

국어과 읽기 영역 선다형 평가를 위한 자동 문항 생성 방안 연구

박고운 (한국교원대학교 국어교육전공 박사과정)*
최숙기 (한국교원대학교 국어교육과 교수)**

요약

본 연구는 국어과 읽기 영역 선다형 평가 문항의 자동 생성을 위한 생성형 AI 기반 자동 문항 생성(AIG) 기술의 가능성을 탐색하였다. 이를 위해 제로샷(Zero-shot), 퓨샷(Few-shot), Chain-of-Thought(CoT) 세 가지 프롬프트 전략의 효과성을 비교 분석하였다. 각 전략으로 생성된 문항은 문항 타당성, 문항의 형식, 문항 내용을 기준으로 8명의 전문가 집단이 질적 평가를 실시하였다. 연구 결과, CoT 프롬프트가 가장 우수한 성과를 보였다. CoT 프롬프트는 AI의 단계적 사고를 유도함으로써 사실적 이해부터 창의적 이해에 이르는 다양한 평가 요소를 균형 있게 반영하였으며, 문항의 논리적 일관성과 평가 목적의 부합성을 효과적으로 구현하였다. 반면 퓨샷 프롬프트는 형식적 완결성은 갖추었으나 창의성과 다양성이 부족했고, 제로샷 프롬프트는 단순 정보 재구성에 치중하여 고차원적 사고를 평가하기에는 한계를 보였다. 이러한 결과는 생성형 AI 기술이 교사의 평가 전문성을 보완하고 문항 설계의 논리성과 타당성을 높이는 보조 도구로서 기능할 수 있음을 시사한다. 특히 CoT 프롬프트는 교사의 문항 설계 부담을 경감하고 평가의 질적 향상에 기여할 수 있는 가능성을 보여주었다. 본 연구는 CoT 프롬프트의 효과성을 실증적으로 입증하고, AIG 기술이 교육 평가의 질적 향상과 교사의 전문성 보조 도구로 활용될 수 있는 가능성을 제시했다는 점에서 의의가 있다.

주제어: 자동 문항 생성(AIG), 생성형 AI, 프롬프트 엔지니어링, 제로샷(Zero-shot), 퓨샷(Few-shot), Chain-of-Thought(CoT), 국어과 읽기 영역 선다형 평가 문항, Human-in-the-Loop(HITL)

* 제1저자, koniholi@gmail.com

** 교신저자, agreement@knue.ac.kr

I. 서 론

국어 교사의 평가 문항 개발은 다양한 시험 평가 상황에서 상당한 시간과 비용, 노동력을 요구할 뿐만 아니라, 개별 교사의 평가 역량에 따라 그 과정과 결과물의 질적 수준에서 큰 편차를 보인다(Haladyna & Rodriguez, 2013). 특히 학교 현장에서 폭넓게 활용되는 국어과 선다형 평가는 신뢰도 측면에서는 강점을 가지고 있으나, 고차원적 사고력을 포함한 독서 평가의 복합적 요구(Afflerbach, 2017)를 충족하기에는 근본적인 한계를 지닌다. 그럼에도 학교 지필평가에서는 학습자의 독해 역량을 간접적으로 측정하기 위해 선다형 평가 문항이 지속적으로 활용되고 있다. 이러한 선다형 문항의 타당성을 확보하기 위해서는 개별 교사의 높은 전문성과 상당한 시간 투자가 필요하다. Kissi 외(2023)는 교사의 선다형 문항 작성 역량이 문항의 내용 타당도(content validity)와 품질에 직접적으로 영향을 미친다고 강조하였다. 또한, Zimmaro(2010)는 선다형 문항이 객관적이고 효율적인 평가 도구로 활용될 수 있지만, 고품질 문항을 작성하기 위해서는 세심한 설계와 검토 과정이 필수적이라고 지적하였다. 특히, 효과적인 선다형 문항을 개발하려면 학습 목표에 부합하는 명확한 문항 설계와 다양한 난이도를 고려한 정교한 출제 과정이 요구된다.

이러한 문제를 해결하기 위한 방안으로 자동 문항 생성(Automated Item Generation, 이하 AIG) 기술이 주목받고 있다. AIG는 인지 및 심리 측정 이론을 기반으로 컴퓨터 기술을 활용하여 평가 문항을 자동으로 생성하는 기술을 의미한다(Hornke & Habon, 1986; Embretson & Yang, 2007; Gierl & Haladyna, 2012). 이는 평가 문항의 생산성을 높이는 동시에 신뢰성과 공정성을 확보할 수 있는 도구로서, 특히 교사들의 문항 개발 전문성 부족으로 인한 문항 질 저하 문제를 해결할 수 있는 대안으로 제시되어 왔다(Tarrant et al., 2006). AIG 기술의 발전은 교육적 활용 가능성을 크게 확장시키고 문항 개발에 소요되는 시간을 줄이며, 효율적으로 난이도나 측정 요소에 대응하는 문항 개발의 가능성을 확장시키는 데 기여한다(Kurdi et al., 2020).

그러나 기존의 AIG 연구는 주로 템플릿 기반 접근법(Gierl & Lai, 2013)에 의존해 왔으며, 이는 생성된 문항의 유형이 제한적이고 여전히 인간 전문가의 초기 작업에 의존해야 하는 한계를 보였다(von Davier, 2018; Kosh et al., 2019). 이러한 한계를 극복할 수 있는 새로운 가능성으로 생성형 AI 기술이 주목받고 있다. AI 기반 자연어 처리(NLP) 기술의 발전, 특히 GPT-3 및 그 후속 모델(GPT-4 등)과 같은 대규모 언어 모델의 등장은 템플릿 없이도 고품질 문항을 생성할 수 있는 혁신적인 전환점이 되었다. 이러한 모델들은 대량의 텍스트 데이터를 사전 학습한 후 특정 작업에 대한 소량의 예

시만으로도 원하는 과제를 수행할 수 있는 ‘퓨샷 학습(few-shot learning)’ 능력을 보여주며(Brown et al., 2020), 교육 평가 영역에서 획기적인 변화를 가져오고 있다.

생성형 AI 기술은 기존 템플릿 기반 접근법의 한계를 효과적으로 극복하며 다음과 같은 세 가지 강점을 보여준다. 첫째, 인간 전문가가 작성한 템플릿 없이도 다양한 유형의 문항을 자유롭게 생성할 수 있다. 둘째, 문항 작성 속도를 획기적으로 높이면서도 일정 수준 이상의 품질을 안정적으로 유지할 수 있다. 셋째, 사전 학습된 방대한 데이터를 기반으로 다양한 언어와 문화적 맥락에 적합한 문항을 생성할 수 있어 평가의 공정성과 접근성 향상에 기여한다(von Davier, 2018; Shin, 2021).

이러한 생성형 AI의 가능성은 교육 평가, 특히 독해 평가 분야에서 구체적인 성과로 입증되고 있다. von Davier(2019)는 GPT-2 모델을 활용하여 의도 면허 시험과 유사한 복잡한 문항을 생성하는 데 성공했으며, 이는 문항 작성의 초기 단계를 자동화함으로써 문항 작성 시간을 대폭 단축하면서도 평가 목표에 부합하는 고품질 문항 생성이 가능함을 보여주었다. 또한 Hommel 외(2021)의 GPT-3를 활용한 성격 평가 문항 개발 연구는, GPT-3가 인간 전문가의 초안을 기반으로 문항을 확장하거나 수정함으로써 평가의 신뢰성을 유지하면서도 문항 개발 과정을 효율화할 수 있음을 입증하였다.

또한 국제 독해력 평가 프로그램인 PIRLS(Progress in International Reading Literacy Study)에서도 생성형 AI의 활용 가능성이 확인되고 있다. Brown 외(2020)는 GPT-3를 활용하여 PIRLS의 문학적·정보적 텍스트 특성을 반영한 독해 자료를 생성하고 이에 기반한 문항을 개발하였다. 연구진들은 생성된 문항의 난이도, 맥락 적합성, 교육적 유용성에 대한 전문가 검토를 통해 생성형 AI가 인간 작성 문항과 비견할 만한 수준의 품질을 유지할 수 있음을 입증하였다. 특히 주목할 만한 것은 최근의 연구들이 보여주는 다양한 접근 방식이다. Shin(2021)은 템플릿 기반 및 비템플릿 기반 기법을 결합하여 문학 작품 기반의 독해 추론 테스트 문항을 개발하였고, Fung 외(2020)는 T5 모델을 활용하여 독해 문항과 문법 관련 다지선다형 문항을 생성하는 웹 시스템을 구축하였다. 또한 Duolingo의 영어 시험(DET)은 GPT-3를 활용하여 설명문과 서사문을 포함한 독해 평가를 자동으로 생성하는 성과를 보여주었다(Attali & Burstein & Tetreault, 2022).

생성형 AI를 활용한 AIG 연구는 국제적으로 활발히 진행되고 있는 반면, 국내 연구는 아직 초기 단계에 머물러 있다. 오규설(2022)이 사실적 독해 문항 모형을 중심으로 자동문항생성 기법의 구체적 적용 방안을 제시한 것을 시작으로, 이하늘, 이용상(2024)은 외국어로서의 한국어 읽기 평가를 위한 AIG 활용 방안을 탐색하였다. 최근 허동석 외(2024)는 GPT-4 모델 API와 커스텀 데이터베이스를 활용한 ‘신속문답생성기’ 시스템을 개발하여 실제적 적용을 시도하였다. 그러나 이러한 연구들은 실제 교육 현장에서 교사들이 직면하는 복합적인 문제들, 즉 지문과 연계된 다수 문항의 생성, 문항 간 간접

현상 방지, 문항 오류 최소화 등의 과제(장성민, 2024)를 해결하는 데는 아직 한계를 보이고 있다.

이러한 맥락에서 본 연구는 생성형 AI를 활용한 AIG 기술의 확장 가능성을 탐색하고자 한다. 특히 제로샷(Zero-shot), 퓨샷(Few-shot), CoT(Chain-of-Thought) 세 가지 프롬프트 유형별로 국어과 읽기 영역 선다형 세트(사실, 추론, 비판, 적용 및 창의) 문항 개발의 가능성을 탐색하고, 현직 국어 교사 및 국어교육 전문가 평가를 통해 생성된 문항의 질을 검증하고자 한다. 본 연구의 구체적인 연구 문제는 다음과 같다.

첫째, 생성형 AI 기반 AIG 기술을 통해 국어과 읽기 영역 평가의 특성을 반영한 세트 문항을 생성할 수 있는가?

둘째, 국어과 읽기 영역 선다형 평가 세트 문항 생성을 위한 CoT, 제로샷, 퓨샷 프롬프트 중 가장 적합한 전략은 무엇인가?

셋째, 생성형 AI 기반 AIG 기술이 국어과 읽기 영역 선다형 평가 세트 문항 생성 과정에서 드러난 주요 한계는 무엇인가?

이 같은 연구를 통해 학교 현장에서 교사의 문항 개발 전문성을 보조할 수 있는 생성형 AI의 도구적 가능성을 확인하고, 향후 국어과 읽기 영역 선다형 평가 문항 개발의 새로운 방향을 제시할 수 있을 것으로 기대한다.

II. 이론적 배경

자동 문항 생성(AIG)의 체계적 접근은 von Davier(2018)의 신경망 기반 연구에서 시작되었다. 그는 AIG 과정을 사전 준비, 문항 생성, 검토 및 수정, 문항 최적화의 단계로 체계화했는데, 이는 이후 GPT 모델을 활용한 의로 인증 문항 생성 연구(von Davier, 2019)를 통해 더욱 구체화되었다. 사전 준비 단계에서는 평가의 목적, 응시자의 연령 및 능력 수준, 평가하고자 하는 학습 목표를 명확히 정의하고, 대규모 언어 모델에 입력할 프롬프트를 설계한다. 문항 생성 단계에서는 독해 지문을 생성하고 이에 따른 문항과 선택지를 작성하며, 검토 및 수정 단계에서는 전문가 검토를 통해 평가 목표 부합도, 문법적 정확성, 내용 적합성, 논리적 일관성, 난이도 조정, 편향 제거 등을 점검한다. 마지막 문항 최적화 단계에서는 검토 결과를 반영하여 평가 목적에 맞게 문항을 확정한다.

이러한 AIG 절차에서 프롬프트 설계는 생성형 AI 기반 문항 생성의 핵심 요소로 부각된다. Brown 외(2020)는 GPT-3의 few-shot 학습 능력을 입증하며, 프롬프트 입력 방식이 모델 출력 품질에 중요한 영향을 미칠 수 있음을 시사한다. 프롬프트는 AI 모델

에게 특정 작업을 수행하도록 지시하는 입력 형태로, 문항 유형, 난이도, 평가 목표 등을 명확히 정의함으로써 AI의 출력을 제어한다. 적절하지 않은 프롬프트 설계는 생성되는 문항의 적합성, 일관성, 난이도가 평가 목적과 부합하지 않을 위험을 초래할 수 있어, AIG의 성공적 구현과 평가 신뢰성을 확보하기 위해 정교한 프롬프트 설계가 필수적이다.

