영어 객관식 문제 자동 정답 생성기 개발

34팀 Re:Fresh 강형원 박규태 주연학

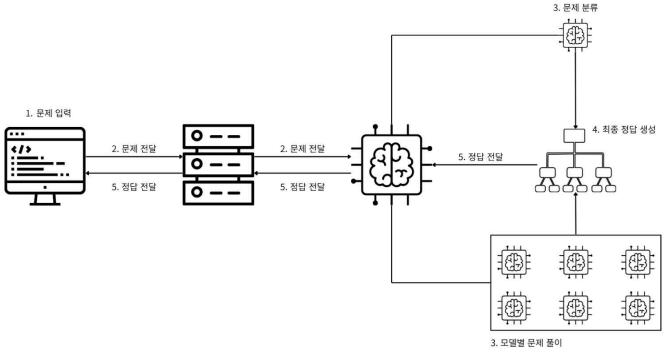
연구 배경

- 현대인의 삶에 많이 녹아 들게된 AI, 그리고 그 대표인 LLM
- 앞으로 LLM을 활용하는 능력은 현대인에게 필수 능력이 될 것이며, 이를 활용하여 문제를 더 정확하게 해결할 수 있는 서비스를 만드는 것은 중요한 기술이 될 것
- 본 연구는 일상 생활에서 마주칠 수 있는 문제점 중 하나인 “영어 객관식 문제“ 풀이에 특화된 AI 모델을 제안하고자 함

LLM

- 언어 생성 AI 모델로, 우리가 흔히 접하는 ChatGPT가 대표적인 LLM
- 사전에 학습되어 있던 데이터의 크기에 따라 기본 모델의 정확도 및 성능이 달라지게 됨
  - 사전학습된 데이터의 양이 작으면 기본 정확도는 낮아지지만 이를 개량하는 속도는 빠르며, 이식성이 뛰어남
  - 사전학습된 데이터의 양이 크면 기본 정확도가 높아지지만 학습 비용이 비싸지며, 모델을 PC에 올려놓는 비용도 비싸 이식성이 떨어짐
- LLM을 활용한 서비스를 개발할 때는 현재 개발이 진행되는 환경과 필요한 정확도, 사용될 환경 등을 고려하여 모델을 선정할 필요가 있음

전체 시스템 구성도


1. 문제 입력
2. 문제 전달
5. 정답 전달
2. 문제 전달
5. 정답 전달
3. 문제 분류
5. 정답 전달
4. 최종 정답 생성
3. 모델별 문제 풀이

데이터셋 선정

- RACE
  - 28000 여개의 지문과 10만여개의 질문으로 구성된 규모가 큰 영어 객관식 문제 데이터셋
- CLOTH
  - 영어 빈칸 문제들에 대해 제공하는 데이터셋
- 교육청에서 제공하는 한국 수능/검정고시 문제들

- 위의 3개를 적절히 혼합하여 데이터셋을 구축함

Base Model 선정

- 파인튜닝이 가능한 HuggingFace 오픈 웨이트 공개 모델 중 파라미터 4B 이하인 모델을 1차적으로 선택
- 각 모델의 기술 문서를 참고

- 최종 Google의 Gemma3-4B-it 모델과 Microsoft의 Phi-4-Mini-instruct를 최종 모델로 선택

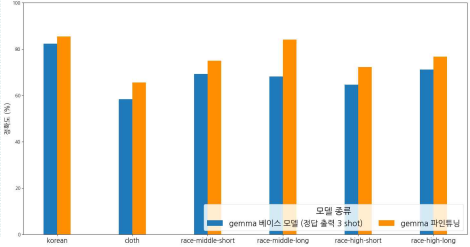
Parameter Effective Finetuning

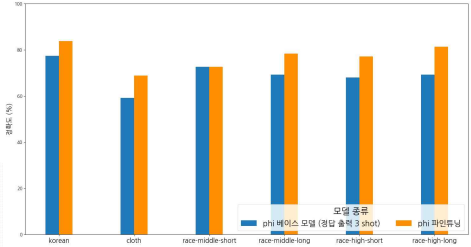
- 사전 학습 모델을 특정 도메인에 맞게 성능 향상시키기 위해 미세조정을 수행
- 전체 파라미터를 업데이트하는 Full fine-tuning은 연산, 메모리 비용이 매우 커 실용성이 떨어짐
- 이에 따라 매개변수 효율 미세조정 (PEFT)을 적용해 적은 비용으로 파인튜닝을 진행

