 요구사항의 누락 여부를 검토하는 것을 목적으로 하였다.

Figure 2. Framework draft prototype

Fig. 2는 ChatGPT Canvas 환경에서 Fig. 1의 6단계 프레임워크를 노트 기반으로 구현한 프레임워크 초안 프로토타입의 모습이다. 교사가 좌측에 성취기준 코드와 난이도를 입력하면 프레임워크 초안 기반으로 정의된 워크플로우에 따라 RAG 모듈이 PDF에서 근거를 검색하고, 최종 문항을 생성하여 우측에 제시한다.

프레임워크 초안의 시연 결과 6단계의 기본 구조는 논리적이며 타당하다는 긍정적 평가를 받았다. 또한 개방형 설문문을 통해 프레임워크 초안을 개선하기 위한 6가지 핵심 기능적 요구사항이 도출되었다.

첫째, 난이도 간 수준 차이가 불명확하며, 문항유형이 출제도를 반영하기 어려워 관련 데이터셋의 기준을 정교화할 필요성이 있다.

둘째, 데이터셋 자동 생성 및 검증 기능의 강화가 필요하며, 이는 데이터셋 A-D의 자동화와 오류 검증 절차를 통해 교사의 데이터 구축 부담을 경감시키는 방향으로 개선되어야 한다.

셋째, 출처 기반 문항근거 제시 기능이 강화되어야 하며, 교과서와 NCS 문서의 페이지 정보를 자동으로 표기하여 문항의 신뢰성과 투명성을 확보할 수 있다.

넷째, 피드백 기반 재생성 기능의 고도화가 요구되며, 교사의 수정 및 피드백이 실시간으로 반영되어 문항을 재생성할 수 있는 구조가 필요하다.

다섯째, PDF 추출 정확도 향상을 위해 문서 구조 인식, 요약 보기, 오류 수정 기능을 보완해야 한다.

마지막으로, 자동 검증 및 매뉴얼 지원 기능을 강화하여 교사용 매뉴얼 제공과 함께 오류 검증 절차를 자동화함으로써 프레임워크의 신뢰성과 사용성을 높일 필요가 있다.

이 결과는 프레임워크가 기술적 자동화를 넘어 교사의 개입과 설명 가능한 구조를 지닌 교육과정 기반 시스템으로 발전해야 함을 시사한다. 프레임워크 초안의 시연 결과 도출된 개선사항의 내용은 Table 6과 같다.

Table 6. Summary of Framework Draft Improvements

Improvements	Major changes	Expected effect
Difficulty, Question Type Revised Standards	Proposal to modify the difficulty level and question types to suit the characteristics of specialized subjects	Improvement of difficulty classification and job reflection
Automatically generate datasets	Automatically Proposing Dataset A~D Schemas Based on PDF Structure Recognition	Reducing the burden of building teacher data
LLM Rule Interpretation Monitoring	Added Chain-of-Thought log tracing feature	Ensuring consistency in question creation
Automatic mapping of evidence	Automatically generate evidence_map and display it on the page	Strengthening source reliability
Feedback-based regeneration	Immediate regeneration based on teacher feedback	Continuous improvement of question quality
Multi-format storage	Supports export to XLSX, DOCX, PDF, etc.	Improving the practicality of teaching and assessment

전문가 의견 일치도를 정량 지표로 산출한 결과는 Table 7과 같다. 내적 합치도 계수 Cronbach's α 는 -0.07, ICC(2,1)은 -0.02, Kendall's W는 0.17($p < .001$)이었다. 표준 신뢰도 계수가 낮게 나타난 것은 평가 점수의 분산이 극히 작고(대부분 4-5점에 집중), 항목 간 변별이 거의 없었던 데에 기인한다. 절댓값은 낮게 산출되었으나, 이는 평가 점수가 상위 구간(4-5점)에 집중하여 분산이 작았기 때문으로, 낮은 변동성 하에서의 수렴 경향으로 해석할 수 있다. 실제 합의 수준을 보조적으로 확인하기 위해 '전문가 4인의 점수가 완전히 동일한 항목'의 비율과 '최대-최소 차이가 1점 이내인 항목'의 비율을 산출한 결과, 각각 60.3%, 80.8%로 나타났다. 즉, 통상적 신뢰도 지수는 낮게 산출되었으나, 실무적 관점의 합의 지표(정확 합치: ± 1 합치)로 보면 전문가 판단이 높은 수준으로 수렴했음을 확인할 수 있다.