이와 같은 맥락에서 Attali 외(2022)는 Transformer 기반 언어 모델을 활용하여 독해 평가 문항을 자동 생성하는 Interactive Reading Task(IR Task)를 개발하였다. 이 연구는 대형 언어 모델과 심리측정학적 접근법을 결합한 컴퓨테이셔널 심리측정학(computational psychometrics)을 통해 독해 문항 생성을 체계적으로 탐구하였다. 연구진은 반복적인 파일럿 테스트와 대규모 파일럿 연구를 통해 생성된 문항의 품질을 인간 전문가와 심리측정학적 관점에서 평가하였으며, 이를 통해 자동 생성된 문항이 인간 작성 문항에 비견할 만한 수준의 품질을 유지할 수 있음을 입증하였다.

특히 Attali 외(2022)는 프롬프트 설계를 통해 독해 지문과 문항 유형 간의 적합성을 높이는 방법론을 제시하였다. 이들은 Transformer 모델이 긴 맥락을 처리하고 텍스트 내 장거리 어휘 관계를 효과적으로 활용할 수 있다는 점에 주목하여, 독해 지문과 연계된 다지선다형(MC) 문항 및 속성 질문(attribute questions)을 생성하였다. 이러한 접근은 기존 템플릿 기반 접근법보다 더 복잡하고 정보가 풍부한 문항 생성을 가능하게 하였다.

한편, Bezirhan & von Davier(2023)는 PIRLS 평가 지문 생성을 위한 체계적인 프롬프트 설계 연구를 수행하였다. 이들은 목표 설정, 텍스트 유형 명시, 난이도 수준, 주제, 어휘 수준, 문장 길이 등 프롬프트 구성 요소를 체계화하며 AI 기반 자동 문항 생성(AIG)이 국제 학업성취도 평가에서 실질적으로 활용될 가능성을 보여주었다.

프롬프트 설계와 더불어 학습 방식 또한 AIG 성과에 중요한 영향을 미친다. Brown 외(2020)는 퓨샷 학습이 다양한 과제에서 우수한 성능을 보임을 입증하였으며, 이는 독해 평가 문항 생성에서도 높은 활용 가능성을 보여준다. 제로샷 방식은 지시 사항만으로 문항을 생성하는 반면, 퓨샷 방식은 예시를 함께 제공하여 모델이 문항의 구조와 스타일을 학습하도록 한다. Attali 외(2022) 역시 퓨샷 학습 방식을 활용하여 다양한 유형의 독해 평가 문항 생성을 성공적으로 구현하며 그 효과성을 입증하였다.

더 나아가 최근의 프롬프트 연구는 Chain-of-Thought(CoT) 전략이라는 새로운 접근법을 보여주고 있다. Wei 외(2022)와 Huang 외(2022)는 CoT 프롬프트 전략이 대규모 언어 모델의 성능을 획기적으로 향상시킬 수 있음을 발견하였다. 중간 추론 단계를 명시적으로 제공하여 모델이 단계별 논리적 사고를 수행하도록 하는 CoT 방식은 본 연구의 프롬프트 설계에도 중요한 시사점을 제공한다. <표 1>은 이러한 최근 프롬프트 연구 동향을 정리한 것으로, 구체적인 CoT 프롬프트 구조와 설계 방안은 다음 장(Ⅲ. 연구 방법)에서 상세히 다루도록 하겠다.

<표 1> 프롬프트 연구 동향

연구자	연구 주제	주요 내용	핵심 기법	시사점
Das et al. (2021)	AI 사고 과정 추적	AI의 논리적 진행 과정을 단계별로 추적	단계적 추적	연구자의 개입 기회 제공
Wei et al. (2022)	Chain-of-Thought (CoT) 프롬프트	복잡한 문제 해결 과정에서 언어 모델 성능 향상	중간 추론 단계 제공	문제 해결의 정확성과 일관성 증대
Huang et al. (2022)	Inner Monologue	언어 모델의 중간 피드백 반영	중간 피드백 메커니즘	정교한 계획 및 추론 수행
Shah (2024)	AI 문항 생성 검증	AI 문항 생성의 질적 수준 향상	다단계 검증	AI 문항 생성의 체계적 접근

III. 연구 방법

1. 프롬프트 설계

본 연구는 AI를 활용한 국어과 읽기 영역의 선다형 문항 생성에 적합한 프롬프트 유형을 탐색하기 위해 ChatGPT 4o를 활용하여 세가지 프롬프트 기법을 비교 분석하였다. 이를 위해 제로샷(Zero-Shot), 퓨샷(Few-Shot), CoT(Chain-of-Thought) 프롬프트의 효과를 평가하고, AI 기반 문항 생성(AIG)에서 보다 효과적인 프롬프트 전략을 탐색하는 것을 그 목적으로 한다.

선행 연구에 따르면 문항의 질적 향상을 위해 제로샷과 퓨샷 프롬프트만을 비교하는 것은 한계가 있을 수 있으며, 전문가가 만족할 만한 결과를 얻기 어려울 가능성이 제기되었다. 또한 단순히 프롬프트를 결합하는 방식은 문항의 질에 부정적인 영향을 미칠 수 있어, 초기부터 오류 가능성을 줄일 수 있는 프롬프트 설계가 필요하다는 점이 논의되었다(허동석 외, 2024). 프롬프트의 정교한 조정이 문항 생성의 품질 향상에 기여할 가능성이 있음을 여러 연구에서 언급한 바 있다(신동광, 2023, 2024; 최숙기, 박종임, 2024; 허동석 외, 2024; 임상묵 외, 2024; 이하늘 외, 2024). 이에 본 연구에서는 파일럿¹⁾ 단계를 거쳐 다양한 프롬프트를 적용한 문항 생성을 실험하였다.

1) 파일럿 과정에서 제로샷과, 일부 학습 자료를 제공하는 퓨샷을 비교한 결과, 두 가지 방법만으로는 기존 선행 연구들과 차별성을 확보하기 어렵다는 점이 확인되었다. 이에 따라, 다른 연구에서 본격적으로 활용된 사례가 적은 CoT(Chain-of-Thought) 프롬프트를 추가하여 본 연구의 차별성을 확보하고자 하였다. CoT 프롬프트는 일반적인 교사들의 출제 과정상의 인지 과정을 반영한 7단계 구조로 설계되었으며, 이를 통해 AI가 교사를 보조할 수 있는 보다 체계적이고 맥락적인 문항 생성 방식을 탐색할 수 있도록 설계의 목적을 두었다.

본 연구는 국어과 읽기 영역 문항 생성의 복잡성을 고려하여 CoT 전략을 활용함으로써, 계산 비용이 증가하더라도 논리적이고 체계적인 접근이 가능할 가능성이 있음을 검토하였다(Huang et al., 2022). 이는 설명 가능한 AI(XAI)의 구현을 위한 실험적 접근의 일환이며, 생성된 문항의 타당성을 평가하고 문제 요소를 명확히 분석하는 데 기여할 것으로 기대된다. 본 연구에서 비교된 각 프롬프트 유형은 AI의 문항 생성 능력을 다각도로 평가하고 보다 효과적인 프롬프트 전략을 탐색하는 데 활용될 수 있을 것으로 보인다.

제로샷 프롬프트는 AI가 사전 학습 없이 문항을 생성할 수 있도록 최소한의 정보를 제공하는 방식으로 설계되었다. 구체적으로, ① ‘10년 차 국어 교사 및 국어 독서 평가 출제 위원’이라는 페르소나를 설정하고, ② 문항 생성 조건으로 평가 기본 정보, 문항 구성 요건, 지문을 포함하였다. 이를 통해 AI가 국어과 읽기 영역 평가 문항을 생성하는 기본적인 수행 과정을 인지하도록 하였다.

퓨샷 프롬프트는 제로샷 프롬프트의 요소를 포함하면서, 추가적으로 ③ 기출 문항과 해설을 추가 제공하였다. 이는 AI가 기존 평가 문항의 구조적 특징과 출제 패턴을 학습하고 이를 모방하여 새로운 문항을 생성할 수 있도록 돕는다. 이러한 퓨샷은 기존 문항 패턴을 학습하여 재현하는 데 초점을 둔다.

CoT 프롬프트의 특이점은 퓨샷과 동일한 요소를 활용하고 있으나 ③을 제공하는 목적 면에서 차이가 있다. 이는 학습자의 문제 해결 사고 과정을 AI가 직접 수행하도록 단계적 과정에 활용되었다. 파일럿 실험을 통해 AI가 실제 고등학생들의 문제를 풀 때 틀리는 문제가 발생하였다는 점, 이것에 대한 원인을 스스로 분석하여 제시하였을 때 일반적인 학습자들의 어려움과 비슷하기 때문이다(예: 본문 내용의 일부를 꼼꼼하게 읽지 못했습니다. / 제가 ~한 문장을 오해해서 오답을 도출했습니다.). 이러한 반응을 고려하여 이 단계를 AI가 스스로 문제 해결을 학습자의 관점에서 경험하고, 이를 문항 생성 과정에 반영하는 전략으로 활용할 수 있을 것이라 판단하였다. 이는 현장의 숙련된 교사들 역시 출제를 위해 수행하는 선행 과정이다. 이러한 과정은 학습자의 수준을 예측하고 매력적인 오답을 형성하거나 문제 오류를 방지하기 위해 적극적으로 반영하는 인간의 전략이기 때문에 AI 역시 CoT 과정에서 반영함으로써 효과성을 높일 수 있을 것이다. 결과적으로, CoT 프롬프트는 퓨샷과 달리 ③의 요소는 AI가 문제 해결의 논리적 사고 절차를 직접 수행하고 재현하는 것에 초점을 둔다.

또한 CoT 프롬프트는 퓨샷 프롬프트와 동일한 구성 요소를 포함하면서도, 추가적으로 AI의 논리적 사고와 추론 능력을 극대화하기 위한 고차원적 사고 과정을 설계에 반영하였다. CoT 프롬프트는 단순한 데이터 모방 수준에서 벗어나 AI가 스스로 사고 과정을 단계별로 구현하도록 유도하였다. 이를 위해 ①, ②, ③의 요소는 단계별로 나누어

제공되었으며, AI가 출제를 위한 여러 맥락을 이해하고 활용할 수 있는 기반을 마련하는데 중점을 두어 설계하였다. <표 2>에서 CoT 프롬프트의 단계적 내용을 구체화하였으며, 그 구체적 구현 양상의 일부를 [그림 2]에서 제시하였다.

<표 2> 각 유형별 프롬프트 구성의 공통 및 차이 요소

프롬프트 상세 내용	제로 샷	퓨 샷	CoT
① [페르소나 설정] 당신은 고등학교 독서 수업을 담당하는 10년 차 국어 교사이며, 평가원에서 대단위 평가 출제 위원으로 활동하고 있는 평가 전문가입니다.	○	○	○
② [평가 기본 정보] 1. 평가 목적: 고등학교 2학년 학생들의 독해 능력 평가 측정 (지필평가 시험 맥락) 2. 평가 형태: 5지 선다형 (정답 1개, 오답 4개) [문항 구성 정보] 1. 하나의 지문에 대해 4개의 연계 문항 구성 2. 평가 요소별 문항 배분 및 난이도 지정 - 1번: 사실적 이해(하) - 2번: 추론적 이해(중) - 3번: 비판적 이해(상) - 4번: 적용 및 창의적 이해(상) 3. 문항 형식:[문항 번호] [문항 내용] 선지①②③④⑤ 정답: [정답 번호] [지문 제공] 1. 제공한 지문을 활용하여 문제를 출제할 것. 2. 지문의 특성을 고려하여 서로 중복되지 않은 문제를 출제할 것.	○	○	○
③ [기출 문항 및 해설 제공] 1. 동일한 문제를 출제하지 말 것. 2. 다양한 문형을 활용하여 문제를 출제할 것.		○	○

[그림 1]은 퓨샷과 차별되는 CoT 프롬프트의 세부 단계 및 의미를 시각화한 것이고, [그림 2]는 CoT 과정에서 나타난 일부 반응을 제시한 것이다. 본 연구의 CoT 프롬프트 설계는 메타인지적 지식과 경험의 상호작용을 강조한 Flavell(1979)의 메타인지 이론과 계획-수행-자기성찰의 순환적 자기조절 과정을 제시한 Zimmerman(1990)의 자기조절 학습 모델을 기반으로, AI의 문항 생성 과정에 인간의 고차원적 사고 과정을 구현하고자 하였다. 이러한 이론적 토대 위에서 AI가 전체 진행 과정을 체계적으로 이해하고 실행할 수 있도록 단계별 설계를 구현하였다.

우선 CoT 프롬프트의 초기 단계에서는 교육과정 성취 기준 정보를 제공하여 AI가 이

를 분석하도록 하였고, 이해한 내용을 향후 문제 분석 및 출제 과정에 적용하도록 지시하였다. 또한 출제할 지문을 제공하여 텍스트의 중심 내용과 구조를 파악하도록 하였다. 이후 기출 문제를 활용하여 학습자 관점에서 문항을 이해하고, 기출 문제 해설서를 통해 출제 의도를 비판적으로 검토하도록 하였다. 문항 생성 단계에서는 ‘선지 간섭 방지’, ‘지문 내용의 직접 활용 지양’, ‘복수 정답 가능성 제거’, ‘선지 길이의 균형 유지’ 등 세부 지시 사항을 포함하여 AI가 창의적이고 정교한 문항을 설계할 수 있도록 하였다. 자체 검증 단계에서는 생성된 문항의 타당성을 점검하고 필요시 개선하며, 최종 확정 단계에서는 해설과 함께 문항 세트를 완성하도록 구성하였다. 이러한 CoT 전략의 핵심은 단계적으로 진행한 내용들이 향후 수행에 일관성 있게 적용되게 하는 데 있다. 이는 파일럿 과정에서 발견한 AI의 한계를 종합하여 잠정적인 해결안으로써 정리한 내용을 단계에 반영한 것이다.

