JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 25, No. 11, pp. 3309-3317, Nov. 2024

Check for updates

생성형 AI기술을 활용한 컴퓨터 사이언스 문제 생성 모델 개발 및 평가

김 정 순[1] · 김 성 희[2*]
[1]동의대학교 인공지능학과 박사과정
[2]동의대학교 산업ICT기술공학과 교수

Development and Evaluation of a Computer Science Question Generative Model Using Generative AI Technology

Jung-Soon Kim[1] · Sung-Hee Kim[2*]

[1]Ph.D. Course, Department of Artificilal Intelligence, Dong-eui University, Busan 47340, Korea
[2]Associate Professor, Department of ICT Industial Engineering, Dong-eui University, Busan 47340, Korea

[요 약]

대학생들과 취업준비생들은 취업 경쟁력을 높이기 위해 다양한 종류의 자격증을 취득한다. 본 연구에서는 메타에서 공개한 오픈소스 LLaMA-3.1기반의 사전학습 모델과 Unsloth기반의 파인튜닝 모델로 컴퓨터 관련 자격증 문제를 생성하는 알고리즘을 개발했다. 사전 모델과 파인튜닝한 모델에서의 생성한 문제가 차이가 있을 것이라고 가정하고, 이를 테스트하기 위해 성능 비교를 위한 지표를 정의하고, 사전학습 모델과 파인튜닝한 모델의 의미있는 지표를 도출하였다. 실험은 20명의 참가자가 100개의 문제를 평가하는 설문에 참여하였다. 실험 후 수집된 자료는 통계적 검증을 위한 유의수준 0.05로 설정하여 분석하였다. 총 5가지 지표 중 4개에서 통계적으로 유의미한 차이가 나타났다. 유창성, 지문 연관성, 과목 일관성, 유일성이라는 종속 변수들이 통계적으로 유의미한 차이를 보였다. 이 결과는 파인튜닝이 모델 성능을 실질적으로 향상시킬 수 있다는 것을 입증하였다.

[Abstract]

University students and job seekers obtain various types of certifications to enhance their competitiveness in the job market. In this study, we developed an algorithm to generate questions for computer science related certification exams using a pre-trained model based on Meta's open-source Llama 3.1 and a fine-tuned model based on Unsloth. We hypothesized that there would be differences in the questions generated by the pre-trained and fine-tuned models. To test this, we defined performance comparison metrics and derived meaningful indicators from both models. The experiment involved 20 participants who completed a survey evaluating 100 questions. The collected data were analyzed with a significance level of 0.05 for statistical validation. Among the five metrics, four dependent variables, namely, fluency, passage relevance, subject consistency, and uniqueness, exhibited statistically significant differences. The results demonstrated that fine-tuning can significantly improve model performance.

색인어 : 시험 문제, 파인튜닝, Llama 3.1, 성능비교, 문제생성
Keyword : Exam Questionnaire, Fine-Tuning, Llama 3.1, Performance Comparison, Problem Generation

http://dx.doi.org/10.9728/dcs.2024.25.11.3309

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-CommercialLicense(http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 23 October 2024; **Revised** 14 November 2024
Accepted 20 November 2024

***Corresponding Author; Sung-Hee Kim**

Tel: +82-51-890-2366
E-mail: sh.kim@deu.ac.kr

Ⅰ. 서 론

대학생들과 취업준비생들은 취업 경쟁력을 높이기 위해 다양한 종류의 자격증을 취득한다. 특히 IT 관련 자격증은 취업 시 개인의 직무 수행 능력 판단에 플러스 요인으로 이용되고 있어 많은 학생들이 관련 자격증을 취득하고 있다[1].

2024 국가기술자격통계연보에 따르면 국가인증자격증 종류 중 컴퓨터활용능력2급 필기는 6,610,757명이 가장 많았고, 컴퓨터 활용능력 1급 필기 접수 인원은 4,010,508명으로 두번째로 많이 응시하였다[2].

이러한 이유로 컴퓨터 자격증 취득을 위한 사용자 프로파일을 이용한 실시간 평가 시스템[3], 웹 기반 학습 시스템과 같은 연구가 진행되었다[4]. 이들의 연구는 자동 채점을 용이하게 하고, 학생들에게 더 빠른 피드백을 제공하며, 다양한 유형의 문제를 낼 수 있게 한다. 하지만 데이터베이스 기반의 문제로 동일한 문제가 지속적으로 나올 가능성이 있고, 새로운 기술 관련 문제를 출제하려면 출제자는 많은 시간이 필요하고, 관련 기관에서는 많은 비용이 든다.

이를 해결하기 위해 규칙기반 연구에서 최근 생성형 AI를 활용한 연구가 활발히 이루어졌다. 그 중 전통적으로는 규칙과 템플릿을 사용하여 질문을 생성하는 연구[5]와 심층 신경망의 인기가 높아짐에 따라, 신경망 인코더-디코더 아키텍처[6], [7]와 대규모 트랜스포머를 이용한 연구가 있었다[8]-[11]. 이후 ChatGPT 3.0 출시 이후부터 대규모 언어 모델(LLM, Large Language Model)이 널리 대중에게 알려지고 사용되면서 다양한 교육분야에서도 문제 생성 연구가 활발한 연구가 이뤄지고 있다. 생성된 문제를 기반으로 출제자는 효율적인 출제가 가능하고, 이는 결국 좋은 문제를 만들어내는 과정에서의 총 비용 절감으로 이어질 것이다. 이러한 대규모 언어 모델 기반의 연구에는 ChatGPT와 Prompt Engineering 기반의 한국어능력시험(TOPIK)[12], 수능 국어 맞춤형 문제생성시스템[13], LLaMA2 기반의 영어 문제생성 모델[14]이 있다.

