LLM을 활용한 한국어 학습자 대상 추론적 읽기 문제 자동 생성 시스템

임성원°, 류법모
부산외국어대학교, 인공지능융합학과
wonirosoida33@gmail.com, pmryu@bufs.ac.kr

An Automated Inferential Reading Question Generation System for Korean Language Learners Using Large Language Models

Sung-won Lim°, Pum-mo Ryu
Busan University of Foreign Studies, Department of Artificial Intelligence Convergence

요 약

본 연구는 한국어능력시험(TOPIK)을 기반으로 학습자의 어휘·문법 수준에 맞춘 추론적 읽기 문제 자동 생성 시스템을 개발하고 평가하였다. LLM 모델을 사용하여 문제를 생성하였으며, 자동 평가, 학습자 평가, 전문가 평가를 통해 문제의 어휘·문법 적절성, 일관성, 실제성을 분석하였다. 평가 결과, 중급 문제에서 전반적으로 높은 결과를 보였으나 초급 문제에서는 난이도 조정과 선택지 구성의 개선이 필요함을 확인하였다. 본 연구는 LLM 기반 자동 문제 생성 시스템의 교육적 활용 가능성을 제시한 점에서 의의를 갖는다.

주제어: 한국어 학습 문제 생성, LLM, 추론적 읽기

1. 서론

한국에 대한 글로벌 인식이 변화하면서 한국으로 대학·대학원을 진학하는 학문 목적 학습자의 비중이 크게 증가하였다. 이런 학문 목적 학습자는 고도의 읽기 능력이 필수적이다[1]. 고도의 읽기 능력 중 특히 '추론 능력'을 필요로 하는 학습자의 요구를 알 수 있었다[2].

'추론적 읽기'는 박수자(2006)는 '추론적 읽기'를 텍스트 내에 명시적으로 제시되지 않은 내용을 파악하고 이해하는 과정으로 정의하였고 이윤자(2016)는 텍스트의 정보나 필자의 의도를 명확히 파악하는 것이라고 정의되어 왔지만[3][4], 본 연구에서는 한국어능력시험(TOPIK)에서 제시하는 추론적 읽기 문제 유형을 분석하여, '추론적 읽기'를 텍스트 속 숨겨진 의미와 의도를 파악하고 이해하는 능력을 평가하는 것으로 정의하였다[5].

최근 인공지능(AI) 기술은 일상생활에 자연스럽게 통합되고 있으며, 한국어 교육에서도 영향을 미치고 있다. 세종학당의 '세종학당 AI 선생님' 도입과 국립국어원의 'AI말평' 경진대회는 이러한 기술 적용의 사례이다. 그러나 여전히 추가적인 연구와 연구 자원이 필요하다.

본 연구에서는 한국어능력시험 기출 문제를 기반으로 한국어 학습자를 위한 추론적 읽기 문제를 자동 생성하는 시스템을 개발하는 것을 목표로 한다. 한국어능력시험은 국가기관인 국립국제교육원에서 주관하는 공인된 시험으로, 전 세계 한국어 학습자들에게 한국어 능력을 인증하고 기준이 된다. 특히, 학문 목적 학습자는 이 시험을 통해 대학·대학원 진학에 필수적인 요소로 작용한다. 이러한 배경으로 본 연구는 한국어능력시험의 문제 형식을 기반으로 문제를 자동 생성함으로써, 실제 시험 준비에 효과적으로 사용할 수 있는 학습 자원을 제공하고자 한다.

LLM을 활용한 시스템 설계는 한국어능력시험에서 제공하는 등급별 어휘 및 문법 목록을 활용하여 학습자의 등급을 고려한 맞춤형 문제를 생성했다. 이러한 접근 방식은 학습자가 실질적으로 사용할 수 있는 문제를 생성하며, 실제성을 갖춘 학습 문제 데이터셋을 구축하는 데 기여할 것이다.

2. 관련 연구

2.1. 한국어능력시험

한국어능력시험(TOPIK)은 외국인 학습자들의 한국어 능력을 평가하는 국가 공인 시험으로 초급부터 고급까지 다양한 등급을 평가하는 문제로 구성되어 있다. TOPIK 1, TOPIK 2로 분류되어 있고 듣기, 읽기, 쓰기 3가지를 평가하며 말하기 시험은 별도로 진행한다. 또한, level에 따른 읽기 영역 문항 수와 전체 등급 구성은 표1과 같다[6].