본 연구에서 설계한 CoT 프롬프트에 반영된 이론적 맥락은 다음과 같다. CoT 프롬프트는 여러 선행 연구에서 보고된 바와 같이 복잡한 논리적 사고를 요구하는 문항에서 AI의 추론 능력을 극대화하는 데 있다(Wei et al., 2022; Huang et al., 2022). 특히 자기 설명(self-explanation) 메커니즘의 일환인 내적 독백(Inner Monologue) 기법은(Huang et al., 2022) 현재 AI가 사고하는 지점과 문제점을 밝혀내고 중간에 인간(교사)이 개입(Das et al., 2021).하여 맥락을 재구성하는 데 도움을 준다. 이를 위해 본 연구에서 CoT의 모든 단계는 AI가 각 단계에서 이해한 내용과 문제 해결을 위해 반영한 사고 과정을 언어로 명확히 표현하도록 지시하였다. 이는 현재의 행위가 선행 단계와 꾸준히 연계성을 지니고 있는지 인간(교사)이 확인하는 데 도움을 준다.

더불어 다단계 검증(Shah, 2024)을 통해 AI 응답의 논리적 일관성을 확보하고자 하였다. 이는 복잡한 처리 과정에서 발생할 수 있는 맥락 상실과 일관성 부족 문제(Funk et al., 2023; Maity et al., 2024)를 보완하기 위한 전략이다. 이를 통해 AI가 문항 생성 과정에서 복잡한 사고 과정을 투명하게 구현하고, 인간 참여자가 AI의 성능과 한계를 체계적으로 평가할 수 있는 기반을 마련하였다.

메타인지 프레임 구축	핵심 구성요소	메타인지 구축 및 활성화
	세부 내용	페르소나 설정 및 전 단계 안내 문항 생성 조건 제공
	CoT 단계적 의미	문제 해결을 위한 인지적 준비 상태 구축
↓		
(1) 평가 요소 체계화	핵심 구성요소	교육과정 성취기준을 통한 평가 요소 추출
	세부 내용	중요 평가 요소 식별, 문항 설계 지침 수립
	CoT 단계적 의미	문제 해결을 위한 구조적 프레임 워크 형성
↓		
(2) 텍스트 구조 분석	핵심 구성요소	출제할 지문 분석
	세부 내용	중심 내용 파악, 논지 전개 방식 분석, 핵심 어휘 및 개념 정리
	CoT 단계적 의미	텍스트의 구조와 의미 이해
↓		
(3) 기출문제 해결 전략 분석	핵심 구성요소	지문에 따른 기출 문제 풀이
	세부 내용	기출 문제 직접 풀고 답안 선택 근거 제시.
	CoT 단계적 의미	기출 문제 풀이를 통한 지문 이해도 점검 및 학습자 수준 파악
↓		
(4) 출제 맥락 통합 분석	핵심 구성요소	기출문제 풀이에 대한 정·오답 및 출제 경향 파악.
	세부 내용	인간 출제자의 출제 의도 및 문항 분석.
	CoT 단계적 의미	인간 출제자의 출제 맥락을 고려하여 다음 단계 수행 전략 수립
↓		
(5) 맥락 기반 문항 설계	핵심 구성요소	조건에 따른 문항 생성
	세부 내용	문항 생성 조건 및 이전 단계에서 도출한 맥락과 정보 및 선지 구성 조건을 고려하여 창의적인 문제 생성
	CoT 단계적 의미	메타인지를 기반으로 이전 단계의 정보를 선별하고 맥락을 고려하여 목표 지향적 문제 해결
↓		
(6) 메타인지 기반 문항 평가	핵심 구성요소	문항 타당성 검토
	세부 내용	이전 지시 사항을 고려하여 스스로 평가를 위한 루브 릭을 생성하고 비판적 점검 및 문항 수정.
	CoT 단계적 의미	이전 단계와 이 단계의 연계성을 파악하고 자기 모니 터링 실시
↓		
(7) 통합적 문항 최적화	핵심 구성요소	최종 문항 완성
	세부 내용	해설 작성, 최종 점검표 구성.
	CoT 단계적 의미	전체 문제 해결 과정에 대한 메타 인지적 성찰

[그림 1] CoT 프롬프트 세부 단계 및 의미

1단계: 교육과정 성취기준 학습 및 분석
지시사항

다음은 비문학 문항 출제를 위한 2022 독서 영역 교육과정 성취기준 자료입니다. 문항 설계에 앞서 이를 학습하여 출제에 반영하세요.
성취기준에서 중요한 평가 요소 추출: 교육과정 성취기준에서 평가해야 할 주요 요소와 핵심 내용을 파악하세요.
문항 설계 지침 정리: 학습한 성취기준을 토대로 문항 설계에 반영할 수 있도록 정리하세요.
목적
이 성취기준 분석은 이후 모든 문항 설계 단계에서 참조 자료로 활용됩니다. 성취기준에 대한 명확한 이해를 바탕으로, 일관성 있고 학습자 수준에 맞는 문항을 생성할 수 있도록 합니다.

주요 평가 요소 및 핵심 내용

- 정보 파악과 추론: 글에 드러난 정보의 이해뿐만 아니라 드러나지 않은 정보의 추론 능력을 평가
- 비판적 평가: 글의 관점, 표현 방법, 필자의 의도 등을 평가하는 능력
- 조직 방법과 표현 전략 탐색: 다양한 글의 조직 방식과 표현 전략을 분석하고 이를 글쓰기에 활용하는 능력
- 사회문화적 맥락 이해: 사회적·역사적 현상이나 정점 등 맥락을 고려하여 비판적으로 읽는 능력
- 매체의 유형과 특성 고려: 다양한 매체의 특성을 이해하고 이에 따라 정보를 수집하고 전달하는 능력

문항 설계 지침

- 평가 목적과 기준: 성취기준에 맞춰 글의 정보를 정확하게 파악하고 비판적 사고를 발휘할 수 있도록 문항을 구성
- 문항 유형 다양화: 각 성취기준에 따른 다양한 문항 유형을 설계해 정보 이해, 비판적 평가, 추론 능력 등을 종합적으로 평가
- 맥락 반영: 사회적·문화적 맥락을 고려한 질문을 통해 학습자들이 비문학 지문을 폭넓게 이해하도록 유도
- 표현 방식과 전략의 분석 요구: 글의 표현 전략과 내용 조직 방식을 분석하는 질문을 포함하여, 이를 통해 학습자들이 논리적 사고를 발휘할 수 있도록 함

이 분석을 바탕으로 이후 단계에서 일관성 있고, 성취기준을 충실히 반영한 문항 설계가 가능할 것입니다.

교육과정 분석 및 AI 반응

앞서 제공한 출제를 위한 기본 정보들을 다시한번 숙지합니다.

출제를 위한 기본 정보를 다시 한 번 정리하여 숙지하였습니다.

평가 기본 정보

- 평가 목적: 고등학교 2학년 학생들의 독서 학습 능력 진단 (출발평가)
- 평가 대상: 고등학교 2학년 비문학 독해 수업을 받은 학생 250명
- 평가 형태: 5지 선다형 (정답 1개, 오답 4개)

문항 구성 요건

- 연계 문항 구성: 하나의 지문에 대해 4개의 연계 문항을 구성
- 평가 요소별 문항 배분:
 - 1번: 사실적 이해

문항 출제 전, 메모리의 한계를 고려하여 현재 수행 과정에 대한 중간 점검 지시와 AI 반응

확정한 문항에 대한 최종 점검표

[그림 2] CoT 프롬프트 세부 단계 예시

3단계: 기출문제 풀이 및 문항 구조 이해
지시사항

다음은 해당 지문에 대한 기출문제입니다. 기출문제를 풀어보고, 정답과 해설을 통해 구조와 평가 요소를 이해합니다.
기출문제 풀이: 스스로 문제를 풀고 선택한 답의 이유를 간단히 설명하세요.
정답 및 해설 검토: 정답과 해설을 참고하여 본인의 답과 비교합니다.
문항 구조 및 평가 요소 분석: 각 문항이 어떤 평가 요소(사실적 이해, 추론적 이해 등)를 평가하는지 확인하고, 선지 구성 방식을 분석하세요.
주의사항
이 단계에서는 학습자 입장에서 문제를 이해하는 데 주력하세요. 학습자가 흔히 겪는 오류와 어려움을 파악하는 것이 목표입니다.
이후에 정답지를 제공하겠습니다.

30번 문제: 적절하지 않은 내용 선택

- 선택지 풀이:
 - ① 틀린 선택지: 요역지가 승역지를 배타적으로 이용할 수 있다는 내용은 본문과 일치하지 않음.
 - ②~⑤ 적절한 선택지: 모두 본문 내용과 일치함.
- 정답: ①
- 평가 요소: 사실적 이해 - 지문에 명시된 내용을 정확하게 파악하는 능력을 평가.

31번 문제: 토지 소유권자에 대한 이해로 가장 적절한 선택

- 선택지 풀이:
 - ④ 정답 선택지: 지상물이 없는 토지라도 지상권 계약을 맺을 수 있음을 본문에서 설명.
 - 다른 선택지는 본문 내용과 다르거나, 오해의 소지가 있는 표현 포함.
- 정답: ④
- 평가 요소: 추론적 이해 - 지문 내용을 토대로 응용하고 추론하는 능력 평가.

이와 같이 문제를 풀어보고 해설을 검토하며, 평가 요소와 문항 구조를 이해할 수 있었습니다. 이러한 분석을 통해 학생들이 주로 겪을 수 있는 오류와 어려움을 파악하여 이후 문항 설계 시 참고하겠습니다.

기출문제 풀이 수행 후 AI 반응

최종 점검 결과

평가 항목	평가 내용	결과
교육 목표 부합성	평가 요소별 성취기준을 충실히 반영하여 목표에 부합하는 문항입니다.	우수
내용 적절성	문항이 지문의 핵심 내용에 충실하며 교육과정 성취기준에 부합하여 학습자의 평가에 적절합니다.	우수
난이도 조절	각 문항의 난이도 조절이 적절하게 이루어졌고, 오답 선지 구성이 정교하여 학습자 간의 변별이 용이합니다.	우수
표현의 명확성	지문의 내용을 패러프레이징하여 다양한 어휘로 표현하였으며, 선지 구성이 명확하고 직관적입니다.	우수
참의성 및 실재성	실생활에 연계된 사례와 자료를 활용한 문항 구성을 통해 학습자의 흥미와 참여도를 높였습니다.	우수

최종적으로, 이 문항 세트는 평가 요소와 난이도 조절이 적절히 반영되어 있으며, 지문과의 일관성을 유지하면서 다양한 사고 과정을 유도하도록 구성되었습니다.

2. 지문 및 문항 세트 선정

본 연구는 국어과 읽기 영역 평가에서 특정 분야의 지문을 선정하여 자동 문항 생성(AIG)의 적용 가능성을 심층적으로 분석하는 데 목적이 있다. 읽기 평가는 학습자의 비판적 사고력과 이해력을 확장하는 핵심 과정으로, 수능 국어 시험은 국가 수준의 국어 교육과정을 바탕으로 대학 학습에 필요한 독해 능력을 종합적으로 평가하는 도구로 활용된다(한국교육과정평가원, 2024). 특히 독서 영역은 다양한 분야의 글을 활용하여 학습자의 폭넓은 독해력과 추론 능력을 향상시키도록 설계되어 있으며, 이는 단순한 지식 습득을 넘어 고차원적 사고와 문제 해결력을 요구한다. 인문·예술, 사회·문화, 과학·기술 등 다양한 분야의 글은 각각 학습자의 분석적 사고, 역사적 통찰, 논리적 추론 능력을 강화하는 데 기여한다. 따라서 국어과 독서 평가에서 다양한 분야의 텍스트를 활용하는 것은 평가의 타당성과 공정성을 확보하는 중요한 전략이다.