2018년 등장한 BERT, GPT는 파라미터의 수가 1억개 이상이고, 2019년 발표된 GPT-2는 최대 15억개, 2020년 GPT-3는 1750억개로 기하급수적으로 증가하였다. 이와 같이 파라미터의 확장은 어려운 추론으로 발생할 수 있는 시간과 메모리 비용이 많이 발생한다. 이에 최근 LLM을 연구하는 스타트업들은 LLM을 보다 작은 모델로 만들어 이를 다룰 수 있도록 하는 경량화 작업에 집중하고 있다.

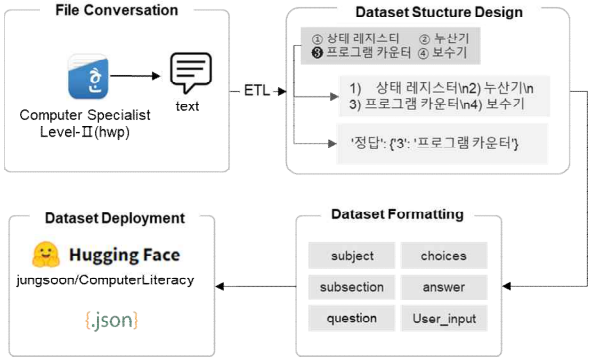
자격증 시험의 경우, 연습으로 풀 수 있는 문제가 한정적이어서, 생성형 AI기술을 활용하면 다양한 문제를 빠른 시간 안에 생성할 수 있다는 장점이 있다. 또한, 이 방법론을 다양한 자격증 분야에 활용할 수 있을 것으로 본다. 그래서 본 연구에서는 메타에서 공개한 오픈소스 LLaMA-3.1기반의 사전학습 모델과 Unsloth/Meta-Llama-3.1-8B[15]를 활용한 경량화 및 양자화(Quantization)가 가능한 파인튜닝 모델로 컴퓨터 사이언스 관련 자격증 문제를 생성하는 알고리즘을 개발하였다. 우리는 LLaMA-3.1 기반의 사전학습 모델과 파인튜닝한 모델에서의 생성한 문제가 차이가 있을 것이라고 가정하고 실험을 실시하였다. 이를 테스트하기 위해 각 문제별 5가지 지표 중 의미있는 지표를 식별하는 것을 목표로 한다.

Ⅱ. 본 론

2-1 모델 설계 개요 및 실험 설계

객관식 문제 생성에 사용되는 용어를 정리하고자 한다. 과목은 자격증 시험에서의 과목을 뜻하며, 예를 들어서 컴퓨터 활용능력2급의 필기 과목은 '컴퓨터 일반'과 '스프레드시트 일반'이 있다. 세부과목(Subsection)은 각 과목 내에서 세부 항목을 뜻하며, '컴퓨터 일반'에는 '운영체제 사용', '컴퓨터 시스템 설정 변경' 등이 포함된다. 문항은 문제의 질문(Question)을 의미하고, 보기는 문항에 대한 선택지(Choices), 답은 문제와 보기에 알맞은 답안(Answer)을 의미한다. 문제는 문항, 보기, 답안을 포함한 개념이다.

본 연구에서는 사전 모델로 LangChain 기반의 Ollama 환경 사전 학습된 Llama-3.1 기본 모델에서 추론을 통해 문제생성하였다. 이후 자격증에 특화된 모델 개발을 위해 Hugging Face 에 데이터셋 (jungsoon/ComputerLiteracy)을 구축하였고, 사후 모델은 Unsloth/Meta-Llama-3.1-8로 파인튜닝을 실시하여 모델을 저장하였고, Unsloth에서 제공하는 토크나이저와 디코더를 통해 문제를 생성하였다. 사전 모델과 사후 모델에서 생성한 문제의 성능 비교를 위해 실험을 실시하기 위한 측정 지표는 표 2에 설명하고 있다. 실험은 2명의 평가자가 사전 모델과 사후 모델 각각에서 생성된 10개의 문제를 평가하는 방식으로 진행하였고, 총 20명의 참가



*The existing dataset consists of certification exam questions written in Korean, and it is presented in Korean to prevent any potential distortion of the original meaning or context during translation.

그림 1. 데이터 전처리 및 허깅페이스 배포 프로세스
Fig. 1. Data preprocessing and hugging face deployment process

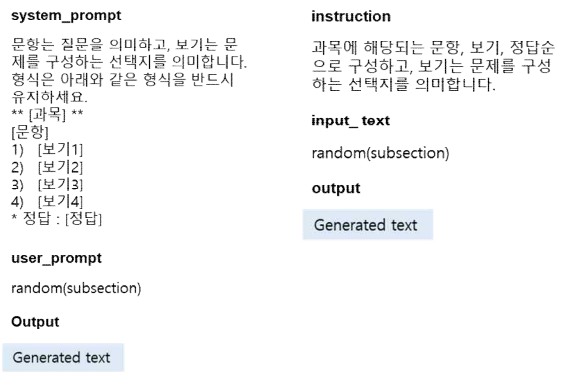
자가 100개의 문제를 평가하는 설문에 참여하였다. 예를 들어, 1에서 10번까지의 문제를 서로 다른 시간대에 2명의 참가자가 동일한 문제에 대해 평가하였다. 사전모델은 Windows 환경에서 Jupyter lab를 통해 개발하였고, 사후모델은 Google Colaboratory 환경에서 T4 GPU를 사용하여 개발하였다. 실험 후 수집된 자료는 통계적 검증을 위한 유의수준 0.05로 설정하여 SPSS 26.0 for windows 프로그램을 이용하여 분석하였다. 모델 설계에 관한 자세한 내용은 2-2장에서 2-5장까지 설명하고, 실험 설계에 관한 자세한 내용은 2-6장에서 2-9장까지 설명한다.

