https://dx.doi.org/10.20880/kler.2025.60.4.129

GenAI-HITL 기반 ‘독서와 작문’ 연계 서술형 평가 과제 개발 및 타당성 검토

박고운 한국교원대학교 국어교육 박사과정

I. 서론

미래 사회는 예측하기 어려운 문제 상황에 창의적으로 대응할 수 있는 고차적 사고 능력을 갖춘 인재를 요구하고 있다(World Economic Forum, 2024). 이에 따라 2022 개정 교육과정을 비롯한 국내외 교육계는 단순한 지식 암기를 넘어 분석, 추론, 비판적 사고 등 복합적 사고 능력의 함양을 핵심 목표로 삼고 있다(교육부, 2022). 이러한 교육 목표의 전환은 학습자의 성장과 사고 과정을 타당하게 측정할 수 있는 평가 방식의 혁신을 필연적으로 요구한다.

기존의 선다형 문항 중심 평가는 학습자의 사고를 제한된 선택지 내에서 측정한다는 점에서, 복합적 사고 과정을 총체적으로 평가하기에는 한계가 있다(김경희, 2020; 박도순, 2025). 이러한 문제의식 속에서 교육과정 문서와 학계는 서술형 평가와 수행평가의 비중을 확대할 필요성을 꾸준히 강조해 왔다(교육부, 1992, 2022; 김선·반재천, 2023; 곽선영, 2025). 그러나 정책적 담론에서 '서술형·논술형 평가'라는 용어가 혼용되면서, 두 평가 유형의 개념적 경계가 모호하게 인식되는 경향이 있다. 학문적으로는 서술형 평가

는 제한된 분량과 명시적 조건 속에서 답을 구성하는 '제한 반응형(restricted response essay)'에 가깝고(Kubiszyn & Borich, 2013; 성태제, 2019), 논술형 평가는 논리적·창의적 사고를 확장적으로 조직하여 표현하는 '확장 반응형(extended response essay)'에 해당한다(McMillan, 2014; Miller, Linn, & Gronlund, 2013). 따라서 두 평가 유형은 목적과 설계 원리가 상이하므로 문항 개발 접근도 달라야 한다.

서술형 평가의 타당성은 지문과 문항의 유기적 연계에 의해 결정된다(김선·반재천, 2023; 박종임, 2024). 지문은 학습자의 사고를 촉발하는 맥락을 제공하고, 문항은 그 맥락 속에서 사고를 구체적으로 구조화하도록 요구한다(박혜영·김성숙·김경희·이명진·김광규·김지영, 2019). 그러나 교육과정 성취기준에 정합적인 지문을 선정하고, 이를 기반으로 서술형 문항과 채점 기준을 통합적으로 설계하는 과정은 교사의 전문성과 시간적 부담을 크게 요구한다(권태현, 2021; 김형성, 2023; 함은혜·박소영·이병윤·김기동·이대형, 2024). 이러한 제약 속에서 생성형 인공지능(Generative AI, GenAI)이 새로운 대안으로 주목받고 있다. 초기 연구가 자동 채점에 초점을 두었다면(박종임·이상하·송민호·이문복·이민정·최숙기, 2022), 최근에는 지문과 문항을 통합적으로 생성해 평가도구의 일관성과 정합성을 확보하려는 시도가 활발히 이루어지고 있다(최숙기·박종임, 2024; Memarian & Doleck, 2024; Shah, 2024).

최근의 GenAI-인간 협력(Human-in-the-Loop, HITL) 기반(GenAI-HITL), 관련 연구들은 AI가 단독으로 평가 문항을 생성하는 것이 아니라, 인간 전문가의 판단과 상호작용을 통해 평가의 정합성과 교육적 타당성을 높이는 방향으로 발전하고 있다(박고운·최숙기, 2025ㄴ; Eager & Brunton, 2023; Zanzotto, 2019). 이러한 연구들은 AI의 계산적 효율성과 교사의 평가 전문성을 결합하여, 평가 문항 개발의 품질과 효율성을 동시에 확보하고자 하는 시도로 볼 수 있다. 본 연구는 이러한 일반적 GenAI-HITL 접근을 토대로 하되, 서술형 평가의 핵심 요건—사고의 단계성, 조건의 명료성,

지문-문항 정합성—을 중심으로 한 설계 절차를 적용하여 서술형 과제를 생성하고, 그 결과물의 교육적 질과 활용 가능성을 전문가 평정에 기반해 검토하였다. 특히 기존 연구들이 선다형 문항이나 단일 발문 수준의 자동생성에 초점을 맞춘 것과 달리, 본 연구는 서술형 평가의 특성에 맞춰 '지문-문항-루브릭-해설'이 통합된 구조를 재설계하였다. 이를 통해 GenAI가 교사의 평가 설계 전문성을 보완하는 협력적 시스템으로 기능할 수 있음을 탐색하고자 한다.

따라서 본 연구의 목적은 GenAI를 활용하여 2022 개정 국어과 '독서와 작문' 성취기준에 부합하는 고차적 사고력 평가용 서술형 문항을 지문과 함께 통합적으로 생성하고, 그 결과물의 교육적 타당성과 질적 수준을 전문가 평가를 통해 검토하는 데 있다. 이를 위해 다음과 같은 연구 문제를 설정하였다.

첫째, GenAI는 2022 개정 국어과 '독서와 작문' 성취기준에 부합하면서, 교육과정 정합성과 사고 단계의 체계성을 갖춘 서술형 과제를 지문과 함께 생성할 수 있는가?

둘째, 전문가 평정을 통해 확인된 GenAI-HITL 기반 서술형 과제의 질적 특성과 한계는 무엇이며, 이러한 평가 결과는 서술형 과제 자동생성의 교육적 타당성 확보에 어떤 시사점을 제공하는가?

본 연구는 기술적 구현이 아니라 교육적 타당성을 중심으로 GenAI-HITL 기반 과제 생성의 질적 수준과 교육적 활용 가능성에 대한 전문가 준거 기반의 예비적 타당화 근거를 탐색하였다. 이러한 접근을 통해 GenAI가 교사의 과제 설계 전문성을 보완하는 협력적 도구로 작동할 수 있음을 보여주고자 한다. 연구 결과는 서술형 평가의 질적 향상과 AI 기반 평가 설계의 실천적 토대를 제공함으로써, 향후 교육평가 패러다임의 전환 가능성을 탐색하는 데 기여할 수 있다.

II. 이론적 배경

1. 2022 개정 국어과 〈독서와 작문〉 교육과정과 서술형 평가

서술형 평가는 단순히 정답을 서술하는 행위를 넘어, 학습자가 주어진 텍스트와 상호작용하며 의미를 재구성하는 고차적 문식 실행(literacy practice)의 과정을 가시화하는 평가 방식이다(김경희, 2020; 성태제, 2019). 이러한 평가의 핵심은 지문이 제공하는 풍부한 맥락과 문항이 요구하는 인지적 활동의 유기적 통합에 있다(김선·반재천, 2023; 박종임, 2024). 지문은 사고를 촉발하고 방향을 제시하는 '인지적 비계(cognitive scaffold)' 역할을 수행하며, 문항은 이 비계를 활용하여 분석, 추론, 비판, 재구성 등 복합적인 사고 기능을 수행하도록 유도한다(박종임, 2024; 송슬기; 2024, 최숙기, 2023).

이러한 평가 방식의 지향점은 국제적 평가 담론에서도 명확히 확인된다. 특히 독일의 아비투어(Abitur)는 자료 기반 서술형 평가의 대표적 사례로, 단순한 텍스트 이해를 넘어 주어진 자료를 바탕으로 '재현·이해(reproduction)', '분석·적용(analysis/application)', '비판·평가(evaluation/judgment)'라는 위계적 요구 영역(Anforderungsbereiche)을 통합적으로 수행하도록 요구한다(Kultusministerkonferenz, 2002). 이는 학습자가 하나의 과제 안에서 텍스트의 표면적 정보를 이해하는 것부터 시작하여, 이를 분석하고 새로운 상황에 적용하며, 궁극적으로는 자신만의 비판적 관점을 형성하는 전 과정을 총체적으로 평가하려는 의도를 담고 있다(장성민, 2021, 2024).

이러한 평가 패러다임은 2022 개정 국어과 〈독서와 작문〉 과목의 교육목표와 일치한다(최숙기, 2021; 최숙기·박종임, 2023). 〈독서와 작문〉은 읽기와 쓰기를 분리된 기능이 아닌, 공동의 인지적 자원을 활용하는 통합적 '문어 의사소통(written communication)' 행위로 규정한다(최숙기·박종임, 2023; Fitzgerald & Shanahan, 2000). 이는 학습자가 텍스트를 수동적으로

수용하는 독자에서 벗어나, 텍스트를 비판적으로 분석하고 자신의 지식과 결합하여 새로운 의미를 창출하는 '의미 구성의 주체'가 되어야 함을 의미한다(박혜영 외, 2019; 정민주·서수현·남민우·최숙기·이상일·남가영, 2022).

따라서 〈독서와 작문〉의 평가는 '읽기 따로, 쓰기 따로'의 분절적 방식을 지양하고, '읽고-이해하여-재구성하고-표현하는 통합적 사고 과정'을 온전히 담아낼 수 있는 '지문-문항 통합형' 과제를 필수적으로 요구한다(남민우·이상일·최숙기·서수현·남가영·정민주, 2022; 최숙기, 2023). 이는 Messick(1989)의 통합적 타당도 관점에서도 지지된다. 잘 설계된 지문-문항 통합 과제는 평가의 내용(content)이 측정하려는 고차적 사고 능력(construct)과 긴밀히 연계되도록 보장하며, 이를 통해 긍정적인 교육적 결과(consequence)를 유도하는 가장 타당한 평가 방식이라 할 수 있다(김경희, 2020). 본 연구에서 GenAI를 통해 지문과 문항을 함께 생성하려는 시도는 이러한 교육과정의 요구와 평가학적 타당성을 동시에 만족시키기 위함이다. 이를 위해서는 AI의 생성 메커니즘에 대한 이해가 선행되어야 한다.

2. GenAI 기반 서술형 과제 생성의 메커니즘

본 연구에서 제안하는 GenAI 기반 서술형 과제 개발 프로토콜은 최신 자동 문항 생성(Automatic Item Generation, AIG) 연구의 이론적 토대 위에 구축된다(Attali, Runge, LaFlair, Yancey, Goodwin, Park et al., 2022; Circi, Hicks, & Sikali, 2023). 이는 단순한 기술의 조합을 넘어, 교육적 의도를 AI의 연산 과정에 체계적으로 반영하기 위한 접근이다(Qian, 2025).

본 연구에서 지향하는 지문-문항 통합 생성의 핵심은 컨텍스트 엔지니어링(context engineering)이라기보다, 교육과정 성취기준을 평가 설계의 준거로서 정확히 '고정'하고 이를 AI가 일관되게 참조하도록 입력 조건을 체계화하는 컨텍스트 설계 절차에 있다(Shah, 2024; White, Fu, Hays, Sandborn, Olea, Gilbert et al., 2023). 이 단계에서 수행된 작업은 성취기준

을 임의의 다른 표현으로 변환하거나 재서술하는 것이 아니라, 성취기준 원문과 평가 설계에 필요한 조건(평가목표, 지문 장르·난이도, 산출물 구성요건 등)을 명시적으로 제시하여 생성 과정의 이탈을 최소화하는 데 목적이 있다(Bozkurt, 2024). 특히 본 연구는 성취기준의 요구를 보다 분명히 '위계화하여 인지'하도록, Bloom의 개정 분류 체계(Anderson & Krathwohl, 2001)와 Webb의 지식의 깊이(DOK) 준거(Webb, 2009)를 보조 정보로 함께 제공하였고, 이때 성취기준 해석의 적절성 및 위계 부여의 타당성은 인간 전문가가 확인·보정하는 HITL 절차를 통해 보완하였다.

복잡하고 다단계적인 '지문 생성 후 문항 생성' 과업을 안정적으로 수행하기 위해서는 AI의 추론 과정을 명시적으로 제어하는 기제가 필수적이다(Bozkurt, 2024; U.S. Department of Education, 2023). 이에 따라 사고연쇄(Chain of Thought, CoT)는 AI가 최종 결과물만 내놓는 '블랙박스' 모델의 한계를 극복하고, '① 성취기준 분석 → ② 지문 주제 및 구조 계획 → ③ 지문 생성 → ④ 생성된 지문 분석 및 평가 요소 추출 → ⑤ 문항 및 채점 기준 생성'과 같이 인간의 문제 해결 과정과 유사한 단계적 추론 경로를 따르도록 유도한다(박고운·최숙기, 2025ㄱ; Wei, Wang, Schuurmans, Bosma, Ichter, Xia et al., 2022; Wang, Xu, Lan, Hu, Lan, Lee et al., 2023). 이러한 추론의 외현화는 결과물의 예측 가능성을 높일 뿐만 아니라, 오류 발생 시 어느 단계에서 문제가 발생했는지 진단하고 수정하는 것을 용이하게 한다(Brown, 1987; Lightman, Kosaraju, Burda, Edwards, Bake, Lee et al., 2023).