그러나 본 연구는 AIG의 적용 가능성을 보다 면밀히 검토하기 위해 특정 읽기 영역 분야를 집중적으로 연구할 필요가 있다고 판단하였다. 이를 통해 문항 개발 과정에서 AIG가 가지는 실제적 효용과 한계를 명확히 분석하고자 한다. 이를 위해 본 연구는 ‘2024학년도 10월 고2 전국연합학력평가’ 국어 영역의 30~34번 문항 지문을 연구 대상으로 선정하였다(〈표 3〉 참조). 지문 선정 과정에서는 실제 평가 환경에서의 활용 가능성을 우선적으로 고려하였으며, 전문가 집단(현직 교사 3인)과의 사전 인터뷰를 통해 활용의 적절성을 검토하였다. 전문가들은 해당 지문이 실제 지필 평가에서 적용성과 적절성을 갖춘다고 판단하였다. 일부 전문가가 수능 기출 지문의 활용을 제안하기도 했으나, 연구 참여 교사들이 모두 고등학교 2학년을 담당하고 있다는 점과 수능 지문의 난도가 고2 학습자 수준에 비해 상대적으로 높다는 점을 고려하여 이를 배제하였다. 또한 본 연구는 활용된 지문과 문항을 절대적인 질적 우수성보다는 실제 교육 현장에서의 적용 가능성을 중심으로 선정하였다. 현직 교사들의 문항 출제 실태를 고려할 때, 교사들은 대체로 기존 공인된 지문을 활용하여 기출 문항을 변형하거나 평가 문항 유형을 참고하는 경향이 두드러진다. 이에 따라 본 연구는 기존 지문을 제공하는 방식을 채택하였으며, 이는 실제 교육 현장의 평가 실태를 반영하는 실용적인 접근이라고 판단하였다.²⁾

선정된 〈표 3〉의 지문은 모든 프롬프트에 공통적으로 제공되었으며, [그림 3]의 기출 문제는 퓨샷과 CoT 프롬프트의 학습 자료로 활용되었다. 앞서 설명한 바와 같이 퓨샷

2) 지문과 문항 생성의 필연적 연계성은 매우 중요하다. 이를 고려하여 연구 수행 당시(2024.11.)의 GPT-4.0를 활용한 연구 파일럿 단계에서 지문 생성과 문항 생성을 병행하는 방안을 실험적으로 검토하였으나, 생성된 지문이 출처가 불명확하거나 부정확한 정보를 포함하는 문제가 빈번하게 발생하였다. 이에 따라, 본 연구는 지문과 문항 생성을 개별적으로 수행하는 것이 보다 타당하다고 판단하였으며, 지문 생성 가능성 여부와 이를 문항에 연계하는 부분에 대한 요소는 향후 후속 연구에서 보다 정교하게 다뤄질 필요가 있다.

은 별도의 지시사항 없이 기출 문제를 제공 받았지만, CoT 프롬프트에서는 AI가 기출 문제를 참고하되 동일한 내용의 문제를 출제하지 않도록 명시적으로 지시하였다. 본 연구는 AI가 독해 능력에 대한 기본적 이해를 바탕으로 사실적 이해, 추론적 이해, 비판적 이해, 적용 및 창의적 이해를 반영한 문항을 생성하도록 설계하였다. 특히 생성형 AI의 주요 한계점인 응답 일관성 문제를 고려하여, Funk 외(2023)가 제안한 명시적 지시 사항 제공, 반복 검증 방식, 응답 형식의 표준화 방안을 적용하였다. 구체적으로 2024년 10월 한 달 동안 일정한 시간 간격을 두고 각 프롬프트 유형별로 5회 이상 반복하여 문항을 생성하였다. 이를 통해 각 프롬프트가 보여주는 응답 패턴의 특성을 확인하여 전문가 집단(국어 박사 과정 1명, 국어 교육 독서 전공 교수 2명)의 검토를 거쳐 최종 평가를 받을 각 프롬프트 별 대표 문항 세트를 선정하였다.

<표 3> 프롬프트 유형별 국어과 읽기 영역 선다형 평가 생성 문항 지문

[제시 지문]
<ul style="list-style-type: none"> • 민법에서 토지는 일정 범위의 지면과 정당한 이익이 있는 범위 내에서의 그 공중과 지하를 포함하는 것을 말한다. 토지 소유권자는 자신의 토지를 자유롭게 이용하면서 이를 통해 이익을 얻고, 매매, 상속 등의 처분을 할 수 있다. 그렇다면 건물을 짓거나 인공 구조물을 설치하는 등 토지를 이용하기 위해서는 반드시 그 토지에 대한 소유권이 필요할까? 한정된 자원인 토지를 효율적으로 이용하기 위해 민법에서는 다른 사람의 토지를 일정한 목적을 위해 이용하고, 이를 통해 이익을 얻을 수 있는 권리를 규정하고 있다. 지상권과 지역권 등이 이에 해당한다. • 지상권과 지역권은 모두 토지 소유권과는 별개로 설정할 수 있는 독립된 권리이다. 이들 권리는 당사자 간에 계약을 맺은 후, 권리관계를 다른 사람이 알 수 있도록 등기부에 목적과 범위 등의 필요한 내용을 기재하는 등기를 해야 효력이 발생한다. 하지만 두 권리는 권리를 설정한 목적과 그 특징에서 분명한 차이가 있다. • 먼저 지상권은 다른 사람의 토지에서 건물, 인공 구조물, 수목 등의 지상물을 소유하기 위해 그 토지를 이용할 수 있는 권리이다. 토지와 지상물은 별개의 부동산이므로 지상물을 안정적으로 소유하고 유지하기 위해서는 지상권 설정이 필요하다. 지상권의 목적은 지상물을 소유하기 위한 토지의 이용에 있으므로, 이 권리는 토지에 지상물이 현재는 없더라도 성립할 수 있다. • 지상권을 설정하기 위해서는 지상물을 소유하고자 하는 사람이 해당 토지의 소유권자와 지상권에 관한 계약을 맺고, 지상권자와 지상권설정자인 토지 소유권자가 공동으로 등기를 해야 한다. 이를 통해 지상권자는 지상권이 설정된 존속기간 동안 해당 토지를 배타적으로 점유하고 이용할 수 있다. 이때, 토지 소유권이 제삼자에게 이전되더라도 해당 토지에 설정된 기존의 지상권은 영향을 받지 않으며, 지상권자는 토지 소유권자의 의사와 관계없이 자신의 지상권을 다른 사람에게 이전할 수 있다. • 다음으로, 지역권은 소유한 토지의 편익을 높이기 위해 다른 사람의 토지를 출입로로 이용하거나, 다른 사람의 토지에서 물을 끌어오는 등의 특정 목적으로 이용할 수 있는 권리이다. 즉, 지상권과 달리 지역권은 편익을 제공하는 토지만만 아니라 편익을 제공받는 토지가 필요하며, 이들을 각각 승역지와 요역지라고 한다. 이때, 두 토지는 반드시 경계를 인접할 필요는 없다. 승역지를 이용하는 목적은 승역지를 통해 요역지에 필요한 접근성이나 자원을 확보하여 요역지의 가치를 높이는 데 있어야 한다. 가령, 요역지의 접근성을 높이려는 목적으로 승역지의 일부를 출입로로 이용하거나, 요역지의 토양을 비옥하게 하기 위한 목적으로 승역지의 물을 요역지로 끌어오는 것이다. • 지역권을 설정하기 위해서는 두 토지 소유권자가 지역권에 대한 계약을 맺고, 공동으로 등기를 해야 한다. 지역권이 설정되면 지역권자는 승역지를 이용하면서 필요에 따라 인공 구조물을 설치할 수 있으며, 지역권설정자인 승역지 소유권자 역시 해당 인공 구조물을 함께 이용할 수 있다. 즉 지역권이 설정되더라도 지역권자가 해당 토지를 배타적으로 이용하는 것이 아니라, 지역권설정자는 지역권자의 지역권을 침해하지 않는 범위 내에서 자신의 토지를 그대로 이용할 수 있다. 이때 승역지의 소유권이 제삼자에게 이전되더라도, 기존에 설정된 지역권은 영향을 받지 않지만, 요역지의 소유권이 제삼자에게 이전될 경우 지역권도 함께 이전된다. 그러나 요역지의 소유권과 분리하여 지역권만을 양도할 수 없다.

30. 뒷글에서 알 수 있는 내용으로 적절하지 않은 것은?

- ① 요역지의 소유권자는 승역지를 배타적으로 이용할 수 있다.
- ② 지상권과 지역권은 모두, 토지 소유권과 별개로 설정할 수 있다.
- ③ 지상권과 지역권은 모두, 계약의 주체가 공동으로 등기를 해야 한다.
- ④ 지상권자는 토지 소유권자의 의사와 관계없이 자신의 지상권을 다른 사람에게 이전할 수 있다.
- ⑤ 한정된 자원으로 토지를 효율적으로 이용하기 위해 민법에서는 지상권과 지역권을 규정하고 있다.

32. <보기>는 뒷글을 이해하기 위한 학습지의 일부이다. 뒷글을 바탕으로 <보기>에 대해 보인 반응으로 적절하지 않은 것은? [3점]

< 보 기 >

A 씨는 토지 Y를 이용하기 위해 B 씨와 지역권을, B 씨는 토지 X를 이용하기 위해 A 씨와 지상권을 각각 30년을 존속기간으로 설정하였다. 2023년 10월 20일 A 씨는 토지 X를 C 씨에게 매도하고, 당일 등기를 하였다. 아래는 2024년 10월 현재 토지 X의 등기부 중 일부를 발췌한 것이다.

【을 구】 (소유권 이외의 권리에 관한 사항)				
순위 번호	등기 목적	접수	등기 원인	권리자 및 기타 사항
1	요역지 지역권			승역지 토지 Y 목적 동행 범위 동측 50㎡ 2014년 10월 20일 등기
2	지상권 설정	2015년 10월 20일 등기	2015년 10월 15일 설정 계약	목적 철근 콘크리트 건물 범위 토지 전부 지상권자 B 씨

- ① A 씨는 토지 X의 지상권설정자로 2015년 10월 20일 이후, 토지 X의 토지 전부를 이용할 수 없겠군.
- ② A 씨는 토지 Y의 지역권자로서 2014년 10월 21일에는 토지 X에 동행하는 데 필요한 인공 구조물을 토지 Y의 동측 50㎡에 설치할 수 있겠군.
- ③ B 씨는 토지 Y의 지역권설정자로 2014년 10월 20일 이후, A 씨가 토지 Y를 동행하는 데 방해가 되지 않는 범위 내에서 토지 Y를 이용할 수 있겠군.
- ④ B 씨는 토지 X의 지상권자로서 토지 X에 지은 철근 콘크리트 건물을 안정적으로 유지하기 위해 2023년 10월 20일 이후, C 씨와 새로운 지상권을 설정해야겠군.
- ⑤ 2023년 10월 20일 이후, C 씨는 토지 Y의 지역권자로서 B 씨와의 별도 계약 없이 동행을 목적으로 토지 Y의 동측 50㎡를 이용할 수 있겠군.

31. 토지 소유권자에 대한 이해로 가장 적절한 것은?

- ① 토지 소유권자의 권리는 자신이 소유한 토지의 지면에 한정된다.
- ② 토지 소유권자가 자신의 토지 위에 있는 지상물을 소유하기 위해서는 지상권이 필요하다.
- ③ 승역지의 토지 소유권자라도 요역지의 소유권자가 승역지에 설치한 인공 구조물은 이용할 수 없다.
- ④ 지상물이 없는 토지의 소유권자라도 해당 토지를 이용하여 지상물을 소유하고자 하는 사람과 계약을 맺을 수 있다.
- ⑤ 토지 소유권자는 자신의 토지가 지역권이 설정되어 편익을 제공 하는 토지라면, 제삼자에게 토지를 처분할 수 없다.

33. 뒷글의 ㉠과 <보기>의 ㉡을 비교한 내용으로 가장 적절한 것은?

< 보 기 >

민법에서 규정하는 ㉡상권권은 경계가 인접한 토지의 소유권자들이 토지를 원만하게 이용하기 위해 상호 간에 토지의 이용을 조절할 수 있도록 하는 권리이다. 이 권리는 토지 소유권자 간의 공동 이익을 위한 최소한의 규제로서 토지 소유권을 가지면 자동으로 발생하기 때문에 별도의 등기가 필요 없다. 예를 들어 서로 인접한 두 토지의 소유권자 중 한쪽이 다른 한쪽의 토지를 이용하지 않고는 도로에 접근할 수 없는 부득이한 상황이라면, 상권권을 통해 인접한 토지의 이용을 요구하여 도로에 접근할 수 있다.

- ① ㉠은 ㉡과 달리 토지 소유권자 간의 원만한 토지 이용을 목적으로 한다.
- ② ㉠은 ㉡과 달리 소유한 토지에 필요한 접근성을 확보할 수 있게 한다.
- ③ ㉡은 ㉠과 달리 대상이 되는 토지들이 반드시 경계를 인접할 필요는 없다.
- ④ ㉠과 ㉡은 모두, 토지 소유권자 간에 토지의 이용에 대한 계약을 맺고, 등기를 해야 효력이 발생한다.
- ⑤ ㉠과 ㉡은 모두, 토지 소유권자 간의 토지 이용에 대해 한쪽의 권리는 제한하는 반면, 다른 한쪽의 권리는 확장하는 것으로 볼 수 있다.

34. ㉢와 문맥적 의미가 가장 유사한 것은?

- ① 예술에 대한 안목을 높였다.
- ② 담을 높여 도둑이 들어오지 못하게 했다.
- ③ 회사에서 그의 직급을 과장으로 높여 주었다.
- ④ 자동차 타이어의 압력을 지나치게 높이면 위험하다.
- ⑤ 민간단체들이 정부의 시책에 대해 찬성의 목소리를 높였다.