2-2 데이터 전처리 및 데이터셋 구성

Llama 3.1[16]와 같은 첨단 대형 언어 모델(LLM)은 다국어 지원, 코드 생성, 추론, 도구 사용 등 다양한 사례를 지원한다. 대부분의 특정 목적에 맞춰 모델을 더 잘 정렬하고 맞춤화하기 위해 파인튜닝을 수행한다. 하지만 유해하게 미세조정된 모델이 모든 모델 중에서 유용하지 않고 견고하지 않으며, 잘못된 맥락이 제공되었을 때 가장 낮은 정확도 점수를 보였고, 불확실성 지표에서도 이를 입증하였다[17]. 한편 자격증 시험의 특성상 세부과목 정보는 데이터셋에 필수적이지만, 기존 MQAG(Multiple-choice Question Answering)에서 많이 활용하고 있는 KMMLU(Measuring Massive Multitask Language Understanding in Korean) 데이터셋에는 문제별 세부과목이 포함되어 있지 않았다. 그래서 본 논문에서는 자격증 중에서 가장 많이 응시하는 컴퓨터 활용능력 2급 기출문제 10개의 문제를 이미지가 포함된 문제를 제거하여 총 386개의 세부과목이 포함된 자체 데이터셋를 구축하였다. 이러한 세부과목 정보가 포함된 자체 구축한 데이터셋은 모델이 특정 과목 내의 고유한 패턴을 학습할 수 있도록 하였다. 또한 자격증 문제에 특화된 문제와 문장 구조를 반영하여 데이터셋의 일관성을 높이고, 실제 시험 문제와 유사한 문제 생성을 가능하게 하였다. 파인튜닝을 하기 전에 Hugging Face 포맷에 맞춰 데이터를 전처리하였다. 기출문제는 hwp(한글 파일 확장자) 파일로 문항, 보기로 구성되어 있고, 보기에는 정답이 마킹되어 있다. 전처리를 위해 text 파일로 변환하여 문항, 보기, 정답으로 분할하였다. 이후 엑셀파일로 저장하여 세부과목을 ChatGPT few-shot 프롬프팅 방법으로 과목을 매칭하여 세부과목을 엑셀파일에 수기로 수정하였다. 전체 처리 과정은 그림 1에 도식화하였고, 최종 Hugging Face 데이터셋의 필드는 표 1과 같이 구성하였다. 데이터 전처리는 Python 언어를 사용하였고, Pandas, re, huggingface_hub 라이브러리를 사용하였다.

표 1. 허깅페이스 데이터셋 필드 구조
Table 1. Hugging face dataset field configuration

Item	Explain
subject	Certification Name + Subject Name
subsection	Subsections by Certification
question	Question (Problem Statement)
choices	Choices for the Question
answers	Answer to the Problem and Question
user_input	Field for User Input for Fine-Tuning. Composed of Empty Cells.



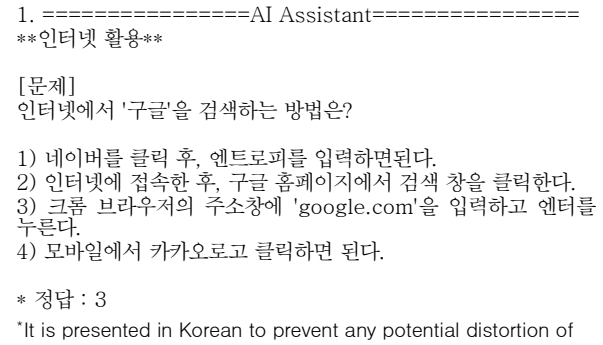
(a) Pre-model (b) Post-model

*It is presented in Korean to prevent any potential distortion of the original meaning or context, as the questions were generated for a Korean certification exam.

그림 2. Llama 3.1 기반의 추론을 위한 프롬프트 형식
Fig. 2. Prompt format for inference based on Llama 3.1

2-3 Ollam 기반 Llama 3.1의 문제 생성 개요

LangChain은 LLM의 기능을 효율적으로 활용하고 사용자에게 편리한 사용 경험을 제공하는 것을 목표로 한다[18]. 또한 타 모델과의 통합을 사용하여 애플리케이션을 구축할 수 있다. 한편, Ollama는 저성능 환경에서 LLM을 신속하게 배포할 수 있는 능력으로 주목받고 있다[19]. 그래서 사전 학습된 Llama 3.1 기본 모델로 문제생성은 LangChain 기반의 Ollama 환경에서 python 언어로 모델을 구현하였다. langchain의 promptTemplate에는 input_variables와


1. ==========AI Assistant==========
인터넷 활용

[문제]
인터넷에서 '구글'을 검색하는 방법은?

1) 네이버를 클릭 후, 엔트로피를 입력하면된다.
2) 인터넷에 접속한 후, 구글 홈페이지에서 검색 창을 클릭한다.
3) 크롬 브라우저의 주소창에 'google.com'을 입력하고 엔터를 누른다.
4) 모바일에서 카카오로고 클릭하면 된다.

* 정답 : 3

*It is presented in Korean to prevent any potential distortion of the original meaning or context, as the questions were generated for a Korean certification exam.

그림 3. Ollama 기반의 질문 생성 결과
Fig. 3. Question generation results based on Ollama

1.==========AI Assistant==========

**과목: 인터넷 활용
문제:다음 중 인터넷을 이용하여 다른 컴퓨터에 있는 파일을 자신의 컴퓨터에서 열 수 있는 기능으로 옳은 것은?
보기:
1) FTP
2) HTTP
3) SMTP
4) POP
정답:1: FTP

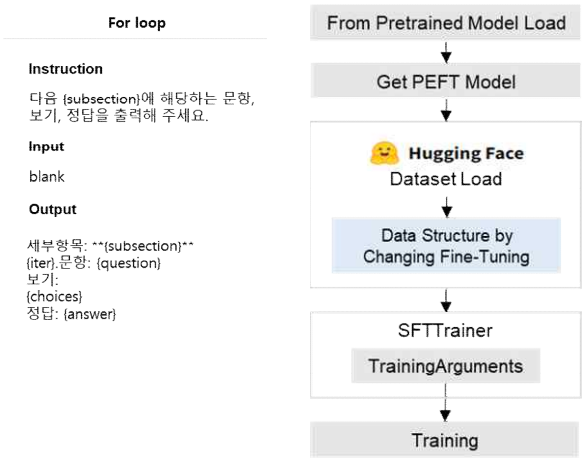
*It is presented in Korean to prevent any potential distortion of the original meaning or context, as the questions were generated for a Korean certification exam.