국어과 평가에서 지문의 내용적 타당성과 사실적 정확성은 평가의 타당도와 신뢰도에 모두 영향을 줄 수 있다(김경희, 2020; 권태현, 2021). 검색 증강 생성(Retrieval-Augmented Generation, RAG) 및 웹 기반 그라운딩은 생성 모델이 내부 지식에만 의존할 때 발생할 수 있는 사실 오류(환각)를 완화하기 위해, 외부 문서 검색 결과를 생성 과정에 결합하는 접근으로 제안되어 왔다(Lewis, Perez, Piktus, Petroni, Karpukhin, Goyal et al., 2020; Ji, Lee, Frieske, Yu, Su, Xu et al., 2023). 본 연구에서는 ChatGPT의 웹 검색

기능과 Gemini의 Google Search grounding을 활용하여, 지문 생성 시 핵심 사실(인물·사건·수치·연도 등)에 대해 검색 근거를 확인하고 출처를 기록·대조하는 방식으로 지문의 사실성을 점검하였다. 다만 이러한 절차는 사실 오류 가능성을 낮추는 데 기여할 수 있으나, 검색 결과의 한계나 해석 오류 가능성을 완전히 제거하지는 못하므로, 최종 검토는 인간 전문가의 확인 절차를 통해 보완하였다.

AI가 생성한 초안의 품질을 담보하기 위해, 프로토콜은 AI의 내재적 피드백 루프(internal feedback loop)를 포함한다. 자기 정교화(self-refine)와 자기 일관성(self-consistency)과 같은 기법은 AI가 생성한 지문과 문항에 대해 스스로 비판적인 질문을 던지고('이 지문은 성취기준의 핵심 요소를 충분히 담고 있는가?', '문항이 요구하는 정보가 지문 내에 명확히 제시되어 있는가?'), 여러 추론 경로를 생성하여 다수결로 가장 일관된 답을 찾는 방식으로 결과물을 반복적으로 수정하도록 하는 메커니즘이다(Wang, Wei, Schuurmans, Le, Chi, Narang et al., 2023; Madaan, Tandon, Gupta, Hallinan, Gao, Wiegreffe et al., 2023). 또한, 검증 사슬(Chain of Verification, CoVe)은 생성된 결과물을 단계적으로 검토하며 결과물의 논리적 정합성과 완성도를 크게 향상시킨다(Dhuliawala, Komeili, Xu, Raileanu, Li, Celikyilmaz et al., 2024). 이는 인간의 초고 작성 및 퇴고 과정과 유사한 자기 성찰적 과정을 모사한다(Brown, 1987).

그러나 이러한 AI의 자기 검토 노력에도 불구하고, LLM은 학습 데이터에 내재된 편향을 재현하거나 사실과 다른 내용을 생성할 위험을 지닌다(Bender, Gebru, McMillan-Major, & Mitchell, 2021; Ji et al., 2023). 교육 평가 도구의 공정성과 신뢰도는 무엇보다 중요하므로(Kane, 2013; 박도순, 2025), AI 기반 평가의 교육적 타당성은 GenAI-HITL을 통해 강화·보완될 수 있으며, 특히 고위험(채점/의사결정) 맥락에서는 인간 전문가의 최종 검토가 핵심 조건으로 요청된다(Zanzotto, 2019; Memarian & Doleck, 2024; 박고운·최숙기, 2025ㄴ). AI는 구조적 측면에서 뛰어난 성능을 보이지만, 학습

자의 발달 단계 적합성, 미묘한 문화적 맥락, 잠재적 편향성 등 교육 전문가의 고차원적 판단이 요구되는 영역에서는 한계를 가진다(곽선영, 2025; Eager & Brunton, 2023; U.S. Department of Education, 2023). 따라서 본 프로토콜에서 인간 전문가는 AI 생성물을 단순히 검토하는 수동적 역할을 넘어, 각 생성 단계에 개입하여 피드백을 제공하고 AI의 추론 방향을 재설정하는 능동적 협력자로서 기능한다(Bozkurt, 2024; Shah, 2024). 이는 기술의 효율성과 인간의 전문성이 상호보완적으로 결합하여, 교육적으로 타당하고 현장 적용성 높은 평가 과제를 개발하는 최적의 경로이다(Circi et al., 2023).

III. 연구 방법

1. 연구 참여자

본 연구의 전문가 평가에는 전국 각지에서 근무하는 현직 국어 교사 18인(T1-T18)이 참여하였다. 참여자 패널의 구체적인 정보는 〈표 1〉과 같다.

〈표 1〉 전문가(평정 패널) 정보 개요

번호	성별	지역	교직 경력
T1-T7	여(5), 남(2)	경기(3), 전남(2), 제주(2)	10년 이상
T8-T13	여(4), 남(2)	서울(2), 인천, 충북, 강원, 부산	
T14-T18	여(3), 남(2)	세종, 충남, 울산, 광주, 대구	10년 미만
합계	여(12), 남(6)	전국 13개 시도	18명

전문가 집단은 여성 12인, 남성 6인으로 구성되었으며, 교직 경력 10년 이상의 숙련된 교사가 13인, 10년 미만의 교사가 5인 포함되었다. 특히, 참여

자들은 경기도, 전남, 제주를 비롯하여 서울, 인천, 충북, 강원, 부산, 세종, 충남, 울산, 광주, 대구 등 전국 각지에 분포하여 특정 교육 환경에 편중되지 않은 의견을 수렴하고자 하였다. 참여자 전원은 서술형 평가 과제 개발 경험을 보유하고 있었다. 이 중 5명(T1-T5)은 여러 교과목의 내용을 통합하거나 심층적인 사고를 요구하는 고난도 논술형 과제를 출제한 경험이 있었으며, 13명(T6-T18)은 교과 내용의 이해와 적용을 중심으로 하는 서술형 평가 과제를 출제한 경력이 있었다. 본 연구는 전국 13개 시도에 분포한 18명의 현직 국어교사를 대상으로 하여 지역적 다양성과 교과 전문성을 동시에 확보하였다.

2. 연구 절차 및 자료

본 연구는 GenAI-HITL 방식을 적용하여 서술형 과제 개발의 타당성과 신뢰성을 검토하는 것을 목적으로, 2024년 6월부터 8월까지 총 4단계의 절차를 수행하였다. 연구의 전체 흐름은 ① 기초 자료 수립 및 프로토콜 개발(1단계), ② AI 기반 과제 생성(2단계), ③ 전문가 평가(3단계), ④ 자료 분석 및 결과 도출(4단계)의 순차적 과정으로 구성된다. 특히 본 연구 절차의 핵심은 AI의 자동 생성 능력과 인간 전문가의 검토 능력을 결합하여, 교육과정 성취기준에 부합하면서도 실제 교육 현장에서 활용 가능한 고품질의 서술형 문항을 개발하는 데 있다. 단계별 구체적인 수행 내용과 자료는 다음과 같다.

3. 단계별 연구 절차 및 자료

1단계: 기초 자료 수립 및 프로토콜 개발

먼저 '교육과정 분석 및 기준 자료'를 위해 2022 개정 국어과 교육과정의 '독서와 작문' 영역 성취기준을 심층 분석하였다. 특히 고등학교 3학년 수준에 적합한 고차적 사고력, 비판적 읽기 능력, 자료 기반 논증 능력을 중점

적으로 다루는 성취기준을 선별하였다. 학생평가지원포털(2024)에서 제공하는 평가 도구 중 가장 유사한 성취기준을 다룬 서술형 대표 과제 1세트를 선정하여 AI 학습 데이터로 활용하였다. 2022 개정 교육과정에 특화된 평가 도구가 개발되지 않은 상황에서 가장 적합한 예시 자료를 확보한 것이다. 특히 서술형 과제에 필수적인 루브릭 개발과 해설서 작성 단계를 새롭게 추가하고, 선행 연구를 참고하여 '서술형 과제 개발 핵심 10 요소'[1] 종합한 후 이를 반영한 사전 지침 체계를 구축하였다(김선·반재천, 2023; 서울특별시교육청, 2022; 경기도교육청, 2024).

2단계: AI 기반 문항 생성 실행(2025년 7월 1일~12일)

1) 연구팀 구성

연구자와 AI 문항 생성 경험이 있으며 현재 고등학교 3학년을 대상으로 EBS 수능특강을 활용한 수업을 진행하고 있는 교사 2인을 섭외하여 생성에 참여할 연구팀을 구성하였다.

2) 지문 선정

사전 협의를 통해 EBS 수능특강 및 교과서 지문 등을 포함한 3가지 유형의 지문을 선정하였다. 이중 최종 지문으로는 『2026학년도 수능특강 국어 영역 독서』(한국교육방송공사, 2025)에 수록된 지문을 선정하였다. 물론 EBS 독서 지문이 교육과정의 수준과 범위를 상회하여 평가용 텍스트로서의 교육

1) 서술형 과제 개발 핵심 10 요소: ① 교육과정 정합성: 성취기준-평가요소-문항-루브릭 일치 ② 고차적 사고 촉진: 사실 → 추론 → 평가 → 창안 위계 설계 ③ 발문 명료성: 구체 동사·대상·범위 명시, 혼란 최소화 ④ 비계(보기) 설계: 원문 단락·수치·상반 관점 제공 ⑤ 조건 세분화: 내용·과정·형식 3층위로 구체 지시 ⑥ 평가 타당성·신뢰도: 세분 루브릭·교차 채점 검토 ⑦ 실제성·흥미성: 동시대 이슈·현장 자료 활용 ⑧ 메타인지 유도: 조건은 학습자에게 비계로 작용하도록 구성 ⑨ 공평성·가독성: 난도 조정·편향·어휘 수준 고려 ⑩ 학습 피드백 연계: 수준별 예상 답안 및 개선 포인트 제공

과정의 수준·범위와의 정합성 측면에서 우려가 제기될 수 있다. 그러나 본 연구는 이러한 우려에도 불구하고, 다음과 같은 논리적 근거에 따라 해당 지문이 본 연구의 목적에 가장 부합한다고 판단하였다.

첫째, 본 지문 선택은 교육과정 문서상의 수준을 '완전히 충족'하기 위한 조치라기보다, 고3 현장에서 실제로 활용되는 텍스트를 기반으로 현장 실행 맥락에서의 적용 가능성을 높이기 위한 도구적 선택이다(생태학적 타당성). 현재 고등학교 3학년 교육 현장에서 EBS 연계 교재는 사실상 '준(準) 교과서'로서 기능하며, 실제 교수·학습 활동의 핵심 기제로 작동하고 있다. 따라서 기계적인 교육과정 난이도를 준수한 텍스트보다 실제 학습자들이 학습하고 있는 '실행된 교육과정(implemented curriculum)' 상의 텍스트를 활용하는 것이 연구 결과의 현장 적용 가능성에 대한 예비적 근거를 강화할 수 있다고 판단하였다.

둘째, 고차적 사고력 측정을 위한 도구적 적합성이다. 본 연구가 개발하고자 하는 서술형 과제는 단순한 사실적 독해를 넘어, 텍스트의 정보를 비판적으로 통합하고 추론하는 고차적 사고 과정을 요구한다. 교과서 수준의 평이한 제재보다는, 복합적인 정보와 높은 논리적 밀도를 가진 텍스트가 학습자에게 의도된 '인지적 마찰(cognitive friction)'을 유발하여 심층적인 사고 과정을 가시화하는 데 효과적이다. 따라서 본 연구는 EBS 지문이 이러한 역량 중심 평가 모델을 탐색하기 위한 적절한 도구적 적합성(pragmatic fit)을 갖추었다고 보았다.

3) AI 모델 및 생성 조건

본 연구에서는 2025년 7월 기준 연구자가 접근 가능한 최신 대규모 언어 모델 중 GPT-o3 (OpenAI, 2025)와 Gemini 2.5 Pro (Google DeepMind, 2025)를 예비 비교한 뒤, 서술형 과제 생성 성능이 상대적으로 우수한 GPT-o3를 최종 분석 대상 모델로 선정하였다. 두 모델에는 동일한 지문, 프로토콜 및 사전 지침을 제공하여 서술형 과제 초안을 자동 생성하도록 하였

으며, 각 모델별로 최소 3일의 간격을 두고 평균 3회씩 세션을 실행하여 출력의 안정성을 점검하였다. 모든 모델 호출에는 〈표 2〉에 제시한 동일한 시스템 설정과 프롬프트 조건을 적용하였다.

〈표 2〉 AI 생성 조건

항목	내용
대상 학습자	고등학교 3학년
문항 구성	서술형 2개(각 200~300자 내외)
평가 중점 요소	고차적 사고력, 비판적 읽기 능력, 자료 기반 논증 능력
필수 포함 요소	발문, 조건, 루브릭, 수준별 예상 답안, 해설서

생성된 과제들을 비교 분석한 결과, Gemini 2.5 Pro와 GPT-o3 모두 우수한 성능을 보였으나, 최종적으로 선정된 GPT-o3는 다음과 같은 측면에서 본 연구의 목적에 더 부합하였다. 첫째, 체계적 요구사항 반영을 위해 사전에 제시된 '서술형 과제 개발 핵심 10요소'를 가장 충실히 반영하였다. 특히 '교육과정 정합성(성취기준-평가요소-문항-루브릭 일치)'과 '고차적 사고 촉진' 측면에서 체계적 구조를 보였다. 둘째, 추론 능력 기반 문맥 일관성 측면에서 GPT-o3의 추론 모델 특성을 활용하여 제시된 지문과 발문, 보기 간의 논리적 연계성이 자연스럽고, 200~300자 내외의 답안 분량에 적합한 사고 과정의 단계성을 적절히 설계하였다. 셋째, 실용적 완성도에서 발문 명료성, 조건 세분화, 루브릭 개발, 수준별 예상 답안 제공 등 실제 현장에서 활용 가능한 형태의 완성된 평가 도구를 일관되게 생성하였다. 반면 비교군은 일부 요소(루브릭 세분화, 메타인지 유도 장치 등)에서 누락이나 불완전한 결과를 보여 최종 분석에서 제외하였다.