[그림 3] 2024학년도 10월 고2 전국연합학력평가 국어 영역 30~34 문항

3. 문항 생성 결과 평가 방법

가. 문항 생성 결과 평가 참여자

본 연구에서는 개발된 국어과 읽기 영역 선다형 평가 문항 생성 결과 평가를 위해 <표 4>와 같이 전문가 집단을 구성하였다. 이 집단은 교육 경력 4~13년의 고등학교 국

어 교사 5인과 국어교육 박사과정 연구자 3인으로 구성되었으며, 총 8명의 전문가가 참여하였다. 참여자들은 모두 국어과 읽기 영역 평가 문항의 출제 및 검토 경력을 보유하고 있었고, 연구가 진행되는 시점에서 교사 집단은 고등학교 2학년 독서 과목을 담당하고 있었다. 이러한 배경은 본 연구에서 생성된 문항의 타당성을 평가하는 데 있어 이 집단이 적합한 전문성을 갖추었음을 보여준다. 전문가 집단의 검토와 평가는 연구 결과의 신뢰성과 타당성을 높이는 데 중요한 역할을 하였다.

<표 4> 전문가 정보 개요

번호	성별	분류
T1	남	현직 교사
T2	여	현직 교사
T3	여	현직 교사
T4	여	현직 교사
T5	여	현직 교사
T6	남	국어교육 박사과정
T7	여	국어교육 박사과정
T8	여	국어교육 박사과정

나. 문항 생성 결과 평가 루브릭

본 연구에서는 생성된 문항의 평가를 위해 <표 5>에서 제시된 루브릭을 활용하였다. 해당 루브릭은 문항의 질적 요소를 다각적으로 검토하기 위해 문항 타당성, 문항 형식, 문항 내용의 세 가지 주요 평가 영역으로 구성되며, 각 영역은 두 가지 하위 평가 요소를 포함하고 있다.

먼저, 문항 타당성 평가는 문항이 목표 성취 기준을 적절히 반영하고 있는지, 그리고 학습자의 인지적 수준과 적합한 난이도를 유지하는지를 중심으로 이루어진다. 문항 형식 평가는 문항의 이해 용이성과 기존 기출 문항과의 유사성을 분석하는 데 초점을 둔다. 특히, 기출 문항과의 유사성은 AI가 인간 출제자의 문제 생성 방식과 어느 정도 일치하는지를 판단하기 위한 기준으로 포함되었다. 이는 현직 교사들이 기출 문항을 참고하여 출제하는 경향을 고려하여 설정된 항목으로, AI 기반 문항 생성이 현실적 교육 환경에서 활용될 가능성을 반영한다. 본 연구는 AI가 인간 전문가(교사)의 역할을 대체하는 것이 아니라, 문항 출제 과정에서 보조적 도구로 활용될 가능성을 검토하는 것을 목적으로 한다. 이에 따라, AI가 생성한 문항이 교사의 의도와 출제 전략에 부합하며, 교육적 맥락에서 실제적으로 활용될 수 있는지를 평가하는 것이 연구의 핵심 과제 중 하나이다. 또한 문항 내용 평가는 문항의 정답이 명확하게 설정되어 있는지, 그리고 문항

유형이 다양성을 확보하고 있는지를 분석하는 데 중점을 둔다. 이를 통해 AI가 생성한 문항이 기존의 인간 전문가가 출제한 문항과 비교했을 때 얼마나 다양한 형식으로 구성될 수 있는지를 검토하고자 하였다.

결과적으로, 본 연구에서 활용된 평가 루브릭은 AI 기반 문항 생성이 실제 평가 환경에서 인간의 검토 및 수정을 거친 후 어느 정도 활용 가능할지를 검증하는 데 초점을 두었다. 또한, AI가 생성한 문항이 기존의 인간 출제자가 제작한 문항과 조화를 이루며, 교육적 가치를 유지할 수 있는지를 평가하기 위한 기준으로 설계되었다.

해당 평가는 전문가 집단에 의해 루브릭의 각 항목을 5점 척도(1점: 전혀 아니다 ~ 5점: 매우 그렇다)로 평가되었으며, 평가자 간 신뢰도는 Cronbach's Alpha .924로 나타나 높은 신뢰도를 보였다. 또한, 전문가들은 전체적인 문항 구성을 검토하여 초안 문항으로 활용 가능한 문항 세트를 선택하였다. 이 과정에서 제로샷, 퓨샷, CoT 프롬프트 유형별로 생성된 문항이 평가되었으며, 필요 시 복수 응답 및 미응답이 가능하도록 안내하여 전문가들이 자유롭게 의견을 제공할 수 있도록 하였다.

이와 함께 본 연구에서는 국어과 읽기 영역 선다형 평가 문항의 결과 평가를 위해 FGI(Focus Group Interview)를 실시하였다. FGI는 전문가들이 작성한 서면 의견서와 심층 인터뷰 자료를 활용하여 문항의 장단점과 개선 방향을 도출하는 데 중점을 두었다. 각 문항 세트에 대해 구체적인 피드백을 수집함으로써 문항의 구조적 및 내용적 특징을 파악하였다. 심층 인터뷰를 통해 각 평가자의 평가 기준, 문항에 대한 세부적인 맥락, 추가적인 개선 의견 등이 수집되었다.

<표 5> 문항 생성 결과 평가 루브릭

평가 항목	평가 요소	평가 내용
문항 타당성 (10점)	① 성취 기준 반영 (1-5점)	각 문항이 교육과정 성취 기준을 명확히 반영하고 있는가
	② 학습자 적합성 (1-5점)	문항이 학습자 수준에 맞게 설계되어 있는가
문항 형식 평가 (10점)	③ 문항 이해 용이성 (1-5점)	문항의 문장 구조와 지시 사항이 명확하고 이해가 용이한가
	④ 기출문제와 유사성 (1-5점)	문항이 일반적인 기출문제의 형식과 구조에 비했을 때 유사한가
문항 내용 평가 (10점)	⑤ 정답 명확성 (1-5점)	하나의 정답이 명확하게 드러나는 가(복수 정답 가능성 여부)
	⑥ 다양한 문항 구성 (1-5점)	하나의 문항 세트가 '사실, 추론, 비판적, 적용 및 창의적 이해'와 같은 다양한 평가 요소와 문항 유형을 포함하고 있는가

IV. 연구 결과

1. 프롬프트 유형별 국어과 읽기 영역 선다형 문항 생성 결과

본 연구는 제로샷, 퓨샷, CoT의 세 가지 프롬프트 유형을 사용하여 고등학교 국어과 읽기 영역 선다형 평가 문항을 생성하고 이를 비교하였다. 연구에 사용된 지문은 ‘2024 학년도 고등학교 2학년 전국연합학력평가의 국어 읽기’ 영역 지문으로, 동일한 지문을 각 프롬프트 유형에 제공하여 문항을 생성하였다. 모든 프롬프트는 문항 생성 조건으로 평가 요소를 사실적 이해, 추론적 이해, 비판적 이해, 적용 및 창의적 이해의 네 가지로 설정하고, 각 평가 요소에 맞는 난이도를 조정하여 총 네 문항을 구성하였다.

생성된 문항의 특징을 살펴보면, 제로샷 프롬프트는 주어진 지문 정보를 그대로 활용하여 발문과 선지를 구성하는 단순한 형태를 보였다. 이러한 문항은 간결한 구조를 가지고 있었으나, 학습자의 고차원적 사고를 유도하기에는 한계가 있었다. 퓨샷 프롬프트는 기출문제와 해설을 활용하여 문항의 정교함이 다소 향상되었으며, 관계성을 파악하는 문항을 생성하는 데 효과적이었다. 그러나 여전히 문항의 형식적 다양성과 창의적 요소에서는 부족한 모습을 보였다. CoT 프롬프트는 체계적인 설계를 통해 고차원적 사고와 창의적 문제 해결을 유도하는 문항을 생성하였으며, 사실적 이해에서 창의적 이해에 이르기까지 평가 요소를 균형 있게 반영하였다.

프롬프트 유형별 문항 생성 결과는 CoT 프롬프트가 상대적으로 가장 우수한 성과를 보였으며, 이는 AI 프롬프트 설계의 체계성과 세부 설계가 문항 생성에 미치는 영향을 보여준다. 이 결과를 바탕으로 이어지는 분석에서는 생성된 문항의 질적 특성을 구체적으로 논의할 것이다.

<표 6> 프롬프트 유형별 국어과 읽기 영역 선다형 평가 문항 생성 결과

[생성 문항]						
제로샷		퓨샷		CoT		
사 실	1. “지상권”에 대한 설명으로 옳은 것을 고르시오. ① 지상권은 토지 소유자가 자신의 토지를 자유롭게 이용하고자 할 때 설정할 수 있다. ② 지상권 설정 시 지상권자는 해당 토지에서만 거주할 수 있다. ③ 지상권은 승역지와 요역지 관계를 포함한다. ④ 지상권자는 지상물 소유를 위해 토지를 이용할 수 있다. ⑤ 지상권은 지상물 소유와는 관련이 없다.	1. 지상권과 지역권의 권리 설정과 관련하여 옳은 설명은 무엇인가? ① 지상권자는 토지 소유권자의 의사와 관계없이 지상권을 제삼자에게 이전할 수 없다. ② 지역권은 요역지와 승역지의 경계가 인접해야만 설정할 수 있다. ③ 지상권 설정 시 토지 소유권자와 지상권자 간의 계약을 맺고 등기해야 한다. ④ 지역권이 설정된 승역지는 요역지 소유자의 권한에 의해 배타적으로 이용된다. ⑤ 지상권자는 지상물이 설치된 상태에서만 해당 권리를 설정할 수 있다.	1. 다음 중 지상권과 지역권의 설정 방식에 대한 설명으로 가장 적절한 것은? ① 지상권은 토지 소유자가 단독으로 설정할 수 있다. ② 지역권은 반드시 인접한 토지 사이에서만 설정될 수 있다. ③ 지상권과 지역권 모두, 계약 주체가 공동으로 등기해야 효력이 발생한다. ④ 지상권은 해당 토지 위에 지상물이 있어야만 성립할 수 있다. ⑤ 지역권은 편익을 제공하는 토지만 설정 대상이 된다.			
	정답: ④ 난이도: 하	정답: ③ 난이도: 하	정답 ③ 난이도: 하			
	2. 지상권자가 지상권을 다른 사람에게 이전할 수 있는 상황으로 가장 적절한 것은? ① 토지 소유자가 해당 토지를 포기한 경우 ② 지상권자가 지상물 건설을 마친 경우 ③ 토지 소유권자가 제삼자에게 토지를 매매한 경우 ④ 지상권자가 요역지를 통해 물을 공급받는 경우 ⑤ 토지 소유자가 해당 지상권을 공유하는 경우	2. 다음 중 지상권 설정이 불필요한 경우로 적절한 것은 무엇인가? ① 제삼자가 소유한 토지에 농작물을 재배하기 위해 토지를 장기간 사용할 경우 ② 공용 출입로로 활용하기 위해 타인의 토지를 이용할 경우 ③ 토지의 소유권자는 해당 토지에 지상물을 설치하지 않고 단기적으로 이용할 경우 ④ 일정 기간 동안 다른 사람의 토지에 임시 구조물을 설치할 경우 ⑤ 토지를 안정적으로 이용할 수 있는 방법으로 제삼자의 토지에 지상권을 설정할 경우	2. 지상권이 설정된 토지에서 발생할 수 있는 상황으로 가장 적절한 것은? ① 토지 소유권자가 자신의 결정에 따라 지상권자를 퇴거시킬 수 있다. ② 지상권자는 지상권 설정 후 해당 지상권을 자유롭게 타인에게 이전할 수 있다. ③ 지상권이 설정된 토지는 소유권자가 변동되면 지상권도 자동으로 소멸한다. ④ 지상권자는 승역지를 사용하여 요역지의 가치를 높이는 방식으로 권리를 행사할 수 있다. ⑤ 지상권 설정 기간이 끝나면 지상권자는 지상물을 철거해야 한다.			
추 론	정답: ③ 난이도: 중	정답: ③ 난이도: 중	정답: ② 난이도: 중			
비 판	3. 다음 중 “지역권”과 “지상권”이 가진 공통점으로 적절하지 않은 것은? ① 계약 후 공동 등기를 통해 효력이 발생한다.	3. 다음 중 지역권이 다른 권리와 구분되는 특징으로 가장 적절한 것은 무엇인가? ① 지역권 설정 시 권리 이전이 제한된다.	3. 다음 중 지상권과 지역권의 설정 목적에 대한 비판적 견해로 가장 타당한 것은? ① 지상권은 건물 소유자의 이익을 보호하지만, 토지 소유자의			