Table 7. Expert consensus

characteristic	value	analysis
Cronbach's α	-0.07	Low internal consistency (due to insufficient variance and ceiling effect)
ICC(2,1)	-0.02	Low inter-rater consistency index (affected by insufficient variance)
Kendall's W	0.17	$p < .001$ (low level of agreement)
Accurate match rate	60.3%	Percentage of items where 4 people scored exactly the same
± 1 point agreement rate	80.8%	Proportion of items where the difference between the maximum and minimum scores of the four people is less than 1 point

4.4 데이터셋 정교화

프레임워크 초안에 대한 전문가 검증 결과, 난이도별 수

준 차이가 뚜렷하지 않다는 점과 문항유형의 직관성 및 현장성이 부족하다는 점이 핵심 한계로 지적되었다. 이를 보완하기 위해 교육평가 이론과 NCS 직무 특성을 반영하여 데이터셋 B(난이도 기준)와 C(문항유형)의 구조를 정교화하였다.

먼저 데이터셋 B(난이도 기준)의 정교화는 인지적 요구 수준과 사고의 깊이를 이론적으로 구분하는 데 초점을 두었다. 이를 위해 인지적 요구 수준은 Bloom의 신교육목표 분류학, 사고의 깊이는 DOK 모형을 준거로 삼아 다음과 같이 구체화하였다.

첫째, Bloom의 신교육목표 분류학을 기반으로 인지 과정의 위계를 명확히 하였다. 김성재 외[20]는 2015 및 2022 개정 교육과정의 성취기준을 분석하며, 기억·이해→적용·분석→평가·창안의 위계가 학습자의 인지적 요구 수준을 세분화하는 기준으로 기능함을 밝혔다. 이를 반영하여 본 연구에서는 문항의 인지적 요구를 기억·이해(하), 적용·분석(중), 평가·창안(상)으로 구분하였다.

둘째, Webb의 DOK 모형[21]을 적용하여 과제 복잡성과 사고의 깊이를 난이도 설계에 반영하였다. 문현우 외[18]는 2022 개정 교육과정 성취기준의 데이터 리터러시 내용을 DOK 관점에서 분석한 결과, 대부분의 성취기준이 DOK 1-2 수준에 집중되어 있어 고차적 사고를 촉진하는 DOK 3 이상의 과제 개발이 필요하다고 지적하였다. 이에 따라 본 연구는 DOK 1(회상)을 하, DOK 2(기술·개념 이해)를 중, DOK 3(전략적 사고)를 상 난이도로 정의하고, 각 수준별로 요구되는 인지 복잡성을 명확히 구분하였다.

또한 정은선[22]은 고등학교 문학 교과서의 고전시가 학습활동을 DOK 기준으로 분석하며, 사고의 깊이를 반영한 학습활동이 학습자의 사고 확장을 유도한다는 점을 확인하였다. 본 연구는 이러한 결과를 참고하여 문항 난이도 구분이 단순 지식 수준이 아니라 사고의 깊이와 과제 복잡성에 근거하도록 설계하였다.

다음으로 데이터셋 C(문항유형)은 직업계고 전문교과의 직무 맥락을 반영하도록 재구조화하였다. 전문가 피드백을 토대로 성태제 등[23]이 제시한 문항 형식을 참고하여 다음 네 가지 유형을 신설하고 정교화하였다.

- 개념 이해 확인형: 핵심 개념과 원리의 이해를 평가한다.
- 절차 확인형: 직무 수행 절차와 방법의 이해를 평가한다.
- 상황 판단형: 실제 현장 시나리오 속에서 최적의 판단을 선택하도록 하여 응용력을 평가한다.

복합 수행형: 도면, 데이터 등 복수 정보를 분석·종합하여 DOK 3 수준의 고차적 사고를 요구한다.

또한 국내 AIG 연구의 문항 모델을 참고하여 문항의 변별력을 높이기 위한 오답 생성 전략을 데이터셋 C에 명시적으로 추가하였다[10, 11].

