Brain, Digital, & Learning

pISSN : 2384-2474
eISSN : 2586-7490

RESEARCH ARTICLE

A Stage-wise Analysis of the Effects and Limitations of Small-Scale SFT in Korean CSAT-Style Question Generation

Giljae Kim*
MHResearch

수능형 국어 문항 자동 생성에서 소규모 SFT의 효과와 한계에 대한 단계별 분석

김길재*
엠에이치리서치

*Corresponding Author: Giljae Kim, advisor-ai@mhresearch.net

ABSTRACT

This study examines the effects and limitations of small-scale supervised fine-tuning (SFT) in Korean CSAT-style reading question generation. We propose a two-stage pipeline in which Stage 1 generates expository passages from keywords and Stage 2 generates multiple-choice questions and distractors from the generated passages. To identify stage-specific effects, we separately evaluate passage quality and item quality using a pairwise comparison framework with LLM-based judges and bootstrap-based statistical estimation.

Experiments compare Qwen2.5-7B-Instruct and EXAONE 3.5 7.8B under base, QLoRA-based SFT, expanded-training-set, and JSON-output conditions. Results show that small-scale SFT does not affect all stages equally. In the Qwen family, fine-tuned models repeatedly underperform the base model in item-level evaluation, particularly in distractor plausibility and discriminative quality, while passage-level quality shows no consistent decline. This suggests that small-scale SFT may improve surface-level conformity to exam-style formats without improving substantive item quality. In contrast, the EXAONE family shows a different pattern, with some fine-tuned conditions yielding improvement signals in item quality.

These findings indicate that the effectiveness of small-scale SFT is stage-dependent and model-dependent, and that passage generation and item generation should be evaluated separately in high-constraint educational generation tasks.

Keywords: Korean CSAT; automatic question generation; distractor generation; large language models; educational assessment

Brain, Digital, & Learning
2026, Vol. 16, No. 1, 95–116
https://doi.org/10.31216/BDL.2026.16.1.7

Received: March 27, 2026
Revised: March 31, 2026
Accepted: March 31, 2026

Introduction

대학수학능력시험(이하 수능) 국어 영역 비문학 문항은 출제 비용이 매우 높은 평가 항목이다. 수능은 고등학교 교육과정의 내용과 수준에 맞추어 출제되며, 기본 개념과 원리에 충실하면서도 추리, 분석, 종합, 평가 등의 사고력을 측정하도록 설계되는 대표적 고부담 국가시험이다(Korea Institute for Curriculum and Evaluation, 2025). 최근에는 이러한 수능 출제 체계를 보다 안정적이고 효율적으로 지원하기 위한 제도적 논의도 본격화되고 있다. 교육부는 2026년 2월 수능 출제 체계 개선 방안을 발표하면서 교육평가·출제지원센터 설립과 함께 인공지능(AI) 활용 영어 지문 생성 시스템 개발을 추진하고, 향후 AI를 난이도 예측과 유사 문항 검토에도 활용하겠다고 밝혔다(Ministry of Education, 2026). 이는 수능 출제 영역에서도 AI 기반 출제 지원 체계의 도입이 정책적으로 검토되기 시작했음을 보여준다. 그러나 실제 수능형 문항 개발은 여전히 전문 출제자와 검토자의 고도한 협업에 의존하고 있으며, 단일 지문당 2~3개의 문항을 오답 선지의 매력도를 유지하면서 설계하는 데 상당한 시간과 비용이 소요된다. 또한 기출 문항과의 중복 검증이 필요하여 출제 풀의 지속적 확장에도 구조적 한계가 존재한다.

대규모 언어 모델(LLM)의 발전은 자동 문항 생성(Automatic Question Generation, AQG) 분야에 새로운 가능성을 제시하였다(Du et al., 2017; Kurdi et al., 2020). 최근에는 교육 맥락에서도 LLM을 활용한 문항 자동화 가능성이 활발히 논의되고 있으며, 국내에서도 사회과 자동문항생성을 위한 LLM 활용 가능성 탐색(Lim et al., 2024), 국어과 읽기 영역 선다형 평가 문항 자동 생성 방안 연구(Park & Choi, 2025a), GAI-HITL 기반 독서 문항 자동 생성의 심리측정학적 타당성 분석(Park & Choi, 2025b), LoRA 파인튜닝을 활용한 개방형 GPT 기반 수능 국어 문항 생성 품질 비교(Hong & Lee, 2025)와 같은 연구가 보고되고 있다. 이러한 흐름은 국내 교육·평가 맥락에서도 자동문항생성이 더 이상 주변적 주제가 아님을 보여준다.

그러나 선행 연구의 대부분은 영어 데이터셋에 집중되어 있거나, 국내 연구의 경우에도 특정 교과에서의 활용 가능성 탐색, 프롬프트 전략 비교, 혹은 단일 단계 문항 생성에 초점을 두는 경우가 많다. 수능과 같이 고도로 정형화된 문체, 엄격한 선지 구성 규칙, 추론 및 전개방식 파악 등 다양한 인지 수준을 요구하는 고부담 시험 환경에서의 체계적 검증 연구는 여전히 부족한 실정이다.

최근 QLoRA(Dettmers et al., 2023)를 비롯한 효율적 파인튜닝 기법의 등장으로 소규모 도메인 데이터를 활용한 LLM 특화가 가능해졌으나, 파인튜닝이 실제로 생성 품질을 향상시키는지에 대한 실증적 검증은 여전히 제한적이다. 국내에서도 LoRA 기반 수능 국어 문항 생성 품질을 비교한 연구가 보고되었으나(Hong & Lee, 2025), 지문 생성과 문항·선지 생성을 분리하여 평가하거나, 지문 품질과 문항 품질을 독립적으로 검증한 연구는 드물다. 특히 수능과 같이 형식적 정형성과 내용적 깊이를 동시에 요구하는 과제에서, 소규모 SFT(Supervised Fine-Tuning)가 형식 모방에 그치고 내용 품질을 오히려 저하시킬 수 있다는 가능성은 충분히 검토되지 않았다.

본 연구는 이러한 문제의식에 기초하여, 수능형 국어 비문학 문항 자동 생성에서 소규모 SFT의 효과와 한계를 단계별로 검증하고자 한다. 이를 위해 키워드로부터 수능 문체의 비문학 지문을 생성하는 Stage 1과, 생성된 지문으로부터 발문과 5지선다형 선지를 생성하는 Stage 2를 통합한 엔드-투-엔드 파이프라인을 구축하였다. 또한 생성 품질을 단일 척도로 환원하지 않고, 지문 품질과 문항 품질을 분리하여 평가하는 이중 평가 프레임워크를 설계하였다. 본 연구가 다루는 핵심 질문은 다음과 같다.

연구질문1. 두 단계 파이프라인에서 소규모 QLoRA 기반 SFT는 지문 품질과 문항·오답 선지 품질 각각에 어떤 영향을 미치는가.

연구질문2. 소규모 SFT는 수능 문항 생성의 내용 품질을 향상시키는가, 아니면 형식 모방에 그치는가.

연구질문3. 확장 학습데이터 조건과 출력 형식 변경은 파인튜닝의 효과를 어떻게 변화시키는가.

연구질문4. 한국어 특화 베이스 모델에 동일한 도메인 SFT를 결합할 경우, 범용 모델에서 관찰되는 양상과 어떤 차이를 보이는가.

본 연구의 기여는 다음과 같다. 첫째, 키워드 기반 지문 생성과 문항·선지 생성을 연결한 이단계 엔드-투-엔드 파이프라인을 구현하고, 반복적 파이프라인 개선의 효과를 정량적으로 검토한다. 둘째, Qwen2.5-7B-Instruct(Qwen Team, 2024)와 EXAONE 3.5 7.8B(LG AI Research, 2024)를 중심으로 학습 데이터 조건, 출력 형식, 파이프라인 버전을 체계적으로 비교함으로써 소규모 SFT의 효과와 한계를 분석한다. 셋째, 지문 품질과 문항 품질을 분리 측정하는 이중 평가 프레임워크를 제안하고, position bias 보정과 bootstrap 신뢰구간을 결합한 비교 체계를 제시한다. 넷째, 수능형 국어 문항 자동 생성이 단순한 질문 생성 문제를 넘어, 고부담 평가 문항 설계와 생성형 AI 적응 전략이 교차하는 복합적 연구 과제임을 실증적으로 검토한다.

Related Work

1. Development of Automated Item Generation Research

문항 자동화 연구는 대체로 교육측정학 중심의 자동 아이템 생성(automatic item generation, AIG)과 자연어처리 중심의 자동 문항 생성(automatic question generation, AQG)이라는 두 흐름 속에서 전개되어 왔다. AIG는 인지모형(cognitive model)과 item model을 기반으로 문항의 심리측정적 속성을 사전에 통제하면서 다수의 병렬 문항을 체계적으로 생성하는 접근으로, 실제 평가에 투입 가능한 문항 생산을 목표로 한다 (Embretson & McIntyre, 2011; Gierl et al., 2012; Gierl & Lai, 2013). 이 관점에서는 문항 수급의 효율화 자체보다도 내용 타당도, 신뢰도, 측정학적 안정성이 핵심 과제가 된다.

국내 연구에서도 이러한 흐름은 분명하게 나타난다. Kim(2022)은 CAFA 시스템을 중심으로 수학교과에서 자동문항생성 기반 디지털 평가의 활용 가능성을 검토하였고, Chung & Kim(2023)은 다변량 일반화가능도 이론을 적용하여 자동문항생성 기반 평가의 신뢰도 탐색 가능성을 제시하였다. L2학습자를 대상으로 한국어 읽기 평가에서의 자동 문항 생성 도입을 연구하거나(Lee & Lee, 2024), 국어과 읽기 영역의 선다형 평가에서 문항 자동 생성 방안 연구(Park & Choi, 2025a)도 이어지고 있다.

반면 AQG는 주어진 텍스트나 지식자원으로부터 질문을 생성하는 자연어 생성 문제로 출발하였다. 초기에는 규칙 기반 구문 변환과 통계적 순위화가 중심이었으나(Heilman & Smith, 2010), 이후 seq2seq 기반 신경망 모델이 도입되면서 문맥을 반영한 질문 생성으로 연구의 무게중심이 이동하였다(Du et al., 2017). Kurdi et al.(2020)의 체계적 문헌고찰은 교육 목적 AQG 연구가 빠르게 성장했지만, 여전히 질문 난이도 통제, 평가 방법의 표준화, 보고 방식의 일관성 측면에서는 한계가 크다고 지적하였다. 결국 AIG가 측정학적 타당성과 문항은행 확장에 강점을 지닌다면, AQG는 문맥 이해와 자연어 생성의 유연성에 강점을 가진다. 고부담 시험형 문항 생성은 이 둘 가운데 어느 하나만으로 충분히 설명되기 어렵고, 측정학적 통제와 자연어 생성 품질을 동시에 고려하는 통합적 관점이 필요하다.