[생성 문항]						
제로샷		퓨샷		CoT		
② 제삼자에게 소유권이 이전되어도 권리에 영향이 없다. ③ 다른 사람의 토지에 특정 목적을 가지고 접근할 수 있다. ④ 지역권은 요역지를 위해 승역지를 사용하는 특징이 있다. ⑤ 두 권리 모두 목적과 특징에 있어 명확한 차이를 가진다.		② 지역권은 요역지를 소유한 사람이 승역지의 편익을 위해 특정 목적을 가지고 사용한다. ③ 요역지와 승역지의 소유권자 간에 상호 계약이 필요하지 않다. ④ 승역지를 배타적으로 점유할 수 있는 권리가 포함된다. ⑤ 요역지 소유자는 승역지의 권리를 제한할 수 있다.		권리를 지나치게 제한할 수 있다. ② 지역권은 승역지 소유자가 더 큰 경제적 이익을 얻도록 설계된 제도이다. ③ 지상권은 토지와 지상물을 일체화해 소유자의 자산 가치를 높이는 목적이 있다. ④ 지역권은 승역지 소유자가 요역지의 사용을 제한하는 역할을 한다. ⑤ 지상권과 지역권 모두 사적 재산권의 보장을 최우선으로 설정된다.		
정답: ④ 난이도: 상		정답: ② 난이도: 상		정답: ① 난이도: 상		
적용 및 창의	4. 지상권과 지역권의 주요 차이점을 바탕으로, 두 권리의 역할과 효용성을 비교한 설명으로 가장 적절한 것은? ① 지상권과 지역권 모두 독립된 권리로 설정되며, 다른 사람의 토지 소유권을 제한하는 데 목적이 있다. ② 지상권은 승역지와 요역지 개념을 포함하지 않으며, 독립적으로 건축물을 소유하고자 하는 사람에게 적합하다. ③ 지역권은 건축물 소유에 직접적인 목적이 있어, 지상권과 유사한 형태로 제한된 토지에서만 설정된다. ④ 지상권은 요역지에서 공급받는 자원을 이용하는 권리로, 지역권에 비해 제약이 적다. ⑤ 지역권과 지상권 모두 승역지와 요역지 간의 관계를 필요로 한다.		4. 다음 중 지상권과 지역권의 권리 설정에서 나타날 수 있는 문제 상황으로 적절하지 않은 것은 무엇인가? ① 지상권자가 토지 소유자의 의사와 관계없이 지상권을 제삼자에게 이전하여 토지 사용을 확대한다. ② 승역지를 소유한 사람이 요역지 소유자의 지역권을 방해하지 않도록 제한된 범위 내에서 이용한다. ③ 지상권이 설정된 토지를 제삼자가 매입했을 때 기존 지상권이 영향을 받지 않고 유지된다. ④ 요역지의 소유권자가 바뀌었을 때 지역권이 자동으로 이전되지 않고 새로 설정되어야 한다. ⑤ 요역지 소유자는 승역지 소유자의 동의를 얻어 물을 끌어오는 목적으로 승역지를 이용한다.		4. 다음은 A 씨와 B 씨가 토지 소유권, 지상권, 지역권을 설정한 예입니다. 이 상황을 바탕으로 분석한 내용 중 적절하지 않은 것은? <div><div><보기> A 씨는 자신의 토지 X에서 건물을 소유하기 위해 B 씨의 토지 Y에 대해 지상권을 설정하였고, 이후 B 씨에게 이를 양도하였습니다. B 씨는 자신의 토지 Z를 출입로로 활용하기 위해 A 씨와 지역권을 설정하였습니다. 이후 A 씨는 토지 X를 C 씨에게 매도하였습니다.</div><div>① C씨는 A씨의 지상권을 인정하고 유지할 의무가 있다. ② B씨는 지역권을 통해 A씨의 토지 Z를 요역지로 활용할 수 있다. ③ C씨가 A씨의 토지를 매수한 경우, B씨와의 계약은 소멸된다. ④ 지상권은 A씨가 매도했더라도 C씨에게 유효하다. ⑤ B씨는 A씨와의 지역권 설정 계약에 따라 출입로로 토지 Z를 사용할 수 있다.</div></div>	
	정답: ② 난이도: 상		정답: ④ 난이도: 상		정답: ③ 난이도: 상	

2. 프롬프트 유형별 국어과 읽기 영역 선다형 문항 생성 결과에 대한 기술통계

각 프롬프트 유형별로 생성된 국어과 읽기 영역 선다형 평가 문항에 대한 전문가 평가 결과를 정리하면 <표 7>과 같다. 먼저 CoT 방식은 문항 개발에서 가장 높은 성과를 보였다. CoT는 30점 만점에 평균 점수 27.57점으로 모든 프롬프트 유형 중 가장 높은 점수를 기록하였으며, 제로샷 프롬프트의 경우는 가장 낮은 점수를 기록했다. 이를 통해 국어과 읽기 영역 선다형 문항 평가 생성에서 CoT 방식이 가장 높은 성능을 보이는 프롬프트인 것으로 나타났다.

<표 7> 프롬프트 유형별 국어과 읽기 영역 평가 문항 평가 결과

	제로샷	퓨샷	CoT
평균	17.57	21.86	27.57
표준편차	4.31	2.59	3.44
최솟값	8	16	18
최댓값	23	24	28

이러한 프롬프트 유형별로 생성된 국어과 읽기 영역 평가 문항의 상대적 질적 차이에 대한 통계적 유의성을 검증하기 위해 일원분산분석(ANOVA)을 실시하였으며, 결과는 <표 8>과 같다. F값은 42.076($p < .001$)로 나타나 통계적으로 세 프롬프트 유형간 생성된 국어과 읽기 영역 문항의 상대적 질적 차이가 있는 것으로 확인되었다. 일원분산분석 이후 다중 비교를 수행하기 위해서 Scheffé 사후검정 분석을 실시하였다. 그 결과, CoT가 제로샷과 퓨샷에 비해 가장 높은 효과를 보이는 것으로 나타났다. CoT은 제로샷과의 평균 차이(1.458), 퓨샷과의 평균 차이(0.833)에서 모두 통계적으로 유의미한 차이를 보였으며, Scheffé 통계량에서도 가장 큰 값을 기록하였다. 따라서, CoT의 성능이 가장 우수하다고 할 수 있다.

<표 8> 프롬프트 유형별 국어과 읽기 영역 평가 문항 평가 차이(분산 분석)

프롬프트 유형	N	평균	표준편차	F	유의확률
제로샷	48	2.563	0.897	42.076	<0.001
퓨샷	48	3.188	0.673		
CoT	48	4.021	0.758		
집단비교	평균차이	Scheffé 통계량	기준값	유의확률	
제로샷 vs 퓨샷	0.625	3.86	1.21	<0.001	
제로샷 vs CoT	1.458	8.61	1.21	<0.001	
퓨샷 vs CoT	0.833	5.70	1.21	<0.001	

전체 지표값에 대한 통계적 유의성 검증에 이어, 세부 평가 항목별로 국어과 읽기 영역 평가 문항과 문항의 질에 대한 심층 분석을 수행하였다. <표 9>는 각 세부 항목의 기초 통계량과 분산분석 결과를 제시하고 있으며, 항목별로 일원분산분석(ANOVA)을 통해 그룹 간 차이를 검토한 후, Scheffé 사후 검정을 통해 그룹 간 세부 차이를 추가적으로 분석하였다.

분석 결과, 3. 문항 이해 용이성을 제외한 모든 항목에서 F값이 통계적으로 유의미한 차이를 보였으며, CoT가 전반적으로 가장 우수한 성능을 나타냈다. 특히, Scheffé 사후 검정 결과에 따르면 CoT는 모든 항목에서 다른 그룹(제로샷, 퓨샷)과 명확한 성능 차이를 보여주었으며, 이는 CoT가 초고 문항으로써 인간 전문가(출제 교사)와 협력하면 국어과 읽기 영역 평가 문항의 질적 향상에 긍정적인 영향을 미칠 수 있음을 시사한다. 한편, 3. 문항의 이해 용이성 항목에서는 제로샷과 퓨샷 간의 성능 차이가 통계적으로 유의하지 않은 것으로 나타났으나, 나머지 항목에서는 CoT가 다른 그룹에 비해 일관되게 우수한 성능을 보였다. 이러한 결과는 CoT 접근 방식이 국어과 읽기 영역 평가 세트 문항 생성에 상대적으로 우수한 성과를 보이며, 향후 AIG 기술 발전을 위한 프롬프트 전략으로 심화된 탐색을 수행할만한 것임을 설명한다.

<표 9> 프롬프트 유형에 따른 세부 평가 항목별 국어과 읽기 영역 평가 문항 생성 결과 평가

		문항 타당성(10)		문항 형식 평가(10)		문항 내용 평가(10)	
		1. 성취기준 적합성	2. 학습자 적합성	3. 문항 이해 용이성	4. 기출과 유사성	5. 정답의 명확성	6. 문항 다양성
제로샷(A)	평균	2.50	2.50	2.88	2.88	2.25	2.38
	표준편차	0.93	1.20	0.64	0.64	1.04	0.92
	최솟값	1	1	2	2	1	1
	최댓값	4	5	4	4	4	4
퓨샷(B)	평균	3.00	3.38	3.13	3.63	3.00	3.00
	표준편차	0.53	0.74	0.83	0.52	0.76	0.53
	최솟값	2	3	2	3	2	2
	최댓값	4	5	4	4	4	4
CoT(C)	평균	3.75	4.00	3.75	4.63	3.88	4.13
	표준편차	0.71	0.76	0.71	0.52	0.83	0.83
	최솟값	3	3	3	4	2	3
	최댓값	5	5	5	5	5	5
F값		5.78***	5.34**	3.03*	19.55***	6.79***	10.36***
사후검증 (Scheffé)		A<B	A<B	A<C	A<B	A<B	A<B
		A<C	A<C	B<C	A<C	A<C	A<C
		B<C	B<C		B<C	B<C	B<C

3. 프롬프트 유형별 국어과 읽기 영역 선다형 문항 질에 대한 심층 분석

본 연구는 생성형 AI를 활용한 프롬프트 기반의 AIG 기술을 활용하여 국어과 읽기 영역 평가 문항 생성의 자동화 가능성을 명확히 보여주었다. 연구 결과, 생성형 AI가 생성한 문항은 평가 문항 초안으로 활용될 가능성을 가지고 있으며, 이는 연구에 참여한 전문가들 사이에서도 일정 부분 합의된 내용이었다. 특히 CoT 프롬프트는 문항 세트 생성 과정에서 가장 높은 평가를 받았으며, 사실적 이해와 추론적 이해 문항에서 비교적 우수한 성과를 보여 초안 작성 도구로서의 가능성을 입증하였다. CoT 방식은 AI의 단계적 사고 유도 설계를 통해 발문과 선지 간의 논리적 일관성을 유지하고 지문을 효과적으로 반영하여 문항을 구성하는 데 적합한 전략으로 평가되었다.

그러나 CoT 프롬프트도 모든 평가 항목에서 일관된 성과를 보인 것은 아니었다. 비판적 이해 문항에서는 성취 기준을 형식적으로 반영하는 데 그쳤으며, 발문과 선지가 충분히 구체적이지 않다는 평가를 받았다. 전문가 평가에서는 비판적 사고를 유도하는 문항 설계에서 맥락 제공 부족과 사고 유도의 제한이 CoT 방식의 주요 한계로 지적되었다. 예컨대 T7은 발문 뒤에 대화나 토론과 같은 추가적인 맥락을 제공할 경우, 학습자의 사고를 더욱 깊이 자극할 수 있을 것이라고 언급하며, AI가 국어과 읽기 영역 평가 문항의 복합적 요구를 완전히 충족하지 못했음을 지적하였다.

제로샷 프롬프트와 퓨샷 프롬프트는 AI 기반 문항 생성의 가능성을 보여주었으나, 세트 문항 생성을 위한 전략으로는 한계를 드러냈다. 제로샷 프롬프트는 단순 정보 확인에 치중된 문항을 생성하며, 학습자의 사고를 충분히 유도하지 못했다는 평가를 받았다. T1과 T2는 제로샷 문항이 지문 내 정보를 단순히 재구성하는 경향이 강하며, 난이도가 지나치게 낮아 학습자 수준과 맞지 않는다고 평가하였다. 이러한 결과는 제로샷 방식이 효율적인 초안 작성 도구로 활용될 수 있으나, 고차원적 독해력을 평가하는 문항 설계에는 적합하지 않음을 시사한다.

퓨샷 프롬프트는 기출문제와 해설을 학습 자료로 활용하여 안정적이고 형식적으로 완결된 문항을 생성할 수 있었지만, 창의적 평가 요소를 구현하거나 복합적 문제 해결을 유도하는 데는 한계를 보였다. 전문가 T5는 퓨샷 문항에서 선지 간섭 현상이 발생하여 문항 구조의 완결성이 저해된 사례를 언급하며, 퓨샷 방식이 기존 데이터를 모방하는 데 의존적이라고 평가하였다. 이로 인해 퓨샷 프롬프트는 특정 유형의 안정적 문항 생성을 지원하는 데 유용하지만, 세트 문항 설계와 같은 복잡한 평가 요구를 충족하기에는 부적합하였다.

연구 결과, 생성형 AI를 활용한 프롬프트 활용을 통해 자동 문항 생성(AIG) 기술이 보다 향상될 가능성을 보여준다. 이를 위해 반영해야 할 프롬프트는 CoT 전략이다. 이

는 사실적 이해와 추론적 이해 문항에서 가장 높은 평가를 받았다는 점과 세트 문항 생성을 위해 효과적인 프롬프트 전략임을 확인하였다는 점을 주목할 필요가 있다. 그러나 CoT 방식도 특정 평가 요소에서는 한계를 드러냈으며, 인간 전문가의 개입이 필요하다는 점을 연구 결과 전반에 걸쳐 강조되었다. 비판적 이해와 같은 고차원적 사고를 평가하는 문항에서는 인간 전문가가 필수적으로 문항의 맥락과 구조를 보완할 필요가 있다.