표 1 TOPIK 읽기 영역 문항수와 전체 등급 구성

level	읽기	등급
TOPIK 1	40문항	1급: 80~139점 2급: 140~200점
TOPIK 2	50문항	3급: 120~149점 4급: 150~189점 5급: 190~229점 6급: 230~300점

TOPIK에서 읽기 문제는 사실적 이해, 추론적 이해, 어휘, 논리적 이해, 구조적 이해로 구분되어 출제된다. 본 연구는 표2와 같이 TOPIK Ⅰ, Ⅱ 읽기 유형별 문항 수 분석을 토대로 추론적 문제를 분류하였다[7][8].

표 2 TOPIK 읽기 유형별 문항수 분석표

출제 의도	등급	문항 번호	합계
사실적 문제	TOPIK Ⅰ	31, 32, 33, 52, 40, 41, 42, 43, 44, 45, 50, 54, 56, 60, 66, 68, 70, 46, 47, 48	22 (55%)
	TOPIK Ⅱ	5~8, 9~12, 16~18, 20, 24, 28~31, 32~34, 43, 45, 47, 49, 22, 25~27, 35~38, 44	33 (66%)
추론적 문제	TOPIK Ⅰ	49, 51, 53, 55, 61, 65, 67, 69, 63	9 (22.5%)
	TOPIK Ⅱ	23, 24, 50, 48	4(8%)
어휘·문법·표현 능력 평가	TOPIK Ⅰ	34, 35, 36, 37, 38, 39	6(15%)
	TOPIK Ⅱ	1, 2, 3, 4, 19, 21	6(12%)
논리적 이해 능력 평가	TOPIK Ⅰ	57, 58	2(5%)
	TOPIK Ⅱ	13~15	3(6%)
구조적 이해 능력 평가	TOPIK Ⅰ	59	1(2.5%)
	TOPIK Ⅱ	39~41, 46	4(8%)

표 3. 4를 통해 사실적 문제와 추론적 문제를 비교할 수 있다.

표 3 제83회 TOPIK Ⅰ 읽기 43번 문제

문제 유형	사실적 문제
context	저는 요리를 잘 못합니다. 그래서 음식을 보통 사 먹습니다. 오늘 저녁은 집 근처 식당에서 불고기를 먹을 겁니다.
question	다음을 읽고 내용이 같은 것을 고르십시오.
answers	① 저는 요리를 자주합니다. ② 집 근처에 식당이 없습니다. ③ 저는 오늘 불고기를 먹을 겁니다. ④ 저는 오늘 집에서 저녁을 먹습니다.

표 4 제83회 TOPIK Ⅱ 읽기 55번 문제

문제 유형	추론적 문제
context	얼마 전 만화 박물관이 문을 열었습니다. 이 박물관에는 1960년부터 지금까지 나온 여러 가지 만화책이 많이 있습니다. 유명한 만화 영화도 즐길 수 있어서 특히 아이들이 좋아합니다. 그리고 이 만화 박물관에는 (㉠) 만화책을 오랜만에 다시 읽으러 오는 어른들도 많습니다.
question	㉠에 들어갈 말로 가장 알맞은 것을 고르십시오.
answers	① 요즘에 나온 ② 어릴 때 읽은 ③ 아이들이 만든 ④ 박물관에서 빌려 온

김정현(2020)에서 정의하는 토픽 등급별 추론적 읽기 문제 지시문은 표 5과 같다[9].

표 5 TOPIK 등급별 추론적 읽기 문제 유형

등급	유형	지시문
초급	상황이나 맥락 활용	㉠에 들어갈 말을 고르십시오.
	글의 기능, 목적 파악	왜 이 글을 썼는지 맞는 것을 고르십시오.
중급	인물의 심정 파악	[ ] 속에 나타난 나의 심정으로 알맞은 것을 고르십시오.
	필자의 태도 파악	필자의 태도로 알맞은 것을 고르십시오.
	글의 목적 파악	이 글을 쓴 목적을 고르십시오.