2. Multiple-Choice Item and Distractor Generation

다지선다형 문항의 질은 정답의 명확성만으로 결정되지 않는다. 실제 평가 상황에서는 지문, 발문, 정답, 오답 선지가 하나의 구조를 이루며, 특히 오답 선지는 피험자의 오개념이나 피상적 독해를 자극할 정도로 그럴듯하면서도 객관적으로는 배제 가능해야 한다(Taslimipoor et al., 2024). 이러한 이유로 오답 선지 생성은 질문 생성의 부속 문제가 아니라, 다지선다형 평가의 핵심 하위 과제로 간주된다. Alhazmi et al.(2024)은 오답 선지 생성 전반을 정리한 대규모 연구에서 그럴듯하지만 틀린 선택지를 생성하는 문제는 여전히 독립적 난제로 남아 있으며, 단순한 의미 유사도나 문법적 자연스러움만으로는 좋은 오답 선지를 정의하기 어렵다고 보았다. Benedetto et al.(2025)은 기존의 자동 평가 지표는 참조 오답과의 일치 여부에 의존하기 때문에, 실제 시험 환경에서 오답 선지가 얼마나 효과적으로 학습자를 혼동시키는지와 같은 핵심적인 특성을 충분히 반영하지 못한다. 이러한 한계는 오답 선지의 질을 평가할 때 학생 응답 기반의 지표나 보다 정교한 평가 방식이 필요함을 시사한다.

국내 연구도 이 문제의 중요성을 점차 분명히 하고 있다. Jo et al.(2025)은 일차방정식 문항을 대상으로 오개념 기반 오답지 자동 생성을 시도함으로써, 오답 선지가 단순한 임의 오류가 아니라 학습자의 개념적 오류 구조와 연결되어야 함을 보였다. Park & Choi(2025a)는 국어과 읽기 영역 선다형 평가 문항 생성에서 제로샷, 퓨샷, Chain-of-Thought(CoT) 프롬프트를 비교한 결과, CoT가 문항 타당성, 형식, 내용 측면에서 가장 우수한 평가를 받았다고 보고하였다. 또한 Park & Choi(2025b)는 GAI-HITL 기반 독서 문항 자동 생성의 심리측정학적 타당성을 분석함으로써, 생성 문항 연구가 단순 생성 성공 여부를 넘어 실제 평가도구로서의 적절성을 검증해야 함을 보여주었다. 결국 고부담 읽기 문항 생성에서 핵심은 문항 생성 가능성이 아닌 측정 가능한 문항 생성 가능성에 있으며 오답 선지의 질은 중심 변인이 된다고 할 수 있다.

3. LLM-based Educational Question Generation

대규모 언어모델의 등장 이후 교육용 문항 생성 연구는 질적으로 새로운 국면에 들어섰다. Scaria et al.(2024)은 인도 고등학교 사회과 교육과정을 대상으로 여러 LLM이 Bloom 수준이 다른 문항을 생성할 수 있는지를 검토하였고, 인간 전문가 평가를 통해 생성 문항의 상당수가 교육적으로 유의미한 품질을 보인다고 보고하였다. Ghanem et al.(2022)은 읽기 이해 평가에서 질문 유형 통제가 중요하다는 문제의식 아래, “어떻게 물을 것인가”와 “무엇을 물을 것인가”를 분리한 이단계 질문 생성 모델을 제안하였다. Poon et al.(2024) 역시 PIRLS 기준의 읽기 이해 범주를 고려한 few-shot LLM prompting이 고차적 읽기 이해 문항 생성에 유의미한 가능성을 가진다고 보고하였다. 이러한 연구들은 LLM이 단순히 문법적으로 자연스러운 질문을 만드는 도구가 아니라, 인지 수준이나 읽기 기능과 같은 평가 목적을 어느 정도 반영할 수 있는 생성 도구가 될 수 있음을 보여준다.

국내에서도 관련 연구가 빠르게 축적되고 있다. Lim et al.(2024)은 사회과 자동문항생성을 위해 텍스트, 이미지, 표, 그래프를 포함한 다양한 문항 유형에서 LLM의 활용 가능성과 한계를 탐색하였다. Lee & Lee(2024)은 한국어 읽기 평가에서 자동 문항 생성의 가능성을 검토하였고, Shin(2024)은 LLM 기반 맞춤형 자동문항 생성기의 활용 가능성을 탐색하였다. 국어 과목 연구에서 Park & Choi(2025a)는 국어과 읽기 영역 선다형 평가 문항 생성에서 프롬프트 전략 간 차이를 비교하였고, Hong & Lee(2025)은 LoRA 파인튜닝을 적용한 개방형 GPT 모델 기반 수능 국어 문항 생성 품질을 비교하였다. 이처럼 최근 국내 연구는 사회과, 국어과, 영어교육, 수학교육 등 다양한 교과로 확장되고 있으며, prompt engineering, Human-in-the-Loop, LoRA 기반 적응 등

방법론도 다양화되고 있다. 그러나 상당수 연구가 단일 단계의 문항 생성, 프롬프트 효과 비교, 혹은 전문가 인식 조사 수준에 머물러 있고, 수능형 비문학처럼 고부담·고제약 시험 환경에서 지문 생성과 문항·선지 생성을 연쇄적으로 다루는 연구는 여전히 드물다. 또한 생성 결과를 지문 품질과 문항 품질로 분리하여 검증한 사례도 많지 않다.

4. Fine-Tuning and Evaluation with Large Language Models

고성능 언어모델을 교육평가 도메인에 적용할 때 현실적인 제약은 계산 자원과 데이터 규모이다. 이러한 맥락에서 LoRA와 QLoRA는 제한된 자원 환경에서도 대규모 모델을 특정 도메인에 적응시킬 수 있는 대표적 방법으로 자리잡았다. Hu et al.(2021)은 LoRA를 통해 전체 파라미터를 재학습하지 않고도 효율적인 적응이 가능함을 제시하였고, Dettmers et al.(2023)은 4-bit 양자화와 결합한 QLoRA가 제한된 메모리 환경에서도 고성능 fine-tuning을 가능하게 함을 보였다. 이러한 흐름은 고품질 라벨 데이터를 대량으로 확보하기 어려운 교육 도메인에서 특히 매력적이다.

그러나 소규모 SFT가 항상 실질적 품질 향상으로 이어진다고 보기는 어렵다. 형식이 강하게 제약된 생성 과제에서는 모델이 시험 문체나 출력 형식을 빠르게 모방할 수 있지만, 그것이 곧 지문-문항 간 의미 정합성이나 오답 선지의 변별력 향상으로 이어지는지는 별개의 문제이다. 따라서 고제약 문항 생성에서는 형식 적합성과 내용 품질을 구분하여 평가할 필요가 있다. 이러한 맥락에서 생성 결과 평가 방법 또한 중요해진다.

최근에는 LLM-as-Judge가 자동 평가를 위한 유망한 패러다임으로, 인간 전문가와 지표 기반 자동평가에 비해 높은 확장성과 적응성을 제공하며 다양한 평가 작업에 활용되고 있다(Gu et al., 2025). Zheng et al.(2023)은 MT-Bench와 Chatbot Arena를 통해 LLM 평가 결과가 인간 평가와 높은 상관을 보이며 인간평가보다 일관성이 높음을 확인하였다. 또한, pairwise 비교 방식이 가장 안정적이라고 보고함과 동시에 먼저 제시한 답이 유리해지는 현상(position bias)이 존재함을 밝혔다. Bavaresco et al.(2025)은 20개 NLP 평가 과제에 걸친 대규모 분석을 통해, LLM 평가자는 일부 과제에서는 인간 판단과 잘 정렬되지만 과제 유형, 평가 속성, 인간 평가자의 전문성, 텍스트의 생성 주체에 따라 변동성을 보인다고 지적하였다. 이 결과는 교육평가용 문항 생성처럼 미세한 의미 차이와 오답 선지의 기능적 적절성을 판단해야 하는 과제에서, LLM-as-Judge를 사용할 경우 다중 judge, 인간 평가를 병행해야 함을 시사한다.

Research Method

1. Research Design

본 연구는 수능형 국어 비문학 지문·문항 자동 생성에서 소규모 supervised fine-tuning의 효과와 한계를 단계별로 검증하기 위한 실험연구이다. 전체 생성 과정은 두 단계로 구성하였다. Stage 1에서는 주어진 키워드로부터 수능 문체의 비문학 지문을 생성하고, Stage 2에서는 생성된 지문을 바탕으로 발문과 5지선다형 선지를 생성하였다. 본 연구의 핵심 목적은 소규모 QLoRA 기반 SFT의 효과가 지문 생성과 문항·오답 선지 생성에 동일하게 나타나는지, 혹은 특정 단계에 국한되어 나타나는지를 실증적으로 확인하는 데 있다. 이를 위해 생성 품질을 단일 지표로 평가하지 않고, 지문 품질과 문항 품질을 분리하여 측정하는 이중 평가 프레임워크를 적용하였다. 또한 Qwen2.5-7B-Instruct와 EXAONE 3.5 7.8B를 비교 모델로 설정하고, 파이프라인 버전, 학습 데이터 수량 조건, 출력 형식의 차이를 반영한 ablation 설계를 적용하였다.

2. Data Construction

학습 데이터는 2021학년도부터 2026학년도 수능 국어 비문학 기출 문항을 PDF에서 파싱하여 구축하였다. 파싱 파이프라인은 pdfminer를 사용하여 지문-문항-선지 구조를 추출하고, 규칙 기반 후처리를 통해 선지 번호, 정답 정보, 문항 유형을 레이블링하였다. 이 과정에서는 문학 지문을 제외하고 비문학 지문만 포함하였다. 최종적으로 지문 완결성과 선지 구조 정합성 기준을 충족한 46개 지문을 기본 SFT 학습용 원천 데이터로 확정하였다.

학습 샘플은 각 기출 문항을 하나의 instruction-response 쌍으로 재구성하는 방식으로 구축하였다. instruction에는 해당 지문과 문항 유형을 명시하고, response에는 실제 수능 기출 문항의 발문과 5개 선지를 포함하였다. 이 방식으로 46개 지문에서 총 198개의 instruction-response 쌍을 구성하였다. 기본 데이터세트의 유형 분포는 추론 76개, 사실확인 63개, 보기연계 51개, 전개방식 8개였다. 확장 학습데이터 조건은 모의고사 지문 50개를 추가하여 구축하였으며, 최종적으로 95개 지문과 300개의 instruction-response 쌍으로 구성하였다. 확장 학습데이터의 유형 분포는 추론 158개, 사실확인 127개, 전개방식 15개였고, 보기연계 유형은 포함하지 않았다. 따라서 본 연구에서 198 조건과 300 조건의 차이는 단순한 샘플 수 증가가 아니라, 사용한 지문 범위와 문항 유형 구성이 함께 변한 확장 학습데이터 조건으로 해석하였다.