결론적으로, AI 기술만으로는 고차원적 평가 요소를 완전히 충족시키기 어렵기 때문에, 문항 설계 과정에서 인간 전문가와의 협업이 요구됨을 확인할 수 있다. 이러한 연구 결과는 향후 국어과 읽기 영역 선다형 평가 문항 생성을 보다 정교하고 효과적으로 발전시키기 위한 AI와 인간의 상호 보완적 접근의 필요성을 명확히 보여준다.

<표 10> 프롬프트 유형별 문항 생성 결과에 대한 전문가 인터뷰

프롬프트 유형	전문가 인터뷰
제로샷	T1: “문항의 난이도가 전반적으로 쉬운 편에 속하여 교사의 많은 수정이 필요할 듯”
	T2: “문항의 난이도가 전반적으로 쉬운 편에 속하여 교사의 많은 수정이 필요할 듯. 학습자에 따라 애매하거나 중의적으로 해석될 여지도 보임”
	T7: “선지가 너무 간단하게 서술이 되어 있어서 평가 문항으로서 부족”
퓨샷	T3: “제로샷 보다는 좀 나으나 여전히 문제 유형의 다양성 면에서는 부족해 보임. 유형이 유사”
	T5: “선지 간섭 등으로 인해 5지 선다형이라 보기 힘들. 수정이 필요”
	T8: “문항의 구조가 단순하고 학습자의 수준에 비해 다소 쉬운 편이지만 제로샷과 비교했을 때는 상대적으로 적절한 난이도로 보임”
CoT	T1: “제로샷에 비해 문항 난이도는 올라갔으며, 퓨샷에 비해서 발문의 자연스러움이 높아진 것으로 보여 기출문제와 유사하게 느껴짐. 그러나 여전히 고난이도 문제는 없음”
	T2: “미세하지만 1번 문항에서 지문 그대로가 아니라 변형된 문항을 사용하고 있어 난이도가 더 높아보이며 지문을 직접적으로 활용하기 보다는 다양한 페르프라이징을 시도함으로써 문항 난이도를 올리는 데 영향을 미침”
	T7: “지금 이대로 비판적 견해를 제시해도 좋고 개인적으로는 발문 뒤에 글 상자로 ‘사람들의 대화(토론) 내용’ 후에 선지가 제시되는 방식은 어떨지 추상적이지만 생각해 보았음”
	T8: “사실적 이해 관련 문항은 적절해 보이지만 비판적 이해 관련 문항은 성취 기준을 반영하고 있지 못함. 적용 및 추론 문항의 발문 “~의 예입니다.”를 제외하면 대체로 적절한 형식임”

V. 결론 및 제언

본 연구는 국어과 읽기 영역의 선다형 평가 문항 자동 생성에서 AI 프롬프트 전략(제로샷, 퓨샷, Chain-of-Thought(CoT))의 효과를 비교 분석하였다. 이를 위해 세 가지

프롬프트 전략을 적용하여 생성된 문항을 전문가 평가를 통해 검증하였으며, 문항의 형식적 적절성, 논리적 타당성, 평가 목적과의 적합성을 중심으로 분석하였다(Min et al., 2022). 연구 결과, CoT 프롬프트는 AI의 단계적 사고 과정을 유도함으로써 문항의 논리적 일관성을 높이는 데 가장 효과적인 전략으로 확인되었다. CoT 기반 문항은 단순 정보 재구성을 넘어 개념 간의 관계를 논리적으로 조직하여 평가 목표에 부합하는 문항을 생성하는 데 기여하였다(Chen et al., 2023). 이에 반해, 제로샷 프롬프트는 평가 목표와의 부합성이 낮으며, 고차원적 사고를 평가하는 데 한계를 보였다. 퓨샷 프롬프트는 형식적 완결성이 높았으나, 사고 수준의 다양성을 반영하는 데 제한이 있었다(Cai et al., 2023). 이러한 결과는 AI 기반 문항 생성의 실효성을 극대화하기 위해서는 단순한 데이터 입력 방식이 아닌, AI의 사고 과정을 조정하는 정교한 프롬프트 설계가 필수적임을 시사한다. 또한, 문항 생성 과정에서 교사의 개입을 통해 문항의 평가 적합성을 보완하는 Human-in-the-Loop(HITL) 접근이 필요할 것으로 보인다(Min et al., 2022).

본 연구는 AI 기반 자동 문항 생성(AIG)의 가능성을 탐색하는 데 초점을 맞추었으나, 다음과 같은 한계를 가진다. 첫째, 전문가 평가단이 8명으로 제한되어 있어 평가 결과의 일반화 가능성이 낮다. 보다 다양한 전문가 집단이 참여하였다면 연구 결과의 신뢰성과 타당성을 높일 수 있었을 것이다. 둘째, 연구에서는 특정 지문과 평가 상황에서 생성된 문항을 분석하였기에, 다양한 평가 맥락과 학습자 수준에서의 적용 가능성을 충분히 검증하지 못하였다. 셋째, AI가 생성한 문항에 대한 학습자 반응 데이터를 수집하지 못했기 때문에, 실제 평가 상황에서의 효과성을 확인하는 데 한계가 있었다. 향후 연구에서는 학습자의 응답 데이터를 수집하여 AI 기반 문항 생성이 평가 결과에 미치는 영향을 보다 면밀하게 분석할 필요가 있다. 이러한 한계를 고려할 때, 생성형 AI 기술은 독자적으로 완벽한 평가 문항을 생성하는 것이 아니라 교사의 평가 전문성을 보조하는 도구로서 보다 효과적으로 활용될 가능성이 높다. 본 연구에서 확인한 바와 같이, CoT 전략을 활용한 문항 생성은 교사의 검토와 수정 과정을 거칠 때 평가의 논리적 일관성과 타당성을 높이는 데 기여할 수 있다. 이는 AI 기술이 단순한 평가 자동화 도구를 넘어 교사의 평가 역량을 보완하고, 교육 평가의 질적 향상을 지원하는 방향으로 발전해야 함을 시사한다.

이를 고려하여 AI 기반 자동 문항 생성(AIG) 기술의 실효성을 높이고 연구의 한계를 극복하기 위해 다음과 같은 방안을 제안한다.

첫째, AI 기반 문항 생성 기술의 신뢰성과 타당성을 극대화하기 위해 사고 과정의 연계를 체계적으로 반영한 프롬프트 설계가 필수적이다. 정민주 외(2022)의 연구는 국어과 평가 문항의 질적 특성을 분석하며, 문항의 평가 적합성과 신뢰도를 높이는 데 필요한 기준을 제시하였다. 이를 반영하여, AI 기반 문항 설계에서는 국어과 평가 문항이 요

구하는 사고 과정뿐만 아니라, 교육과정에서 강조하는 핵심 역량(예: 비판적 사고, 논리적 사고 등)을 반영하는 방식으로 접근해야 한다. 이러한 기준은 AI 기반 문항 설계의 평가 적합성을 보장하는 중요한 요소로 작용해야 한다. 또한 Flavell(1979)은 메타인지적 사고 과정이 평가 문항의 질을 결정하는 핵심 요인임을 강조하였으며, AI 기반 문항 설계에서도 이러한 요소가 적절히 반영될 필요가 있다. 특히, AI 기반 문항 생성은 단순한 정보 확인을 넘어, 추론 및 비판적 분석을 포함한 다양한 사고 과정을 유도하는 방향으로 설계될 필요가 있다. 이를 위해 본 연구에서 CoT(Chain-of-Thought) 프롬프트를 활용하여 문항 간 논리적 연계를 보장하는 방안을 실험적으로 검토한 결과, 프롬프트 설계가 문항의 사고 과정 연계에 중요한 영향을 미친다는 점이 확인되었다(Gierl & Lai, 2013; Burstein, 2023; Zimmerman, 1990).

그러나 본 연구에서 구축한 프롬프트 설계는 보완해야 할 측면을 내포하고 있으며, 향후 연구에서는 이를 개선하기 위한 구체적인 전략이 필요하다(Attali et al., 2022; Maity et al., 2024). 예를 들어, 논설문 기반 세트 문항에서는 첫 번째 문항이 핵심 주장과 근거를 식별하도록 하고, 이후 문항들이 논증의 타당성 평가 및 반박 구조 분석으로 확장될 수 있도록 설계해야 한다. 설명문 기반 문항에서는 개념 간 관계를 점진적으로 파악할 수 있도록 계층적 사고를 유도하는 방식이 요구된다(Chen et al., 2023). 이러한 설계를 실현하기 위해 평가 전문가의 피드백을 반영한 동적 프롬프트 튜닝(dynamic prompt tuning) 기법을 적용해야 한다. 이를 통해 AI가 생성하는 세트 문항이 개별적으로 완결성을 가지면서도, 평가 목표와의 정합성을 유지할 수 있도록 해야 한다. 궁극적으로, AI 기반 문항 생성의 효과성을 극대화하기 위해서는 평가 문항의 형식적 완결성과 사고 과정 연계성을 동시에 확보하는 연구가 지속적으로 이루어져야 한다.

둘째, AI 기반 문항 생성의 신뢰성과 타당성을 확보하기 위해 Human-in-the-Loop(HITL) 모델을 체계적으로 구축해야 한다. AI가 생성한 문항은 교육과정 및 평가 기준과의 정합성을 유지해야 하며, 이를 위해 교사와 평가 전문가가 적극적으로 개입하는 HITL 모델이 필수적이다(Attali et al., 2022; Embretson & Yang, 2006; Kissi et al., 2023; 최숙기, 박종임, 2024). 따라서 AI 기반 문항 생성을 교육과정과의 정합성을 고려하여 HITL(Human-in-the-Loop) 모델을 적용해야 한다. 예를 들어, 설명문 기반 문항에서 선택지 간 의미적 유사성이 지나치게 높거나 정답과 오답의 변별력이 낮은 경우, 교사가 이를 수정하는 절차를 명확히 마련해야 한다. 또한, 교사의 피드백을 AI가 학습하여 반복적으로 품질을 개선하는 보정 모델(adaptive refinement model)을 적용할 수 있다. 이 과정에서 AI는 단순히 문항을 생성하는 도구에 머무르지 않고 지속적으로 발전하는 평가 지원 시스템으로 기능할 수 있다. 특히, 추론형 문항에서는 학습자의 기존 반응 데이터를 활용해 정답 선택률과 오답 패턴을 분석하고, 이를 기반으로 변별력이 높

은 문항을 생성하도록 조정해야 한다(Gierl & Haladyna, 2013). 이러한 HITL 모델은 교사와 AI 간 협업 체계를 강화하며, AI가 교육 현장에서 신뢰받는 도구로 자리 잡는데 기여할 것이다.

셋째, 학습자 반응 데이터를 활용한 실증 연구를 통해 AI 기반 문항 생성 기술의 효과성을 검증해야 한다(Kurdi et al., 2020). 본 연구는 AI가 생성한 문항의 형식적 적절성과 논리적 일관성은 주로 전문가 평가에 의존하고 있다. 그러나 학습자 반응 데이터를 활용하여 실제 평가 상황에서 AI 문항이 독해력 측정에 미치는 영향을 분석하는 연구가 필요하다(Kurdi et al., 2020; Funk et al., 2023). 예를 들어, 논설문 문항이 논증 구조를 정확히 반영하는지 또는 설명문 문항이 개념 간 관계를 적절히 평가하는지를 학습자의 정답 및 오답 패턴과 비교하여 검증할 수 있다(Bezirhan & von Davier, 2023). 또한, 특정 인지 과정(예: 추론, 비판적 사고 등)을 효과적으로 유도하는지를 확인함으로써 AI 기술의 교육적 가치를 보다 명확히 입증할 수 있다(Flavell, 1979; Shah, 2024).

넷째, AI가 신뢰성 있는 문항을 생성하기 위해서는 다양한 난이도와 평가 요소를 반영한 대규모 데이터셋 구축이 필수적이다(Kosh et al., 2019; von Davier, 2018; Gierl & Haladyna, 2013; Embretson & Yang, 2006). 이러한 데이터셋에는 단순히 텍스트 자료뿐만 아니라 문항의 메타데이터(metadata)를 포함해야 한다. 메타데이터는 문항의 난이도, 변별도, 평가 목표, 텍스트 유형(예: 논설문, 설명문), 학습자의 인지 수준 등을 상세히 기술한 정보를 의미한다. 예를 들어, 독해력 수준별로 분류된 문항 데이터셋과 함께 각 문항의 난이도 수준(쉬움, 보통, 어려움), 평가 기준(사실 확인, 추론, 비판적 분석), 그리고 학습자가 해당 문항에서 보인 정답률과 오답 패턴 같은 학습자 반응 데이터를 포함해야 한다. 특히, 공신력 있는 기출문제 데이터(예: 대학수학능력시험 및 국가 공인 시험)와 함께 이러한 메타데이터를 포함하면 AI가 문항 생성 시 평가 목표와 학습자 수준에 더욱 정밀하게 부합하는 결과를 도출할 수 있다(Bezirhan & von Davier, 2023; National Academies of Sciences, Engineering, and Medicine, 2022). Embretson & Yang(2006)의 연구는 자동 문항 생성(AIG)의 신뢰성을 높이기 위해 문항 속성과 평가 모델 간의 체계적인 연계를 강조했으며, 이를 기반으로 AI 평가 도구의 신뢰성을 확보하는 방안이 필요하다.