- 개념적 오류: 학습자가 흔히 오해하는 개념을 반영한 오답
- 절차적 오류: 순서 혼동, 핵심 단계 누락 등 절차상의 오류
- 비효율적 선택지: 상황 판단형 문항에서 ‘차선책’을 매력적인 오답으로 제시하여 판단력 수준을 변별

마지막으로 정교화된 데이터셋 B와 C를 통합하여 난이도

수준(DOK 1-3)-문항유형-오답전략 간의 연계 구조를 확립하였다. 이를 통해 생성 문항은 성취기준의 인지적 요구 수준과 실제 직무 상황의 복잡성을 동시에 반영하도록 하였으며, 교육적 타당성과 실무적 타당성을 겸비한 AIG 프레임워크의 데이터셋 구조를 Table 8과 같이 완성하였다.

Table 8. Dataset refinement results

Difficulty (DOK)	Key question types (Dataset C)	Key Features (Dataset B)	Distractor strategy
lower (DOK 1)	Concept Understanding Confirmation Type	Do you remember and understand key terms, facts, and principles?	conceptual error
middle (DOK 2)	Procedural verification type	Can you follow and apply work procedures?	procedural error
upper (DOK 3)	Situational judgment type Complex performance type	Can you analyze problems and make optimal decisions in realistic scenarios?	Inefficient/ Suboptimal Complex Errors

4.5 최종 프레임워크 설계

이상의 요구사항 분석(4.1절), 초기 설계(4.2절), 전문가 피드백(4.3절), 데이터셋 정교화 및 구조 보완(4.4절)을 거쳐 Fig. 3과 같은 최종 AIG 프레임워크를 완성하였다. 본 프레임워크는 직업계고 전문교과의 성취기준을 기반으로 하며, 교사의 전문적 판단이 AI의 자동화 과정에 순환적으로 개입하는 설명 가능한 반자동(Explainable Human-in-the-loop) 구조로 설계되었다.

프레임워크는 외부 루프(outer loop)와 내부 루프(inner loop)가 결합된 이중 순환 구조를 가진다. 외부 루프는 문항이 자동으로 생성되는 전체 절차를 담당하며, 내부 루프는 교사의 피드백을 반영하여 문항 품질을 개선하는 순환 구조로 작동한다.

먼저, 외부 루프는 다섯 개의 주요 모듈로 구성된다.

입력(Input) 단계에서는 교사가 성취기준, 난이도, 문항 유형, 교과서(PDF) 등 원본 자료를 입력한다. 시스템은 입력된 자료를 분석하여 데이터셋 A-D의 스키마를 자동 생성하고, 4.4절에서 제시한 정교화된 DOK/역량 기반 난이도(B)와 직무유형(C) 규칙을 문항생성 규칙으로 적용한다.

분석(Requirement Analysis) 단계에서는 입력된 메타 데이터를 기반으로 성취기준-행동동사-국가직무능력표준(NCS) 능력단위를 매핑하고, 난이도 및 문항유형 규칙을 해석한다. 또한 LLM의 규칙 적용 과정을 추적하기 위해 CoT 모니터링 기능이 작동한다.

정보검색(Information Retrieval) 단계에서는 분석된 결과에 따라 교과서와 NCS 문서에서 근거 문장을 탐색한다. 이때 RAG 메커니즘을 적용하여 관련 근거를 검색하고, 문항별 근거 문서와 페이지를 evidence_map 형태로 자동 매핑한다.

문항생성(Question Generation) 단계에서는 RAG 기반으로 확보된 근거 문장과 규칙(B, C)을 결합하여 문두(stem), 선지(option), 정답·오답 해설, 근거(출처)를 포함한 문항을 자동 생성한다.

저장(Storage) 단계에서는 생성된 문항을 XLSX, DOCX, PDF 등 다중 포맷으로 저장하고, 루브릭 평가 결과와 메타 데이터를 함께 보존하여 향후 학습 및 품질 개선에 활용할 수 있도록 한다.

한편, 내부 루프는 교사의 피드백을 중심으로 문항 품질을 지속적으로 개선하는 구조로 작동한다. 외부 루프의 ‘저장’ 단계 이후, 교사는 생성된 문항을 검토하고 교사 점검 및 피드백을 수행한다. 이 단계에서는 문항의 성취기준 적합성, 난이도 적절성, 오답 타당성, 실무 반영도 등을 평가하고 수정 의견을 제시한다. 이 피드백은 문항 재생성(Regeneration) 과정으로 이어지며, 교사의 평가 결과가 새로운 입력 조건으로 반영되어 외부 루프의 분석 단계로 다시 환류된다. 즉, 교사 피드백은 단순한 수정 지시가 아니라, AI의 문항생성 규칙과 데이터셋 해석 기준을 갱신하는 역할을 수행한다.

