생성AI기반 TOPIK 문제 자동생성 및 모의학습 IBT 플랫폼 구현

TOPIK auto generates question based Gen AI and Implements a mock IBT study platform

2024 전기 졸업과제 착수보고서 <훈민정컴>

제출일	2024. 05. 13	전공	정보컴퓨터공학부
팀장	201924497 신병근	팀명	훈민정컴
팀원	202055517 김범수 201924612 허진영	지도교수	조준수

목차

1. 연구 배경……..3

2. 문제 상황 ……4

3. 목표……..4

4. 시스템 구조 및 흐름…..……5

5. 개발 환경 및 사용 기술……7

6. 일정 및 역할 분담……9

7. 참고 문헌……10

1. 연구 배경

1.1. TOPIK(Test of Proficiency in Korean):

1.1.1. 외국인을 비롯한 한국어를 모어로 하지 않는 사람들을 대상으로 한 1997년부터 시행된 한국어 능력 시험.

1.1.2. TOPIK I(1,2급) TOPIK II(3~6급)으로 나뉘며, 쓰기(=작문), 읽기, 듣기가 각각 100점 만점이다.

1.1.3. 대한민국을 포함하여 87개 국가, 323개 지역에서 시험을 실시하며 2019년 통계 기준으로 TOPIK I는 101,617명, TOPIK II는 274,254명이 지원하는 등 국제적으로 상당한 인지도를 지닌 언어자격시험 중 하나이다.

1.2. **LLM**의 발전**:**

지난 몇 년 동안 LLM(Large Language Models)의 크기가 급격히 증가했으며, GPT-3와 같은 모델은 1750억 개의 매개변수를 갖추고 있음.

많은 GPU/TPU에 걸쳐 병렬화와 모델 아키텍처의 혁신과 같은 기술적 기법은 이러한 거대한 모델을 훈련하는 것을 가능하게 함. LLM의 확장성은 급격한 능력 향상을 가능케 하였으며, 이로써 pretrain 단계에서 얻은 방대한 지식을 활용하여 훈련받지 않은 작업을 수행할 수 있게 됨. 이는 새로운 문제를 만드는 데 유망한 요소가 됨.

1.3. **LLM** 문제 생성의 성능**:**

여러 연구에서 LLM이 문제를 자동 생성하는 데 잠재력을 보임.

1.3.1. [1]2022년 연구에서 연구자들은 의학 교육을 위해 GPT-3를 사용하여 고품질의 객관식 질문을 생성했는데, 의학 전문가들의 맹목적 평가에서 인간이 작성한 질문과 비교 가능한 수준의 품질을 보임.

1.3.2. [2]또 다른 2022년 논문에서는 미세 조정된 LLM이 수학 단어 문제를 고품질로 생성할 수 있다는 것이 선생님들의 판단으로 입증됨. 반복적인 개선을 통해 생성된 문제는 종종 인간이 작성한 문제와 구별하기 어려운 수준이었음.

1.3.3. [3]유기 화학과 같은 더 특화된 분야에서도 연구자들은 LLM이 새롭고 유효하며 비즈니스적인 문제를 생성할 수 있다는 것을 보였으며, 이는 강력한 일반화 능력을 시사함.

---

[1] Qiu et al. (2022). Automatic generation of multiple choice questions for medical education using a large language model.

[2] Yu et al. (2022). Automatic Generation of High-Quality Math Word Problems with Iterative Refinement by Large Language Models.

[3] Matteson et al. (2022). Can GPT-3 Generate Novel, Non-Trivial and Diverse Organic Chemistry Problems? CopyRetryClaude can make mistakes. Please double-check responses.

2. 문제 상황

2.1. 국내상황

작년 한국어 시험 응시자가 41만명으로 최고치를 기록했고, 한국어 TOPIK 시험이 2023년 10월부터 PBT(Paper Based Test)에서 IBT(Internet Based Test)로 전환하였으나, 한국어 학습관련 교재 및 자료는 대학 출판사나 소규모 출판사에 의존하고 있음. 또한 국내 이민청 신설에 따른 이민자 수의 증가가 예상됨.