그림 4. Unsloth 기반의 질문 생성 결과
Fig. 4. Question generation results based on Unsloth

template로 파라미터가 있다. 이곳에 Llama에서 기대하는 응답의 prompt를 설정하고, system_prompt는 시스템에서 처리했으면 하는 형식을 작성하였고, user_prompt에는 응답으로 출력했으면 하는 과목을 입력하였다. 100문제를 생성하기 위해 반복문을 통해 한 문제씩 생성하였고, 이때 세부과목은 리스트로 정의하고 랜덤하게 user_prompt에 포함되도록 코딩하였다. Ollam 기반의 Llama 3.1에서 system_prompt와 user_prompt 구현은 그림 2(a)에 도식화했고, 생성한 문제 예시는 그림 3에 작성되었다.

2-4 Unsloth/Meta-Llama-3.1-8기반 파인튜닝 모델의 문제 생성 개요

Unsloth는 Llama, Mistral, Phi-3, Gemma 등과 같이 대규모 언어 모델을 파인튜닝하는 속도를 2배 더 빠르게 하고, 메모리를 70% 적게 사용하며, 정확도는 기존 모델과 동일하다라고 설명하고 있다[15]. Unsloth에서 서비스 하고있는 모델 중 Meta-Llama-3.1-8 모델을 빠른 속도로 파인튜닝이 가능하다. 또한 LLM의 크기를 줄이면서 성능을 유지하거나 향상 시키기 위한 기술. 즉, 양자화(Quantization), LoRA(Low-Rank Adaptation), DPO(Distillation and Pruning Optization), PPO(Proximal Policy Optimization) 등의 옵션들을 제공하여 다양한 응답을 할 수 있는 모델이다. 파인튜닝 절차는 (1) Pretrained Model (unsloth/Meta-Llama-3.1-8B)을 불러오고, (2) unsloth에서 제공하는 FastLanguageModel 을 활용하여 PEFT (Parametric Efficient Fine-Tuning)의 Lola_alpha, gradient_checkpoint, random_state 값을 설정하여 메모리 사용량을 30% 줄이고, 배치 크기를 2배로 늘릴 수 있는 옵션을 설정했다. (3) Hugging Face의 데이터셋을 로드하고, (4) 문제에 대한 정답을 출력하기 위해 파인튜닝 전 데이터 구조를 unsloth에서 요구하는 형식(instruction, input, output)으로 세팅하고(그림 5(a)), (5) 데이터셋에 프롬프트 포맷팅 함수를 적용하여 데이터셋을 업데이트한다. (6) SFTTrainer 메서드로 하이퍼 파라미터를 조정하여 학습할 준비를 하고,



(a) Data structure for fine-tuning (b) Fine-tuning Process

*It is presented in Korean to prevent any potential distortion of the original meaning or context during the inference process of generating questions for a Korean certification exam.

그림 5. 파인튜닝 데이터 구조 및 프로세스
Fig. 5. Data structure and process for fine-tuning

(7) 학습을 시작한다(전체 프로세스: 그림 5(b)). (5)의 instruction은 세부과목(Subsection)에 따른 문제생성, Output은 공정한 비교 및 문제 생성할 때 형식을 유지하기 위해 사전 모델의 system_prompt(그림 2(a)와 비슷한 데이터 구조를 사용했다.

학습을 완료하고 추론을 위해 학습하기 전 데이터 구조(instruction, input, output)를 설정하고(그림 2(b)), unsloth에서 제공하는 토크나이저를 사용하여 디코딩 후 문제를 생성하였다. 이를 실험을 위한 100문제 생성을 위해 반복하였다. 파인튜닝 모델에서 생성한 문제 예시는 그림 4에

표 2. 학습 파라미터 설정
Table 2. Training parameter setting

Item	Setting	Explain
batch_size	2	Learning more detailed patterns
gradient_accumulation_steps	4	Accumulating gradients over 4 Steps for increased training stability
warmup_steps	5	Stably learning certification Question patterns
num_train_epochs	1	Full fine-tuning
max_step	100	Maximum training steps
learning_rate	2e-4	Balancing training speed and stability
optimizer	adamw_8bit	Maximizing memory efficiency
weight_decay	0.01	Preventing overfitting
lr scheduler type	linear	Using a linear learning rate
seed	3407	Ensuring consistency and reproducibility in experiments
trainable parametes	41,943,040	Number of trainable parameters

서 확인할 수 있다.

2-5 성능 향상을 위한 학습 파라미터

배치크기를 2로 설정하고 그래디언트를 4번 누적함으로써, 모델은 세밀한 그래디언트 업데이트를 통해 자격증 문제의 주요 패턴을 효과적으로 학습할 수 있었다. 이러한 설정은 데이터셋 내에 존재하는 세부 패턴을 반영하는 데 유리하며, 학습 안정성을 높여 과적합을 방지하는 데에도 기여했다. 또한 학습률 2e-4로, 가중치 감소를 0.01로 설정하여 모델이 데이터에 과도하게 적응하지 않으면서도 자격증 문제의 핵심 패턴을 잘 반영할 수 있도록 하였다. 이러한 설정은 학습 과정에서 모델의 성능을 안정적으로 향상시키는 데 도움이 되었다.