3단계: 전문가 평가(2025년 7월 14일~26일)

18인의 전문가 패널을 대상으로 구글 설문지를 활용한 온라인 평가를

실시하였다. 평가 자료는 개별적으로 제공되었으며, 전문가들은 평균 2일에 걸쳐 다음과 같은 절차로 평가를 진행하였다. ① 평가 자료를 직접 인쇄하여 학생 입장에서 문제 해결 시도 ② 루브릭과 해설서를 기반으로 문항의 구조적 특성 점검 ③ 3영역 12항목 5점 리커트 척도를 활용한 양적 평가 완료 ④ 개별 의견서 작성(오류, 개선점, 구체적 견해 포함). 전체 평가 과정은 2025년 7월 14일에 시작되어 26일까지 완료되었다. 양적 평가 완료 후, 2025년 7월 26일까지 결과지에 따라 추가 비대면 개별 면담을 실시하여 세부적 견해를 수집하였다. 면담에서는 과제 난이도의 적절성, 채점 편의성, 사고 과정 설계의 타당성, 교수학습 연계 가능성 등을 중점적으로 탐색하였다.

4단계: 자료 분석 및 결과 도출(2025년 8월 1~2주)

전문가 평정 결과를 바탕으로 생성된 과제의 타당성과 신뢰성을 다각도로 분석하였다. 정량 분석을 통해 과제의 전반적 품질 수준을 수치화하고, 정성 분석을 통해 그 의미를 심층적으로 해석함으로써 GenAI 기반 과제 개발의 가능성과 한계를 종합적으로 평가하였다.

4. 과제 생성 프로토콜

본 연구는 GenAI-HITL에 기반한 과제 생성 절차(Madaan et al., 2023; Wang et al., 2023; 박고운·최숙기, 2025ㄴ)를 바탕으로, 서술형 평가의 특성에 맞게 절차적 정교화 및 교육적 확장을 시도하였다. 생성형 AI의 비결정적(nondeterministic) 특성을 고려하여, 본 연구는 결과물의 동일성보다 절차의 일관성과 재현 가능성에 중점을 두었다(Ganguli, Lovitt, Kernion,, Askell, Bai, Kadavath et al., 2022). 기존의 HITL 기반 연구들이 주로 AI의 자동 생성 효율성과 교사의 일차 검토 기능에 초점을 두었다면(Eager & Brunton, 2023; U.S. Department of Education, 2023), 본 연구는 서술형 문항의 사고 위계, 루브릭 기반 검토, 교육적 환류 가능성을 통합함으로써 GenAI-HITL

구조를 교육평가 맥락에 적합하게 재구성하였다. 전체 프로토콜은 [1단계: 기반 설정] → [2단계: 핵심 생성] → [3단계: 정교화 및 검토]의 세 단계로 구성되며, 각 단계는 AI의 생성 기능과 교사의 전문적 판단이 상호작용하는 순환적 협력 구조(cyclic HITL framework)로 설계되었다(〈표 3〉 참조).

〈표 3〉 서술형 과제 생성 프로토콜 개요

단계	핵심 기능	주요 투입 자료 및 맥락	HITL 개입 지점
1. 기반 설정	- AI 역할 및 페르소나 설정 - 교육과정 성취기준 및 성취 수준 분석 - 예시 문항 학습 및 평가 설계 전략 형성	- 평가 설계자 역할 프롬프트 - 2022 개정 국어과 성취기준 - Bloom/Webb 사고 수준 체계 - 예시 문항 분석	역할 수행 및 기준 해석 검토
2. 지문-문항 핵심 생성	- CoT 기반 지문 분석 및 새로운 지문 (다) 생성 - 발문·조건·보기 구조화 - 단계적 사고 위계 반영	- 원본 지문(EBS 수능특강 등) - 다중 페르소나 협업 규칙 - 발문 및 조건 설계 지침	지문 품질 및 문항 논리성 검토
3. 정교화 및 검토	- 인간 전문가 피드백 기반 문항 수정 - 루브릭 개발 및 평가 기준 설계 - 해설 개발	- 전문가 수정·보완지시 채점 기준표 구성 - 해설서 작성	최종 품질 판정 및 완성도 확보

1단계: 기반 설정

1단계는 GenAI가 이후 단계에서 생성할 지문·문항·루브릭이 교육과정 성취기준을 준거로 일관되게 구성되도록, 초기 입력 조건을 구성하고 검토하는 절차이다. 먼저 GenAI에 '국어과 평가 설계자' 역할을 부여하여, 산출이 단순 생성이 아니라 평가 설계 관점의 산출물로 조직되도록 하였다(Kharrufa & Johnson, 2024). 다음으로 성취기준(및 성취수준) 원문과 평가 조건을 제시하고, 각 성취기준의 요구 수준을 개정 Bloom 분류 체계(Anderson & Krathwohl, 2001)와 Webb의 DOK 준거(Webb, 2009)에 교차 매핑한 정보를 함께 제공하여, 모델이 성취기준의 인지 과정×지식 깊이 위계를

선행적으로 인지하도록 하였다. 또한 고품질 서술형 과제 사례를 예시 자료로 제공하여, 발문 구조·조건 진술·답안 유도 방식 등 형식적 요소를 참조할 수 있도록 하였다. 마지막으로 HITL 단계에서 인간 전문가가 성취기준 해석과 Bloom×DOK 매핑의 적절성을 점검하고, 필요 시 "추론·비판을 포함하도록 사고 요구를 조정하라"와 같은 구체적 수정 지침을 부여한 뒤 재생성을 반복함으로써 초기 생성 방향의 이탈을 보정하였다.

2단계: 지문–문항 핵심 생성

이 단계는 AI가 실제 과제의 중심 요소인 지문과 발문을 생성하는 과정으로, 사고 과정의 위계성과 논리적 연결성을 구현하는 것이 핵심이다. 우선 CoT 기반의 사고 연쇄 추론을 적용하여, 지문의 중심 논지와 학습자가 수행해야 할 사고 경로를 명시적으로 설계하였다(Wei et al., 2022). 기존의 선다형 문항 생성과 달리, 본 연구에서는 서술형 문항의 인지적 특성을 반영하여 '이해 → 추론 → 비판'의 3단계 사고 위계를 중심으로 문항 구조를 구성하였다. 또한 EBS 수능특강 등 실제 학습 텍스트를 기반으로 지문을 재구성함으로써, 평가 상황의 현실성과 학습 맥락의 적합성을 제고하고자 하였다. 발문 설계 과정에서는 지문 정보가 반드시 답안 구성에 활용되도록 조건을 설정하였으며, 다중 페르소나 기반 토론 규칙을 활용하여 논리적 비약이나 정보 왜곡이 최소화되도록 하였다. HITL은 AI가 제시한 초안에 대해 지문과 발문 간의 논리적 연계성, 사고 위계의 적절성, 조건 진술의 명확성을 점검하며, 필요한 경우 구체적인 수정 지침을 제시하였다. 예컨대 이해 수준 문항이 지문 외부 배경지식에 과도하게 의존하거나, 비판 수준 문항이 단순 의견 진술에 머무르는 경우, 인간 평가자는 "지문 내 단서 인용을 의무화하는 조건 추가", "비판 기준을 두 가지 이상 제시하도록 요구"와 같이 세부 피드백을 제공하고 이를 반영한 재생성을 수행하였다. 이 단계는 단순히 과제를 산출하는 과정이 아니라, AI의 사고 과정을 교사 피드백을 통해 정교화하는 상호작용적 학습의 단계로 기능한다.

3단계: 정교화 및 검토

최종 단계는 AI가 생성한 과제 초안을 교육 현장에서 활용 가능한 수준으로 다듬는 과정이다. 본 단계의 핵심은 인간 평가자의 질적 판단을 중심으로 과제의 내용 타당도와 실용성을 점검하는 것이다. 먼저 전문가 피드백을 반영하여 과제를 수정하고, 평가의 명확성을 높이기 위해 루브릭을 개발하였다. 루브릭은 성취기준에 근거한 세부 평가 요소(입장 제시, 근거 제시, 글의 구조, 표현 적절성 등)와 각 점수 수준의 수행 특성을 포함하도록 구성되었으며, 과제의 채점 가능성과 피드백 제공의 기반으로 활용되었다.

또한 본 단계에서 GenAI-HITL 구조는 두 가지 검토 루프를 중심으로 작동하였다. AI 피드백 루프는 자동화된 재생성·자기검토를 통해 과제의 논리적 일관성과 조건 충족 여부를 점검하는 단계이며, 교사 검토 루프는 교육과정 부합성과 학습자 적합성을 판단하여 과제의 교육적 타당성을 확보하는 단계이다. 예를 들어 교사 검토 루프에서는 실제 고3 수업 경험을 바탕으로 "지문과 발문의 연계성, 지문 조건과 성취기준 요소 간의 연결 정도, 루브릭이 수업 피드백에 활용될 만큼 구체적인지"를 점검하고, 필요 시 과제의 분량·언어 수준·평가 요소를 축소·조정하였다. 두 루프는 상호보완적으로 작동하며, AI의 자기검토 결과가 교사 검토의 근거로 활용되고, 교사의 피드백은 다시 AI의 재생성 과정으로 반영되어 과제의 완성도를 향상시킨다. 이러한 순환적 구조를 통해 AI와 인간 전문가 간의 협력적 검토 체계가 구축되며, 이는 단순한 기술적 반복이 아니라 '교육적 검토-자동화 검토-인간적 조정'이 병행되는 통합적 품질관리 과정으로 기능한다.

이후 루브릭에 따라 해설서를 작성하여, 학습자 피드백의 방향성과 평가 기준 간 정합성을 확인하였다. 이러한 루프 구조는 AI 결과물을 반복적으로 검토하고 개선하는 순환적 구조로 작동하며, 결과적으로 과제의 품질과 타당도를 동시 확보한다. 본 연구의 GenAI-HITL 프로토콜은 이처럼 생성-검토-피드백의 순환을 통해, 서술형 평가 과제 개발에서 요구되는 교육적 정합성과 절차적 재현 가능성을 동시에 달성하였다.

5. 평가 루브릭

본 연구는 GenAI가 생성한 서술형 과제의 품질을 다각적으로 평가하기 위해, 국내외 서술형 평가 및 루브릭 개발 연구를 바탕으로 3개 영역 12개 항목의 5점 리커트 루브릭을 개발하여 활용하였다(〈표 4〉 참조).

〈표 4〉 문항 생성 결과 루브릭(3영역 12항목)

평가 영역	하위 기준	주요 평가 내용
1. 교육과정 정합성 및 내용 타당도	1-1. 성취기준 부합도	문항이 목표한 2022 개정 교육과정 성취기준의 핵심 요소를 충실히 반영하고 있는가?
	1-2. 텍스트 기반성	문항이 답을 구성하기 위해 지문에 제시된 정보(명시적/함축적)를 반드시 활용하도록 요구하는가?
	1-3. 통합적 사고 요구	문항이 지문 내용을 바탕으로, 〈보기〉의 관점이나 외부 자료와 통합하여 종합적으로 사고하도록 요구하는가?
	1-4. 맥락적 연관성	문항이 묻는 내용이 지문의 핵심 주제나 논지에서 자연스럽게 파생되었으며, 인위적이거나 지엽적이지 않은가?
2. 문항 설계의 구조적 체계성	2-1. 발문의 명료성	발문이 평가 의도를 명확히 드러내고, 학습자가 수행할 과제를 구체적으로 이해할 수 있도록 진술되었는가?
	2-2. 답안 조건의 구체성	요구하는 답안의 분량, 형식, 포함 요소 등이 명시적으로 제시되어 사고의 경로를 안내하는가?
	2-3. 사고 과정의 단계성	문항 1에서 문항 2로 이어지는 사고의 흐름이 사실적 이해 → 추론·적용 → 비판·평가의 위계를 반영하여 설계되었는가?
3. 현장 적용성 및 실용성	3-1. 난이도 적절성	문항에 사용된 어휘, 문장 구조, 요구하는 사고의 복잡성이 목표 학년(고3) 학습자의 평균적인 수준에 부합하는가?
	3-2. 자료 구성의 실제성	지문과 〈보기〉 자료가 신뢰할 수 있는 출처를 기반으로 하며, 학습자의 동기를 유발할 수 있는 실제적인 내용인가?
	3-3. 채점 기준의 타당성	채점 요소가 평가 목표를 정확히 반영하고, 각 점수 수준별 수행 특성(앵커)이 구체적으로 기술되어 있는가?
	3-4. 채점 편의성 및 피드백	채점 기준이 명확하여 채점자 간 이견의 소지가 적고 효율적인 채점이 가능하며, 그 결과가 학습자에게 구체적인 피드백을 제공할 수 있는가?
	3-5. 교수·학습 연계성	평가 과제 및 결과가 해당 주제에 대한 심화 학습이나 후속 토론 등 교수·학습 활동과 연계될 가능성이 있는가?