이러한 이중 순환 구조는 AI의 자동화 효율성과 교사의 교육적 판단이 순환적으로 결합된 협력적 설계를 구현한다. AI는 RAG 기반의 데이터 중심 추론을 담당하고, 교사는 HITL을 통해 문항 품질과 교육적 타당성을 검증·보완한다. 완성된 최종 프레임워크의 구조는 Fig. 3과 같다. 결과적으로 본 프레임워크는 데이터셋 기반 규칙성, 출처 기반 신뢰성, 교사 피드백 순환 구조를 통합함으로써 문항의 신뢰성과 교육적 타당성을 동시에 확보하였다.

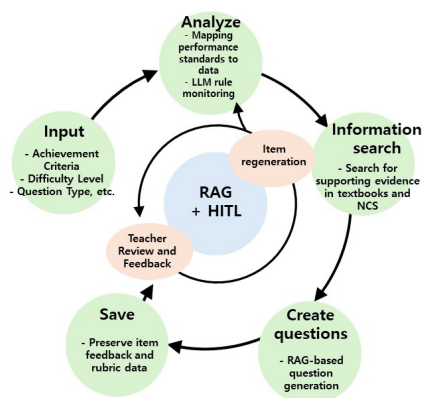


Figure 3. Final Framework

5. 결론 및 제언

5.1 연구 요약

본 연구는 직업계고 전문교과의 성취기준을 기반으로 한 AIG 프레임워크를 설계하고, 그 구조적 타당성과 기능적 요

구사항을 체계적으로 도출하여 개선된 프레임워크를 제시하는 것을 목적으로 한다.

연구는 DSRM에 따라 핵심 요소 도출, 데이터셋 구축, 프로토타입 시연, 전문가 검증의 단계를 거쳐 수행되었다. ChatGPT Canvas 기반의 시연을 통해 프레임워크의 논리 구조와 작동 절차를 검증하였으며, 전문가 피드백을 반영하여 개선 및 정교화를 진행하였다.

최종 프레임워크는 외부 루프(입력-분석-정보검색-문항생성-저장)와 내부 루프(교사 점검 및 피드백-문항 재생성)로 구성된 이중 순환 구조를 갖는다. AI는 RAG 기반의 자동화 기능을 수행하고, 교사는 HITL 피드백을 통해 문항의 품질과 타당성을 검증·보완한다.

이러한 구조를 통해 프레임워크는 데이터셋 자동화, 출처 기반 문항근거 제시, 교사 피드백 환류 기능을 통합한 설명 가능한 반자동 문항생성 시스템으로 완성되었다.

5.2 연구의 의의와 발전 방향

본 연구의 주요 의의는 다음과 같다. 본 연구는 성취기준-NCS-교과서의 삼중 연계 구조를 프레임워크 설계에 내재화함으로써 교육과정의 정합성을 확보하고, 평가의 교육적 타당성과 실무성을 동시에 강화하였다. 또한 AIG의 근거(evidence_map)와 정답·오답 해설을 필수 요소로 포함하여 XAI 기반의 교육평가 모델 구현 가능성을 제시하였다. 아울러 교사의 피드백과 교육적 판단을 반영하는 HITL 설계를 통해 AI와 교사가 협력적으로 작동하는 평가 구조를 정립함으로써, 자동화 중심의 평가를 보완하고 평가의 신뢰성과 현장 활용성을 높일 수 있는 구조적 대안을 제시하였다. 이러한 설계는 성취기준-NCS-교과서 간 연계를 문항생성 절차에 통합함으로써 직업계고의 기초 및 실무 교과 간 학습 연계성을 강화하고, 평가의 일관성과 품질을 동시에 확보할 수 있는 기반을 마련하였다. 나아가 AI의 자동화 기능과 교사의 전문적 판단을 결합한 협력적 평가 모델을 제시함으로써, AI가 교사를 대체하기보다 교육적 판단을 보조하는 방향으로 활용될 수 있음을 보여주었다.