2-6 참가자 모집

실험 대상자는 자격증 문제에 대한 이해도가 있는 전산지식이 있는 IT관련학과 학생들을 모집하였다. 동의대학교 재학생을 대상으로 모집 공고를 통해 20명의 참가자가 참여하였다. 본 연구는 한 학생이 사전모델(Ollama 기반의 LLaMA-3.1)과 사후 모델(Unsloth/Meta-Llama-3.1-8B 기반의 파인튜닝 모델)의 문제를 풀고, 설문을 통해 평가하였다. 유의수준 0.05, 70% 이상의 검정력을 가진 대응표본 검정을 시행하였다.

2-7 실험 절차 및 방법

본 연구는 한명의 참가자가 사전모델(Ollama 기반의 LLaMA-3.1)과 사후 모델 (Unsloth/Meta-Llama-3.1-8B 기반의 파인튜닝 모델)을 비교하는 실험을 수행하였다(그림 6). 동일한 날짜에 사전 모델과 사후 모델 각각에서 학습된 모델에서 100개의 문제를 생성한 뒤, 실험 전에 Google Docs에 복사하였고, 설문을 위해 Google Form도 각각 준비하였다. 실험 절차는 (1) 참가자에게 실험 방법을 충분히 설명한 후 설명서를 제공하였다. (2) 개인정보 동의를 한 참가자는 인구통계학적 조사를 위해 Google Form을 통해 참가자가 직접 설문했다. (3) 사전모델과 사후모델 각각의 10문제를 보고, 한 문제당 5개의 설문 문항에 응답했다. 문제 또는 설문 문항이 애매모호할 때는 인터넷 검색을 허용했고, 설문 조사 관련 문의사항은 실험자에게 문의했다. (4) 설문을 완료한 후 참가자들은 보상을 받고 퇴장하였다. 모든 참가자는 동일한 실험환경에서 3일동안 실험을 실시하였다.

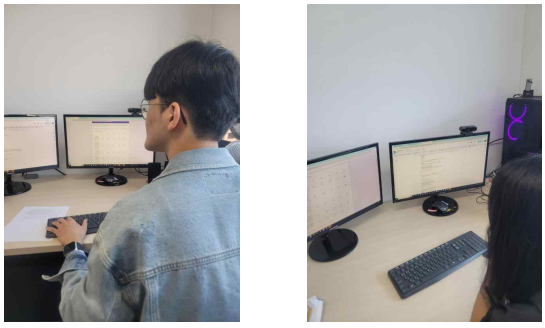


그림 6. 실험 환경
Fig. 6. Experimental environment

2-8 측정지표

문제의 품질을 평가하기 위해 이전 연구[20]에서 제시한 평가 지표를 참고하여 유창성, 지문 일관성, 과목 연관성을 Likert-5점 척도 구성하였으며, 유일성과 정확성은 이분 척도를 사용하여 평가하였다. 마지막으로, 전체 문제에 대한 평가와 추후 개선사항을 서술형으로 작성하도록 요청하였다(표 3). 모든 지표는 성능 측정을 위한 지표로 점수가 낮을 경우 성능이 낮고, 점수가 높을수록 좋은 성능을 뜻한다.

표 3. 사전모델과 사후모델의 성능 비교를 위한 측정 지표
Table 3. Evaluation metrics for comparing the performance of pre-model and post-model

No	Measurements	Explain
1	fluency	Is there any grammatical error in the problem? Is the problem description incomplete or missing any information? Are the choices incomplete or are any of the choices duplicated?
2	stem-choice coherency	Are the provided question and choices logically connected?
3	subsection relevance	Are the subsection and the question logically connected?
4	uniqueness	Is the correct answer among the choices unique?
5	correctness	Is the correct answer accurate?
6	others	Please provide an overall evaluation of the problem and suggest improvements.

2-9 통계 분석 설계

LLM을 정량적으로 평가하고 비교하기 위해서 Evaluation Metric을 주로 이용한다. 대표적인 성능 평가 방법으로 Perplexity, Human Evaluation[21], BLEU(Bilingual Evaluation Understudy)[22], ROUGE(Recall-Understudy for Gristing Evaluation)[23] 등이 사용된다. Perplexity, BLEU, ROUGE와 같은 자동 평가 지표는 주로 언어적 패턴이나 어휘적 일치성을 기반으로 하며, 모델이 생성한 텍스트가 얼마나 원본 데이터와 유사한지를 측정한다. 이러한 자동 지표보다는 자격증 시험 문제에 특화된 모델의 품질을 평가하기 위해서는 전문 지식을 가진 사람이 평가하는 것이 더 적합한 것으로 보인다. 그래서 본 연구에서는 Human Evaluation으로

생성된 응답을 다양한 지표를 표 3과 같이 정의하였다.

Shapiro-Wilk 검정은 정규성 검정 중 하나로, 데이터를 분석할 때 정규분포를 따르는지 확인하는 데 사용된다. 또한 윌콕슨 부호 순위 검정(Wilcoxon Signed-Rank Test)은 비모수 검정의 한 종류로, 대응된 두 집단 간의 차이를 비교할 때 사용한다. 즉, 동일한 그룹에서 두 조건(사전모델과 사후모델)간의 차이를 분석할 때 사용한다. 크론바흐의 알파값(Cronbach's alpha)은 여러 평가자가 동일한 대상을 평가할 때 평가의 일관성을 측정하는지를 평가하는 신뢰도(내적 일관성) 측정 방법이다.

본 논문에서는 사전모델과 사후모델의 차이를 비교하기 위해 측정지표 중 유창성(fluency), 지문 일관성(stem-choice coherency), 과목 연관성(subsection relevance)은 데이터가 정규분포를 따르는지 확인하기 위해 Shapiro-Wilk 검정을 수행하였고, 정규분포를 따르지 않기 때문에 비모수 검정 방법인 윌콕슨 부호 순위 검정(Wilcoxon Signed-Rank Test)을 실시하여 두 모델 간의 유의미한 성능 차이를 확인하고, 신뢰도 측정을 위해 크론바흐의 알파값을 확인하였다. 그리고 유일성(uniqueness)과 정확성(correctness)은 명목형 변수(1:예, 2:아니오)로 카이제곱 검정(Chi-square test)을 사용하여 두 모델간의 차이를 비교하였다.