출력 형식은 텍스트 형식과 JSON 형식의 두 조건으로 구성하였다. 텍스트 형식에서는 발문과 5개 선지를 자연어 서술 형식으로 그대로 출력하도록 하였고, JSON 형식에서는 question, choices, answer, points 필드를 갖는 구조화 포맷을 사용하였다. 이때 answer 필드는 학습 시 null로 설정하고, points 필드에는 배점 정보를 기재하였다. 모델별 입력 형식을 일치시키기 위해 Qwen 계열에는 ChatML 형식을, EXAONE 계열에는 EXAONE 전용 chat template을 적용하였다.

본 연구의 비교 설계는 파이프라인 개선 효과와 SFT 효과를 가능한 한 분리하여 해석할 수 있도록 구성하였다. 구체적으로 Qwen-base(v1)와 Qwen-base(v2)의 비교는 파이프라인 개선 효과를, Qwen-base와 Qwen-SFT-198의 비교는 SFT 기본 효과를, Qwen-SFT-198과 Qwen-SFT-300의 비교는 확장 학습 데이터 조건의 효과를, Qwen-SFT-198과 Qwen-SFT-JSON의 비교는 출력 형식 효과를 측정하도록 설계하였다. EXAONE 계열 역시 base, SFT-198, SFT-300, SFT-JSON 조건으로 구성하여 동일한 논리의 비교를 적용하였다. 다만 EXAONE 계열은 단일 파이프라인 환경에서만 평가되었으므로, Qwen과 EXAONE의 직접 비교에는 파이프라인 버전이 완전히 통제되지 않은 상태가 포함된다는 점을 설계 한계로 명시하였다.

3. Evaluation Data and Gold Standard

평가 데이터는 학습 데이터와 직접 중복되지 않도록 별도로 설계하였다. 이를 위해 총 19개의 평가용 키워드를 과학, 사회, 인문·철학, 예술, 기술의 5개 분야에 걸쳐 구성하였다. 분야별 분포는 과학 4개, 사회 4개, 인문·철학 4개, 예술 4개, 기술 3개였다. 학습-평가 분리는 두 기준에 따라 이루어졌다. 하나는 평가 키워드와 학습 데이터에 사용된 입력 키워드 간 문자열의 완전 일치 여부를 수동으로 확인하는 절차였고, 다른 하나는 동일 개념의 유의어 또는 변형 표현이 학습 데이터에 포함된 경우 해당 분야의 유사 키워드를 평가 데이터에서 제외하는 절차였다. 이를 통해 평가 데이터가 학습 데이터의 일반화 능력을 측정하도록 설계하였다.

또한 생성 결과의 절대적 품질 상한을 비교하기 위해 실제 수능 기출 데이터를 Gold Standard로 구성하였다. 이를 위해 평가 키워드와 분야 분포가 유사한 기출 지문 19개를 연구자가 수동 선정하였다. 선정 기준은 분야

분포의 매칭, 실제 수능 비문학 지문의 통상적 길이 범위, 사실확인·추론·전개방식 문항의 포함 여부, 그리고 생성 데이터와 비교 가능한 난이도와 문체를 갖추었는지에 대한 수동 검토였다.

4. End-to-End Generation Pipeline

본 연구의 생성 파이프라인은 지문 생성과 문항·선지 생성의 두 단계로 이루어진다. 입력 키워드는 외래어 조어 등 할루시네이션 유발 가능성이 높은 항목을 사전에 차단하는 키워드 필터를 거친 뒤 생성 함수에 전달된다. 입력 키워드 k가 주어지면 Stage 1 생성 함수가 수능형 비문학 지문 p를 생성하고, 이어서 Stage 2 생성 함수가 지문 p를 입력으로 받아 문항 유형 t에 따라 발문과 5지선다 선지를 생성한다. 전체 흐름은 키워드 입력 → 지문 생성 → 지문 품질 필터 → 문항·선지 생성 → 선지 품질 필터 → 정답 분포 균등화 → JSONL 출력의 순서로 설계하였다. 문항 유형은 사실확인, 추론, 전개방식의 세 범주로 고정하였다.

Stage 1의 지문 생성 프롬프트는 도입-원리-분류/대비-적용의 4단계 단락 구조를 명시적으로 요구하도록 구성하였다. 생성 지문은 1,300~1,700자 범위의 한국어 학술 산문체를 목표로 하였으며, 최소 길이, 한국어 비율, 영어 단어 과다 여부, 반복 패턴, 특정 문장 시작의 과도한 반복을 기준으로 품질 필터를 적용하였다. 구체적으로는 900자 미만, 한국어 비율 50% 미만, 허용 목록 외 영어 단어 3개 초과, 15자 구간 n-gram 7회 이상 반복, '이는/이런/이를'로 시작하는 문장이 4회 이상 연속될 경우 생성물을 폐기하고 최대 6회까지 재생성하였다. 재생성 정책은 first-pass 채택 방식으로, 품질 필터를 통과한 첫 번째 생성물을 최종 지문으로 사용하였다. 또한 모든 모델에 동일한 재생성 규칙을 적용하였고, vLLM 추론 서버의 난수 시드는 0으로 고정하였다.

Stage 2에서는 사실확인, 추론, 전개방식의 세 유형에 대해 각각 문항과 선지를 생성하였다. 각 유형별 프롬프트에는 단계별 설계 절차와 선지 작성 규칙을 포함시켰고, 선지는 30~80자 범위의 한국어 서술형을 기본으로 하였다. 선지 유효성 검사는 선지 수의 정확성, 플레이스홀더 포함 여부, 완전 중복, 구조 중복, 최소 길이 충족 여부를 기준으로 수행하였다. 특히 앞 15자가 동일한 선지가 3개 이상 반복될 경우 구조 중복으로 간주하여 폐기하였다. 생성된 문항의 정답 번호는 누적 분포 기준으로 가장 덜 사용된 번호로 회전시키는 정답 분포 균등화 함수를 적용하여 균등성을 확보하였다.

5. Model and Fine-Tuning Configuration

모든 Qwen 어댑터는 Qwen2.5-7B-Instruct를, EXAONE 계열 어댑터는 EXAONE 3.5 7.8B를 베이스 모델로 사용하였다. 모든 파인튜닝은 QLoRA 방식으로 수행하였으며, 4-bit NF4 양자화와 LoRA를 결합하였다. 양 모델 공통으로 LoRA rank는 16, alpha는 32, target modules는 q_proj와 v_proj, dropout은 0.05로 설정하였다. 학습률은 2e-4, 스케줄러는 cosine, warmup ratio는 0.05, epoch 수는 5, 최대 시퀀스 길이는 2048, optimizer는 paged_adamw_8bit, 연산 정밀도는 BF16으로 통일하였다. 유효 배치 크기는 두 모델 모두 16으로 유지하였으나, Qwen은 per-device batch size 2와 gradient accumulation 8을, EXAONE은 per-device batch size 1과 gradient accumulation 16을 사용하였다.

EXAONE 학습에는 두 가지 구현상 주의사항이 있었다. EXAONE의 응답 포맷에서는 completion-only loss 적용이 안정적으로 작동하지 않아 full sequence cross-entropy loss를 사용하였다. 또한 gradient checkpointing은 4-bit 모델 초기화 이후 별도로 재활성화하는 순서를 따라야 했다. 일관성을 위해 Qwen 계열에도 동일하게 full sequence CE를 적용하였다. Qwen 대비 EXAONE 계열은 배치 크기를 절반으로 줄이고 gradient

accumulation을 두 배로 늘려 유효 배치 크기를 맞추었으나, 이러한 배치 구성 차이가 gradient 추정 분산에 미세한 차이를 유발할 가능성은 설계 한계로 고려하였다.

본 연구의 학습 및 추론 실험은 단일 GPU 환경에서 수행하였다. 구체적으로 RTX 4090 GPU 24GB 1대, Intel CPU Ultra 5 225F, RAM 32GB가 장착된 시스템에서 WSL 기반 Ubuntu 22.04 환경을 사용하였다. 이러한 자원 제약은 본 연구에서 7B–8B 규모의 오픈웨이트 모델과 QLoRA 기반 파인튜닝 전략을 채택한 실질적 배경이 된다. 본 연구의 모델 선택은 단순 성능 비교를 위한 것이 아니라, 단일 범용 GPU 환경에서도 수능형 국어 문항 자동 생성의 가능성과 한계를 검증할 수 있는 현실적 조건을 반영한 것이다.

6. Generation Settings

모든 생성은 vLLM 기반 서빙 환경에서 수행하였다. LoRA 어댑터는 동적으로 로딩하였으며, max_model_len은 8192, temperature는 0.7, max_tokens는 1024, top_p는 0.95로 설정하였다. 평가용 19개 키워드 각각에 대해 지문 1개와 문항 3개를 생성하였다. 생성 과정에서 품질 필터를 통과하지 못한 경우에는 중복 선지, 영어 단어 과다, 지문 반복 패턴, 선지 유효성 검사 실패 등에 대해 조건별 재시도 정책을 적용하였다. 본 연구에서 보고하는 성능은 모두 품질 필터를 통과한 최종 생성물을 기준으로 한다.

7. Evaluation Methods and Statistical Analysis

평가는 지문 품질과 문항 품질을 분리한 이중 평가 프레임워크로 수행하였다. Method A는 문항 품질 평가로, 동일 키워드에 대해 생성된 두 결과의 지문, 발문, 5개 선지를 함께 제시한 뒤 지문연계성, 오답매력도, 수능문체, 변별력, 언어품질의 다섯 차원에서 A/B/동점을 판정하도록 설계하였다. Method B는 지문 품질 평가로, 문항과 선지를 제외한 지문만 제시한 뒤 내용정확성, 수능문체, 논리일관성, 난이도적절성의 네 차원에서 A/B/동점을 판정하도록 설계하였다. 이러한 분리 설계는 문항 품질이 지문 평가에 혼입되는 것을 방지하고, SFT의 효과가 두 단계 중 어느 지점에 집중되는지 확인하기 위한 것이다.

본 연구는 Zheng et al.(2023)의 방법론을 참조하여, 응답 순서를 무작위로 교환한 pairwise 비교 평가와 bootstrap 기반 신뢰구간 추정 방식을 적용하였다. 이러한 접근은 인간 평가와 높은 수준의 일치도를 보이면서도, 대규모 비교 실험에서 일관성과 확장성을 확보할 수 있는 평가 방법으로 제시된 바 있다.

LLM-as-Judge에는 Claude Haiku와 GPT-4o-mini를 사용하였다. 두 judge 모두 동일한 한국어 프롬프트를 적용하였고, 출력은 JSON 형식으로 강제하여 파싱 안정성을 확보하였다. 또한 position bias를 통제하기 위해 A/B 제시 순서를 50% 확률로 무작위 교환한 뒤, 판정 결과를 원래 순서 기준으로 역변환하였다. 모든 비교쌍에 대해 bootstrap 재표집 10,000회를 수행하여 95% 신뢰구간을 산출하였으며, 신뢰구간이 50%를 포함하지 않을 때 유의한 차이로 해석하였다.