향후 연구에서는 본 설계안을 실제 교과에 적용하여 문항 품질, 학습자 반응, 신뢰도 등을 포함한 실증적 검증 연구로 확장할 필요가 있다. 아울러 본 연구는 설계과학연구방법론에 기반한 탐색적 설계 연구로서, 소수의 고경력 전문가를 중심으로 프레임워크의 구조적 타당성과 현장 적합성을 검토하였다는 한계를 지니므로, 향후 연구에서는 보다 다양한 교과 및 지역의 교사를 포함한 확장된 표본을 통해 전문가 검증을 보완할 필요가 있을 것이다. 더불어 다양한 전문 교과에 본 프레임워크를 적용하여 교과 간 일반화 가능성을 검증하고, 학교 단위의 문항은행 시스템과 연계함으로써 AI 기반 평가 생태계(Educational AI Ecosystem)로 발전시킬 수 있을 것이다. 결국 본 연구의 프레임워크는 교육과정 정합성, 출처 기반 신뢰성, 교사 피드백 순환 구조를 통합한 설명 가능한 AI 모델로서, 직업계고 교육 현장에서 문항 개발의 효율성과 평가의 공정성을 동시에 향상시킬 수 있는 구

조적 대안을 제시하였다. 이는 AIG 기술이 교육 현장의 맥락적 판단과 교사의 전문성을 결합하여 데이터 기반의 질 높은 평가 문화를 구축할 수 있음을 보여주는 실질적 성과라 할 수 있다.

참고문헌

- [1] Korea Research Institute for Vocational Education and Training. (2024). *Case Studies of Guaranteed Minimum Achievement Levels for Vocational High School Credit Systems*(Research Report 2024-02). Korea Research Institute for Vocational Education and Training. <https://www.krivet.re.kr/kor/sub.do?menuSn=12&pstNo=E120240111>
- [2] Lee, S., Kim, N., & Lee, S. (2023). *Guidebook for specialized subject evaluation activities according to the 2022 revised vocational high school curriculum* (Research Report 2023-37). Korea Research Institute for Vocational Education and Training. <https://www.krivet.re.kr/kor/sub.do?menuSn=12&pstNo=E120240061>
- [3] Woo, W. (2019). A survey on the perception of subject teachers regarding the introduction of the high school credit system in vocational high schools. *The Journal of Korean Institute of Industrial Educators*, 44(2), 65-85. <http://dx-doi-org.proxy.knue.ac.kr/10.35140/kiiedu.2019.44.2.65>
- [4] Kim, N., Lee, S., Park, M., & Lee, S. (2024). Validation of the concept and constituent domains of professionalism for vocational high school specialized subject teachers. *Journal of Vocational Education & Training*, 43(4), 23-48. <http://doi.org/10.37210/JVER.2024.43.4.23>
- [5] Song, Y., Du, J., & Zheng, Q. (2025). Automatic item generation for educational assessments: A systematic literature review. *Interactive Learning Environments*, 33(9), 5386-5405. <https://doi.org/10.1080/10494820.2025.2482588>
- [6] Circi, R., Hicks, J., & Sikali, E. (2023). Automatic item generation: foundations and machine learning-based approaches for assessments. *Frontiers in Education*, 8, Article 858273. <https://doi.org/10.3389/feduc.2023.858273>
- [7] Tan, B., Armoush, N., Mazzullo, E., Bulut, O., & Gierl, M. (2024). A review of automatic item generation techniques leveraging large language models. *International Journal of Assessment Tools in Education*, 12(2), 317-340. <https://doi.org/10.21449/ijate.1602294>
- [8] Lee, P., Son, M., & Jia, Z. (2025). AI-powered Automatic Item Generation for Psychological Tests: A Conceptual Framework for an LLM-based Multi-Agent AIG System. *Journal of Business and Psychology*, 1-29. <https://doi.org/10.1007/s10869-025-10067-y>
- [9] Kim, S., Yeom, T., & So, K. (2025). A critical study on the discourse of teacher professionalism in high school credit system policy documents. *Korean Teacher Education Research*, 42(1), 463-490. <http://doi.org/10.24211/tjkte.2025.42.1.463>
- [10] Park, G., & Choi, S. (2025). A study on automatic item generation for multiple-choice reading assessment in Korean language education. *Journal of Curriculum and Evaluation*, 28(1), 215-246. <http://doi.org/10.29221/jce.2025.28.1.215>
- [11] Moon, H., Son, J., Lee, S., & Lee, Y. (2024). An analysis of the Depth of Knowledge (DoK) of data literacy reflected in the 2022 revised elementary school curriculum achievement standards. *Journal of The Korean Association of Computer Education*, 27(5), 89-100. <https://doi.org/10.32431/kace.2024.27.5.009>
- [12] Oh, G. (2022). A preliminary study on the development of reading assessment using automatic item generation techniques: Focusing on factual reading item models. *Cheongnam Korean Language Education*, 87, 7-34. <http://doi.org/10.26589/jockle.87.202205.7>
- [13] Lee, H., & Lee, Y. (2024). Exploring the possibility of automatic item generation in Korean reading assessment. *Journal of Education and Culture*, 30(3), 659-686. <http://doi.org/10.24159/joec.2024.30.3.659>
- [14] Shin, D. (2024). Exploring the applicability of an LLM-based customized automatic item generator: Focusing on the Poe AI chatbot builder. *English Subject Education*, 23(2), 181-206. <http://doi.org/10.18649/jkees.2024.23.2.181>
- [15] Bae, J. (2023). A study on the AI Act proposal and the basic principles of Explainable AI (XAI). *Journal of the Korea Institute of Knowledge Information Technology*, 18(2), 439-448. <http://doi.org/10.34163/jkits.2023.18.2.017>
- [16] Han, G., Park, Y., Jang, Y., & Kim, H. (2024). Evaluating the evidence generation capability of large language models for Explainable AI. *Proceedings of the KIISE Conference*, 510-512. <https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE11861895>
- [17] Park, H. (2025). AI-based assessment: Theoretical foundations, concrete models, and classroom application strategies. *Journal of Educational Comprehensive Research*, 23(1), 81-111. <http://dx-doi-org.proxy.knue.ac.kr/10.31352/JER.23.1.81>
- [18] Song, M., Noh, E., & Seong, G. (2016). Analysis of the accuracy of a sentence-level automatic scoring program for scoring short-answer items in large-scale assessments. *Journal of Curriculum and Evaluation*, 19(1), 255-274. <https://doi.org/10.29221/jce.2016.19.1.255>
- [19] Ahn, H. (2024). A framework for building a customized art museum guide system using LLM. *Journal of Arts Management*, (70), 5-34. <https://doi.org/10.52564/JAMP.2024.70.5>
- [20] Kim, S., Lee, S., & Choi, H. (2024). A comparative study of middle and high school information subject achievement standards in the 2015 and 2022 revised curricula based on Bloom's new taxonomy of educational objectives. *Journal of The Korean Association of Computer Education*, 27(8), 85-97. <https://doi.org/10.32431/kace.2024.27.8.007>