2.2 LLM을 활용한 문제 생성 연구

인공지능(AI)을 활용한 언어 학습 시스템은 GPT와 같은 대형 언어 모델(LLM)이 개발된 후 급격하게 발전하고 있다. 허동석 외(2023)에서는 수능 국어 문제를 자동으로 생성하는 시스템인 '신속 문답 생성기' 개발을 위해 LLM과 프롬프트 엔지니어링을 사용하였다[10]. 이 시스템을 통해 생성된 문제는 기존 기출 문제와 유사한 수준의 품질을 냈고, 또한 학습자들이 기존 문제와 구분하는 것에 어려움을 겪었다. 물론 실험자 수가 제한적이었으나, LLM을 활용하여 고도의 읽기 문제가 개발 가능하다는 것을 알 수 있었다. 조우성 외(2021)는 BERT와 GPT2를 이용하여 한국사 관련 질의응답과 문제 자동 생성 시스템을 제안하였다[11]. 두 가지 모델을 기반으로 자동 문제 생성 가능성을 입증하였다. 이 연구는 한국사 문제 생성을 위한 역사 데이터와 한국어 데이터를 활용하여 구축하였지만, 이와 달리 본 연구에서는 학습자 등급별 어휘, 문법, 주제 데이터를 먼저 고려하였다.

3. 실험 설계

3.1 데이터 수집

본 연구에서 사용된 데이터는 '한국어능력시험(TOPIK)'에서 제공한 자료를 기반으로 하며, '학습자 등급별 어휘 목록', '학습자 등급별 문법 목록', '학습자 등급별 주제 목록', '한국어능력시험 기출문제'를 사용하였다. 이는 문제의 구조·난이도·주제 등을 분석·활용하는데, 기초 자료로 활용된다.

3.1.1 학습자 등급별 어휘, 문법 목록

학습자 등급별 어휘 목록은 2015년 TOPIK에서 공개된 데이터로, 등급별로 구분된 어휘 목록이다. 목록은 '등급(level)', '어휘(word)', '품사(tag)'로 구분되어 있으며, 초급 어휘는 총 1,560개, 중급 어휘는 총 2,869개로 구성되어 있다. 이 목록을 통해 문제 생성 시 학습자의 언어 등급에 맞는 어휘를 반영하고 난이도 설정의 기준을 제공한다.

표 6 학습자 등급별 어휘 목록 예시

등급(level)	어휘(word)	품사(tag)
초급	학생증	명사
중급	격려	명사

학습자 등급별 문법 목록은 2009년 TOPIK에서 공개한 데이터를 사용하였고, 각 등급에 필요한 문법 항목을 나열한다. 초급 문법은 총 110개, 중급 문법은 총 187개의 데이터로 구성되어 있다. 목록은 '등급(level)', '문법(grammar)', '분류(category)'로 구성된다.

표 7 학습자 등급별 문법 목록 예시

등급(level)	문법(grammar)	분류(category)
초급	-거나	연결어미
중급	-듯이	연결어미

3.1.2 학습자 등급별 주제 목록

본 연구에서는 주제 목록을 활용하여 문제 생성 시 학습자 등급에 적절한 주제를 무작위로 선정하였다. 이를 통해 다양성을 확보하면서도 시험 범위에 벗어나지 않도록 지정하였다. 기존의 한국어 데이터셋은 주제의 다양성이 광범위하여 한국어 교육 목적으로 사용하기에 적절하지 않았다. 따라서 본 연구는 김종숙(2015)에서 분석한 토익 주제 분석 결과를 참고하여 목록화하였다[12]. 초급 주제어는 총 66개, 중급 주제어는 총 155개이다. 이 주제 목록은 학습자의 등급에 맞는 주제를 반영함으로써, 문제의 신뢰도와 교육적 타당성을 강화하였다.

표 8 학습자 등급별 주제 목록 예시

등급(level)	주제(subject)	주제어(keywords)
중급	문화	영화
초급	문화	한국 명절

3.1.3 한국어능력시험 기출 문제

본 연구는 한국어능력시험의 기출 문항 중, 읽기 영역의 '추론 문제'를 대상으로 데이터를 수집하였다. 수집한 데이터는 총 5회 분량(52회, 60회, 64회, 83회, 91회)의 기출 문제이며, 2017년부터 2024년까지 공개된 문제를 활용하였다. 추론 문제 문항의 총개수는 65개이다.

한국어능력시험 중 추론적 읽기 문제는 표 5과 같이 추론적 문제 유형과 지시문이 정형화되어 있다. 이때, 초급과 중급 문제 유형 중 '상황이나 맥락 활용', '인물의 심정 파악'은 문제의 지문과 질문에 '㉠'과 '[ ]'과 같은 특수문자가 포함된다. 하지만 자동 문제 생성 시 표 4와 같이 특수문자가 포함된 지문을 인식하지 못하여 특수문자가 생략되는 현상을 관찰하였다. 따라서 표 9와 같이 파일을 3가지로 분류하여 연구를 진행하였다.