- Low-Rank Adaptation(LoRA)은 PEFT 기법 중 하나로 기존 베이스 모델의 가중치를 동결 후 별도로 구분된 어댑터를 학습하는 방식
- 모델 추론 시에는 기존 모델의 출력과 어댑터의 출력을 더해 미세조정의 효과를 냄
- LoRA 기법을 통해 Full fine-tuning에 비해 최대 3배 까지 하드웨어 성능 요구를 낮출 수 있음

파인튜닝 모델

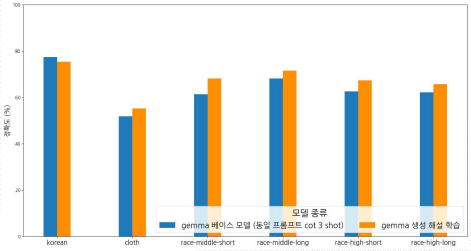
- 모델에 데이터셋의 정보를 바탕으로 문제, 지문, 선지와 정답을 하나의 프롬프트로 가공 후 학습 데이터로 제공
- 정답에는 별도의 해설 데이터를 제공하지 않고 선지 문자 하나만을 제공

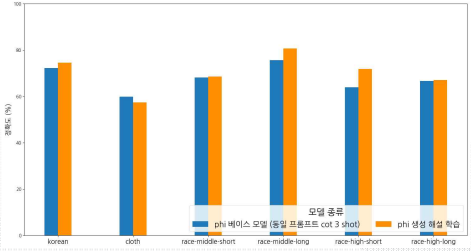




모델 데이터 생성

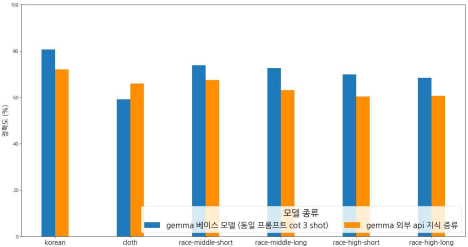
- STaR(Self-Taught Reasoner)는 베이스 모델의 추론을 통해 문제의 해설을 생성후 정답 여부를 필터링
- 모델이 맞추지 못한 문제의 경우 모델에 정답 힌트를 주고 다시 해설을 생성하게 하는 합리화 과정을 통해 데이터 셋에 대한 해설을 다시 생성
- 이렇게 만들어진 해설 포함 데이터로 모델을 학습 시키고, 이 과정을 반복

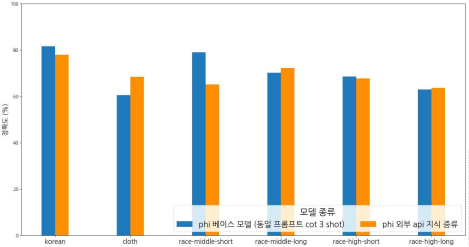




지식 증류

- 큰 규모의 Teacher 역할의 모델 응답 데이터를 바탕으로 작은 규모의 모델인 Student를 파인튜닝해 Teacher에 비해 빠르고 경제적이지만 기존 소형 모델에 비해 높은 성능을 지니는 모델을 제작
- Teacher 역할로 외부 API인 Gemini API를 통해 문제의 해설을 생성해 Student의 학습에 활용





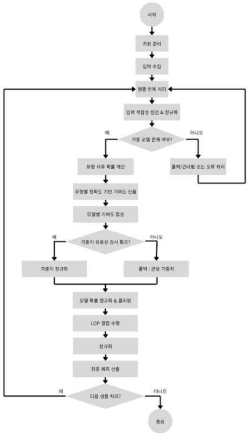
Ensemble

- 단일 모델은 주어진 데이터와 학습 과정에서 형성된 특정 패턴에만 집중하기 쉬우며, 그 결과 편향과 분산으로 인해 정답 예측에 한계가 있음
  - Ensemble은 이러한 한계를 보완하기 위해 여러 모델을 결합하는 방법
  - 이렇게 하면 각 모델의 강점은 살리고 약점은 상호 보완 가능
- 대표 방식으로 Bagging, Boosting, Stacking, Voting이 있으며, 본 연구는 Voting을 중심으로 논의
  - 여러 학생의 답안을 종합하면 각자의 강약이 보완되어 단일 답안보다 정답에 가까워질 수 있음