영역 1은 ‘교육과정 정합성 및 내용 타당도’로, 과제가 교육과정에서 의도한 학습 목표를 충실히 반영하고 있는지를 평가한다. 구체적으로 ‘성취기준 부합도’는 2022 개정 교육과정과의 연계성을 평가하는 기준으로, 이는 국내 교육 상황에서의 평가 타당성을 확보하는 데 필수적이다(서울특별시교육청, 2022; 경기도교육청, 2024; 김경희, 2020). 또한 ‘텍스트 기반성’, ‘통합적 사고 요구’, ‘맥락적 연관성’은 문항이 지문 정보에 기반하여 외부 자료와 통합적으로 사고하도록 요구하는지를 평가하는 기준으로, 이는 국제적 대규모 평가에서 강조되는 구성형 응답 과제의 핵심 요건과 일치한다(권태현, 2021; College Board, 2019).

영역 2는 ‘문항 설계의 구조적 체계성’으로, 문항의 형식적·구조적 완성도를 평가한다. ‘발문의 명료성’과 ‘답안 조건의 구체성’은 학습자가 과제를 명확히 인지하고 수행하도록 안내하는지를, ‘사고 과정의 단계성’은 문항이 사실적 이해에서 비판적 사고로 나아가는 인지적 위계를 반영하는지를 평가하는 기준이다(최숙기, 2023; Kan & Bulut, 2014).

영역 3은 ‘현장 적용성 및 실용성’으로, 과제가 실제 교육 현장에서 효과적으로 활용될 수 있는지를 평가한다. ‘난이도 적절성’은 목표 학년의 학습자 수준을 고려했는지를, ‘채점 기준의 타당성’과 ‘채점 편의성’은 채점의 일관성과 효율성을 담보하는지를, ‘교수·학습 연계성’은 평가 결과가 수업 개선과 연계될 수 있는지를 평가하는 기준으로 구성되었다. 이는 과정 중심 평가에서 강조되는 요소를 반영한 것이다(박혜영 외, 2019; 서울특별시교육청, 2022; McCaffrey, Casabianca, Ricker-Pedley, Lawless, & Wendler, 2022).

이러한 루브릭 구성은 과제가 ‘무엇을’(내용 타당도), ‘어떻게’(구조적 체계성), ‘왜’(현장 적용성) 평가하는지를 종합적으로 검토하기 위함이며, 구성형 응답(constructed response) 채점의 국제적 사례를 참조하여 평가의 타당도와 평정 신뢰도 확보를 도모하였다(김형성, 2023; McCaffrey et al., 2022) .

6. 분석 방법

본 연구는 GenAI-HITL 방식으로 개발된 서술형 과제의 교육적 타당성에 대한 전문가 준거 기반의 예비적 근거를 확인하고, 생성 과제의 질적 특성을 기술하기 위하여 정량 분석과 정성 분석을 병행하였다. 정량 분석은 전문가 평정 결과를 기초로 네 가지 단계로 진행되었으며, 이들 단계는 평가 도구의 신뢰성 검토과 생성 과제의 질 평가라는 이중 목적을 지녔다. 먼저, 생성된 과제에 대한 전문가 평정 결과의 전반적 분포를 파악하기 위해 항목별·영역별 평균, 표준편차, 최솟값, 최댓값 등 기술통계치를 산출하였다. 다음으로, 18명의 평가자가 동일한 AI 생성 과제를 평정하였기에 평정자 간 신뢰도를 추정하기 위하여 Intraclass Correlation Coefficient(ICC)를 산출하였으며(Koo & Li, 2016; Shrout & Fleiss, 1979), 특히 ICC(3,k) 모델을 주지표로, ICC(2,k) 모델을 보조지표로 병행 제시하였다. 이어서, 평가에 활용된 12개의 루브릭 항목이 'AI 생성 과제의 질'이라는 단일 구인을 내적으로 일관되게 측정하는지를 확인하기 위해 Cronbach's α 계수를 산출하였다(Tavakol & Dennick, 2011).

정량 분석과 병행하여, 평정 점수만으로 해석하기 어려운 맥락적 의미를 파악하기 위해 심층 면담을 실시하였다. 면담에서는 평가자들이 AI 생성 과제의 난이도 적절성, 채점 편의성, 사고 과정 설계의 타당성, 교수·학습 연계 가능성 등에 관해 경험한 구체적 장점과 한계를 탐색하였다. 최종적으로 본 연구는 기술통계, ICC, Cronbach's α를 통해 평가 도구의 신뢰성을 입증하고, 전문가 평정 결과를 통하여 AI 생성 과제의 질을 다각도로 검토하였다.

IV. 연구 결과

1. GenAI 생성 과제의 구조적 특성 분석

본 연구에서 GenAI-HITL 방식을 통해 생성된 서술형 과제의 구체적 특성과 품질을 지문 구성, 문항 설계, 루브릭 개발의 세 측면에서 체계적으로 분석하였다.

1) 지문 구성 분석

생성된 지문은 기준 지문 (가), (나)와 생성 지문 (다)로 구성되었다. 기준 지문은 『2026학년도 EBS 수능특강 국어영역 독서』 234-235쪽에서 선정한 (가)와 (나)를 활용하였다. 이는 국제 정치학의 핵심 이론을 다루는 사회·문화 분야 지문으로, 2022 개정 교육과정의 [12독작01-08] 성취기준에 부합하는 것으로 나타났다(〈표 5〉 참조).

〈표 5〉 (가), (나)에 기초한 생성 지문 (다)[2)]

(가)

국제 정치학에서 국가의 힘과 국가 간 힘의 분포 혹은 힘의 균형에 관한 논의는 오랜 논쟁의 내용 중 하나이다. 국가 간 힘의 이해와 관련된 것 중에서도 특히 강대국 간에 힘이 균형을 이루고 있는지 여부와 관련하여 국제 체제의 안정과 평화가 주로 논의되는데, 대표적인 이론으로 세력 균형론과 패권 안정론이 있다.

세력 균형론에서 세력 균형은 국가 간의 힘이 배분된 측면과 정책적 측면으로 나누어 그 의미를 파악해 볼 수 있다. 전자의 측면에서는 무정부 상태의 국제 관계에서 국가 간의 관계가 수평적인 것을 전제로 특정의 어떤 한 국가가 자국의 의지를 타국에 일방적으로 강요할 수 없는 상태, 동맹 등을 통해 국가 간의 힘이 균형 있게 분포되어 있는 상태를 의미한다. 한편 후자의 측면에서 세력 균형의 의미는 타국의 힘이 지나치게 강해지는 것을 막기 위해 자국이 수립하거나 실시하는 정책,

2) (가), (나)는 『2026학년도 EBS 수능특강 국어영역 독서』 234-235쪽 원문; (다)는 GenAI가 생성한 연계 지문

즉 생존의 차원이자 평형의 창출 또는 유지를 위한 정책을 의미한다. 특정 국가의 국력, 특히 군사력이 지나치게 비대해지는 것은 다른 나라들의 존망에 큰 영향을 줄 수 있기 때문이다. 국가는 세력 균형을 위해서 내부적으로 경제 발전, 군비 증강, 전략 개발이라는 수단을, 국가 외부적으로는 동맹이라는 수단을 추구한다. 이 이론에서는 세계의 모든 나라를 관리하거나 지배하는 정부가 없는 상황, 즉 무정부적인 세계이면서 스스로의 힘으로 자국을 구제해야 하는 체계하에서 세력 균형이 일어난다고 보았는데, 실제로 국제 정치사에서 나타났던 세력 균형의 형태로는 균형자형, 비스마르크형, 냉전형 등이 있다.

첫 번째로 균형자형은 서로 힘이 다른 두 개의 세력이 존재하고, 제3의 세력으로서의 균형자가 개입하여 힘이 약한 쪽에 균형자의 힘을 더해서 균형을 유지해 나가는 것이다. 두 세력의 힘은 변화될 수 있기 때문에 ㉠균형자는 그 어느 국가와도 영구적인 동맹 관계를 맺어서는 안 된다. 두 번째는 비스마르크형으로 예상되는 침략국이 존재하고, 이 국가를 둘러싼 여러 국가가 공통적 이익에 따라 몇 개의 복합적 동맹을 맺어 예상 침략국을 고립시켜 견제하는 것이다. 가령 F 국가로부터 침략을 받을 것 같은 A국이 공통의 특정 이익으로 B국, C국과 동맹을 맺고, A국은 또 다른 공통의 특정 이익으로 D국과도 동맹을 맺어 F 국가를 고립시키는 것이다. 세 번째, 냉전형은 서로 적대하는 당사국 간의 힘의 균형이 생겨 있는 상태이다. 자본주의 국가들과 공산주의 국가들이 적대 관계를 형성하면서 세력 균형을 이룬 상태가 대표적 예이다.

한편 패권 안정론은 강대국 중에서도 특정의 한 강대국이 국제 질서에서 패권국이 되고, 이 국가를 중심으로 한 국제 체제의 안정이 만들어진다고 본다. 이 입장에서는 국제 관계를 무정부 상태로, 국가 간의 관계는 강대국과 중강국 그리고 약소국으로 구성되어 있는 수직적 관계로 본다. 패권국은 자국을 중심으로 한 국제 질서를 형성하고 유지하기 위해 정치, 경제, 군사적 원조 등과 같은 긍정적 자극을 타국에 주고, 안전 및 소유권의 보호라는 국제적인 공공재를 제공하거나 여기에 필요한 비용을 분담한다. 패권국이 제공하는 일종의 공공재는 현상 타파를 기도하려는 국가의 출현을 억제해 국제 사회 전체의 안정과 평화를 가져오는 동시에 장기적으로 자국의 이익을 최적화할 수 있는 국제 체제를 창출한다. 패권 안정론에서는 패권국을 국내 정치에서의 정부처럼 권력을 가진 권위체로 본다. 따라서 ㉡패권국은 다른 국가들을 통제하는 것이 가능하며 이로 인해 패권국은 국제 체제의 안정을 가져온다.

※ 출처: 「2026학년도 EBS 수능특강 국어영역 독서」, 234쪽

(나)

각 국가를 통할하는 공권력이 없는 국제 사회에서 자국의 안보를 확보 또는 유지하는 것이 곤란한 국가가 심각한 정도의 위협을 느낄 때 채택하는 전형적인 안보 정책의 수단이 동맹이다. 동맹은 국가 간 힘의 결합이며 상호 군사적 지원의 약속이다. 동맹은 안보상의 위협뿐만 아니라 공통의 이익을 위해 결성되기도 한다. 그러나 동맹은 어디까지나 자국을 위한 것이기 때문에 이념에 따른 동맹의 결속 약화보다는 이익에 따른 동맹의 결속 약화가 더 급격히 이루어질 수 있다. 이러한 문제에도 불구하고 동맹은 약소국의 안보 불안을 해소하는 데 도움이 된다는 것에는 이론의 여지가 없다.

일반적으로 동맹 전략에는 '균형 전략'과 ㉢'편승 전략'이 있다. 균형 전략은 우월한 힘을 가지고 있는 나라 또는 패권국이 될 가능성이 있는 국가에 대항하여, 힘이 덜 강한 국가들이 하나의 동맹을 맺어 국제사회에서 힘의 균형을 맞추고, 패권국이 등장하지 않게 하는 것이다. 하지만 제2차 세계대전 이후 강한 힘을 가진 미국에 서구 유럽이 협력을 하는 국제 정치 상황이 나타났다. 이는 힘이 강한 국가에 힘이 덜 강한 국가들이 동맹을 맺은 것이었으므로 균형 전략의 한계를 보여 주었다. 균형 전략과 달리 편승 전략은 패권국 또는 힘이 강한 국가와 동맹을 맺고, 이 국가가 가진 힘을 통해 국익을 극대화하는 전략이다.

슈웰러는 편승 전략을 다양한 국익의 확보와 증진이라는 차원에서 방어적 편승, 자칼식 편승, 영합적 편승으로 나누었다. 방어적 편승은 강대국이나 패권국과 대립하는 약소국이 무모한 전쟁의 비용을 피하기 위해 강대국이나 패권국과 동맹을 맺는 것이다. 자칼식 편승은 수정주의 국가와 세력

권에 관한 협정을 체결하고 자국의 이익을 추구하는 것이다. 수정주의 국가는 패권국에 도전하여 패권국이 되려는 국가로, 이를 위해 국가의 권력과 군사력을 확장한다. 영합적 편승은 전쟁의 결과가 거의 확정적일 때 승리의 배당을 얻기 위해 승자 측에 편승하는 것이다.

동맹의 개념과 전략을 통해서 동맹은 하나의 공통된 목표로 맺어지는 것이 아님을 알 수 있다. 동맹 관계에서는 영원한 패권국이나 강대국도, 영원한 동지도, 영원한 적도 없고 오직 자국의 국익만이 존재한다. 또한 동맹을 맺은 국가는 동맹을 맺은 국가 간의 호의적인 관계가 영원히 유지될 것이라고 믿지 않는다. 그러므로 동맹 전략은 자국의 상황과 국제 정치의 상황에 대한 깊이 있는 이해를 바탕으로 수립돼야 하는 것이다.