2.2. 해외상황

해외에서 한국어 시험을 응시하기 위해 대도시로 1년에 2회 정도 이동하고, 숙소가 마땅치 않아서 길에서 노숙하며, 한 달 급여를 단 한번의 시험을 위해 지불하는 상황이지만, 시험 낙제율은 50%이상에 달함.

2.3. 교육환경

교사/강사를 위한 디지털 교보재나 교육 자료가 부족한 상황이며, 학생들도 한국어 모의고사를 칠 수 있는 여건이 취약함.

3. 목표

3.1. 따라서, 본 과제에서는 TOPIK에 대한 학습자의 접근성을 높이고, 학습의 질을 향상 시킬 뿐 아니라 저렴하게 모의고사를 연습할 수 있는 프로덕션 레벨 플랫폼 개발을 제시함.

3.2. 교사/강사의 부담을 줄여 주고 학생 맞춤형 수업을 지원하는 교육 자료 디지털화

3.3. 세부 기능

① 모의, 기출문제를 활용해 모의고사 시험지 자동 제작
(예) 모의고사 회분 + 기출문제 1회분 = 새로운 모의고사 형성
② AI기반 문제 유형별 자동 생성
문법, 빈칸, 주제찾기 등의 총 45개 유형별 기출 데이터를 통해 프롬프트 엔지니어링, RAG기술로 새로운 문항 생성
③ 한국어 단어 어휘 테스트 기능
④ 모의고사 출력과 컴퓨터에서도 바로 응시할 수 있도록 웹 기반 모의고사(PBT, IBT) (컴퓨터에서 자동 채점이 가능)
⑤ 학습자의 학습 데이터 분석을 기반으로 새 문항 추천
⑥ 자동 질의응답 (Chat GPT기반)
⑦ 연관된 동영상 강의 및 학습 콘텐츠 추천

4. 시스템 구조 및 흐름

4.1. 시스템 구조

4.1.1. 사용자 인터페이스 **(Front-end)**

- Vite + React JS + tailwind를 사용해 사용자 친화적인 웹 기반 IBT(Internet-Based Test) 시스템을 구현
- React의 상태 관리 및 Virtual DOM을 통해 사용자 경험을 향상시키고, 컴포넌트의 재사용성을 높임.

4.1.2. 서버 측 애플리케이션 **(Back-end)**

- AWS 클라우드 서비스를 활용하여 확장성과 유연성을 높이고 비용을 절감
- Amazon EC2 인스턴스에서 여러 개의 Docker 컨테이너를 구성하여 애플리케이션을 배포
- NGINX Webserver를 사용하여 로드 밸런싱과 리버스 프록시를 구현
- Spring 프레임워크로 백엔드 서버를 구성하여 Restful API통신과 트래픽에 따른 멀티스레딩을 활용

4.1.3. 데이터베이스

- 기존 기출문제들을 크롤링하여 Amazon RDS로 PostgreSQL의 Master와 Slave 인스턴스를 구성해 저장하고, 백엔드 서버와의 고가용성 통신을 지원함.
- Redis를 캐시 서버로 사용하여 자주 접근되는 데이터를 인-메모리에 로드하여 애플리케이션의 성능(반응 시간)을 높임.