다섯째, AI 기반 문항 생성의 신뢰성과 타당성을 제고하기 위해 지문 생성 방식의 구조적 개선이 필수적이다. 국어과 평가에서 활용되는 지문은 단순한 기출 자료의 반복이 아니라, 평가 목표에 맞춰 변형 및 재구성되어야 하며, 이는 학습자의 독해 역량을 진단하고 평가의 변별력을 확보하는 핵심 요소로 작용한다. 따라서 AI 기반 자동 문항 생성에서는 지문이 단순한 재활용 방식이 아닌, 평가 목표와 연계된 체계적 설계를 바탕으로 이루어져야 한다. 즉, AI 기반 지문 생성은 기존 기출 지문의 변형을 넘어, 국어과 교육

과정이 요구하는 독해 및 비판적 사고 역량을 반영한 새로운 텍스트를 생성하는 방향으로 발전해야 한다. 이를 위해 AI 기반 적응형 지문 생성(adaptive passage generation) 기법을 도입하는 것은 필수적이다. Gierl & Haladyna(2013)는 문항과 지문의 유기적 정합성이 평가의 타당성을 결정하는 핵심 요소임을 강조하였으며, AI가 평가 목표에 부합하는 지문을 생성하는 연구가 필요함을 제안하였다. 이에 따라 기존 기출 지문을 그대로 활용하는 방식에서 벗어나, 국어과 교육과정이 요구하는 독해 능력을 반영하여 새로운 지문을 설계하는 방향으로 나아가야 한다(Bezirhan & von Davier, 2023). 따라서 AI 기반 자동 지문 생성이 단순한 평가 자동화 도구를 넘어 학습자의 사고 과정과 평가 목표에 부합하는 방향으로 발전하기 위해서는 교육과정의 학문적 요구를 반영한 지속적인 연구가 필수적이다.

여섯째, AI 기반 평가 도구가 교육 현장에서 성공적으로 자리 잡기 위해서는 교육학, 언어학, 인공지능 연구 간 융합적 협력이 필요하다(Attali et al., 2022; Kissi et al., 2023). 예를 들어, 학습자의 읽기 전략(예: 핵심어 찾기, 논리적 연결 파악 등)과 AI가 생성한 문항 간 연계를 검토하는 연구가 이루어져야 한다. 또한 컴퓨터 지원 협력 학습(Computer-Supported Collaborative Learning, CSCL) 모델과 인간-컴퓨터 상호작용(Human-Computer Interaction, HCI) 관점을 접목하여 교사와 AI 간 협업 환경을 최적화해야 한다. 이를 통해 교사의 평가 과정과 자연스럽게 통합되는 시스템 설계를 구현하고 파일럿 연구를 통해 실질적인 효과와 활용 가능성을 검증해야 한다.

본 연구의 의의는 AI가 교사의 역할을 대체하기보다, 교사의 평가 전문성을 보완하고 학습자 중심의 평가 환경을 조성하는 ‘협력적 도구’로 기능할 가능성을 체계적으로 제시한 데 있다. 특히, AI 기반 문항 생성에서 제언된 정교한 프롬프트 설계와 Human-in-the-Loop(HITL) 접근은 평가의 공정성과 신뢰도를 높이는 동시에 교사의 업무 부담을 줄이는 방향으로 활용될 수 있음을 시사한다(Attali et al., 2022; Kissi et al., 2023). 이는 AI 기술이 단순히 기존 문항을 ‘자동화’하는 범위를 넘어, 교사와 함께 동적·적응적으로 문항을 생성하고 점검함으로써 개별 학습자의 독해 및 사고 과정을 보다 정교하게 평가할 수 있는 토대를 마련한다는 점에서 의의가 크다. 나아가 교육학, 언어학, 인공지능 연구 간 융합적 협력을 통해 학습자 반응 데이터와 실제 교실 수업 맥락을 반영한 대규모 데이터셋이 구축된다면, 교사가 주도하는 AI 평가 시스템의 신뢰성과 타당성은 한층 제고될 것이다. 궁극적으로, 본 연구가 제시하는 방향성은 교사-AI 협업 체계가 학습자 개개인의 역량과 요구를 반영하는 ‘맞춤형 평가’의 실질적 구현에 이바지할 수 있음을 보여주며, 이는 향후 교육 평가 패러다임 전환에 중요한 시사점을 제공한다.

참고문헌

- 신동광(2023). AI 도구를 활용한 중등교사 영어 출제 연수 사례 연구: ChatGPT를 중심으로. *어학연구*, 59(1), 21-42.
- 오규설(2022). 자동문항생성 기법을 활용한 읽기 평가 개발에 대한 시론: 사실적 독해 문항 모형을 중심으로. *청람어문교육*, 87, 7-34.
- 이하늘, 이용상(2024). 한국어 읽기 평가에서 자동 문항 생성의 가능성 탐색. *교육문화연구*, 30(3), 659-686.
- 임상묵, 조혜원, 이정우, 이현숙(2022). 사회과 자동문항생성을 위한 대규모 언어모델의 활용 가능성 탐색. *교육정보미디어 연구*, 30(3), 1035-1060.
- 장성민(2024). 학교 지필평가 출제 과정에서 느끼는 국어 교사의 실천적 어려움 탐색: 고등학교의 독서 선다형 평가 상황을 중심으로. *리터러시연구*, 15(2), 89-110.
- 정민주, 남가영, 남민우, 서수현, 최숙기, 이상일(2024). 좋은 국어과 평가 문항 특성에 관한 질적 연구-국어과 평가 문항 양호도 분석틀 개발 연구(2). *청람어문교육*, 89, 43-78.
- 최숙기, 박종임(2024). 생성형 AI를 활용한 현직 국어교사의 서·논술형 평가 문항 개발 양상 분석. *청람어문교육*, 97, 243-270.
- 허동석, 김기태, 송형우, 서봉원(2024). 프롬프트 개발을 통한 수능 국어 맞춤형 문제 생성 시스템 제안. *한국HCI학회 발표 논문집*, 183-189.
- 한국교육과정평가원(2024). 2025학년도 대학수학능력시험 학습 방법 안내. CAT 2024-2-1. 한국교육과정평가원.
- Afflerbach, P. (2017). *Understanding and Using Reading Assessment*, K-12 (3rd Edition). ASCD.
- 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.
- Bezirhan, U., & von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *TIMSS & PIRLS International Study Center, Boston College*.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information*

- Processing Systems*, 33, 1877–1901.
- Burstein, J. (2023). Duolingo English Test: Responsible AI standards. Retrieved from <https://duolingopapers.s3.amazonaws.com/other/DET+Responsible+AI+033123.pdf>.
- Cai, Z., Chang, B., & Han, W. (2023). Human-in-the-Loop through Chain-of-Thought. *arXiv preprint arXiv:2306.07932*.
- Chen, A., Phang, J., Parrish, A., et al. (2024). Two failures of self-consistency in the multi-step reasoning of LLMs. *OpenReview*.
- Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). Unleashing the Potential of Prompt Engineering in Large Language Models: A Comprehensive Review. *arXiv preprint arXiv:2310.14735*.
- Das, B., Majumder, M., Phadikar, S., & Sekh, A. A. (2021). Automatic question generation and answer assessment: A survey. *Research and Practice in Technology Enhanced Learning*, 16(1), 1–15.
- Embretson, S., & Yang, X. (2006). Automatic item generation and cognitive psychology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics: Psychometrics, Volume 26*, 747–768. North Holland, UK: Elsevier. [https://doi.org/10.1016/S0169-7161\(06\)26023-1](https://doi.org/10.1016/S0169-7161(06)26023-1).
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive –developmental inquiry. *American Psychologist*, 34(10), 906–911.
- Funk, P. F., et al. (2023). ChatGPT’s response consistency: A study on repeated queries of medical examination questions. *European Journal of Investigation in Health, Psychology and Education*, 14(3), 657–668.
- Gierl, M. J., & Haladyna, T. (2013). *Automatic item generation: Theory and practice*. New York, NY: Routledge.
- Gierl, M. J., & Lai, H. (2013). Using automated processes to generate test items. *Educational Measurement: Issues and Practice*, 32(3), 36–50. <https://doi.org/10.1111/emip.12018>.
- Hommel, B. E., Wollang, F. J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2022). Transformer-based deep neural language modeling for construct-specific automatic item generation. *Psychometrika*, 87(2), 749–772.
- Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P., Zeng, A., Tompson, J.,

- Mordatch, I., Chebotar, Y., ... & Ichter, B. (2022). Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*.
- Kissi, P., Baidoo-Anu, D., Anane, E., & Annan-Brew, R. K. (2023). Teachers' test construction competencies in examination-oriented educational systems: Exploring teachers' multiple-choice test construction competence. *Frontiers in Education, 8*. <https://doi.org/10.3389/educ.2023.1154592>
- Kosh, A. E., Simpson, M. A., Bickel, L., Kellogg, M., & Sanford-Moore, E. (2019). A cost-benefit analysis of automatic item generation. *Educational Measurement: Issues and Practice, 38*(1), 48-53. <https://doi.org/10.1111/emip.12237>.
- Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education, 30*(1), 121-204.
- Maity, S., et al. (2024). The Future of Learning in the Age of Generative AI : Automated question generation and assessment with large language models. *arXiv preprint arXiv:2410.09576v1*.
- Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.
- National Academies of Sciences, Engineering, and Medicine. (2022). *A pragmatic future for NAEP: Containing costs and updating technologies*. Washington, DC: The National Academies Press. <https://doi.org/10.17226/26427>.
- Rodriguez, M. C., & Haladyna, T. M. (2013). Objective testing of educational achievement. In K. F. Geisinger et al. (Eds.), *APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology*, 305-314. American Psychological Association. <https://doi.org/10.1037/14047-018>.
- Shah, C. (2024). From prompt engineering to prompt science with human in the loop. *arXiv preprint arXiv:2401.04122*.
- Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education in Practice, 6*(6), 354-363. <https://doi.org/10.1016/j.nepr.2006.07.002>.

- von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika*, 83, 847–857. <https://doi.org/10.1007/s11336-018-09660-3>.
- von Davier, M. (2019). Training Optimus prime, MD: Generating medical certification items by fine-tuning OpenAI's GPT2 transformer model. *arXiv preprint arXiv:1908.08594*.
- Wei, J., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*.
- Yang, X., & Embretson, S. E. (2007). Construct validity and cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications*, 119–145. Cambridge University Press. <https://doi.org/10.1017/CBO9780511611186.005>.
- Zimmaro, D. M. (2010). Writing good multiple-choice exams. *Center for Teaching and Learning, University of Texas at Austin*. Retrieved from <https://ctl.utexas.edu/sites/default/files/writing-good-multiple-choice-exams-fic-120116.pdf>
- Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist*, 25(1), 3–17.

논문접수 : 2025. 1. 3. / 수정본접수 : 2025. 2. 4. / 게재승인 : 2025. 2. 11.

ABSTRACT

A Study on Automatic Item Generation for Multiple-Choice Reading Comprehension Assessment in Korean Language

Goun Park³⁾

Doctoral candidate, Korean Education, Korea National University of Education

Sookki Choi⁴⁾

Professor, Korean Education, Korea National University of Education

This study investigates the potential of generative AI-based Automatic Item Generation (AIG) technology for creating Korean language reading comprehension multiple-choice items. Three prompt strategies—Zero-shot, Few-shot, and Chain-of-Thought (CoT)—were analyzed, with items generated by each strategy evaluated by a panel of eight experts (five Korean language teachers and three doctoral researchers) based on educational validity, format appropriateness, and content fidelity. The CoT prompt demonstrated superior performance by guiding AI through step-by-step reasoning, effectively generating items that balanced comprehension levels from factual understanding to creative and evaluative reasoning while maintaining logical coherence and alignment with assessment objectives. The Few-shot prompt achieved structural completeness but lacked creativity and diversity in certain item types. The Zero-shot prompt primarily focused on basic information reconstruction, proving inadequate for assessing higher-order thinking skills. These findings suggest that generative AI technology can serve as a valuable supplementary tool for teachers' expertise in assessment design, particularly in enhancing the systematic development and validity of test items. The CoT prompt specifically showed promise in reducing teachers' workload while maintaining assessment quality. This study highlights the potential of AIG technology as a practical tool for improving educational assessment quality and supporting teachers' professional practice in item development.

Key Words: Automatic Item Generation (AIG), Generative AI, Prompt Engineering, Zero-shot, Few-shot, Chain-of-Thought (CoT), Multiple-choice evaluation questions in the Korean language department reading area, reading Assessment Items, Human-in-the-Loop (HITL)

3) First Author, koniholic@gmail.com

4) Corresponding Author, agrement@knue.ac.kr