※ 출처: 「2026학년도 EBS 수능특강 국어영역 독서」, 235쪽

(다)

미·중 전략 경쟁과 한국의 선택

한반도는 역사적으로 **대륙 세력과 해양 세력** 사이의 완충 지대였고, 그 영향은 오늘날에도 이어진다. 미국이 추진하는 **'자유·개방 인도·태평양'** 전략이 공세적 성격을 띠는 반면, 중국은 경제 보복을 통해 대응하며 경쟁 구도를 심화시키고 있다. 한국은 이 틈바구니에서 **'안보는 미국, 경제는 중국'**이라는 안미경중 노선으로 ㉡**'이중 헤징(double hedging)'** 전략을 유지해 왔다.

그러나 2025 년 이후 미·중 기술 경쟁이 가속화하면서 한국 반도체·배터리 산업에 대한 **'선택 압박'**이 커지고 있다. 외교 당국은 편승[2]과 균형[3] 사이에서 전략적 모호성을 유지한다고 설명하지만, 국내 산업계는 시장 다변화와 기술 자립을 위한 '전략적 분산'을 요구한다.

주석-

1. double hedging :안보와 경제를 서로 다른 강대국에 의존해 리스크를 분산하려는 전략.
2. 편승: 강대국과 협력해 방어 비용을 최소화하고 안전을 확보하려는 전략; NATOization: 개별 지원을 NATO 지휘 체계로 통합·표준화하는 과정.
3. 균형: 특정 세력의 팽창을 억제하기 위해 상대 세력과 동맹·군비를 통해 힘의 균형을 맞추는 전략.

※ 출처: 2023년 9월 1일·프리드리히 에버트 재단(FES) Asia 연구 보고서

AI가 생성한 연관 지문 (다) "미·중 전략 경쟁과 한국의 선택"은 기준 지문의 이론적 개념을 현대적 맥락으로 연결하는 특성을 보였다. "한국은 이 틈바구니에서 '안보는 미국, 경제는 중국'이라는 안미경중 노선으로 '이중 헤징(double hedging)' 전략을 유지해 왔다"는 서술을 통해 (나) 지문의 편승 전략 개념과 직접 연계되었다. 이러한 구성은 추상적 이론을 구체적 상황에 적용하는 사고 과정을 유도하는 구조를 갖추었다.

2) 문항 설계 분석

생성된 두 문항은 '발문-보기-조건'의 삼원구조로 설계되었다(〈표 6〉 참조).

〈표 6〉 서술형 1, 2 생성 결과

1. (서술형) (가)의 ㉠·㉡ 문장을 이해한 뒤, 아래 〈보기〉를 참고하여 세력 균형론과 패권 안정론이 국제 질서를 어떻게 안정화한다고 설명하는지 비교·서술하시오.(4점)	2. (서술형) (나)의 ㉢ 편승 전략 세 분류와 (다)의 ㉣ '이중 헤징' 사례를 참고하여, 한국 전략이 편승 유형 중 어디에 가까운지 판단하고 근거를 서술하시오.(6점)
〈보 기〉 - Henry Kissinger 발언 (The National Interest, 2022-10-07) **"오늘날 국제 질서는 19세기 유럽과 유사한 다극 경쟁**으로 회귀하고 있다. 키신저는 '강대국 간 힘의 대칭성이 유지될 때, 어느 국가도 전면전의 대가를 감당할 수 없어 억제 효과가 발생한다'고 진단한다. 동시에 그는 '미국이 제공해 온 **해상 교역로 보호·국제 금융 안정·안보 보장**과 같은 **패권적 공공재**가 약화될 경우, 체제 전환기의 **불안정이 장기화**될 것'이라고 경고한다." 원문 URL ▶ https://nationalinterest.org/feature/henry-kissingers-imbalance-power-205261	**〈보 기〉** - Brookings 토론 요약 (Brookings Institution, 2025-01-22) "세미나 패널은 **2025년 미·중 전략 경쟁이 기술·경제 전면전으로 확산**되면, 한국의 '안미경중' 두 축 전략이 구조적 한계에 봉착할 것이라고 분석했다. 구체적으로, △ 미국 주도의 **반도체 공급망 재편**(CHIPS Act 후속 규제) 과 △ 중국의 **희토류·배터리 핵심 광물 수출 통제**가 동시에 현실화될 경우, 한국은 '동맹 이행 비용'과 '경제 보복 위험' 사이에서 **전략적 선택을 강제**받게 된다는 것이다." 원문 URL ▶ https://www.brookings.edu/articles/how-will-south-korea-navigate-us-china-competition-in-2025
〈조 건〉 **〔구조·형식〕** ① 두 문단 권장 → 1문단에 두 이론 핵심, 2문단에 차이·비교. ② 첫 문장을 주제문으로 두어 문단의 중심을 드러낼 것. **〔내용〕** - 이론 핵심 제시 ① (가)의 ㉠·㉡을 간접 인용해 '세력 균형론'과 '패권 안정론' 각자의 안정 메커니즘을 1문장씩 정리. ② '세력 균형', '공공재' 용어 필수. - 차이·대비 ① 두 이론이 안정을 구축하는 방식이 "힘 분산"과 "집중 통제"라는 점을 최소 1개 이상 대비어(예: 반면, 그러나)로 연결해 서술. ② 접속어 1회 이상 사용. **평가 요소** ① 개념 파악(2점) → 이론 핵심 문장 ② 전제·차이 추론(2점) → 대비 문장	**〈조 건〉** **〔구조·형식〕** ① 3문장 이상 권장(판단 ▶ 근거 ▶ 시사점). ② 문장 표지어(서론·본론·결론)는 쓰지 않되, 논리 흐름을 접속 부사(따라서, 하지만 등)로 명확히 할 것. **〔내용〕** - 판단 문단 ① (나)의 ㉢ 분류 중 방어적·자칼식·영합적 중 하나를 명시하고, ㉣ 한국 전략과 연결. - 근거 문단 ① 안보 측면 근거 1개: 예) 확장억제·군사비 부담. ② 경제 측면 근거 1개: 예) 교역·공급망 의존 수치. ③ 근거마다 짧은 설명·자료(수치·정책) 첨언. -시사점 문단 ① 위 판단이 한국 외교에 주는 의미·과제를 한 문장으로 제시. **평가 요소** ① 유형 판단(2점) → 판단 문단 ② 근거-안보(2점) + ③ 근거-경제(2점) → 근거 문장 (시사점은 조직·논리 흐름 평가에 포함)

문항 1은 “(가)의 ㉠·㉡ 문장을 이해한 뒤, 아래 〈보기〉를 참고하여 세력 균형론과 패권 안정론이 국제 질서를 어떻게 안정화한다고 설명하는지 비교·서술하시오”라는 발문을 통해 사실적·추론적 독해를 요구하였다. 보기 자료로는 Henry Kissinger의 실제 발언을 활용하여 두 이론의 관점을 제시하였으며, 작성 조건에서는 “’세력 균형’, ‘공공재’ 용어 필수” 사용과 “두 이론이 안정을 구축하는 방식이 ‘힘 분산’과 ‘집중 통제’라는 점을 대비어로 연결해 서술”을 명시하였다. 문항 2는 “(나)의 ㉢ 편승 전략 세 분류와 (다)의 ㉣ ‘이중헤징’ 사례를 참고하여, 한국 전략이 편승 유형 중 어디에 가까운지 판단하고 근거를 서술하시오”라는 발문으로 이론의 현실 적용을 요구하였다. 보기로는 Brookings Institution의 토론 요약을 제시하여 전문적 분석 자료를 제공하였으며, 조건에서는 “판단 → 근거 → 시사점” 3단계 서술 구조를 요구하였다. 이러한 설계는 [12독작01-03]의 “글에 드러난 정보를 바탕으로 글의 내용을 파악하고 글에 드러나지 않은 정보를 추론하며 읽는다”에서 [12독작01-04]의 “글의 내용이나 관점을 평가하며 읽는다”로 이어지는 인지적 위계를 반영하였다.

3) 루브릭 및 예시 답안 개발 분석

생성된 루브릭은 분석적(Analytic) 방식을 채택하여 요소별 세분화 채점 체계를 구성하였다(〈표 7〉, 〈표 8〉 참조).

〈표 7〉 루브릭 생성 결과

루브릭 구조
① 루브릭 표(채점 요소·세부 점배점·수행 특성·)
② 개념·자료 체크리스트(✓/✗)
③ 언어 오류 매트릭스(가중치)
④ 수준별 예시 답안 상·중·하 (채점 근거 주석 포함)

서술형 1 루브릭

1. 분석적(Analytic) 루브릭

채점 요소	배점(0-2)	0점	1점	2점
① 개념 파악	2	두 이론 핵심 용어 부정확·누락	한 이론만 정확	"세력 균형 : 힘 분산 견제 / 패권 공공재 : 패권국 규칙·보장"
② 차이·비교	2	차이 미언급·오류	차이 1개·모호	"분산 억지 vs 집중 통제" 등 2포인트 대비

2. 개념 체크리스트

항목	✓	✗
'힘 분산'·'견제' 언급		
'공공재'·'집중 통제' 언급		
대비 접속어(반면·그러나 등)		

3. 언어 오류 감점(총점 후 적용)

오류 유형	계수
철자·띄어쓰기	0.25
조사·어미	0.5
주어·서술어 호응	1

4. -7점 계수 합 = -0.5, 8점↑ = -1

서술형 2 루브릭

1. 분석적(Analytic) 루브릭

요소	배점	0점	1점	2점
① 유형 판단	2	미지정·오판	유형만 명시	"ⓔ은 방어적 편승: 억지·비용 최소화"
② 근거-안보	2	미언급	단편	"확장억제 협의 + 군사비 GDP 2.5 % → 비용↓"
③ 근거-경제	2	미언급	단편	"對中 수출 25 %·희토류 80 % 의존 → 위험 회피"

형식 감점 – '판단·근거·시사점' 3문장 미달 -1

2. 체크리스트

항목	✓	✗
유형(방어·자칼·영합) 명시		
안보 근거 구체		
경제 근거 수치·사실		
시사점 1문장		

〈표 8〉 예시 답안 생성 결과

서술형 1 예시 답안		
수준	**답안 (≈ 260자)**	**4점 분포**
상	세력 균형론은 여러 강대국에 권력이 분산될 때 상호 견제가 작동해 전쟁 비용을 끌어올려 억지력을 만든다고 본다. 반대로 패권 안정론은 단일 패권국이 해상 교역·금융 질서 같은 공공재를 제공하며 규칙을 집중 통제해 안정을 유지한다고 설명한다. 두 이론 모두 힘의 불균형이 위협이라는 전제를 공유하지만, 안정 장치는 '분산 억지'와 '집중 관리'라는 점에서 정반대다. (268자)	'힘 분산·공공재' 정확 + '분산 억지 vs 집중 관리' 2포인트 대비 → 4/4점
중	세력 균형론은 힘이 여러 국가에 고르게 나누어져 있을 때 서로를 감시·견제해 평화가 유지된다고 주장한다. 패권 안정론은 강대국 하나가 규칙을 만들고 책임을 지면 질서가 잡힌다고 본다. 두 이론 모두 무정부 체제의 불안을 인정하지만, 하나는 다자 균형을, 다른 하나는 단일 주도권을 강조한다. (238자)	공공재 개념 빠짐(개념 2), 차이 1포인트 (비교 1) → 3/4점
하	세력 균형론과 패권 안정론은 둘 다 세계를 안정시키려는 생각이라는 점에서 비슷하다. 어느 이론이든 강한 나라들이 있어서 전쟁을 막는다는 뜻으로 이해했다. 구체적으로 어떻게 다른지는 지문에서 잘 모르겠어서 큰 차이를 느끼지 못했다. 그래서 두 이론을 거의 같은 논리로 보아도 된다고 생각한다. (232자)	핵심 개념·차이 모두 부정확 → 0/4점

서술형 2 예시 답안		
수준	**답안 (≈400자)**	**채점 근거 (주석)**
상	한국의 ㉢ 이중 헤징은 전형적인 방어적 편승이다. 안보 측면에서 한국은 한미 확장억제전략협의체·SM-3 요격체계 등을 통해 미국의 핵우산과 미사일 방어망에 무임승차하면서도 국방비를 GDP 2.7 % 수준으로 묶어 군사비 부담을 최소화한다. 경제 측면에서는 대중 수출 25 %, 배터리 핵심 광물 80 % 의존 구조를 유지해 단기 이익과 공급망 안정성을 확보한다. 이처럼 방어적 편승은 비용 절감 효과가 크지만, 미·중 압력이 심화될수록 전략적 유연성이 축소되는 딜레마를 남긴다. (404자)	• 유형 판단 2점 (방어적 편승 + ㉢ 연결) • 근거-안보 2점 (확장억제·GDP 2.7 % 수치) • 근거- 경제 2점(수출 25 %·광물 80 % 수치) • 시사점 O
중	한국의 ㉢ 이중 헤징은 자칼식 편승에 가깝다. 안보에서 한국은 미사일 방어 체계와 한미 연합훈련을 통해 미국의 억지력에 의존하지만, 국방비는 GDP 2 %대에 묶어 비용을 절감한다. 경제적으로는 중국과의 교역이 여전히 20 % 이상을 차지하고, 배터리 소재 수입도 대체선이 부족해 의존도가 높다. 결국 한국은 두 강대국에 "적극 협력도, 완전 결별도 하지 않는" 길을 택해 단기 이익은 지키지만 구조적 압박에는 취약하다. (404자)	• 유형 판단 2점 • 안보 근거 1점(표피적) • 경제 근거 1점(수치 모호) • 시사점 O
하	한국은 미국과 중국 사이에서 균형을 맞춘다. 안보는 미국 편이고 경제는 중국과 거래가 많아서 이익이 크다. 그래서 편승 전략이라고 할 수 있다. 그러나 기술 전쟁이 심해지면 어려움이 생길 수 있다. (166자 — 규정 미달)	• 유형 언급 불명확(0) • 안보·경제 근거 미흡(각 1) • 분량 미달로 형식 -1