표 9 TOPIK 추론적 읽기 문제 분류

분류	등급	유형
1	초급	㉠ 기호
2	중급	[ ](대괄호) 기호
3	초급, 중급	그 외

이 데이터들의 활용은 추론적 읽기 문제의 자동 생성 과정을 최적화하고 한국어 학습자에게 실용적인 학습 자료를 제공하는 데 중요한 역할을 할 것이다.

3.2 문제 생성 알고리즘

본 연구의 핵심 목적은 한국어능력시험(TOPIK)의 기출 문제를 바탕으로 한국어 학습자의 등급에 맞는 추론적 읽기 문제를 자동으로 생성하는 시스템을 개발하는 것이다. 이를 위해 다음과 같은 단계별 접근 방식을 구성했다.

3.2.1 키워드 선택 및 프롬프트 구성

학습자의 등급에 따라 적절한 키워드를 랜덤으로 선정한다. 이는 학습자 등급별 주제 목록에서 추출되며 문제의 주제와 맥락 설정에 기본적인 역할을 한다. 그 후 각 문제에 대해 학습자 등급과 선택된 키워드를 기반으로 프롬프트를 구성한다. 프롬프트는 학습자 등급에 맞는 어휘, 문법 목록을 최대한 반영하고 TOPIK의 문제 형식을 따르며 문맥, 문제 유형 요소를 포함한다.

표 10 자동 문제 생성용 프롬프트

prompt

```
문제 생성 요청:
- 학습자 등급: {level}
- 사용 키워드: {keyword}
- 모든 문제는 {level}에 맞는 {grammar}과 {vocabulary}를 최대한 반영하세요.

- 참조 문제 형식:
문맥: {example_question['context']}
질문: {example_question['question']}
선택지: {example_question['answers']}
정답: {example_question['correct']}
이 정보를 바탕으로 새로운 문제를 생성하세요.

생성된 문제는 다음과 같은 형식을 유지해야 합니다:
- [context]는 최소 300자에서 최대 900자 내로 작성
- ['question']은 긍정문으로 작성.
{example_question['question']}와 최대한 유사하게 생성.

- ['answers'] 속 ['correct']는 무조건 한 개. 답변 생성 시 지문과 동일한 어휘는 피합니다.
- 맞춤법과 띄어쓰기에 유의

문제 출력 형식:
```
[keyword]: {selected_keyword}
[Level]: {level}
[context]: 지문 내용
[question]: 질문 내용
[answers]: ① 선택지1 ② 선택지2 ③ 선택지3 ④ 선택지4
[correct]: 번호 + 정답
```
```

3.2.2 문제 생성

본 연구는 한국어 학습자를 위한 추론적 읽기 문제를 자동 생성하기 위해 OpenAI의 GPT 모델을 활용하였다. 문제 생성을 위해 gpt-4o 모델과 gpt-4o-mini 모델을 선정하였다. 두 모델 모두 뛰어난 텍스트 생성 능력을 갖추고 있으며 이러한 특성이 교육적 콘텐츠 생성에 적합하다고 판단되었고, 본 연구에서는 초기 모델로서 이 두 모델을 먼저 적용하였다. 각 문제에 대해 프롬프트는 학습자 등급에 맞는 어휘와 문법을 반영하여 TOPIK 형식을 따른다. 또한, temperature가 0.5 이상인 경우, 문제 형식을 벗어나거나 지문 내용이 실제성과 떨어지는 현상을 발견하였다. 따라서 문제 생성의 일관성을 유지하기 위해 'temperature' 파라미터를 0.5로 설정하였다.

이를 통해 최종 생성된 문제는 총 두 가지이며, 특수문자가 삽입된 문제 유형과 일반 문제 유형으로 나누어진다.