Method A의 경우 동일 키워드에 속한 3개 문항이 동일 지문을 공유하므로 완전 독립이라고 보기 어렵다. 이에 따라 5개 차원의 판정을 키워드 단위로 합산한 뒤, 동점 판정을 제외한 결정적 판정을 대상으로 키워드 단위 집계 벡터를 bootstrap 입력으로 사용하였다. 따라서 Method A의 실질적 재표집 단위는 19개 키워드이다. 반면 Method B는 키워드당 지문 1개를 직접 비교하므로 상대적으로 독립성 가정이 강하다. 다만 Method A에서는 문항 간 비독립성으로 인해 신뢰구간이 다소 낙관적으로 추정될 가능성이 있으며, 이는 본 연구의 방법론적 한계로 간주하였다.

Method A의 경우 동일 키워드에 속한 3개 문항이 동일 지문을 공유하므로 완전 독립이라고 보기 어렵다.

이에 따라 5개 차원의 판정을 키워드 단위로 합산한 뒤, 동점 판정을 제외한 결정적 판정을 대상으로 키워드 단위 집계 벡터를 bootstrap 입력으로 사용하였다. 따라서 Method A의 실질적 재표집 단위는 19개 키워드이다. 반면 Method B는 키워드당 지문 1개를 직접 비교하므로 상대적으로 독립성 가정이 강하다. 다만 Method A에서는 문항 간 비독립성으로 인해 신뢰구간이 다소 낙관적으로 추정될 가능성이 있으며, 이는 본 연구의 방법론적 한계로 간주하였다.

본 연구에서는 pairwise 비교 결과를 해석하기 위해 최종 판정 규칙을 다음과 같이 정의하였다. 각 비교쌍에 대해 bootstrap 기반으로 추정된 A 승률의 95% 신뢰구간이 50%를 포함하지 않을 경우, 해당 조건을 유의한 우세(significant preference)로 판단하였다. 구체적으로 신뢰구간 전체가 50%를 초과할 경우 A 우세, 50% 미만일 경우 B 우세로 판정하였다.

반면 신뢰구간이 50%를 포함하는 경우에는 통계적으로 유의한 차이가 없는 것으로 간주하여 미결정(inconclusive)으로 분류하였다. 이때 개별 평가에서 동점으로 판정된 경우는 bootstrap 입력에서 제외되며, 최종 판정에는 영향을 미치지 않도록 처리하였다.

또한 두 LLM judge(Claude Haiku, GPT-4o-mini)의 판정 결과가 서로 상이할 경우에는 judge 간 불일치(disagreement)로 표기하였다. 이는 동일 비교쌍에 대해 하나의 judge에서는 유의한 우세가 나타났으나 다른 judge에서는 반대 결과 또는 미결정이 나타난 경우를 포함한다.

Table 1. Overview of experimental design

Category	Method A	Method B
평가 대상	지문 + 발문 + 5개 선지	지문만
평가 목적	문항 품질 평가	지문 품질 평가
평가 차원	지문연계성, 오답매력도, 수능문체, 변별력, 언어품질	내용정확성, 수능문체, 논리일관성, 난이도적절성
Judge 모델	Claude Haiku, GPT-4o-mini	
편향 보정	A/B 순서 50% 무작위 교환	
통계 처리	Bootstrap 10,000회, 95% CI	
재표집 단위	키워드 단위(19개)	

Research Result

1. Statistical Quality Metrics

생성 결과의 표면적 특성을 비교하기 위해 선지 길이, 선지 편차, TTR(Type-Token Ratio), 한국어 비율, 문항 탈락률을 산출하였다. 실제 수능 기출(표 2의 Real)은 생성 모델에 비해 선지 평균 길이가 짧고, 어휘 다양성은 높게 나타났다. 반면 생성 모델은 전반적으로 더 긴 선지를 산출하였으며, SFT 적용 이후 형식적 균일성은 일부 개선되었으나 어휘 다양성은 감소하는 경향을 보였다. 특히 EXAONE-SFT 계열에서는 base 대비 TTR이 일관되게 낮아졌고, JSON 형식에서는 선지 길이 편차가 가장 크게 나타났다. 또한 문항 탈락률은 SFT 모델 계열에서 더 높게 나타나, 품질 필터가 모델별 성능 측정에 실질적 영향을 미치고 있음을 보여준다.

Table 2에서 보듯이, Qwen-SFT-300은 Qwen 계열 가운데 선지 전체 편차와 TTR 측면에서 실제 수능 형식에 상대적으로 가장 근접한 값을 보였다. 그러나 EXAONE-SFT-300과 EXAONE-SFT-JSON은 표면 지표상으로는 어휘 다양성 감소와 선지 불균형 확대를 보였으며, 이는 후술할 judge 기반 평가와 반드시 일치하지 않았다. 즉, 자동 지표상 형식적 근접성과 실제 내용 품질은 별개임을 알 수 있었다.

Table 2. Comparison of statistical quality metrics

Metric	Real	Qwen-base	Qwen -SFT -198	Qwen -SFT- 300	EXAONE -base	EXAONE -SFT -198	EXAONE -SFT -300	EXAONE -SFT -JSON
총 문항 수	85	51	44	52	51	42	56	45
문항 탈락률(%)	—	10.5	22.8	8.8	10.5	26.3	1.8	21.1
선지 평균 길이(자)	28.4	49.2	40.4	40.9	59.5	56.5	69.7	60.1
선지 전체 편차	8.5	16.9	14.7	11.5	14.4	21.4	30.4	45.0
선지 문항내 편차	1.2	8.9	4.6	5.7	8.1	10.6	9.0	12.1
지문 평균 길이(자)	1,710	1,487	1,447	1,403	1,863	1,899	1,895	1,928
TTR	.598	.391	.335	.399	.445	.376	.271	.336
한국어 비율(%)	90.9	95.9	97.0	96.1	96.6	96.6	94.9	96.5

2. Item Quality for Qwen-SFT-198 (Method A)

문항 품질(Method A)을 기준으로 보면 Qwen 계열에서는 파인튜닝 역설이 나타났다. 실제 수능 문항은 두 생성 조건을 전 차원에서 거의 완전하게 상회하였으며(전체 평균 99.1%), Qwen-base는 Qwen-SFT-198을 평균 70.3% 승률로 앞섰다. 그 중 오답매력도(73.2%)와 지문연계성(71.4%)에서 Qwen-base의 우세가 뚜렷하게 나타났다. 이는 소규모 SFT가 문항 구조의 형식적 정렬에는 일부 기여하였으나, 실제 평가 품질을 향상시키지 못했음을 의미한다.

Table 3. Pairwise win rates (%) for Qwen-SFT-198

Metric	Real > base	Real > SFT-198	base > SFT-198
지문연계성	100.0	100.0	71.4
오답매력도	100.0	100.0	73.2
수능문체	100.0	100.0	71.4
변별력	100.0	100.0	69.0
언어품질	97.8	95.5	66.7
전체 평균	99.6	99.1	70.3

두 judge를 동일 조건으로 적용한 Bootstrap 95% CI 결과는 Table 4와 같다. Qwen-SFT-198은 형식적 정렬 효과를 일부 보였음에도, 문항 품질 차원에서는 base 모델을 넘어서지 못했다. Claude Haiku는 Qwen-base의 유의한 우세(71.2%)를 검출하였으나, GPT-4o-mini는 전 쌍에서 미결정으로 판정하였다. 이는 소규모 SFT가 수능형 문항 생성에서 지문 자체보다 문항·오답 선지 단계에서 더 취약하게 작동할 가능성을 보여준다.

Table 4. Bootstrap 95% CI for Qwen-SFT-198 comparisons (item quality, dual judge). Bootstrap N=10,000, alpha=0.05.

Comparison (A vs B)	Haiku	95% CI	Result	4o-mini	95% CI	Result
Real vs base	99.60%	[98.7%, 100.0%]	A 유의 우세	50.20%	[44.0%, 56.4%]	미결정
base vs SFT-198	71.20%	[64.9%, 77.1%]	A 유의 우세	51.90%	[45.2%, 58.6%]	미결정
Real vs SFT-198	99.10%	[97.7%, 100.0%]	A 유의 우세	47.70%	[41.4%, 54.1%]	미결정

3. Effects of Expanded Training Data and Output Format (Qwen)

확장 학습 데이터 조건에서도 Qwen 계열에서는 SFT 모델이 base를 상회하는 패턴이 나타났다. Qwen-base vs Qwen-SFT-300에서 base의 차원별 승률은 38.5~40.4%로, SFT-300이 전 차원에서 우세하였다. Qwen-

SFT-198 vs Qwen-SFT-300은 44.2~48.8%로 미결정 수준이었으며, 실제 수능 문항은 Qwen-SFT-300을 전 차원 100.0%(전체 98.8%)로 압도하였다. 다만 이 비교는 단순 샘플 수 증가가 아니라 지문 범위와 유형 구성의 변화까지 포함하므로, 순수한 데이터 규모 효과로 해석해서는 안 된다.

출력 형식 변경도 유사한 패턴을 보였다. Qwen-base vs Qwen-SFT-JSON에서 base의 차원별 승률은 32.7~36.7%로, JSON 모델이 전 차원에서 우세하였다. Qwen-SFT-300 vs Qwen-SFT-JSON은 53.3~56.5%로 SFT-300이 소폭 우세하나 통계적 유의성은 없었다. 실제 수능 문항은 Qwen-SFT-JSON을 전 차원 100.0%로 상회하였다.

Table 5. Pairwise win rates (%) for Qwen-SFT-300 and Qwen-SFT-JSON ablation

Matrix	base > SFT-300	SFT-198 > SFT-300	Real > SFT-300	base > SFT-JSON	SFT-300 > SFT-JSON	Real > SFT-JSON
지문연계성	38.5	46.5	100.0	32.7	54.3	100.0
오답매력도	38.5	44.2	100.0	36.7	54.3	100.0
수능문체	38.5	44.2	100.0	34.7	56.5	100.0
변별력	40.4	48.8	100.0	34.7	56.5	100.0
언어품질	48.1	51.2	94.2	46.9	44.2	100.0
전체 평균	40.8	47.0	98.8	37.1	53.3	100.0

SFT-300 및 SFT-JSON ablation 전체에 대해 Claude Haiku와 GPT-4o-mini 두 judge를 동일 조건으로 적용한 Bootstrap 95% CI 결과는 Table 6와 같다. Claude Haiku 기준으로 Qwen-base vs Qwen-SFT-300(A승률 40.8%, B 유의 우세)과 Qwen-base vs Qwen-SFT-JSON(A승률 37.1%, B 유의 우세)에서 SFT 모델이 유의하게 우세한 것으로 나타났다. 이는 차원별 승률(Table 5)과 일관된 방향으로, SFT-300과 SFT-JSON이 base 대비 문항 품질에서 개선 신호를 보였음을 확인한다. 다만 4.2의 SFT-198에서 관찰된 파인튜닝 역설과는 대조적인 결과이며, 확장 학습 데이터 조건(SFT-300)과 JSON 형식(SFT-JSON)이 SFT-198의 품질 저하를 어느 정도 해소하였을 가능성을 시사한다. GPT-4o-mini는 전 쌍에서 미결정으로 판정하였다.