문항 1의 경우 '개념 파악(2점)'과 '차이·비교(2점)'로, 문항 2의 경우 '유형 판단(2점)', '근거-안보(2점)', '근거-경제(2점)'로 구분하여 총 10점 만점 체계를 제시하였다. 개념 체크리스트를 통해 "'힘 분산'·'견제' 언급", "'공공재'·'집중 통제' 언급", "대비 접속어(반면·그러나 등)" 사용 여부를 명시적으로 확인하도록 설계되었다. 수준별 예시 답안은 상·중·하 3단계로 제시되어 채점 기준의 구체화를 도모하였다. 상급 답안의 경우 "세력 균형론은 여러 강대국에 권력이 분산될 때 상호 견제가 작동해 전쟁 비용을 끌어올려 억지력을 만든다고 본다. 반대로 패권 안정론은 단일 패권국이 해상 교역·금융 질서 같은 공공재를 제공하며 규칙을 집중 통제해 안정을 유지한다고 설명한다"(268자)와 같은 모범 사례를 제공하였다.

4) 교육과정 정합성 분석

생성된 과제는 2022 개정 교육과정 '독서와 작문' 과목의 다음 성취기준과 연계되었다. 문항 1은 [12독작01-03] "글에 드러난 정보를 바탕으로 글의 내용을 파악하고 글에 드러나지 않은 정보를 추론하며 읽는다"와 [12독작01-04] "글의 내용이나 관점을 평가하며 읽는다"를 통합적으로 다루었다. 문항 2는 [12독작01-04]와 [12독작01-08] "사회·문화 분야의 글을 자료를 활용하여 설득력 있게 쓴다"의 요구를 반영하였다. 교육과정에서 강조하는 "실제적인 국어 활동 상황에서 지식과 기능을 통합하여 적용하는 능력"과 "실제 언어 생활 맥락에서 학습한 내용을 적용할 수 있는 역량" 평가 방향과의 부합성도 확인되었다. 이상의 분석을 통해 GenAI가 생성한 서술형 과제는 지문-문항-루브릭이 통합된 구조를 갖추고 있으며, 교육과정 성취기준과의 연계성을 보였다. 그러나 실제 교육적 타당성과 활용 가능성은 전문가 평가를 통해 검토될 필요가 있다.

2. 전문가 정량 평가

전문가 18인을 대상으로 실시한 평정 결과, GenAI가 생성한 서술형 과제의 품질은 전반적으로 높은 수준으로 평가되었다(〈표 9〉 참조).

〈표 9〉 서술형 과제 생성 결과 정량 평가 결과 요약(N-18)

평가 영역 · 하위 기준	평균(M ± SD; Min-Max)
1. 평가의 기본 요건 및 타당성	
성취기준 부합도	4.70 (0.48; 4-5)
자료 구성의 적절성	4.30 (0.82; 3-5)
과제 난이도 적절성	3.70 (0.95; 2-5)
지문-문항 논리성	4.30 (0.67; 3-5)
2. 문항 구성 및 설계의 체계성	
발문 명료성	4.40 (0.52; 4-5)
지문-발문-보기 연계성	4.50 (0.53; 4-5)
답안 조건 구체성	4.50 (0.53; 4-5)
사고 과정 단계성	4.40 (0.52; 4-5)
3. 채점 및 활용의 실제성	
채점 기준 타당성	4.30 (0.48; 4-5)
응답 특성 분석	4.30 (0.67; 3-5)
채점 편의성 및 피드백	4.10 (0.74; 3-5)
교수-학습 연계성	4.40 (0.70; 3-5)
전체 평가 요약	
전체 평균 (M ± SD)	**4.32 ± 0.28 (항목평균 기준)**
루브릭 내적 일관성 (Cronbach's α)	**0.894**
평정자 간 합치도 ICC (2,k)/(3,k) ✝	**0.685(보통) / 0.695(보통)**

분석 결과, 전체 평균은 4.32점(SD=0.28)으로, 대부분의 항목이 5점 척도 상에서 '대체로 타당하다(4점)' 이상의 긍정적 수준을 보였다. 평가 도구의 신뢰도를 검토한 결과, 루브릭의 내적일관성(Cronbach's α = .894)은 높은 수준으로 나타나 평가 항목 간 일관된 판단이 이루어졌음을 확인하였다. 평정자 간 합치도는 ICC(2,k) = .685, ICC(3,k) = .695로 나타나, 평가자 간 의견의 일치가 보통(moderate) 수준으로 확보되었다(Shrout & Fleiss, 1979; Koo & Li, 2016).

1) 교육과정 정합성 및 내용 타당도 영역

해당 영역의 평균은 4.25점으로, 생성된 과제가 교육과정의 목표와 내용에 비교적 잘 부합하는 것으로 평가되었다. 특히 '성취기준 부합도'(M=4.70)와 '통합적 사고 요구'(M=4.70)는 매우 높은 평정값을 보여, GenAI-HITL 프로토콜이 2022 개정 국어과 교육과정의 핵심 요소와 고차사고 측정 목표를 성공적으로 반영했음을 시사하였다. 이는 AI가 독자적으로 생성하는 것이 아니라, 인간 전문가가 교육과정 성취기준을 명확히 제시하고 각 단계에서 검토하는 HITL 방식의 유효성을 입증하는 결과이다.

반면, '난이도 적절성'(M=3.70)은 상대적으로 낮은 평가를 받아, AI가 생성한 과제의 인지적 요구 수준이 실제 고3 학습자의 평균 발달 수준과 다소 괴리가 있을 수 있음을 의미하였다. 이는 교사의 후속적 난이도 조정의 필요성을 보여주었다.

2) 문항 설계의 구조적 체계성 영역

해당 영역은 평균 4.45점으로 세 영역 중 가장 높은 평가를 받았다. 이는 AI가 생성한 과제가 형식적·구조적으로 완성도 높은 초안임을 의미하였다. 특히 '지문-발문-보기 연계성'과 '답안 조건의 구체성'(각각 M=4.50)은 루브릭 상위 수준에 해당하는 결과를 보였다. '발문 명료성'(M=4.40)과 '사고 과정 단계성'(M=4.40) 역시 높은 평가를 받았다.

이러한 결과는 본 연구에서 적용한 CoT 기반 프롬프트가 '사실적 이해 → 추론·적용 → 비판·평가'로 이어지는 인지적 위계를 효과적으로 구현했음을 뒷받침하였다. 또한 컨텍스트 엔지니어링을 통해 교육과정 성취기준과 평가 목표를 명시적으로 제공한 것이 문항 구조의 체계성 확보에 기여한 것으로 해석되었다.

3) 현장 적용성 및 실용성 영역

해당 영역의 평균은 4.30점으로, AI 생성 과제가 실제 교육 현장에서 활용될 수 있는 잠재력을 지니고 있음을 보여주었다. '채점 기준의 타당성'(M=4.30)과 '교수·학습 연계성'(M=4.40)은 비교적 높은 평정값을 나타내, 일부 문항이 학습자의 사고를 촉진하거나 수업 후속 활동으로 연결될 가능성을 시사하였다.

반면, '채점 편의성 및 피드백'(M=4.10)은 다소 낮은 수준으로, 채점 효율성과 피드백 활용성 측면에서 보완이 필요함을 보여주었다(McCaffrey et al., 2022). 이는 AI가 예측적으로 생성한 루브릭과 예시 답안이 실제 학습자들의 다양한 응답 패턴을 충분히 반영하지 못했기 때문으로 해석되었다.

전체적으로, 전문가 평가 결과는 영역별로 평균 4점 이상을 보여, GenAI-HITL 기반 과제가 교육과정 정합성, 구조적 체계성, 현장 적용성 측면에서 전반적으로 긍정적인 평가를 받았음을 시사하였다. 다만, 일부 항목에서는 평가자 간 편차가 존재하였으므로, 향후 연구에서는 평가자 교정(anchor calibration)이나 추가적인 합의 절차를 통해 신뢰도 향상을 모색할 필요가 있다.

3. 전문가 정성 평가

전문가 집단을 대상으로 실시한 심층 면담과 질적 분석을 통해, 정량 평가에서 나타난 주요 특징과 그 배경 요인을 구체적으로 확인하였다. 정성 분

석 결과는 세 영역에서 AI가 생성한 과제의 강점과 개선 과제를 함께 드러내며, GenAI-HITL 방식의 실제적 성과와 한계를 종합적으로 조명하였다.

1) 구조적 완성도와 난이도 적절성 간의 긴장

전문가들은 GenAI가 산출한 과제의 구조적 체계성에 대체로 긍정적인 평가를 내렸다. 대부분의 전문가는 문항이 "개념 파악 → 이론 비교 → 현실 적용"으로 이어지는 사고 위계를 반영하고 있어 학습자의 인지 발달 단계를 고려한 구성이라고 보았다. 이는 정량 평가에서 '사고 과정의 단계성'(평균 4.40)이 높은 평정을 기록한 결과와도 일치하였다.

그러나 이러한 구조적 완성도는 '문항 난이도의 적절성'과의 긴장을 드러냈다. 문항의 논리적 구조는 정교하다는 평가를 받았지만, 일부 전문가들(T3, T7, T9)은 전문 용어('이중혜징', '편승 전략' 등)의 사용이 고등학교 수준을 초과한다고 지적하였다. T3는 "'이중혜징'이나 '편승 전략' 같은 용어는 일반 고3 학생들에게 생소할 수 있어, 용어 자체가 사고의 장벽이 될 수 있다"고 우려하였다. 반면 T5는 "2022 개정 교육과정이 강조하는 고차적 사고력 평가를 위해서는 적절한 수준의 학술적 용어 사용이 필요하며, 이는 대학 수학 준비 차원에서도 의미가 있다"고 평가하였다.

문항의 조건 설정에서도 유사한 의견 차이가 나타났다. "판단 → 근거 → 시사점"의 3단계 서술 구조와 안보-경제 두 측면의 근거 제시 요구는 일부 전문가(T4, T9)에게는 인지적 과부하로 인식되었으나, 다른 전문가(T1, T6)는 학습자의 논증 능력을 향상시키는 교육적 장치로 보았다. 이러한 평가의 양분은 정량 평가에서 '문항 난이도의 적절성'이 평균 3.70점으로 상대적으로 낮고 표준편차가 가장 컸던 결과(SD = 0.95)와 일관되었다.

공통적으로 제기된 우려는 과도한 구조화가 학습자의 사고 자율성을 제한할 수 있다는 점이었다. 그러나 이러한 전문가들의 지적은 본 과제가 2022 개정 교육과정의 핵심인 '주제 통합적 읽기' 역량을 측정하기 위해 필수적인 '의도된 인지적 마찰(intended cognitive friction)'을 효과적으로 유발하

고 있음을 역설적으로 보여준다. 즉, 단순한 난이도 조절의 실패가 아니라, GenAI-HITL을 통해 구현 가능한 문항 논리성의 최대치를 검토하는 과정에서 사고의 깊이를 유도하기 위한 정교한 장치들이 전문가들에게 높은 인지적 부담으로 인식된 것으로 해석할 수 있다.

2) 채점 기준의 명료성과 피드백의 실용성

전문가들은 AI가 생성한 과제와 함께 제시된 채점 기준 및 예시 답안 체계에 대해 교육적 활용 가능성과 실용적 한계를 동시에 지적하였다. '채점 기준의 타당성'(평균 4.30)이 비교적 높게 평가된 것은, 루브릭이 구체적인 성취기준을 근거로 하고 각 점수 수준의 수행 특성을 명시적으로 제시했기 때문이었다. 전문가들은 이러한 구성이 교사에게 체계적 피드백의 근거를 제공하고, 학습자에게 명확한 성취 목표를 제시한다는 점에서 교육적 가치를 인정하였다.

그러나 실제 적용 과정에서는 채점 기준의 명료성에 대한 우려가 제기되었다. 예를 들어, '단편적 근거'와 '구체적 근거'의 구분 기준이나 상급-중급-하급 답안 간의 경계 설정이 명확하지 않아 채점자 간 편차가 발생할 가능성이 높다는 점이 지적되었다. 특히 T4와 T6은 상급 답안 예시(268자 분량)가 실제 고등학생 수준에서 현실적으로 달성 가능한 목표인지에 대해 의문을 제기하였다.