표 11 최종 생성: 특수문자 삽입 문제

keyword	식당
level	초급
context	저는 친구와 함께 새로운 식당에 가기로 했습니다. 그 식당은 최근에 오픈했는데, 분위기가 아주 좋다고 들었습니다. 메뉴도 다양하고 가격도 적당하다고 하네요. 그래서 저희는 저녁에 그 식당에 가기로 했습니다. 친구는 매운 음식을 좋아하고 저는 간단한 음식을 좋아합니다. 그래서 (㉠) 시키기로 했습니다. 맛있는 음식을 먹고 즐거운 시간을 보낼 생각에 너무 기대가 됩니다.
question	㉠에 들어갈 말로 가장 알맞은 것을 고르십시오.
answers	① 함께 같은 음식을 ② 서로 다른 음식을 ③ 비싼 음식을 ④ 매운 음식을
correct	② 서로 다른 음식을

표 12 최종 생성: 일반 문제 유형

keyword	식당
level	초급
context	안녕하세요! 저는 친구와 함께 저녁을 먹으러 식당에 갔습니다. 그 식당은 맛있는 음식과 좋은 분위기로 유명합니다. 우리는 메뉴를 보고 여러 가지 음식을 주문했어요. 특히, 파스타와 피자가 정말 맛있었습니다. 그런데, 서비스가 조금 느려서 기다리는 시간이 길었어요. 그래서 다음에 다른 식당에 가보려고 합니다. 맛있는 음식을 먹는 것은 좋지만, 빠른 서비스도 중요하다고 생각합니다.
question	왜 이 글을 썼는지 맞는 것을 고르십시오.
answers	① 식당의 가격을 물어보려고 ② 맛있는 음식을 추천하고 싶어서 ③ 식당의 서비스에 대해 이야기하고 싶어서 ④ 친구와의 만남을 자랑하려고
correct	③ 식당의 서비스에 대해 이야기하고 싶어서

4. 평가

4.1 평가 데이터

본 연구에서 평가 대상으로 사용된 데이터는 초급, 중급 학습자 등급을 기준으로 생성된 추론적 읽기 문제이며, 문제는 표 11, 12과 같이 지문, 질문, 답변 선택지, 정답을 포함하고 있다. 평가 데이터는 gpt-4o 모델을 사용한 초급 데이터 50개, 중급 데이터 50개, gpt-4o-mini 모델을 사용한 초급 데이터 50개, 중급 데이터 50개로 구축되었다.

4.2 자동 평가

본 연구에서는 자동 생성된 한국어 추론적 읽기 문제의 품질을 평가하기 위해 GPT 모델을 활용하여 등급별 어휘와 문법이 적절한지, 지문과 질문의 연관성이 적절한지에 대해 평가를 진행하였다.

4.2.1 등급별 적절성 평가

본 연구에서는 GPT를 활용하여 생성된 문제의 어휘 및 문법 적합성을 평가하는 알고리즘을 설계하였다. 생성된 문제의 지문, 질문, 답변, 학습자의 등급(level)을 gpt-4o 모델에 전달하여, 해당 등급에 적합한 어휘 및 문법을 사용되었는지 평가하였고 평가 요청은 문제의 등급에 따라 다르게 설정하였다.

- 초급 문제: 해당 레벨보다 높은 어휘와 문법이 사용되었는지 평가
- 중급 문제: 해당 레벨보다 낮은 어휘와 문법이 사용되었는지 평가

표 13 등급별 적절성 평가 프롬프트

```
prompt

주어진 지문에서 사용된 어휘와 문법이 {level} 수준에 적합한지
평가하세요.
- 지문: {context}
- 질문: {question}
- 답변: {answers}

다음 조건에 맞게 평가하세요:
- {level} 수준보다 높은 레벨의 어휘와 문법이 사용되었는지
  (초급일 경우)
- {level} 수준보다 낮은 레벨의 어휘와 문법이 사용되었는지
  (중급일 경우)

결과는 다음 형식으로 제공하세요:
- higher_level_mismatch: 사용된 상위 레벨 어휘/문법 개수(초급)
- lower_level_mismatch: 사용된 하위 레벨 어휘/문법의
  개수(중급)
```

평가 결과, 다음 그림과 같다.

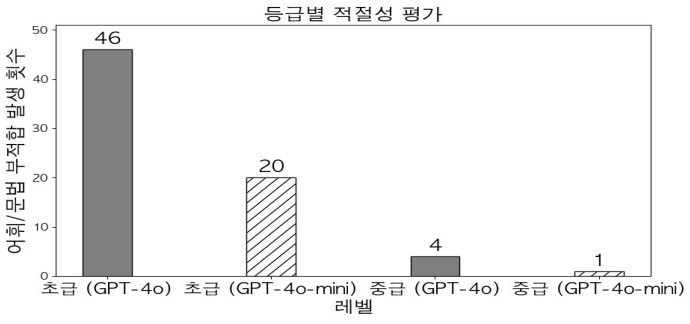

그림 1 등급별 적절성 평가 결과

그림1은 각 모델과 등급에서 어휘와 문법이 적절하게 사용되지 않은 빈도를 나타내고 있다. gpt-4o 모델에서 초급 문제에 중급 이상의 어휘와 문법이 사용된 빈도가 46회로 나타났고, 중급 문제에서 하위 레벨의 어휘와 문법 사용 빈도는 4회에 불과했다. 반면, gpt-4o-mini 모델은 초급 문제에 있어서 중급 이상의 어휘와 문법이 사용된 빈도가 20회로 나타났으며, 중급 문제에서 초급 이하의 어휘와 문법이 사용된 빈도는 1회로 매우 적었다.