Ⅲ. 결 과

3-1 통계 분석

본 연구는 총 20명의 참가자를 모집하였으며, 한 사람이 사전모델(A condition, Ollama 기반의 LLaMA-3.1)과 사후모델(B condition, Unsloth/Meta-Llama-3.1-8B 기반의 파인튜닝 모델)로 생성된 문제를 각 10문제씩 풀고, 설문에 응답했다. 분석 결과에 따르면 평균 연령 23.6세 (SD 1.96)이고, 남성 13명 (65%), 여성 7명 (35%), 자격증 시험 경험이 있는 참가자는 10명 (50%), 경험이 없는 참가자는 10명 (50%) 이었다. 이러한 인구 통계학적 정보는 연령, 성별, 자격증 시험 경험 유무 등의 변수가 파인튜닝한 모델의 전후 문제 생성의 연구를 위한 기초 자료로 제공되었다.

윌콕스 부호순위 검정을 통해 유창성, 지문 연관성, 과목 일관성의 효과를, 카이제곱 검정을 통해 유일성, 정확성을 평가했다. 표 4는 모든 측정값의 대표 통계치로, 결과를 개괄적으로 보여준다. 정의된 측정값 중 4가지(유창성, 지문 연관성, 과목 일관성, 유일성)가 주요 설계 요소에서 유의미한 효과를 나타냈다. 또한 유창성, 지문 연관성, 과목 일관성의 3가지 평가 지표를 사용하여 모델 성능을 평가하였다. 평가 항목 간의 신뢰도를 검증하기 위해 크론바흐의 알파값을 계산한 결과 0.768로 나타났다. 이는 평가 지표 간의 내적 일관성이 비교적 높으며, 연구에서 사용된 항목들이 신뢰할 수 있는 수준임을 시사한다.

표 4. 통계 분석 결과(n=200)
Table 4. Statistical analysis results(n=200)

Variable	Mean		SD		Wilcoxon Signed-Rank Test
	A condition	B condition	A condition	B condition	
fluency	3.22	4.38	1.76	1.22	< .001
tem-choice coherency	3.39	4.11	1.59	1.37	< .001
subsection relevance	3.80	4.17	1.45	1.31	.007
Variable	A condition		B condition		χ^2 (p)
	Yes	No	Yes	No	
uniqueness	76 (38.00%)	124 (62.00%)	100 (50.00%)	100 (50.00%)	5.84* (.02)
correctness	105 (52.50%)	95 (47.50%)	104 (52.00%)	96 (48.00%)	0.01* (.92)

* $p < .05$

Ⅳ. 논 의

윌콕스 부호 순위 검정과 카이제곱 분석 결과, 유창성, 지문 연관성, 과목 일관성, 유일성 4가지 측정값에서 사전모델과 사후모델 간의 명확한 차이가 나타났다. 사후 모델은 사전모델에 비해 5가지 지표 중에서 4가지 면에서 더 우수한 성능을 보여주었다.

유창성은 파인튜닝한 모델이 문법적 오류가 적고, 문제와 문항이 완전성 면에서 우수하다는 것을 나타내고, 지문 연관성은 파인튜닝한 모델이 문항과 보기가 논리적으로 더 잘 연결되었음을 나타내고, 과목 일관성은 파인튜닝한 모델이 세부과목과 문항이 논리적으로 연결된 정도가 우수하다고 해석할 수 있다. 마지막으로 유일성이 유의미하다는 것은 파인튜닝한 모델이 문제의 보기 중 정답이 유일한 정도가 우수하다는 것을 의미하지만 유일하지 않다는 비율이 50% 수준이기 때문에 좋은 성능이라고 하기에는 부족하다.

본 연구에서는 사전 모델과 사후 모델에서 생성한 문제를 경험한 후 참가자들의 의견을 수집하였다. 사전 모델의 장점은 문제 설명이 이해하기 쉽고 보기가 적절하다는 의견이 1명 참가자가 응답하였고, 단점으로는 문항이 없거나 의문형이 아닌 줄글 형태, 주관식 문제, 오타가 많이 보인다는 의견이 있었고, 문항과 보기가 논리적으로 연결되지 않은 문제, 1명의 참가자는 자격증 문제로서 사용 불가능할 정도의 문제였다는 의견도 있었다. 사후 모델의 장점은 자격증 시험 문제와 비슷한 문제, 유창성 측면에서 우수하다는 의견이 6명의 참가자가 있었고, 단점은 8명의 참가자가 보기가 동일한 중복성 문제를 지적하였다. 이를 종합했을 때 사전 모델의 경우 문제

의 유창성이 떨어진다는 것을 알수 있고, 이에 반해 사후 모델은 유창성이 우수하다는 것을 알 수 있다. 표 4에서 유창성의 결과(4.38±1.22)에서도 높다고 할 수 있다.