또한, 일부 전문가는 실제 학습자 응답 데이터의 부재를 근본적 한계로 지적하였다. AI가 생성한 루브릭은 예측적 성격을 지니기 때문에, 학습자의 실제 오답 유형이나 부분 점수 사례를 충분히 반영하지 못한다는 점이 반복적으로 언급되었다. 이러한 우려는 정량 평가에서 '채점 편의성 및 피드백'(평균 4.10)이 상대적으로 낮은 평가를 받은 배경을 설명하였다. 요컨대, 루브릭이 이론적 타당성을 확보하고 있음에도 실제 채점에서는 교사의 추가적 판단이 각자의 교육 환경과 맥락에 맞춰 보다 정교하게 투입되어야 함을 시사하였다.

3) 교육과정 정합성과 현장 적용 가능성의 간극

교육과정 정합성에 대해서는 전문가 간 높은 합의가 형성되었다. 다수의 전문가는 과제가 2022 개정 국어과 교육과정의 성취기준([12독작01-03], [12독작01-04], [12독작01-08])을 충실히 반영하며, 독서와 작문의 통합적 학습 취지에 부합한다고 평가하였다. 이는 정량 평가에서 '성취기준 부합도'(평균 4.70)가 가장 높은 수치를 기록한 결과와 일치하였다.

그러나 교육과정 정합성과 실제 수업 적용 간에는 일정한 간극이 존재하였다. 일부 전문가는 성취기준상으로는 적절하더라도, 실제 수업 시간 내에서 이 수준의 과제를 수행하기에는 시간적 제약이 크며, 학생의 배경지식에 따라 접근성이 제한될 수 있다고 보았다. 특히 본 연구에서 활용된 학습 자료가 고3을 대상으로 한 수능특강 읽기 자료에서 발췌한 지문이라 하더라도 '국제정치학' 개념이나 다중 관점 통합형 지문 구성은 특정 학문적 배경을 가진 학습자에게만 익숙할 수 있다는 점이 문제로 지적되었다.

GenAI-HITL 방식의 적용에 대한 견해는 초안의 품질과 현장 활용의 부담이라는 두 측면에서 나뉘었다. T1, T10 등은 AI가 생성한 과제 초안이 일정 수준의 완성도를 지녀 교사의 과제 개발 부담을 경감할 수 있다고 보았으나, T8, T9 등은 실제 수업에 적용하기 위해서는 교사의 재구성 역량이 요구되며, 이 과정이 오히려 새로운 업무 부담으로 이어질 수 있다고 지적하였다.

전문가들의 질적 평가는 정량 분석 결과와 상호보완적인 양상을 보였다. 즉, AI가 생성한 과제는 구조적 체계성과 교육과정 정합성 측면에서 비교적 높은 품질을 보였지만, 학습자 수준 적합성, 채점 효율성, 현장 적용성 측면에서는 개선이 필요함이 확인되었다. 전문가들은 이 방식이 평가 도구의 자동 완성이라기보다는 교사가 교육적 맥락에 맞게 조정·보완할 수 있는 고품질 초안 제공 도구로 이해되어야 한다는 데 의견을 모았다.

전문가 정성 평가는 정량 분석 결과를 심층적으로 뒷받침하였다. 구조적 완성도와 교육과정 정합성에서의 강점은 GenAI-HITL 프로토콜이 일정

부분 기능했을 가능성을 뒷받침하는 반면, 학습자 수준 적합성과 채점 실용성의 한계는 교사 전문성의 필수성을 확인시켰다. 이러한 결과가 서술형 평가 개발에 주는 함의는 다음 장에서 종합적으로 논의한다.

V. 결론 및 제언

본 연구는 GenAI-HITL 방식을 통해 2022 개정 국어과 '독서와 작문' 성취기준에 부합하는 서술형 평가 과제를 생성하고, 전문가 평정을 통해 그 교육적 타당성을 검토하였다.

첫째, GenAI는 교육과정 성취기준에 부합하는 서술형 과제를 생성할 수 있는가에 대해, 전문가 평정에서는 전반적으로 긍정적 경향이 확인되었다. 성취기준 부합도(M=4.70)와 통합적 사고 요구(M=4.70)가 가장 높은 평정을 받았고, 구조적 체계성 영역 전체는 평균 4.45점을 기록하였다. 이는 컨텍스트 엔지니어링을 통해 교육과정 성취기준을 AI의 연산 가능한 형태로 변환하고, CoT를 활용하여 단계적 추론을 유도한 프로토콜(Wang et al., 2023)이 일정 수준에서 기능했을 가능성을 시사한다. 특히 지문-문항-루브릭의 유기적 통합 구조가 서술형 평가의 핵심 요건인 텍스트 기반성과 사고 단계성을 구현하는 데 기여했을 수 있다(김선·반재천, 2023; 박종임, 2024). 다만 이러한 결과는 체계적 프로토콜의 적용, HITL 방식의 통합, 교사의 후속 조정이라는 세 가지 조건이 함께 충족된 맥락에서 관찰된 것임을 전제할 필요가 있다.

둘째, 생성된 과제의 질적 특성과 한계로서 구조적 완성도(M=4.45)와 학습자 적합성(M=3.70) 간 0.75점의 간극이 확인되었다. AI는 형식적·논리적 일관성 확보에서 상대적 강점을 보인 반면, 학습자 발달 단계 판단에서는 비교적 제약이 나타났다. 전문가들이 지적한 과제의 높은 인지적 부담은 단

순한 난이도 조절의 실패로만 보기보다, 2022 개정 교육과정의 핵심인 '주제 통합적 읽기' 역량을 측정하기 위해 요구되는 '의도된 인지적 마찰(intended cognitive friction)'의 관점에서 해석될 여지도 있다. 이는 속도와 정답 찾기 중심의 기존 선다형 평가와 구분되는 지점으로, 파편화된 지식이 아닌 통합적 사고를 평가하려는 설계 의도와 연관될 수 있다. 따라서 이 간극은 평가의 타당성이 알고리즘적으로 포착 가능한 형식적 차원과 교사의 실천적 지식에 기반한 맥락적 차원으로 구분되어 논의될 수 있음을 시사한다.

구조적 완성도가 상대적으로 높게 나타난 메커니즘은 다음과 같이 정리될 수 있다. 첫째, 컨텍스트 엔지니어링을 통해 성취기준 [12독작01-03], [12독작01-04], [12독작01-08]의 핵심 요소—정보 파악, 추론, 비판적 평가—를 명시적으로 제시함으로써 AI의 생성 방향이 보다 분명해졌을 가능성이 있다. 둘째, CoT가 "성취기준 분석 → 지문 주제 설정 → 지문 생성 → 평가 요소 추출 → 문항 생성"의 단계적 추론을 유도하면서 요소 간 논리적 연결성이 강화되었을 수 있다(Lightman et al., 2023). 이는 최종 결과물만을 요구하는 방식에 비해, 중간 과정을 외현화하는 전략이 일관성 확보에 기여할 수 있음을 시사한다.

반면 학습자 적합성이 낮게 나타난 점은 교육 맥락의 복합성과 관련지어 설명될 수 있다. AI는 '고등학교 3학년'과 같은 범주적 수준은 반영할 수 있으나, 실제 학급 내 배경지식 편차나 개별 학습자의 사전 학습 경험을 정밀하게 고려하는 데에는 한계가 있을 수 있다. 난이도 평가의 표준편차가 컸다는 점은 적절한 난이도가 학습자 집단에 따라 상대적으로 결정될 수 있음을 함의한다. 이는 Brown(1987)이 제시한 메타인지적 판단과 Schön(1983)의 반성적 실천 개념으로 해석될 수 있으며, 교사가 학습자와의 상호작용 속에서 축적한 암묵적 지식을 바탕으로 과제의 적절성을 판단한다는 점을 강조한다(박도순, 2025). 이러한 맥락에서 GenAI-HITL 방식은 형식적 타당성과 실질적 타당성에 관한 타당화 근거를 함께 축적·검토하기 위한 협력 모델로 이해될 필요가 있다.

본 연구의 학술적·실천적 함의는 다음과 같이 정리될 수 있다. 첫째, 생성형 AI를 활용한 국내 서술형 문항 개발 연구가 프롬프트 전략 탐색 또는 AI 생성물과 인간 개발 문항의 비교에 주로 초점을 두어 왔다는 점을 고려할 때, 본 연구는 지문과 문항의 유기적 통합, 다단계 사고 위계의 구현, 분석적 루브릭 기반 평가를 포괄하는 '통합 패키지' 생성 프로토콜을 제안함으로써 방법론적 확장 가능성을 제시하였다. 예컨대 함은혜 등(2024)은 3단계 프롬프트 전략을 제안하였으나, 생성 문항의 과업 명료성이나 국가교육과정 부합도에서 한계가 보고되었고, 최숙기와 박종임(2024) 등은 교사 활용 양상이나 인식 분석에 초점을 두어 구체적 방법론 제안에는 제약이 있었다. 이러한 선행 흐름과 대비하여, 본 연구는 컨텍스트 엔지니어링과 CoT를 활용해 생성 과정의 논리적 결함을 완화하고, HITL 기반 검토를 통해 교육과정 성취기준에 부합하는 과제 설계를 지원할 수 있음을 보여주는 사례가 될 수 있다.

둘째, 본 연구의 프로토콜은 고차적 사고 평가를 위한 하나의 이상적 모형(ideal type)으로서, 현 시점에서 인간-AI 협력을 통해 도달 가능한 논리적 정합성의 수준을 점검하는 데 활용될 수 있다. 또한 문항 구조와 평가 요소의 명시성이 강화될 경우, 향후 AI 기반 자동 채점(AES) 및 서·논술형 수능과 같은 미래형 평가 체제 논의에서 참고 가능한 기초 자료로 기능할 가능성이 있다.

셋째, 실천적 차원에서 GenAI가 교사의 평가 설계 전문성을 대체하기보다는, 일정 수준의 초안을 제공하고 교사가 이를 조정·보완하는 협력적 체계로 작동할 수 있음을 시사한다(Zanzotto, 2019). 특히 학습자 적합성에서 확인된 간극은 교사의 맥락적 판단이 필수적임을 재확인한다. 교육현장에서는 본 연구의 모형을 그대로 적용하기보다, 교사의 전문적 판단에 따라 지문의 수를 줄이거나 조건을 완화하는 등 '난이도 하향 조절(scaling down)'을 통해 유연하게 재구성하는 접근이 제언된다. 이는 AI 산출물을 비판적으로 검토하고 학습자 맥락에 맞게 최적화하는 과정으로서, 교사의 고

유한 전문성이 요구되는 지점이다. 따라서 AI 산출물을 교실 상황에 맞게 변용할 수 있도록 교사의 '교육과정 재구성 및 AI 활용 평가 역량'을 강화하는 제도적·교육적 지원이 병행될 필요가 있다(U.S. Department of Education, 2023).

본 연구는 다음과 같은 제한점을 지니며, 이는 결과 해석에 유의미한 조건을 제공한다. 첫째, 국어과 '독서와 작문' 과목의 고3 수준에 한정되어 교과 간 일반화에는 신중함이 요구된다. 둘째, 전문가 평가만으로 검토하여 실제 학습자 수행 데이터가 부재하다. 후속 연구에서는 생성된 과제를 실제 학습자에게 적용하여 난이도와 변별도를 실증적으로 검토할 필요가 있다. 셋째, 2025년 7월 시점의 모델을 사용하여 기술 발전에 따른 변화를 추적하지 못했다.

종합하면, 본 연구는 GenAI-HITL 방식이 서술형 평가 개발 과정에서 일정 수준의 효율성을 확보하면서도 교육과정 정합성과 구조적 체계성을 유지할 수 있음을 보여주는 근거를 제공하였다. 다만 GenAI가 평가 개발의 구조적 기반을 제시할 수 있다는 점과 별개로, 이를 학습자에게 의미 있는 교육적 경험으로 연결하는 조정 과정은 교사의 맥락적 전문성에 상당 부분 의존하는 것으로 해석된다. 본 연구에서 제안한 프로토콜과 논의는 향후 AI 기반 평가 개발 연구에서 하나의 참고 틀로 활용될 수 있으며, GenAI 시대 교사 전문성의 범위와 역할을 재검토하는 논의에도 제한적이나마 기여할 수 있을 것이다.

* 본 논문은 2025.10.31. 투고되었으며, 2025.11.10. 심사가 시작되어 2025.12.03. 심사가 종료되었음.

참고문헌

경기도교육청(2024),『학생의 사고력과 문제해결력을 키우는 중등 논술형 평가 길라잡이』, 수원: 경기도교육청.

곽선영(2025),「17개 교육청의 서·논술형 평가 지침 비교」,『함께 여는 국어교육』157, 84-97.

교육부(1992),『제6차 교육과정(교육부 고시 제1992-11호)』, 서울: 교육부.

교육부(2022),『2022 개정 국어과 교육과정(교육과정 고시 제2022-33호)』, 세종: 교육부.

권태현(2021),「국어과 평가의 문제점과 체계화 방안 – 수행과 지필 평가의 균형적 접근을 중심으로」,『어문론집』85, 359-394.

김경희(2020),「서·논술형 평가의 평가학적 의미 탐색」,『교육평가연구』33(4), 839-862.