- [21] Webb, N. (1997). Determining Alignment of Expectations and Assessments in Mathematics and Science Education. *Nise Brief*, 1(2), 1-8.
- [22] Jeong, E., & Kim, B. (2023). An analysis of classical poetry learning activities in high school "Literature" textbooks from the perspective of Depth of Knowledge. *Korean Language Education Research*. 52(-), 351-397. <http://dx-doi-org.proxy.knue.ac.kr/10.17313/jkorle.2023..52.351>
- [23] Seong, T. (2019). *Contemporary educational assessment*(5th ed.). Hakjisa.



곽윤미

- 2006년 충북대학교 농공학과(공학사)
- 2022년 한국교원대학교 공업교육전공 (교육학석사)
- 2026년 한국교원대학교 교육과정 및 교육평가전공 (박사과정)
- 2008년~현재 청주농업고등학교 교사

✚ 관심분야 : 전문교과교육, 교육과정평가, 자동문항 생성

✉ ym101092@hamail.net



이시훈

- 2011년 청주교육대학교 컴퓨터교육학과(교육학사)
- 2017년 청주교육대학교 로봇교육전공 (교육학석사)
- 2024년 충북대학교 빅데이터 (공학박사)
- 2024~2025년 청주교육대학교 인공지능로봇교육 연구소 전임연구원
- 2026년~현재 한국교원대학교 뇌AI기반교육대학원 강사

✚ 관심분야 : 인공지능, 빅데이터, 정보교육, 자연어처리, 챗봇

✉ shoon1984@gmail.com