LangChain 기반 Ollama 환경에서는 기존의 사전 학습된 LLaMA-3.1 모델로 자격증 문제와 관련된 특정 패턴을 반영하지 않았지만 일반적인 언어 능력을 보유하고 있다. 반면 Unsloth/Meta-Llama-3.1-8B는 자격증 문제 생성에 최적화하도록 파인튜닝하여 특정한 데이터와 세부 과목 정보를 반영하는 성능을 갖춘 모델이다. 본 논문에서는 동일한 모델 기반의 성능 평가를 위해 정량적 평가보다 Human Evaluation에 집중하였다. 또한 자체 데이터셋의 성능을 정량적으로 판단하기 위해 우리는 KMMLU 데이터셋(haerae-hub/kmmlu)의 컴퓨터 사이언스 카테고리만 포함하여 Unsloth로 파인튜닝을 실시하여 모델의 Loss를 확인하였고, 자체 데이터셋보다 Loss가 더 높았다. KMMLU 데이터셋과 자체 데이터셋의 필드는 세부과목만 다르고, 나머지는 동일하다. KMMLU와 세부과목이 포함된 자체 데이터셋으로 동일한 데이터형식으로 파인튜닝을 실시하였을 때 자체 데이터셋의 Loss가 더 낮다는 것은, 모델이 세부 과목 정보가 포함된 데이터셋에서 더 효과적으로 학습하고 있음을 시사한다. 이는 세부 과목 정보가 포함된 데이터셋이 모델이 문제의 패턴과 구조를 더 잘 이해하도록 돕기 때문일 가능성이 크다. 자체 데이터셋에 세부 과목 정보가 포함되어 있기 때문에, 모델이 각 과목의 고유 패턴을 학습하면서 더 정밀하게 문제를 이해하고 예측할 수 있다. 이로 인해 학습 과정에서 Loss가 더 빠르게 감소하여 성능 향상의 요인이라 할 수 있다. 하지만 실제로 성능 향상되었는지는 명확히 입증하기 위해서는 정확도와 일관성을 확인하는 연구가 추후 필요하다.

연구를 진행하면서 원하는 개수의 문제를 생성하기 위해 다양한 시도를 해 보았다. 프롬프트에 생성할 문제의 개수를 system_prompt와 user_prompt를 여러번 바꿔가면서 시도하였으나 문제 출력 개수가 변동되는 문제를 발견하였다. 그래서 이를 해결하기 위해 반복 구조를 사용하였고, 장점은 내가 원하는 개수만큼 문제를 생성할 수 있었고, 향후 애플리케이션에 적용할 수 있는 실시간으로 문제 생성할 수 있는 장점을 가지고 있다.

참가자들의 인터뷰에서 추후 개선사항으로 난이도 조절, 문제에 대한 최신화, 설명 포함 등의 의견이 있었고, 앞절에서의 파인튜닝한 모델에서의 보기가 동일한 중복성 문제와 답안의 유일성 측면에서 다소 부족한 문제가 있다. 이를 보완하기 위해서는 더욱 다양한 데이터셋과 정확한 응답을 위한 RAG(Retrieval Augmented Generation), DPO(Direct Preference Optimization), PPO(Proximal Policy Optimization)등의 방법을 도입하여 최신 자료나 데이터를 모델이 실시간으로 검색하여 문제 생성에 반영하고, 난이도 조절과 참가자 피드백에 따른 선호도를 반영하여 답안의 유일성을 개선하는 데 도움이 될 것이다. 또한 애매모호한 문항을 줄이고, 반복적인 질문을 거친 프롬프트로 논리적 추론을 거쳐 신뢰할 수 있는 응답을 얻을 수 있을 것이다.

V. 결 론

본 연구는 자격증 문제 생성에 특화된 데이터셋을 구축하고, 세부과목 정보가 포함된 데이터셋을 통해 파인튜닝을 수행하여 기존 연구와 차별화된 접근방식을 제시하였다. 이를 통해 파인튜닝 모델은 유창성, 지문 연관성, 과목 일관성, 유일성 등의 측정값에서 사전모델보다 더 우수한 성능을 보였다. 이는 파인튜닝이 모델 성능을 실질적으로 향상시킬 수 있다는 것을 입증하였다. 특히 자격증 시험과 같은 실제 응용 사례에서 그 효과를 확인한 점을 실질적인 기여라 할 수 있다.

또한, LangChain 기반 Ollama 환경과 Unsloth 환경에서 동일한 모델 아키텍처와 평가 절차를 유지하여 공정한 비교를 위해 노력하였으며, 자격증 문제와 관련된 특성 평가에는 Human Evaluation이 필수적임을 강조하여 연구 결과의 유의미성을 확보하고자 한다. 참가자들의 의견을 수집하여 사전 모델과 사후 모델의 장단점을 분석한 결과는 연구의 신뢰성을 높여준다. 특히 사전 모델의 경우 문제가 이해하기 쉬운 반면, 문항의 부재, 논리적 연결 부족 등의 문제점이 있었던 반면, 사후 모델은 자격증 시험 문제와 유사한 문제를 생성하고, 논리적 연결성에서도 긍정적인 평가를 받은 점은 교육과 시험 출제의 자동화에서 중요한 개선점을 시사한다.

반복 구조를 사용하여 원하는 개수만큼 문제를 생성할 수 있다는 점은 실제 애플리케이션에서 실시간으로 문제를 생성할 수 있는 장점을 제공하고, 시험 출제 및 교육 콘텐츠 개발에서 효율성을 높일 수 있다. 또한 이번 연구는 자격증 시험 문제생성에 특화된 모델을 개발하였지만 다양한 과목의 데이터셋을 구축하면 자격증이 아닌 다양한 문제를 생성할 수 있는 가능성을 보여주었다.

연구결과, 사후 모델이 여러 면에서 우수한 성능을 보였지만 유일성 측면에서는 아직 개선이 필요하다는 점을 지적하였다. 명확한 단일 정답을 요구하는 시험 문제에서는 여전히 한계로 작용할 수 있다. 이는 향후 연구에서 추가적인 개선이 필요한 부분으로 작용할 수 있으며, 본 논문이 제시한 연구의 한계점이다.

감사의 글

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 지역지능화혁신인재양성사업임(IITP-2024-RS-2020-11201791)

참고문헌

[1] J.-C. Yun, “A Study on the IT-Related Certificates Preferred by University Students,” *Journal of Knowledge Information Technology and Systems*, Vol. 17, No. 3, pp. 405-414, June 2022. https://doi.org/10.34163/jkits.2022.17.3.003

[2] Human Resources Development Service of Korea, National Technical Qualification Statistical Yearbook 2024, Author, Ulsan, National Statistics No. 387004, June 2024.