이와 같은 결과는 gpt-4o 모델이 초급 문제에서 중급 어휘와 문법의 사용을 보다 자주 범하는 경향이 있음을 보여주고, gpt-4o-mini 모델도 유사한 경향을 보이나 그 빈도는 낮은 것으로 나타났다. 이는 gpt-4o 모델이 gpt-4o-mini 모델보다 모델 크기와 복잡성이 크기 때문으로 추정된다. gpt-4o 모델의 복잡성은 초급 문제에서 필요 이상으로 복잡한 (중급 수준의) 어휘와 문법을 사용하는 경향을 초래할 수 있다. 따라서 두 모델 사용 시 초급 문제에서 중급 어휘와 문법 사용을 제지하기 위한 문제 난이도 조정에 대한 명시적 지시, 구체적인 학습 데이터 추가가 제공되어야 한다.

4.2.2 일관성 평가

본 연구에서는 지문(context)-질문(question) 간의 일관성을 평가하기 위해 gpt-4o 모델을 사용하였다. 일관성 평가는 문제 생성 과정에서 지문에 제시된 정보와 질문이 논리적으로 연결되어야만, 실제로 사용 가능하고 교육적 가치를 지닌 문제가 될 수 있으므로 중요한 평가 지표이다. 일관성 평가는 두 가지 기준에 따라 이루어졌다.

- 질문이 지문에서 주어진 정보와 일치하는가?
- 질문이 지문의 주요 내용을 반영하고 있는가?

GPT-4o 모델은 주어진 지문과 질문을 분석한 후, 일관성 평가 결과를 0점(일관성이 전혀 없음)에서 5점(매우 일관성이 있음)까지의 점수로 제공하였다.

표 14 일관성 평가 프롬프트

prompt
주어진 지문과 질문이 서로 일관성이 있는지 평가하세요. - 문맥: {context} - 질문: {question} - 다음 기준에 따라 일관성을 평가하세요: 1. 질문이 지문에서 주어진 정보와 일치하는가? 2. 질문이 지문의 주요 내용을 반영하고 있는가? 결과는 다음 형식으로 제공하세요: - consistency_score: 0(일관성 없음)에서 5(매우 일관성 있음)까지의 점수로 제공하세요.

표 15는 gpt-4o, gpt-4o-mini 모델이 생성한 초급과 중급 문제에 대한 평균 일관성 점수를 나타낸다. gpt-4o 모델의 중급 문제에서 평균 일관성 점수는 4.76점으로 매우 높게 나타났으며, 이는 지문과 질문이 논리적으로 잘 연결되어 있음을 의미한다. gpt-4o-mini 모델의 중급 문제 평균 일관성 점수는 4.61점으로 gpt-4o 모델보다 약간 낮은 수준이지만 여전히 높은 일관성을 유지하였다.

표 15 일관성 평가 결과

	일관성 평가	
등급(level)	초급	중급
gpt-4o	4.68	4.76
gpt-4o-mini	4.64	4.61

평가 결과, gpt-4o 모델이 중급 문제에서 더욱 뛰어난 성능을 발휘하며, gpt-4o-mini 모델 역시 일관성 평가에서 높은 성능을 기록했음을 시사한다. 특히, 초급 문제에서는 두 모델 모두 유사한 성능을 보여, 일관성 있는 문제 생성이 가능함을 보여준다.

4.3 학습자 평가

본 연구에서는 자동 평가 외에도 실제 학습자를 대상으로 추가적인 평가를 진행하였다. 어휘 등급 적절성, 문법 등급 적절성, 실제성에 대한 평가를 했고, 이를 위해 자동 평가와 동일한 평가 데이터를 학습자에게 제공하였다. 평가 척도는 5점 Likert 척도를 사용하였으며, 1점은 '전혀 동의하지 않는다', 5점은 '매우 동의한다'를 의미한다. 평가 데이터는 초급 문제 100개, 중급 문제 100개로 진행되었다. 평가에 참여한 학습자는 한국어능력시험(TOPIK)을 3회 응시한 경험이 있는 5급 학습자와 9회 응시한 경험이 있는 6급 학습자로 구성되었다. 이들은 각 문제에 대해 어휘와 문법의 적절성, 문제의 실제성에 대해 평가를 진행하였으며, 그 결과를 바탕으로 자동 생성된 문제를 학습자 인식 측면에서 검증하였다.