Table 6. Bootstrap 95% CI for Qwen ablation comparisons (item quality, dual judge)

Comparison (A vs B)	Haiku	95% CI	Result	4o-mini	95% CI	Result
base vs SFT-300	40.8%	[34.6%, 46.9%]	B 유의 우세	51.9%	[45.8%, 58.1%]	미결정
SFT-198 vs SFT-300	47.0%	[40.5%, 54.0%]	미결정	48.4%	[41.4%, 54.9%]	미결정
Real vs SFT-300	98.8%	[97.3%, 100.0%]	A 유의 우세	51.5%	[45.4%, 57.7%]	미결정
base vs SFT-JSON	37.1%	[31.0%, 43.3%]	B 유의 우세	49.4%	[43.3%, 55.5%]	미결정
SFT-300 vs SFT-JSON	53.3%	[46.7%, 59.9%]	미결정	48.7%	[42.2%, 55.2%]	미결정
Real vs SFT-JSON	100.0%	[100.0%, 100.0%]	A 유의 우세	51.8%	[45.7%, 58.0%]	미결정

4. Item Quality for EXAONE-SFT-198 (Method A)

실제 수능 문항은 EXAONE-base를 전체 82.7%, EXAONE-SFT-198을 전체 84.8%로 압도하였다. EXAONE-base vs SFT-198 비교에서는 SFT-198 쪽이 방향적으로 우세(base 전체 41.0%)하였으나, Bootstrap CI 기준 통계적 유의성에는 이르지 못했다. 수능문체(base 33.3%)에서 SFT-198의 우위가 가장 뚜렷하였고, 오답매력도와 변별력(각 38.5%), 지문연계성(43.6%) 순이었다.

이 결과는 EXAONE-SFT-198이 base 대비 전반적으로 개선 방향의 신호를 보였으나, 그 효과가 충분히 크지 않아 통계적으로 확정되지는 않았음을 의미한다. 또한 실제 수능과의 격차는 EXAONE-base와 EXAONE-

SFT-198에서 유사한 수준으로 유지되어, 198개 규모의 소규모 파인튜닝만으로는 실제 수능 문항과의 품질 격차를 유의하게 축소하지 못하였음을 보여준다.

Table 7. Pairwise win rates (%) for EXAONE-SFT-198 (item quality)

Matrix	Real > base	Real > SFT-198	base > SFT-198
지문연계성	88.2	88.1	43.6
오답매력도	88.2	81.0	38.5
수능문체	88.2	90.5	33.3
변별력	88.2	83.3	38.5
언어품질	60.8	81.0	51.3
전체 평균	82.7	84.8	41.0

두 judge를 동일 조건으로 적용한 Bootstrap 95% CI 결과는 Table 8과 같다. Claude Haiku는 Real이 EXAONE-base와 EXAONE-SFT-198을 각각 유의하게 상회함을 검출하였으나, EXAONE-base와 EXAONE-SFT-198의 직접 비교는 미결정으로 판정하였다. GPT-4o-mini는 모든 비교에서 미결정이었다. EXAONE-SFT-198은 차원별 승률에서 base 대비 방향적 우세를 보였으나, Bootstrap CI 기준으로는 두 judge 모두 미결정이었다. 이는 Qwen-SFT-198이 base 대비 명확한 열세를 보인 결과와 대비되며, EXAONE 계열에서는 소규모 SFT 단계에서도 최소한 품질 저하 없이 개선 가능성이 관찰되었음을 시사한다. 그러나 그 효과는 아직 통계적으로 확정할 수준에는 이르지 못하였다.

Table 8. Bootstrap 95% CI for EXAONE-SFT-198 comparisons (item quality, dual judge)

Comparison (A vs B)	Haiku	95% CI	Result	4o-mini	95% CI	Result
Real vs base	82.7%	[78.0%, 87.5%]	A 유의 우세	50.6%	[44.7%, 56.9%]	미결정
Real vs SFT-198	84.8%	[80.0%, 89.5%]	A 유의 우세	49.5%	[42.9%, 56.2%]	미결정
base vs SFT-198	41.0%	[25.6%, 56.4%]	미결정	48.7%	[33.3%, 64.1%]	미결정

5. Effects of Expanded Training Data and Output Format (EXAONE)

확장 학습 데이터 조건(SFT-300)과 출력 형식 변경(SFT-JSON)에서는 EXAONE 계열이 Qwen과 다른 방향의 결과를 보였다. EXAONE-SFT-300은 EXAONE-base를 전 차원에서 상회하였으며, 특히 지문연계성(73.2%)과 수능문체(75.6%)에서 뚜렷한 우위를 나타냈다. EXAONE-SFT-JSON 역시 base 대비 개선 방향을 보였으나, EXAONE-SFT-300과의 비교에서는 SFT-300이 전체 평균 58.0%로 더 우세하였다. 반면 EXAONE-SFT-198과 EXAONE-SFT-300의 비교는 전체 평균 45.9%로 미결정 수준에 머물렀다.

실제 수능 문항은 EXAONE-SFT-300과 EXAONE-SFT-JSON을 모두 상회하였으나, 그 격차는 각각 전체 평균 65.9%, 77.8%로 나타나 EXAONE-SFT-300이 EXAONE-SFT-JSON보다 실제 수능에 상대적으로 더 근접한 품질을 보였다. 이는 EXAONE 계열에서 데이터 규모 확대가 품질 향상에 실질적으로 기여하였으며, 동시에 출력 형식 측면에서는 JSON보다 텍스트 형식이 더 유리할 가능성을 시사한다.

Table 9. Pairwise win rates (%) for EXAONE-SFT-300 and EXAONE-SFT-JSON ablation

Matrix	base > SFT-300	SFT-198 > SFT-300	base > SFT-JSON	SFT-300 > SFT-JSON	Real > SFT-300	Real > SFT-JSON
지문연계성	26.8	43.2	40.5	62.5	68.2	82.2
오답매력도	36.6	45.9	35.7	55.0	65.9	75.6
수능문체	24.4	48.6	33.3	57.5	68.2	80.0
변별력	34.1	48.6	35.7	57.5	65.9	75.6
언어품질	36.6	43.2	40.5	57.5	61.4	75.6
전체 평균	31.7	45.9	37.1	58.0	65.9	77.8

Bootstrap 95% CI 결과에서도 이러한 경향은 유지되었다. Claude Haiku 기준으로 EXAONE-SFT-300과 EXAONE-SFT-JSON은 모두 base 대비 유의한 우세를 보였고, EXAONE-SFT-300은 EXAONE-SFT-JSON도 유의하게 상회하였다. 반면 GPT-4o-mini는 모든 비교를 미결정으로 판정하였다. 이는 EXAONE 계열의 품질 향상 신호가 judge 모델에 따라 다르게 감지될 수 있음을 보여준다.

또한 이 결과는 자동 지표와 judge 기반 평가가 반드시 같은 방향을 보이지 않음을 보여준다. EXAONE-SFT-300은 TTR 저하와 선지 편차 증가를 보였음에도, judge 기반 문항 품질 평가에서는 EXAONE-base보다 유의하게 높은 값을 보였다. 따라서 표면적 다양성이나 길이 편차와 같은 형식 지표만으로 실제 문항 품질을 충분히 대표하기는 어렵다.

Table 10. Bootstrap 95% CI for EXAONE ablation comparisons (item quality, dual judge)

Comparison (A vs B)	Haiku	95% CI	Result	4o-mini	95% CI	Result
base vs SFT-300	31.7%	[25.4%, 38.0%]	B 유의 우세	49.3%	[42.4%, 56.1%]	미결정
SFT-198 vs SFT-300	45.9%	[38.9%, 53.0%]	미결정	51.9%	[44.9%, 58.9%]	미결정
base vs SFT-JSON	37.1%	[31.0%, 43.8%]	B 유의 우세	47.6%	[41.0%, 54.3%]	미결정
SFT-300 vs SFT-JSON	58.0%	[51.0%, 65.0%]	A 유의 우세	49.5%	[42.5%, 56.5%]	미결정
Real vs SFT-300	65.9%	[59.5%, 72.3%]	A 유의 우세	52.3%	[45.9%, 58.6%]	미결정
Real vs SFT-JSON	77.8%	[72.4%, 83.1%]	A 유의 우세	49.8%	[43.1%, 56.0%]	미결정

6. Cross-Model Comparisons (Method A)

교차 모델 비교로 Qwen과 EXAONE의 모델 계열 간 품질 격차를 측정하였다. EXAONE-base는 Qwen-base(25.1%)와 Qwen-SFT-300(21.6%)을 전 차원에서 유의하게 상회하였다. 나아가 EXAONE-SFT-300은 Qwen-SFT-300(5.6%)과 Qwen-base(6.3%)를 더욱 압도적으로 상회하여, 한국어 특화 사전학습과 도메인 SFT의 결합이 범용 모델 대비 극단적 품질 격차를 만들어냄을 확인하였다. 다만 두 모델 간 파라미터 규모·사전학습 데이터 등 교란 변인이 완전히 통제되지 않았으므로, 순수한 사전학습 효과로 과도하게 일반화해서는 안 된다.

Table 11. Pairwise comparison results (%) between Qwen and EXAONE models

Matrix	Qwen-base vs EXAONE-base	Qwen-SFT-300 vs EXAONE-base	Qwen-base vs EXAONE-SFT-300	Qwen-SFT-300 vs EXAONE-SFT-300
지문연계성	25.5	20.4	2.6	2.3
오답매력도	25.5	18.4	5.3	2.3
수능문체	25.5	24.5	5.3	2.3
변별력	23.5	16.3	7.9	2.3
언어품질	25.5	28.6	10.5	18.6
전체 평균	25.1	21.6	6.3	5.6

두 judge를 동일 조건으로 적용한 Bootstrap 95% CI 결과는 Table 12와 같다. 교차 모델 비교 결과는 한국어 특화 사전학습의 이점을 일관되게 보여준다. Claude Haiku 기준으로 EXAONE-base는 Qwen-base(25.1%)와 Qwen-SFT-300(21.6%)을 유의하게 상회하였고, EXAONE-SFT-300은 이 격차를 더욱 확대하여 Qwen-base(6.3%)와 Qwen-SFT-300(5.6%)을 압도하였다. 특히 Qwen-SFT-300 vs EXAONE-SFT-300에서 Qwen의 승률이 5.6%에 불과한 것은, 동일한 SFT 조건에서도 베이스 모델의 한국어 기반 능력이 결정적 차이를 만들어냄을 시사한다. 실제 수능 문항만이 EXAONE-base를 유의하게 상회하였다(82.7%). GPT-4o-mini는 신뢰구간에서 50%를 포함하여 전 쌍 미결정으로, 두 judge 간 판별력 차이가 교차 모델 비교에서도 동일하게 유지됨을 확인할 수 있다.