[3] Y. L. Kim and R. W. Rhee, “Real-Time Evaluation System Using User Profile for Acquisition of a Computer Certificate of Qualification,” *Journal of the Korea Society of Computer and Information*, Vol. 11, No. 2, pp. 153-158, May 2006.

[4] H. Y. Ryu and E. J. Kim, “Design of a Web-Based Education System for Engineer Test,” in *Proceedings of the Korean Information Science Society Conference*, Daejeon, pp. 679-681, April 2004.

[5] E. Sneiders, Automated Question Answering: Template-Based Approach, Ph.D. Dissertation, KTH Royal Institute of Technology, Stockholm, Sweden, February 2002.

[6] Q. Zhou, N. Yang, F. Wei, C. Tan, H. Bao, and M. Zhou, “Neural Question Generation from Text: A Preliminary Study,” in *Proceedings of the 6th CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC 2017)*, Dalian, China, pp. 662-671, November 2017. https://doi.org/10.1007/978-3-319-73618-1_56

[7] D. Xiao, H. Zhang, Y. Li, Y. Sun, H. Tian, H. Wu, and H. Wang, “ERNIE-GEN: An Enhanced Multi-Flow Pre-Training and Fine-Tuning Framework for Natural Language Generation,” in *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI-20)*, Yokohama, Japan, pp. 3997-4003, January 2021. https://doi.org/10.48550/arXiv.2001.11314

[8] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations,” in *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*, Addis Ababa, Ethiopia, April 2020. https://doi.org/10.48550/arXiv.1909.11942

[9] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, ... and L. Zettlemoyer, “BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 7871-7880, July 2020. https://doi.org/10.18653/v1/2020.acl-main.703

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, ... and V. Stoyanov, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” arXiv:1907.11692, July 2019. https://doi.org/10.48550/arXiv.1907.11692

[11] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, ... and P. J. Liu, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” *The Journal of Machine Learning Research*, Vol. 21, No. 1, pp. 5485-5551, January 2020. https://doi.org/10.48550/arXiv.1910.10683

[12] M.-J. Cho, J.-B. Kim, M.-S. Kang, M.-S. Kang, and D.-I. Jeon, “Research on Developing a TOPIK Reading Comprehension Question Generation System Using Generative AI Technology,” *Journal of the Korea Contents Association*, Vol. 24, No. 9, pp. 55-65, September 2024. https://doi.org/10.5392/JKCA.2024.24.09.055

[13] D. Heo, K. Kim, H. Song, and B. Suh, “Proposal Of Korean CSAT Customized Question Generator System With Prompt Engineering,” in *Proceedings of HCI Korea 2024*, Hongcheon, pp. 183-189, January 2024.

[14] J.-Y. Park, J.-U. Lee, J.-B. Choi, S.-Y. Hong, S.-J. Hong, and E.-S. Jung, “Implementation of a Model for Generating College Scholastic Ability Test (CSAT) English Questions Using a Large-Scale Language Model,” in *Proceedings of 2024 Summer Annual Conference of IEIE*, Jeju, pp. 2580-2584, June 2024.

[15] GitHub. unslothai/unsloth: Finetune Llama 3.2, Mistral, Phi [Internet]. Available: https://github.com/unslothai/unsloth/.

[16] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, ... and Z. Papakipos, “The Llama 3 Herd of Models,” arXiv:2407.21783v2, August 2024. https://doi.org/10.48550/arXiv.2407.21783

[17] S. Kumar, “Overriding Safety Protections of Open-Source Models,” arXiv:2409.19476, September 2024. https://doi.org/10.48550/arXiv.2409.19476

[18] C. Jeong, “A Study on the Implementation of Generative AI Services Using an Enterprise Data-Based LLM Application Architecture,” *Advances in Artificial Intelligence and Machine Learning*, Vol. 3, No. 4, pp. 1588-1618, October 2023. https://doi.org/10.54364/AAIML.2023.1191

[19] J. B. Gruber and M. Weber, “Rollama: An R Package for Using Generative Large Language Models through Ollama,” arXiv:2404.07654, April 2024. https://doi.org/10.48550/arXiv.2404.07654

[20] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, “A Systematic Review of Automatic Question Generation for Educational Purposes,” *International Journal of Artificial*

Intelligence in Education, Vol. 30, No. 1, pp. 121-204, March 2020. https://doi.org/10.1007/s40593-019-00186-y

[21] J. Novikova, O. Dušek, A. C. Curry, and V. Rieser, "Why We Need New Evaluation Metrics for NLG," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 2231-2242, September 2017. https://doi.org/10.18653/v1/D17-1238

[22] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02)*, Philadelphia: PA, pp. 311-318, July 2002. https://doi.org/10.3115/1073083.1073135

[23] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Proceedings of the Workshop on Text Summarization Branches Out*, Barcelona, Spain, pp. 74-81, July 2004.

김정순(Jung-Soon Kim)

2023년 : 동의대학교 대학원 (공학석사)

1994년～1997년: ㈜ 미화당
2021년～2003년: 동의과학대학교 강사
2023년～현 재: 그루협동조합 감사
2023년～현 재: 동의대학교 인공지능학과 박사과정
※관심분야 : Generative Ai, Medical LLM, Chatbot, HCI(Human Computer Interaction), Education 등

김성희(Sung-Hee Kim)

2006년 : 이화여자대학교 B.S.
2008년 : 이화여자대학교 M.S.
2014년 : Purdue University Ph.D

2015년～2017년: 삼성전자 소프트웨어센터 책임연구원
2019년～2023년: 동의대학교 빅데이터인공지능센터 소장
2017년～현 재: 동의대학교 산업ICT기술공학과 부교수
2024년～현 재: 동의대학교 인공지능 대학원 주임교수
※관심분야 : HCI(Human-Computer Interaction), User-centered Artificial Intelligence, Data Visualization 등