김선·반재천(2023),『사고력 함양을 위한 서·논술형 평가 도구 개발 이론과 실제』, 대전: AMEC.

김형성(2023),「국어 교사의 논술형 평가 전문성 검사 도구 개발」,『새국어교육』136, 167-208.

남민우·이상일·최숙기·서수현·남가영·정민주(2022),「국어과 평가 문항의 양호도 분석틀 개발을 위한 기초 연구」,『청람어문교육』86, 71-95.

박고운·최숙기(2025ㄱ).「국어과 읽기 영역 선다형 평가를 위한 자동 문항 생성 방안 연구」,『교육과정평가연구』28(1), 215-246.

박고운·최숙기(2025ㄴ),「GAI-HITL 기반 독서 문항 자동 생성(AIG)의 심리측정학적 타당성 분석 연구」,『교육과정평가연구』28(3), 319-359.

박도순(2025),「서·논술형 평가 시행에 관한 고찰」,『함께 여는 국어교육』157, 242-247.

박종임(2024),「국어과 서·논술형 평가의 도입 현황 및 실행 상의 쟁점 탐색 연구」,『청람어문교육』101, 273-307.

박종임·이상하·송민호·이문복·이민정·최숙기(2022),『컴퓨터 기반 서·논술형 평가를 위한 자동채점 방안 설계(Ⅰ)(RRE 2022-6)』, 진천: 한국교육과정평가원.

박혜영·김성숙·김경희·이명진·김광규·김지영(2019),『수업–평가 연계 강화를 통한 서·논술형 평가 내실화 방안(RRE 2019-6)』, 진천: 한국교육과정평가원.

서울특별시교육청(2022),『서·논술형 평가도구 자료집(국어과)(ORM 2022-150-1)』, 진천: 한국교육과정평가원.

성태제(2019),『교육평가의 기초』, 서울: 학지사.

송슬기(2024),「깊이 있는 학습을 위한 필요조건으로서의 논술형 평가의 특징과 지원 방향에 관한 탐색」,『교육문화연구』30(4), 149-172.

장성민(2021),「대학수학능력시험 서·논술형 평가 도입의 철학적 정당화와 방향 탐색」,『작문연구』51, 117-151.

장성민(2024),「도구 교과로서의 역할을 고려한 표현론적 관점에서의 학문 문식성 구체화 방향 탐색: 수능 서·논술형 문항 설계를 위한 논증 과제 분류를 중심으로」,『작문연구』62, 51-90.

정민주·서수현·남민우·최숙기·이상일·남가영(2022),「좋은 국어과 평가 문항 특성에 관한 질적 분석 연구: 국어과 평가 문항 양호도 분석틀 개발 연구(2)」,『청람어문교육』89, 43-78.

최숙기(2021),「서·논술형 수능 도입을 대비한 2022 개정 국어과 교육과정의 개정 방향 탐색」,『청람어문교육』83, 129-156.

최숙기(2023),「국어과 서·논술형 수능 평가 문항 개발 방안 연구」,『청람어문교육』91, 135-178.

최숙기·박종임(2023),「2022 개정 국어과 교육과정 〈독서와 작문〉 교육과정 개발의 원리와 방향」,『작문연구』57, 165-199.

최숙기·박종임(2024),「생성형 AI를 활용한 현직 국어교사의 서·논술형 평가 문항 개발 양상 분석」,『청람어문교육』97, 243-270.

학생평가지원포털(2024. 12. 31.), 서·논술형 평가 도구 개발의 방법과 사례, 학생평가지원포털, 검색일자 2025. 6. 23., 사이트 주소 https://stas.moe.go.kr/bbs/artcl/artclDtl:EVAL_TASK_DEV_S3?page=0&size=10&redraw=&totalPages=6&sBbsId=EVAL_TASK_DEV_S3&sArtclSeq=500658&sFileKey=&sCprtYn=Y&sCond=ARTCL_TITLE&sWord=

한국교육방송공사(2025),『2026학년도 수능특강: 국어영역 독서』, 서울: 한국교육방송공사.

함은혜·박소영·이병윤·김기동·이대형(2024),「GPT를 활용한 서술형 문항 생성 프로토콜과 문항의 질 평가: 국어과 사례를 중심으로」,『교육학연구』62(8), 63-93.

Anderson, L. W. & Krathwohl, D. R. (Eds.). (2001), *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*, New York, NY: Longman.

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022), "The interactive reading task: Transformer-based automatic item generation", *Frontiers in Artificial Intelligence 5*, 903077.

Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, M. (2021), "On the dangers of stochastic parrots: Can language models be too big?", *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623, New York: Association for Computing Machinery.

Bozkurt, A. (2024), "Tell me your prompts and I will make them true: The alchemy of prompt engineering and generative AI", *Open Praxis 16*(2), 111-118.

Brown, A. L. (1987), "Metacognition, executive control, self-regulation, and other more mysterious mechanisms", In F. E. Weinert & R. H. Kluwe (Eds.), *Metacognition, motivation, and understanding* (pp. 65-116), Hillsdale, NJ: Lawrence Erlbaum Associates.

Circi, R., Hicks, J., & Sikali, E. (2023), "Automatic item generation: Foundations and machine-learning-based approaches for assessments", *Frontiers in Education 8*, 858273.

College Board. (2019), *AP Seminar - End-of-Course Exam Scoring Guidelines*, New York, NY: The College Board.

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2024), "Chain-of-verification reduces hallucination in large language models", *Findings of the Association for Computational Linguistics: ACL 2024*, 3563-3578.

Eager, B. & Brunton, R. (2023), "Prompting higher education towards AI-augmented teaching and learning practice", *Journal of University Teaching & Learning Practice 20*(5), Article 2.

Fitzgerald, J. & Shanahan, T. (2000), "Reading and writing relations and their development", *Educational Psychologist 35*(1), 39-50.

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., Jones, A., Bowman, S., Chen, A., Lieberum, T., DasSarma, N., Drain, D., Li, D., Tran-Johnson, E., Hernandez, D., Kaplan, J., & Clark, J. (2022), "Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned", *arXiv preprint. arXiv:2209.07858*.

Google DeepMind. (2025, June 27), Gemini 2.5 Pro: Model card, Google Cloud Storage. https://storage.googleapis.com/model-cards/documents/gemini-2.5-pro.pdf

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023), "Survey of hallucination in natural language generation", *ACM Computing Surveys 55*(12), Article 248, 1-38.

Kan, A. & Bulut, O. (2014), "Crossed random-effect modelling: Examining the effects of teacher experience and rubric use in performance assessments", *Eurasian Journal of Educational Research 57*, 1-28.

Kane, M. T. (2013), "Validating the interpretations and uses of test scores", *Journal of Educational Measurement 50*(1), 1-73.

Kharrufa, A. & Johnson, I. G. (2024), "The Potential and Implications of Generative AI on HCI Education", *In Proceedings of the 6th Annual Symposium on HCI Education (EduCHI '24)*, Association for Computing Machinery, New York, NY, USA, Article 10, 1-8

Koo, T. K. & Li, M. Y. (2016), "A guideline of selecting and reporting intraclass correlation coefficients for reliability research", *Journal of Chiropractic Medicine 15*(2), 155-163.

Kubiszyn, T. & Borich, G. D. (2013), *Educational testing and measurement: Classroom application and practice*(10th ed.), Hoboken, NJ: John Wiley & Sons.

Kultusministerkonferenz. (2002), *Einheitliche Prüfungsanforderungen in der Abiturprüfung Deutsch*, Bonn: Kultusministerkonferenz.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020), "Retrieval-augmented generation for knowledge-intensive NLP tasks", *Advances in Neural Information Processing Systems 33*, 9459-9474.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., & Cobbe, K. (2023), "Let's verify step by step", arXiv preprint arXiv:2305.20050.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Majumder, B. P., Gupta, S., Hermann, K., Welleck, S., Yazdanbakhsh, A., & Clark, P. (2023), "Self-Refine: Iterative refinement with self-feedback", arXiv preprint arXiv:2305.17651.

McCaffrey, D. F., Casabianca, J. M., Ricker-Pedley, K. L., Lawless, R., & Wendler, C. (2022), "Best practices for constructed-response scoring", *ETS Research Report Series* RR-22-17, Princeton, NJ: Educational Testing Service.

McMillan, J. H. (2014), *Classroom assessment: Principles and practice for effective standards-based instruction* (6th ed.), Boston: Pearson.

Memarian, B. & Doleck, T. (2024), "Human-in-the-loop in artificial intelligence in education: A review and entity-relationship (ER) analysis", *Computers in Human Behavior: Artificial Humans 2*(1), 100053.

Messick, S. (1989), "Validity", In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103), New York, NY: Macmillan.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2013), *Measurement and assessment in teaching* (11th ed.), Upper Saddle River, NJ: Pearson Education.

OpenAI. (2025, April 16), *OpenAI o3 and o4-mini: System card*, San Francisco: OpenAI. https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf

Qian, Y. (2025), "Prompt engineering in education: A systematic review of approaches and educational applications", *Journal of Educational Computing Research 63(7-8)*.

Schön, D. A. (1983), *The reflective practitioner: How professionals think in action*, New York, NY: Basic Books.

Shah, C. (2024), "From prompt engineering to prompt science with human in the loop", arXiv preprint arXiv:2401.04122.

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

Tavakol, M. & Dennick, R. (2011), "Making sense of Cronbach's alpha", *International Journal of Medical Education 2*, 53-55.

U.S. Department of Education. (2023), *Artificial intelligence and the future of teaching*

and learning: Insights and recommendations, Washington, DC: Office of Educational Technology.

Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R. K. W., & Lim, E. P. (2023), "Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models", In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2609-2634, Association for Computational Linguistics.

Webb, N. L. (2009), *Webb's Depth-of-Knowledge Guide: Career and technical education definitions*, Madison, WI: Wisconsin Center for Education Research, University of Wisconsin-Madison.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022), "Chain-of-Thought prompting elicits reasoning in large language models", *Proceedings of the 36th Conference on Neural Information Processing Systems(NeurIPS 2022)* 1800, 24824-24837.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023), "A prompt pattern catalog to enhance prompt engineering with ChatGPT", arXiv preprint arXiv:2302.11382.

World Economic Forum. (2024), *Shaping the future of learning: The role of AI in Education 4.0*, Geneva: World Economic Forum.

Zanzotto, F. M. (2019), "Viewpoint: Human-in-the-loop artificial intelligence", *Journal of Artificial Intelligence Research 64*(1), 243-252.

초록

GenAI-HITL 기반 '독서와 작문' 연계 서술형 평가 과제 개발 및 타당성 검토

박고운

본 연구는 생성형 인공지능(Generative AI)-인간 전문가 협력(HITL) 방식으로 2022 개정 국어과 '독서와 작문' 성취기준에 부합하는 서술형 평가 과제를 개발하고, 전문가 준거 기반 평정을 통해 교육적 타당성에 관한 예비적 타당화 근거를 제시하였다. 컨텍스트 엔지니어링과 사고연쇄(Chain of Thought)를 통합한 3단계 프로토콜을 설계하여 지문-문항-루브릭-해설이 연계된 과제를 생성하였다. 생성 결과물은 전국 13개 시도 현직 국어교사 18인이 3개 영역 12개 항목으로 평가하였다. 정량 분석 결과, 전체 평균은 4.32점으로 나타났으며 성취기준 부합도와 구조적 체계성에서 높은 평정을 받았다. 평가 도구의 내적 일관성과 평정자 간 합치도 역시 양호한 수준으로 확인되었다. 반면 학습자 수준 적합성은 상대적으로 낮아, AI가 형식화 가능한 교육과정 요소는 구현하되 학습자 발달 단계 및 학급 맥락과 같은 비형식적 요소 반영에는 한계가 있음을 시사하였다. 본 연구는 AI 생성 과제가 교사가 조정 가능한 초안으로 기능할 수 있음을 확인하였으며, 형식적 타당성과 실질적 타당성의 균형이 인간-AI 협력을 통해 강화될 가능성을 제시한다.

핵심어 자동 문항 생성, 생성형 AI, 인간-AI 협력, 서술형 평가, 독서와 작문, 2022 개정 교육과정

ABSTRACT

GenAI-HITL Development and Validity Review of Reading-Writing Constructed-Response Tasks

Park Goun

This study developed constructed-response tasks aligned with the 2022 revised Korean Language Arts ("Reading and Writing") achievement standards using a Generative AI-Human-in-the-Loop (HITL) approach and reported preliminary validity evidence from expert-criterion ratings. A three-stage protocol integrating context engineering and Chain-of-Thought generated a linked package of texts, prompts, rubrics, and explanations. Eighteen in-service Korean language teachers from 13 regions rated the outputs on 12 items across three domains. The overall mean was high (M = 4.32), with the strongest ratings for standard alignment and structural coherence; internal consistency and inter-rater agreement were acceptable. Learner-level appropriateness was lower, indicating limits in capturing non-formal factors (developmental stage, classroom context) despite effective operationalisation of formalised curriculum elements. The findings suggest AI outputs can serve as teacher-adjustable drafts, and that human-AI collaboration may strengthen the balance between formal and substantive validity.

KEYWORDS Automatic item generation, generative AI, human–AI collaboration, descriptive assessment, reading and writing