표 16 학습자 평가 결과

	어휘 등급 적절성		문법 등급 적절성		실제성	
level	초급	중급	초급	중급	초급	중급
gpt-4o	4.89	4.79	4.87	4.87	4.81	4.76
gpt-4o-mini	4.68	4.87	4.87	4.9	4.75	4.71

표 16은 어휘 등급 적절성, 문법 등급 적절성, 실제성에 대한 학습자 평가 결과를 나타낸다. gpt-4o, gpt-4o-mini 모델 모두 중급과 초급 문제에 대한 평가가 이루어졌으며, 전반적으로 모든 항목에서 4.7점 이상의 높은 평가를 받았다. 특히 어휘와 문법 적절성에서는 두 모델 간 큰 차이가 없었으나, 실제성 항목에서 gpt-4o가 약간 더 높은 점수를 기록하였다.

학습자들이 남긴 코멘트를 바탕으로 전체적인 문제에 대한 의견을 종합하면, 초급 문제에서는 '일부 초급 문제에서 중급 수준에 해당하는 어휘와 문법이 발견되었다'는 지적이 있었으며, '기본적인 장소나 상황에 대한 질문이 더 추가될 필요가 있다'는 의견이 나왔다. 중급 문제에 대해서는 '주제의 폭이 넓고 다양하다는 점이 긍정적으로 평가되었으나, 추후 중급과 고급 문제의 명확한 구분이 필요하다'고 언급하였다. '주제 중 사회 이슈와 글로벌 문제는 중급 학습자에게는 어렵다'는 점도 지적하였다. 이를 통해 자동 생성된 문제의 개선 방향을 알 수 있었으며, 학습자의 인식과 요구에 기반한 추가적인 문제 구성이 필요함을 인지하였다.

4.4 한국어 전문가 평가

본 연구에서는 자동 생성된 문제에 대한 추가적인 검증을 위해 한국어 교사 경력 10년 이상의 두 명의 전문가를 대상으로 평가를 실시하였다. 평가 항목은 '어휘 적절성', '문법 적절성', '문제의 일관성', '추론 평가의 적절성', '실제성'으로 구성하였다. 각 항목은 5점 Likert 척도(1점: 전혀 동의하지 않는다 ~ 5점: 매우 동의한다)를 사용하여 평가되었다. 평가 대상 문제

는 초급 문제 100개, 중급 문제 100개로 구성되었다.

전문가 평가는 gpt-4o와 gpt-4o-mini 모델로 구분하여 평가되었으나, 두 모델 간의 차이가 거의 없는 것으로 나타났다. 이어 따라 모델별로 구분하지 않고, 문제 등급(level) 구분으로 초급, 중급 문제에 대한 평가 결과를 도출하였다.

평가 결과는 표 17과 같다. 어휘 적절성에서는 중급 문제에 비해 초급 문제가 3.06점으로 낮은 평가를 받았으며, 문법 적절성과 추론 평가 항목도 중급에 비해 초급 문제가 낮게 평가되었다. 반면, 일관성 평가에서는 중급과 초급 문제 모두 높은 평가를 받았으며(각 4.89점, 4.99점), 초급 문제의 실제성이 다소 더 높게 평가되었다.

표 17 전문가 평가 결과

항목	초급	중급
어휘 적절성	3.06	4.91
문법 적절성	4.07	4.77
일관성	4.89	4.99
추론 평가	4.13	4.66
실제성	4.81	4.62

전문가들의 구체적인 코멘트는 표 18과 같다.