4.1.4. 스토리지

- Amazon S3를 사용하여 문제 이미지 및 듣기 음성과 같은 비정형 데이터를 보관

4.1.5. 보안 및 네트워크

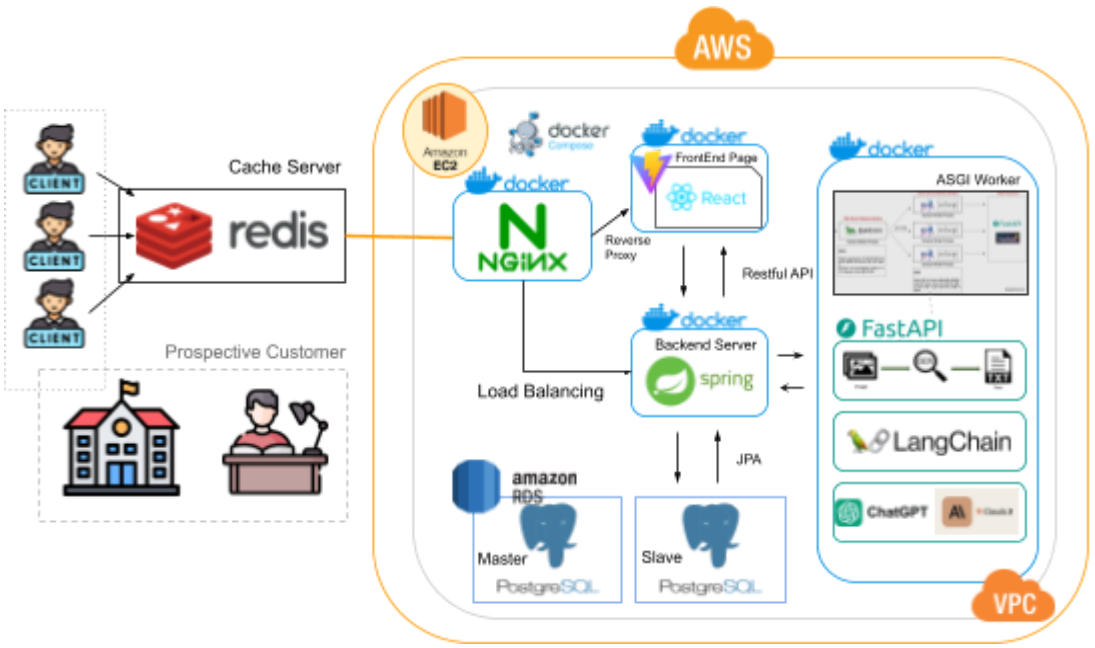
- 전체 애플리케이션은 VPC(Virtual Private Cloud) 내에서 실행해 보안과 네트워크 제어를 강화

4.1.6. 문제 생성 **AI**

- ASGI Worker를 통해 Python Fast API의 병렬 사용을 이용하여, 비동기 통신 API서버를 구현
- LangChain을 사용하여 검색 증강 생성 Rag, LLM(Large Language Model) 프롬프트 엔지니어링 및 체이닝 구현
- 이미지 문제에 대한 OCR인식, 이미지 생성 기술 구현
- 듣기 문제에 대한 STT, TTS 변환 수행

4.2. 전체 구조도

각 서비스들은 모두 Docker를 이용해 마이크로 서비스 아키텍처 형태로 운영되며 각각의 독립적인 작동을 보장함.



[IBT 한국어 문제은행 문항 자동생성 및 모의고사 플랫폼 아키텍처**]**

4.3. 시스템 흐름

- 사용자는 웹 브라우저를 통해 Vite + React JS로 구현된 IBT 페이지에 접속
- 사용자의 요청은 NGINX를 통해 로드 밸런싱되어 적절한 Spring 백엔드 서버로 전달
- 백엔드 서버는 필요한 데이터를 PostgreSQL 데이터베이스 또는 Redis 캐시 서버에서 가져옴.
- 비정형 데이터(이미지, 음성)는 Amazon S3에서 직접 로드됨.
- 생성 AI 기능이 필요한 경우, FastAPI를 통해 Python 서버의 Langchain 시스템과 통신하여 결과를 반환
- 처리된 결과는 다시 사용자에게 전달되어 웹 페이지에 표시됨