Table 12. Bootstrap 95% CI for cross-model comparisons (litem quality, dual judge)

Comparison (A vs B)	Haiku	95% CI	Result	4o-mini	95% CI	Result
Qwen-base vs EXAONE-base	25.1%	[20.0%, 30.2%]	B 유의 우세	52.9%	[46.7%, 59.2%]	미결정
Qwen-SFT-300 vs EXAONE-base	21.6%	[16.7%, 26.9%]	B 유의 우세	49.8%	[43.7%, 56.3%]	미결정
Qwen-base vs EXAONE-SFT-300	6.3%	[3.2%, 10.0%]	B 유의 우세	52.6%	[45.3%, 60.0%]	미결정
Qwen-SFT-300 vs EXAONE-SFT-300	5.6%	[2.8%, 8.8%]	B 유의 우세	50.7%	[44.2%, 57.2%]	미결정

7. Passage Quality Comparison (Method B)

문항을 제외한 지문만을 평가 대상으로 하는 지문 전용 평가(Method B)를 수행하였다. 평가 차원은 내용정확성, 수능문체, 논리일관성, 난이도적절성의 4개이며, Claude Haiku와 GPT-4o-mini 두 judge를 동일 조건으로 적용하였다.

두 모델 계열 모두에서 SFT 전후 지문 품질 차이는 Table 13과 같이 통계적으로 유의하지 않았다. Qwen 계열 5개 비교 쌍과 EXAONE 계열 5개 비교 쌍의 내부 비교에서 Claude Haiku 기준 A승률은 27%에서 63% 범위에 분포하였으며, 모든 비교 쌍에서 95% 신뢰구간이 50%를 포함하여 유의한 차이가 검출되지 않았다(Table 13). GPT-4o-mini 역시 모든 비교에서 신뢰구간에 50%를 포함하여 미결정으로 나타났다. 이는 소규모 SFT가 지문 생성 품질 자체에는 유의한 영향을 미치지 않음을 의미하며, Method A에서 관찰된 파인튜닝 역설(Qwen) 또는 성능 향상 신호(EXAONE)가 지문이 아닌 문항 및 오답 선지 생성 단계에서 발생한 현상임을 뒷받침한다.

Table 13. Bootstrap 95% CI for cross-model comparisons (item quality, dual judge)

Comparison (A vs B)	Haiku(A)	95% CI	4o-mini(A)	95% CI
Qwen-base vs Qwen-SFT-198	58.7%	[31.2%, 81.2%]	50.0%	[25.0%, 75.0%]
Qwen-base vs Qwen-SFT-300	27.8%	[11.1%, 50.0%]	53.7%	[33.3%, 77.8%]
Qwen-base vs Qwen-SFT-JSON	53.9%	[31.6%, 73.7%]	47.4%	[26.3%, 68.4%]
Qwen-SFT-198 vs Qwen-SFT-300	35.9%	[12.5%, 62.5%]	47.9%	[25.0%, 75.0%]
Qwen-SFT-300 vs Qwen-SFT-JSON	50.7%	[27.8%, 72.2%]	48.1%	[27.8%, 72.2%]
EXAONE-base vs EXAONE-SFT-198	36.8%	[11.8%, 58.8%]	51.0%	[29.4%, 76.5%]
EXAONE-base vs EXAONE-SFT-300	45.6%	[23.5%, 70.6%]	51.0%	[29.4%, 76.5%]
EXAONE-base vs EXAONE-SFT-JSON	26.5%	[11.8%, 52.9%]	54.9%	[29.4%, 76.5%]
EXAONE-SFT-198 vs EXAONE-SFT-300	62.5%	[38.9%, 83.3%]	48.1%	[27.8%, 72.2%]
EXAONE-SFT-300 vs EXAONE-SFT-JSON	61.1%	[38.9%, 83.3%]	61.1%	[38.9%, 83.3%]

실제 수능 지문과 모든 생성 모델을 비교한 결과 Table 14와 같이 일관되게 상회하였다. Claude Haiku 기준으로 Real은 Qwen-SFT-300(97.2%), Qwen-SFT-JSON(100.0%), EXAONE-base(94.1%), EXAONE-SFT-300(89.5%), EXAONE-SFT-JSON(88.9%)에 대해 모두 유의한 우위를 보였다(Table 14). 반면 GPT-4o-mini는 모든 비교에서 미결정으로 나타났다. 이는 지문 생성 단계에서도 실제 수능과 생성 모델 간 질적 격차가 유지됨을 보여준다.

Table 14. Passage quality pairwise comparison — Real vs generated models

Comparison (A vs B)	Haiku(A)	95% CI	Result	4o-mini(A)	95% CI
Real vs Qwen-SFT-300	97.2%	[83.3%, 100.0%]	A 유의 우세	51.9%	[27.8%, 72.2%]
Real vs Qwen-SFT-JSON	100.0%	[100.0%, 100.0%]	A 유의 우세	53.4%	[31.6%, 73.7%]
Real vs EXAONE-base	94.1%	[82.4%, 100.0%]	A 유의 우세	42.3%	[17.6%, 64.7%]
Real vs EXAONE-SFT-300	89.5%	[73.7%, 100.0%]	A 유의 우세	48.3%	[26.3%, 68.4%]
Real vs EXAONE-SFT-JSON	88.9%	[72.2%, 100.0%]	A 유의 우세	49.1%	[27.8%, 72.2%]
Qwen-SFT-300 vs EXAONE-SFT-300	5.6%	[2.8%, 8.8%]	B 유의 우세	50.7%	[44.2%, 57.2%]

교차 모델 비교 결과 Table 15와 같이 EXAONE 계열은 Qwen 계열을 지문 품질에서도 일관되게 상회하였다. Claude Haiku 기준으로 Qwen-base의 EXAONE-base 대비 A승률은 2.9%, Qwen-SFT-300의 EXAONE-base 대비 A승률은 4.4%로 나타났으며, EXAONE-SFT-300과의 비교에서도 각각 10.9%와 9.7%로 유의하게 낮은 값을 보였다(Table 15). 이는 한국어 특화 사전학습이 문항 품질뿐 아니라 지문 생성 품질에서도 우위를 제공함을 시사한다. GPT-4o-mini는 모든 비교에서 미결정으로 나타났다.

Table 15. Passage quality pairwise comparison — Cross-model

Comparison (A vs B)	Haiku(A)	95% CI	Result	4o-mini(A)	95% CI
Qwen-base vs EXAONE-base	2.9%	[0.0%, 0.0%]	B 유의 우세	47.1%	[23.5%, 70.6%]
Qwen-SFT-300 vs EXAONE-base	4.4%	[0.0%, 17.6%]	B 유의 우세	47.1%	[23.5%, 70.6%]
Qwen-base vs EXAONE-SFT-300	10.9%	[4.7%, 18.8%]	B 유의 우세	44.9%	[30.6%, 59.2%]
Qwen-SFT-300 vs EXAONE-SFT-300	9.7%	[4.2%, 16.7%]	B 유의 우세	45.5%	[32.7%, 58.2%]

Discussion

1. Differences between High-Stakes Exam Item Generation and AQG

본 연구의 결과는 수능형 국어 비문학 문항 생성이 일반적인 AQG와 질적으로 다른 과제라는 점을 다시 확인시켜 준다. AQG 연구는 규칙 기반 질문 생성에서 신경망 기반 질문 생성으로 발전해 왔고, 최근에는 LLM 기반 질문·선지 동시 생성으로 빠르게 확장되었다. 그러나 기존 AQG 문헌이 주로 다루어 온 것은 질문의 자연스러움, 문맥 관련성, 혹은 일반 교육 장면에서의 사용 가능성이었T다. 반면 본 연구에서 대상으로 한 수능형 비문학 문항은 정형화된 문체, 엄격한 선지 구성 규칙, 추론과 전개방식 파악을 포함한 인지 수준 분화, 그리고 고품질 오답 선지 설계를 동시에 요구한다는 점에서 기존 AQG와 다른 제약 구조를 가진다. 이러한 문제의식은 AIG가 강조해 온 평가 설계 관점과 AQG가 강조해 온 생성 모델 관점이 함께 요구된다는 2장의 정리와도 일치한다. 즉, 본 연구는 수능형 문항 생성이 단순한 질문 생성이 아니라, 고부담 평가 문항 설계의 자동화라는 별도의 연구문제로 다루어져야 함을 실증적으로 뒷받침한다.

이 지점에서 본 연구는 최근 LLM 기반 교육용 문항 생성 연구와도 구분된다. Scaria et al.(2024), Säuberli & Clematide(2024), Ghanem et al.(2022)의 연구는 LLM이 교육적 질문 생성과 읽기 이해 문항 생성에 실용 가능성을 가진다는 점을 보여주었지만, 이들 연구의 중심은 교실 수준 평가나 외국어 독해 과제에 있었다. 본 연구의 결과는 그러한 가능성을 부정하지 않으면서도, 수능형 비문학처럼 고도의 문체적 정형성과 오답 선지의 정밀성이 요구되는 환경에서는 동일한 결론을 쉽게 일반화할 수 없음을 시사한다. 즉, LLM이 "문항을 생성할 수 있는가?"와 "실제 고부담 시험에서 사용할 만한 문항을 생성할 수 있는가?"는 서로 다른 질문이며, 본 연구는 후자의 난도가 현저히 더 높다는 점을 보여준다.

2. The Fine-Tuning Paradox and the Challenge of Distractor Generation

본 연구에서 가장 중요한 발견은 Qwen 계열에서 소규모 SFT의 부정적 효과가 지문 생성 전반에 고르게 나타난 것이 아니라, 주로 문항·오답 선지 생성 단계에서 집중적으로 나타났다는 점이다. 지문 품질(Method B)에서는 대부분 미결정이 반복된 반면, 문항 품질(Method A)에서는 Qwen-base가 Qwen-SFT 조건을 반복적으로 상회하였다. 이 결과는 파인튜닝 역설이 모델 전체의 일괄적 품질 저하라기보다, 고제약 오답 생성 단계에서 특히 두드러지는 현상임을 의미한다.