표 18 전문가 코멘트

평가자	등급	코멘트
1	초급	- 어휘와 문법의 난이도가 과하게 높음 - 선택지 형식 불일치, 모호한 답안 존재
	중급	- 문제 형식과 선택지 구성에서 어색함 - 답안 중복 가능성 존재 - 문항의 자연스러움 개선
2	초급	- 선택지만으로 답을 유추할 수 있는 문제 다수 존재 - 선택지 복잡성에 대한 검토 필요
	중급	- 어휘와 문법 수준 적절 - 화자의 감정을 추론하는 질문에 지문이 적절하지 않음
공통	-	- AI가 문제 출제 방향과 학습자 수준에 맞춘 세밀한 학습이 필요

결론적으로 초급 문제의 경우 어휘나 문법을 학습자 수준에 맞춰 조정하고, 선택지 형식의 일관성을 유지하는 것이 필요하다. 중급 문제는 문제 형식과 선택지 구성의 자연스러움을 개선하고, 질문 의도에 맞춰 지문의 주제를 선정할 수 있어야 한다. 이를 통해 LLM을 활용한 자동 문제 생성이 학습자 수준에 맞는 세밀한 문제 출제 능력을 갖출 수 있도록 향후 연구에서는 어휘, 문법 목록 데이터의 양과 질을 개선하고 문제의 지문-질문-선택지의 향상을 위해 추가적인 조치가 필요하다.

5. 결론

본 연구에서는 한국어능력시험(TOPIK)을 기반으로 LLM 모델(gpt-4o, gpt-4o-mini)을 활용해 추론적 읽기 문제 자동 생성 시스템을 개발하고, 이를 다양한 평가 방법으로 검증하였다.

평가 결과, 두 모델 모두 중급 문제에서는 어휘·문법 적절성 및 일관성에서 높은 성과를 보였으나, 초급 문제에서는 어휘와 문법 난이도가 학습자 수준에 맞지 않은 경우가 있었다. 이는 사전 제공된 어휘, 문법 목록의 세부적인 개선과 충분한 양의 데이터가 필요함으로 보인다. 학습자와 전문가 평가에서도 초급 문제의 난이도 조정과 선택지 구성의 개선이 필요하다는 의견이 있었다.

향후 연구에서는 문제 난이도와 선택지 일관성을 개선하고, 추론적 읽기 외의 다양한 읽기 영역을 포함한 자동 문제 생성 시스템을 확장할 계획이다. 이를 통해 한국어 학습자들에게 맞춤형 학습 자원을 제공하고, 교육적 활용도를 높일 수 있을 것으로 기대된다.

감사의 글

이 연구는 부산광역시가 지원하는 부산빅데이터혁신센터 운영 사업의 지원을 받아 수행되었습니다.

참고문헌

[1] 전수정, "학문 목적 읽기 교육을 위한 한국어 학습자의 요구 분석 연구", 외국어로서의 한국어교육, 29, 203-230, 2004

[2] 초첩, "학문 목적 한국어 학습자의 이해 능력 향상을 위한 교육 방안 연구: 추론적 이해와 비판적 이해를 중심으로", 동국대학교 일반대학원, 석사학위논문, 2022

[3] 박수자,김혜정. "추론적 읽기의 지도 방법 연구." 한국어교육학회 학술발표논문집, 29-50, 2006

[4] 이윤자, "학문 목적 한국어 읽기 교육의 읽기 전략 수업 연구", 숙명여자대학교 대학원, 박사학위논문, 2016

[5] 김종숙, "개편 토픽 읽기 문항 분석." 한국교육논총 36.1: 37-56, 2015

[6] 김종숙, "개편 토픽 읽기 문항 분석." 한국교육논총 36.1: 37-56, 2015

[7] 김종숙, "개편 토픽 읽기 문항 분석." 한국교육논총 36.1: 37-56, 2015

[8] TOPIK, 2021년 개선된 한국어능력시험(TOPIK) 평가틀 및 발문 적용 안내, 한국어능력시험, 2024, https://www.topik.go.kr/TWSTDY/TWSTDY0101.do?bbsId=BBSMSTR00078&nttId=123136&nttC1Code1=ALL&pageIndex=1&searchType=&searchWord=

[9] 김정현. "한국어능력시험(TOPIK1.2) 읽기 영역 분석 및 자기주도적 TOPIK 읽기 학습전략 연구." 국내석사학위논문 고려대학교 세종캠퍼스, 2020

[10] 허동석외3, "프롬프트 개발을 통한 수능 국어 맞춤형 문제 생성 시스템 제안", 한국HCI학회 학술대회, 2024

[11] 조우성외3, "BERT와 GPT2를 이용한 한국사 질의응답 및 문제 생성 시스템". Proceedings of KIIT Conference, 2021

[12] 김종숙, "개편 토픽 읽기 문항 분석." 한국교육논총 36.1: 37-56, 2015