5. 개발 환경 및 사용 기술

5.1. 개발 언어

- Frontend : Html, css, JavaScript
- Backend : Java
- AI Server : Python

5.2. 개발 도구

- Frontend : Vite, React, Tailwind, with ESLint
- Backend : Spring, PostgreSQL, Nginx
- Infra : Docker, Redis, AWS (VPC, EC2, S3, RDS)
- AI Server : ASGI worker, FastAPI, Langchain
- LLM : ChatGPT-3.5-turbo, ChatGPT-4-turbo
- OCR, STT, TTS : Google Cloud platform or Local

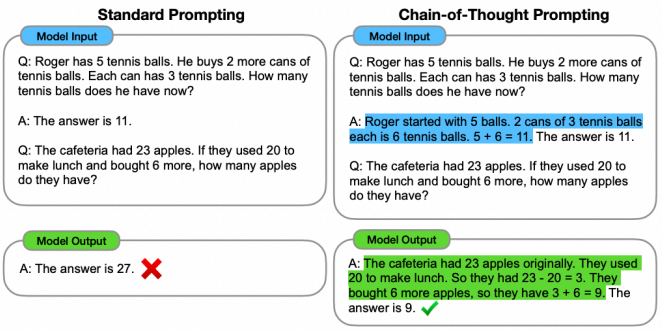
5.3. 사용 기술

5.3.1. **Prompt Engineering**

프롬프트 엔지니어링의 주요 목표는 사용자의 의도와 원하는 결과를 전달하는 프롬프트를 만들어 모델의 성능, 정확성, 유용성을 극대화하는 것이다. 프롬프트 엔지니어링이 필요한 이유는 현재 LLM의 동작 방식의 한계와 인간과 컴퓨터의 상호 작용을 위해 자연어를 사용되기 때문인데, 대표적으로 Auto-Regression 의 한계가 있다.

LLM은 단어의 순서를 비롯한 프롬프트의 작은 변화에 따라 응답의 품질이 상당히 다를 수 있다. LLM은 명령(Instruction)과 예제(Example)의 미묘한 패턴을 감지하여 답변을 조정한다. 따라서 LLM 의 답변 결과는 프롬프트(i.e. 프롬프트에 포함된 특정 명령뿐만 아니라 단어의 선택, 단어의 순서)에 따라 민감하게 달라질 수 있다.

대표적인 LLM인 GPT 모델이 바로 Auto-Regression 모델에 해당한다. Auto-Regression LLM은 이전 단어를 보고 가장 높은 확률의 단어를 다음 단어로 예측하므로 단어의 순서에 따라 얼마든지 다른 답변을 출력할 수 있다.