선행연구는 오래전부터 다지선다형 문항의 핵심 난제가 정답 생성이 아니라 그럴듯하지만 명확히 틀린 오답 선지의 설계라고 보아 왔다. Alhazmi et al.(2024)은 오답 생성을 독립적 난제로 규정하였고, Benedetto et al.(2025)는 자동 평가 지표가 실제 시험 맥락에서 요구되는 매력도와 근접성을 충분히 포착하지 못한다고 지적하였다. Feng et al.(2024)과 Yu et al.(2024)도 LLM이나 지식증강 기법이 오답 선지의 표면적 품질을 개선할 수는 있지만, 실제 학습자의 오류 구조나 시험 상황에서의 기능적 적절성을 재현하는 데는 한계가 있다고 보았다. 본 연구에서 Qwen-SFT가 오답매력도와 변별력에서 반복적으로 열세를 보인 것은 이러한 선행연구의 우려를 수능 비문학 맥락에서 재확인한 결과로 해석할 수 있다. 다시 말해, 소규모 SFT는 수능 문체를 더 잘 흉내 낼 수는 있지만, 수험생의 피상적 독해를 유도하는 정교한 오답을 설계하는 능력까지 안정적으로 학습하지는 못했다.

또한 이 결과는 국내 국어과 읽기 문항 생성 연구와도 연결된다. 선행 국내 연구는 프롬프트 전략이나 GAI-HITL 절차가 문항 타당성과 형식 적절성에 의미 있는 차이를 만든다고 보고해 왔다. 그러나 본 연구는 그 논의를 한 단계 더 밀고 나가, 동일한 수능형 출력 형식을 학습했더라도 오답 선지의 질이 독립적으로 무너질 수 있음을 보여주었다. 이는 한국어 고부담 시험형 MCQ 생성에서 가장 어려운 하위 과제가 무엇인지, 즉 지문을 바탕으로 어떻게 틀리면서도 그럴듯한 선택지를 설계할 것인가라는 문제를 더 중심에 놓아야 함을 시사한다.

3. Divergence between Formal Alignment and Content Quality

본 연구의 Qwen 결과는 형식 정렬과 내용 품질 향상이 동일한 현상이 아니라는 점을 분명하게 보여준다. 파이프라인 개선과 SFT 이후 정답 분포는 균등해졌고, 선지 길이와 구조의 균일성도 어느 정도 향상되었다. 그러나 이러한 표면적 정돈은 오답매력도와 변별력 향상으로 이어지지 않았다. 오히려 선지들이 더 균질하고 예측 가능한 형식으로 정리되면서, 수험생을 실제로 혼동시킬 수 있는 의미적 긴장이 약화된 것으로 볼 수 있다. 본 연구의 결과는 수능형 문항 생성에서 수능처럼 보이는 출력과 실제로 수능다운 문항을 구분해야 함을 보여준다.

이러한 해석은 소규모 SFT 한계에 관한 최근 문헌과 정합적이다. Hu et al.(2021)와 Dettmers et al.(2023)가 제시한 LoRA와 QLoRA는 제한된 자원에서도 대형 모델을 실용적으로 적응시키는 길을 열었지만, 적응 가능성과 실제 과업 향상은 동일한 문제가 아니다. Ren et al.(2024)은 instruction fine-tuning이 새로운 지식을 주입한다기보다, 모델 내부의 표현을 더 잘 드러내도록 정렬시키는 측면이 있다고 보았다. Gudibande et al.(2023)은 모방형 fine-tuning이 표면적 스타일은 빠르게 학습하지만, 실질적 능력 향상까지 보장하지는 않는다고 지적했다. 본 연구에서 Qwen 계열의 SFT가 수능형 형식 재현에는 일정 부분 기여하면서도 내용 품질을 오히려 약화시킨 것은 바로 이 논점을 구체적 도메인에서 실증한 것으로 볼 수 있다. 즉, 소규모 SFT가 형식 모방을 높일지라도 시험형 오답 설계에 필요한 미세한 의미 변별은 오히려 약화될 수 있다.

이 지점은 국내 자동문항생성 연구에도 중요한 함의를 가진다. 국내 연구가 지금까지 프롬프트 비교, 문항 생성 가능성, 혹은 전문가 인식 수준에서 긍정적 결과를 제시한 것은 의미가 있지만, 본 연구는 그 긍정성이 형식 수준의 성공일 가능성과 내용 수준의 성공일 가능성을 분리해서 검토해야 한다는 점을 보여준다. 이는 이후 연구에서 문항 타당성, 오답 근접성, 실제 학생 반응 기반 변별력을 별도 차원으로 다뤄야 함을 시사한다.

4. Effects of Expanded Training Data and Output Format

본 연구에서 확장 학습 데이터 조건과 JSON 출력 형식은 모두 Qwen 계열의 파인튜닝 역설을 해소하지 못했다. 이는 단순히 데이터가 더 많아도 소용없다는 의미로 해석하기보다, 고제약 과제에서 어떤 데이터가 어떤 구조로 제공되는지가 더 중요하다는 점을 시사한다. 본 연구의 결과는 적어도 수능형 문항 생성과 같은 고제약 과제에서는 데이터 수량의 확장만으로는 내용 품질 저하를 해결하기 어렵다는 경험적 근거를 추가한다. 특히 본 연구의 198 조건과 300 조건은 단순 샘플 수 증가가 아니라 지문 범위와 유형 구성의 변화까지 포함하므로, 결과는 순수한 규모 효과라기보다 확장 학습 데이터 조건의 한계를 보여주는 사례로 해석하는 것이 타당하다.

출력 형식 변경의 효과 역시 선행연구와 연결해 볼 수 있다. Ghanem et al.(2022)의 질문 유형 통제 연구는 구조화된 제약이 질문 생성의 방향성을 제공할 수 있음을 보였고, 여러 교육용 LLM 연구는 구조화된 prompting이 생성 안정성에 기여한다고 보았다. 그러나 본 연구에서 JSON은 파싱 안정성과 형식 일관성에는 기여했을 수 있으나, 문항의 실질적 품질을 끌어올리는 조건으로 작동하지 않았다. 이는 구조화된 형식이 생성의 자유도를 제한하면서, 특히 오답 선지처럼 미세한 의미 변별과 표현 다양성이 필요한 하위 과제에서는 오히려 부정적일 수 있음을 시사한다. EXAONE에서 텍스트 형식이 JSON 형식보다 우세했던 결과도 같은 방향의 해석을 지지한다. 다시 말해, 구조화는 후처리와 시스템 구현에는 유리하지만, 그 자체가 평가 문항 품질 향상의 충분조건은 아니다.

5. Model Dependency of Small-Scale SFT Effects

본 연구에서 Qwen과 EXAONE이 다른 양상을 보였다는 점은 단순한 모델 순위 문제가 아니라, 소규모 SFT 효과가 베이스 모델의 언어 기반 능력과 상호작용할 가능성을 보여준다. Qwen 계열에서는 파인튜닝 역설이 세 차례 반복된 반면, EXAONE 계열에서는 적어도 Claude Haiku 기준으로 도메인 SFT가 향상 신호를 보였다. 또한 파인튜닝되지 않은 EXAONE이 파인튜닝된 Qwen 조건을 상회한 비교도 보고되었다. 이는 범용 다국어 모델에 소규모 도메인 SFT를 적용하는 전략과 한국어 특화 기반 모델 위에 동일한 도메인

이 결과는 한국어 특화 모델 논의와도 연결된다. EXAONE 3.5는 한국어-영어 환경에서의 실제 활용 가능성을 강조하는 모델군으로 소개되었고, 이러한 모델군의 등장은 범용 모델과 한국어 특화 모델의 실증 비교를 가능하게 한다고 보았다. 본 연구의 결과는 이론적으로 한국어 기반 능력이 충분한 모델일수록, 도메인 SFT가 형식 재현을 넘어 내용 품질 향상으로 이어질 가능성이 있다는 가설을 지지하는 방향의 신호를 제공한다. 물론 본 비교는 파이프라인 직교성이 완전히 통제된 것은 아니므로, 이를 순수한 사전학습 효과로 단정할 수는 없다. 그럼에도 본 연구는 최소한 수능형 문항 생성에서 베이스 모델의 언어 적합성이 주변적 요소가 아니라 중심 변수일 수 있음을 보여준다. 이 점은 국내 수능 국어 생성 연구와의 직접 비교에서도 중요한 차이를 만든다. 즉, 단순히 LoRA를 적용했다는 사실보다, 어떤 한국어 기반 모델 위에 LoRA를 적용했는지가 결과를 크게 좌우할 수 있다.

6. Heterogeneity in LLM-as-Judge Evaluations

본 연구의 또 다른 중요한 논점은 judge 간 이질성이다. Qwen 계열의 역설은 Claude Haiku에서 일관되게 더 분명하게 드러났고, EXAONE 계열의 향상 신호 역시 주로 Claude에서 관찰되었다. 반면 GPT-4o-mini는 동일한 비교쌍에 대해 전반적으로 더 보수적인 미결정 판정을 내렸다. 본 연구의 결과는 고제약 평가 과제일수록 단일 judge 의존이 위험하며, 다중 judge와 불확실성 보고가 필수적이라는 점을 지지한다.

더 나아가 본 연구는 LLM-as-Judge의 방법론적 설계 방식에 대해서도 시사점을 제공한다. 본 연구는 두 judge 모두에 대해 pairwise 비교, position bias 보정, bootstrap 신뢰구간 산출을 적용하였고, Method A와 Method B를 분리하여 판단 차원을 통제하였다. 이는 기존 LLM-as-Judge 연구가 주로 개방형 응답 전반의 선호 평가에 초점을 두었던 것과 달리, 교육평가용 문항 생성처럼 미묘한 의미 차이와 기능적 적절성이 중요한 과제에 맞춘 구체적 적용 사례라고 볼 수 있다. 본 연구는 단지 judge 간 이질성을 보고하는 데 그치지 않고, 고부담 시험형 생성 과제에서 어떤 평가 설계가 필요한지에 대한 방법론적 함의도 제시한다.

7. Implications for Educational Assessment and System Desig

AIG 전통은 자동화의 목적을 단순한 생성 효율화가 아니라, 타당도와 신뢰도를 유지한 상태에서의 문항 수급 안정화로 보아 왔다. 본 연구의 결과는 이 관점을 다시 환기시킨다. 현재 수준의 소규모 SFT 기반 수능형 문항 자동 생성은 완전 자동 출제 도구라기보다, 후보 문항 생성, 초안 제시, 지문 생성 보조, 오답 선지 아이디어 확장 등 인간 출제자를 지원하는 보조도구로 활용하는 것이 더 현실적이다. 이는 자동 생성 문항의 심리측정학적 타당성을 강조해 온 국내 연구와도 같은 방향이다. 즉, 생성 성공 자체보다 평가 문항으로서의 기능적 적절성을 우선시해야 한다는 점이 다시 확인된다.

동시에 본 연구는 향후 어떤 시스템 설계가 필요한지도 보여준다. 첫째, SFT 데이터는 단순 문항-정답 쌍보다 오답 설계 근거, 오개념 유형, 해설 정보를 함께 포함하는 방향으로 고도화될 필요가 있다. 둘째, 지문 생성과 문항 생성을 하나의 동일 모델 적응 문제로 보기보다, 단계별로 다른 데이터와 다른 적응 전략을 적용하는 편이 더 적절할 수 있다. 셋째, 한국어 특화 기반 모델을 우선 선택하고 그 위에 도메인 적응을 수행하는 전략이, 범용 모델에 제한된 양의 한국어 SFT를 얹는 전략보다 더 유망할 가능성이 있다. 넷째, 향후에는 학생 응답 데이터를 활용한 후속 검증이 결합되어야 한다. 이러한 방향은 AIG의 측정학적 통제, AQG의 생성 유연성, 그리고 최근 LLM 기반 교육용 생성 연구를 통합하는 설계 방향으로 이해할 수 있다.

8. Limitations and Future Work

본 연구의 한계는 선행연구와의 비교 속에서 더 분명해진다. 첫째, 본 연구의 확장 학습 데이터 조건은 순수한 규모 효과라기보다 지문 범위와 유형 구성 변화까지 포함한 조건이므로, 데이터 규모의 효과를 정교하게 분리한 실험과 직접 동일선상에 놓기는 어렵다. 둘째, Gold Standard 선정에는 연구자의 수동 판단이 개입되어 있어, 자동 평가 기준의 객관성이라는 측면에서 한계가 있다. 셋째, 생성 성능은 모두 필터를 통과한 최종 산출물을 기준으로 보고되었으므로, 원천 생성 품질과는 구분해서 해석해야 한다. 넷째, judge 간 이질성이 존재하여 일부 결론, 특히 EXAONE 계열의 향상 신호는 보수적으로 해석할 필요가 있다. 다섯째, 본 연구는 LLM judge를 중심으로 평가했지만, 전문가 평가 및 실제 학생 반응 데이터를 통한 변별도 검증까지는 수행하지 못했다. 이러한 한계는 곧 후속 연구 과제로 이어진다.

Conclusions

본 연구는 수능형 국어 비문학 문항 자동 생성에서 소규모 supervised fine-tuning의 효과와 한계를 단계별로 검증하였다. 이를 위해 키워드 기반 지문 생성과 문항·선지 생성을 결합한 엔드-투-엔드 파이프라인을 구축하고, 생성 품질을 지문 품질과 문항 품질로 분리하여 평가하였다. 그 결과, Qwen 계열에서는 소규모 SFT가 지문 품질보다 문항·오답 선지 품질에서 더 취약하게 작동하였으며, 파인튜닝 역설은 파이프라인 개선과 확장 학습 데이터 조건에서도 반복적으로 나타났다. 반면 EXAONE 계열에서는 적어도 일부 judge 기준에서 도메인 SFT의 향상 신호가 관찰되어, 소규모 SFT의 효과가 모델 의존적일 수 있음을 확인하였다. 이는 고부담 시험형 문항 생성에서 형식적 정렬과 내용 품질 향상을 동일시할 수 없으며, 베이스 모델의 한국어 기반 능력이 도메인 적응의 전제 조건일 수 있음을 시사한다.

본 연구의 학술적 의의는 세 가지이다. 첫째, 수능형 비문학 지문 생성과 문항·오답 선지 생성을 분리한 이중 평가 프레임워크를 제안함으로써, 생성 단계별 품질 차이를 정밀하게 분석할 수 있는 틀을 제시하였다. 둘째, 소규모 SFT의 효과를 단일 성능 향상이 아닌 단계별·모델별 상호작용으로 파악함으로써, 고제약 생성 과제에서 fine-tuning의 한계를 실증적으로 드러냈다. 셋째, 한국어 고부담 시험 문항 생성이라는 맥락에서 범용 모델과 한국어 특화 모델의 차이를 비교함으로써, 도메인 적응 전략 설계에 중요한 실증적 근거를 제공하였다.

실무적 측면에서 본 연구는 현재의 소규모 SFT 기반 자동문항생성 시스템이 완전 자동 출제 도구라기보다, 인간 출제자를 지원하는 보조도구로 활용되는 것이 더 적절함을 보여준다. 향후에는 오답 설계 근거와 오개념 정보를 포함한 데이터 구축, 단계별 전용 적응 전략, 학생 반응 데이터를 활용한 심리 측정 검증, 더 큰 규모의 한국어 특화 모델 비교가 필요하다. 이러한 후속 연구는 생성형 AI를 교육평가 영역에 보다 안전하고 타당하게 적용하기 위한 핵심 과제가 될 것이다.

Conflicts of Interest

The authors declared no conflicts of interest

References

Alhazmi, E., Sheng, Q. Z., Zhang, W. E., Zaib, M., & Alhazmi, A. (2024). Distractor generation in multiple-choice tasks: A survey of methods, datasets, and evaluation. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 14437–14458). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.799

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., Ghaleb, E., Giulianelli, M., Hanna, M., Koller, A., Martins, A., Mondorf, P., Neplenbroek, V., Pezzelle, S., Plank, B., Schlangen, D., Suglia, A., Surikuchi, A. K., Takmaz, E., & Testoni, A. (2025). LLMs instead of human judges? A large-scale empirical study across 20 NLP evaluation tasks. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 238–255). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-short.20

Benedetto, L., Taslimipoor, S., & Buttery, P. (2025). A survey on automated distractor evaluation in multiple-choice tasks. In E. Kochmar, B. Alhafni, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, & Z. Yuan (Eds.), Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025) (pp. 55–69). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.bea-1.5

Chung, J. M., & Kim, S. (2023). Exploring the reliability of an assessment based on automatic item generation using the multivariate generalizability theory. Journal of Science Education, 47(2), 211–224. https://doi.org/10.21796/jse.2023.47.2.211

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. arXiv. https://arxiv.org/abs/2305.14314

Du, X., Shao, J., & Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. In R. Barzilay & M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1342–1352). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1123

Embretson, S. E., & McIntyre, H. H. (2011). Automatic item generation: An innovation for developing complex cognitive tests. Innovation in Social Research Methods, 543

Ghanem, B., Lutz Coleman, L., Rivard Dexter, J., von der Ohe, S., & Fyshe, A. (2022). Question generation for reading comprehension assessment by modeling how and what to ask. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), Findings of the Association for Computational Linguistics: ACL 2022 (pp. 2131–2146). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.168

Gierl, M. J., & Lai, H. (2013). Evaluating the quality of medical multiple-choice items created with automated processes. Medical Education, 47(7), 726–733. https://doi.org/10.1111/medu.12202

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., & Guo, J. (2025). A survey on LLM-as-a-judge. arXiv. https://doi.org/10.48550/arXiv.2411.15594

Heilman, M., & Smith, N. A. (2010). Good question! Statistical ranking for question generation. In R. Kaplan, J. Burstein, M. Harper, & G. Penn (Eds.), Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 609–617). Association for Computational Linguistics. https://aclanthology.org/N10-1086/

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. arXiv. https://arxiv.org/abs/2106.09685

Ikhyeon Hong & Yongsang Lee. (2025). Quality of CSAT Korean Items from LoRA-Tuned Open-Weight GPT Models. Journal of Educational Evaluation, 38(4), 975-998.https://doi.org/10.31158/JEEV.2025.38.4.975

Jo, S., Yoo, Y., & Shin, I. (2025). Automatic generation of misconception-based distractors for linear equation items using large language models. The Mathematical Education, 64(3), 297–313. https://doi.org/10.63311/mathedu.25.6436

Kim, S. (2022). The utility of digital evaluation based on automatic item generation in mathematics: Focusing on the CAFA system. The Mathematical Education, 61(4), 581–595. https://doi.org/10.7468/mathedu.2022.61.4.581

Korea Institute for Curriculum and Evaluation. (2025, August 21). 2026 college scholastic ability test implementation guidelines [PDF]. https://csatcdn.kice.re.kr/resources/pdf/guideline.pdf

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. https://doi.org/10.1007/s40593-019-00186-y

Lee, Haneul & Lee, Yongsang. (2024). Exploring the Feasibility of Automated Item Generation for Korean Language Assessment. Journal of Education & Culture, 30(3), 659-686.https://doi.org/10.24159/joec.2024.30.3.659

LG AI Research. (2024). EXAONE 3.5: Series of large language models for real-world use cases. arXiv. https://arxiv.org/abs/2412.04862

Lim, S., Cho, H., Lee, J., & Yi, H. S. (2024). Exploring the feasibility of using large language models for automated item generation in social studies. Journal of Korean Association for Educational Information and Media, 30(3), 1035–1060. https://doi.org/10.15833/KAFEIAM.30.3.1035

Ministry of Education. (2026, February 11). We will fundamentally improve the CSAT item-development system for stable administration of the College Scholastic Ability Test. https://www.moe.go.kr/boardCnts/viewRenew.do?boardID=294&boardSeq=105330&lev=0&m=020402&opType=N&page=1&s=moe&searchType=null&statusYN=W

Park, G., & Choi, S. (2025a). A study on automatic item generation for multiple-choice reading comprehension assessment in Korean language. The Journal of Curriculum and Evaluation, 28(1), 215–246. https://doi.org/10.29221/jce.2025.28.1.215

Park, G., & Choi, S. (2025b). Psychometric validity analysis of GAI-HITL-based automatic item generation (AIG) for reading comprehension items. The Journal of Curriculum and Evaluation, 28(3), 319–359. https://doi.org/10.29221/jce.2025.28.3.319

Poon, Y., Lee, J. S. Y., Lam, Y. Y., Suen, W. L., Ong, E. L. C., & Chu, S. K. W. (2024). Few-shot question generation for reading comprehension. In K.-F. Wong, M. Zhang, R. Xu, J. Li, Z. Wei, L. Gui, B. Liang, & R. Zhao (Eds.), Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10) (pp. 21–27). Association for Computational Linguistics. https://aclanthology.org/2024.sighan-1.3/

Qwen Team. (2024). Qwen2.5: A party of foundation models. https://qwenlm.github.io/blog/qwen2.5/

Säuberli, A., & Clematide, S. (2024). Automatic generation and evaluation of reading comprehension test items with large language models. In R. Wilkens, R. Cardon, A. Todirascu, & N. Gala (Eds.), Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024 (pp. 22–37). ELRA and ICCL. https://aclanthology.org/2024.readi-1.3/

Scaria, N., Chenna, S. D., & Subramani, D. (2024). How good are modern LLMs in generating relevant and high-quality questions at different Bloom's skill levels for Indian high school social science curriculum? In E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, & Z. Yuan (Eds.), Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024) (pp. 1–10). Association for Computational Linguistics. https://aclanthology.org/2024.bea-1.1/

Shin, D. (2024). Exploring the potential of LLM-based customized test item generators: Focusing on Poe AI as a chatbot builder. Journal of the Korea English Education Society, 23(2), 181–206. https://doi.org/10.18649/jkees.2024.23.2.181

Taslimipoor, S., Benedetto, L., Felice, M., & Buttery, P. (2024). Distractor generation using generative and discriminative capabilities of transformer-based models. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 5052–5063). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.452/

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv. https://arxiv.org/abs/2306.05685

Author Information

Giljae Kim: MHResearch, 1st Author. https://orcid.org/0009-0002-8783-101X