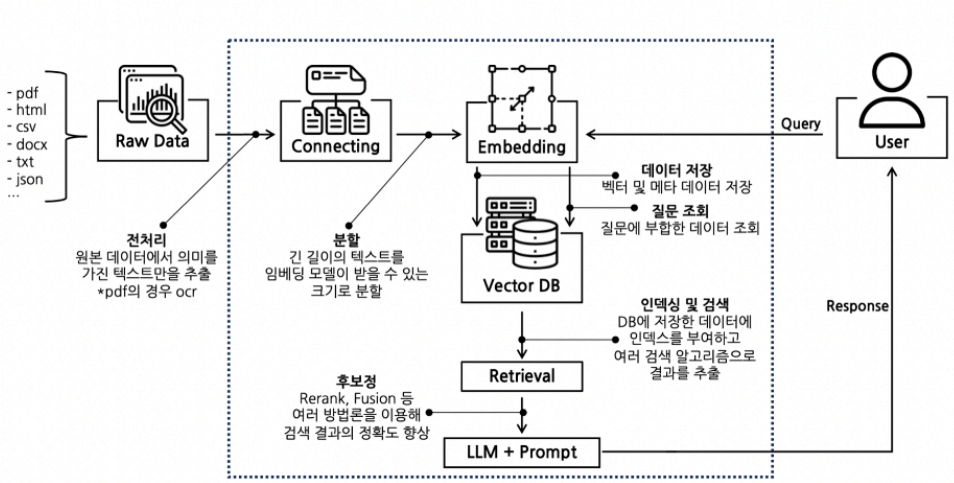


5.3.2. RAG(Retrieval Augmented Generation)

검색 증강 생성(RAG)은 프라이빗 또는 독점 데이터 소스의 정보로 텍스트 생성을 보완하는 기술이다. 대규모 데이터 세트 또는 지식 기반을 검색하도록 설계된 검색 모델에 해당 정보를 가져와 읽을 수 있는 텍스트 응답을 생성하는 대규모 언어 모델(LLM)과 같은 생성 모델을 결합한다.

검색 증강 생성은 추가 데이터 소스의 컨텍스트를 더하고 훈련을 통해 LLM의 원래 지식 기반을 보완함으로써 검색 경험의 정확도를 개선할 수 있다. 따라서 모델을 다시 훈련할 필요 없이 대규모 언어 모델의 출력이 향상되고, 보다 특화된다.

RAG는 생성형 AI 시스템이 외부 정보 소스를 사용하여 보다 정확한 상황 인식 응답을 생성할 수 있도록 해주기 때문에 질문 답변 및 콘텐츠 생성과 같은 작업에 유용한데, 현재 제작할 문제 생성 AI에는 기존 기출 문제 데이터셋을 기반으로, 시맨틱 검색이나 벡터 유사도 검색과 같은 검색 방법을 구현하여 문제 생성에 대한 사용자 의도에 응답하고 보다 신뢰성 있는 결과를 제공하는 것이 목표이다.



5.3.3. ChatGPT-4-Turbo

챗GPT 3.5(ChatGPT)'가 2022년 11월 30일에 처음 세상에 공개된 이후, 23년 11월에 공개된 모델이다. 23년까지의 데이터를 추가 학습해 3.5가 21년까지의 데이터만을 가지고 훈련된 한계를 극복하고 기존 토큰 4096개 대비 3만2768개 토큰을 컨텍스트로 기억할 수 있으며, 가격이 2.75배로 대폭 인하되었다. 또한 각계의 전문가를 통한 학습데이터 검수 및 증가로 더욱 정확한 답변을 생성한다. 그럼에도, GPT-4대비 3.5차이는 20배로 많은 비용이 발생하기에, 본 과제에서는 입출력 토큰이 많이 발생하는 생성 모델에 3.5-turbo를 활용하고 이를 검증하고 보완하는 모델로 GPT-4-Turbo를 도입하여 비용 최적화와 정확도 향상을 목표로 한다.

6. 일정 및 역할 분담

6.1. 개발 일정

6.2. 역할 분담

이름	역할
김범수	Spring, 메인 백엔드 서버 개발 및 인프라 구성, 메인 프론트 개발
신병근	문제 생성 AI 개발 및 PM, 문제 관련 프론트 개발
허진영	데이터 구축 및 문제 생성 연구, 데이터 관련 FastAPI 서버 개발

7. 참고 문헌

- K컬처 인기에 한국어능력시험 대약진…日 JLPT 맹추격 - 매일경제 (mk.co.kr)
- ‘2019~2023 한국어능력시험 시행 현황’, 교육부 국립국제교육원
- ‘한국어능력시험(TOPIK) IBT 시스템(3단계)결과분석 및 지능형 평가 플랫폼 구축 제안요청서’, 교육부 국립국제교육원
- Qiu et al. (2022). Automatic generation of multiple choice questions for medical education using a large language model.
- Yu et al. (2022). Automatic Generation of High-Quality Math Word Problems with Iterative Refinement by Large Language Models.
- Matteson et al. (2022). Can GPT-3 Generate Novel, Non-Trivial and Diverse Organic Chemistry Problems? CopyRetryClaude can make mistakes. Please double-check responses.