LOP

- Ensemble 기법을 쓰기 위해서는 각 모델이 도출한 결과를 하나로 합쳐야 함
  - 본 연구에서는 합치는 기법 중 'LOP' 를 사용
- **각 분포의 로그를 가중합해 지수화하는, 가중 기하평균 형태의 결합**
- 모델마다 분류한 문제 유형에 따라 성능이 다르다는 점에 주목
- 문제 유형을 분류했다면 적절한 가중치를 구해야 함
  - 사전에 확인한 각 모델의 유형별 정확도를 기준으로 삼음

LOP



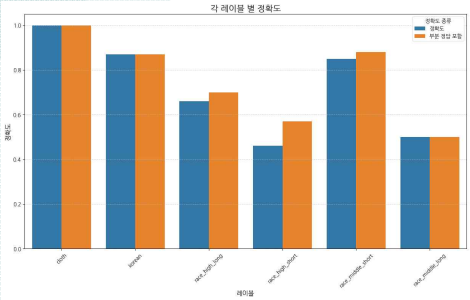
▶ 조건부 가중치(게이팅) 산출

- 각 기본 모델 i와 문항 유형 t에 대한 사전 측정 정확도를 활용해, 해당 문항에 사용할 가중치를 계산

▶ LOP로 확률 결합

- 각 모델이 산출한 선지 분포를 입력으로 받아, LOP로 최종 분포를 계산
- 일부 모델이 실패하면 제외하고 남은 모델 가중치를 재정규화

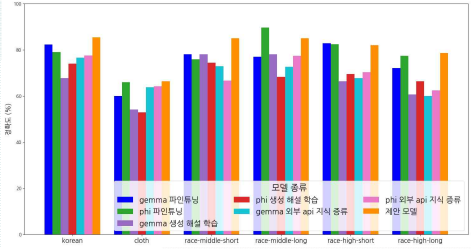
LOP



▶ 분류 모델의 경우

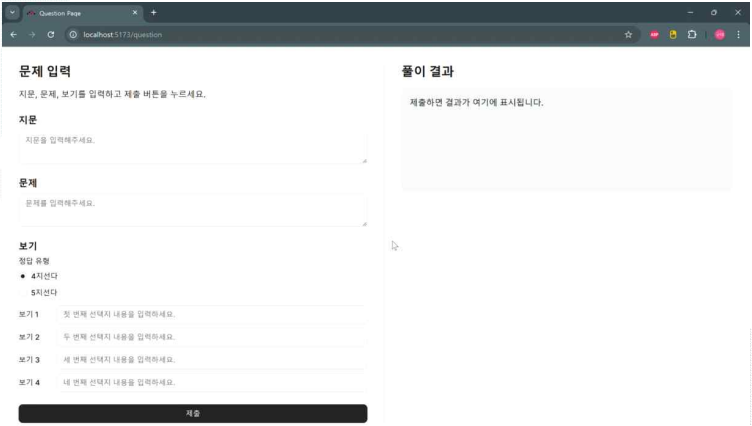
- 600개의 데이터셋을 사용했으며, 이 중 434개를 올바르게 분류해 약 72.33%의 정확도를 기록
- 기준을 좀 더 완화해, race라는 대분류와 길이(long, short)까지는 맞은 것을 부분 정답으로 인정한다면 정확도는 약 75.33%로 상승

최종 결과



- 최고 우측이 앙상블을 통해 제작한 최종 결과물
- 최종 정확도가 사용한 모델에 비해 약간 높거나 그에 수렴하는 형태로 나타남
- 일부 영역에서는 더 뛰어난 단일 모델이 있었는데, 이는 모델별 편차 차이가 커서 발생한 것
- 가중치 부여 시 정확도가 낮은 모델의 가중치를 더 낮추거나 아예 배제한다면 더 좋은 정확도를 낼 수 있을 것으로 기대

시연


Question Page
localhost:5173/question
문제 입력
지문, 문제, 보기를 입력하고 제출 버튼을 누르세요.
지문
지문을 입력해주세요.
문제
문제를 입력해주세요.
보기
정답 유형
4지선다
5지선다
보기 1
첫 번째 선택지 내용을 입력하세요.
보기 2
두 번째 선택지 내용을 입력하세요.
보기 3
세 번째 선택지 내용을 입력하세요.
보기 4
네 번째 선택지 내용을 입력하세요.
제출
풀이 결과
제출하면 결과가 여기에 표시됩니다.

결론

- 모델 개발 및 정확도를 올리기 위한 다양한 방법을 적용하였고, 베이스 모델에 비해 문제를 좀 더 잘 풀어내는 모델을 개발하여 문제 풀이에 접목
- 모델의 성능을 향상 시키는 방법은 다양함
  - 다양한 방법을 추가 접목 시켜 성능을 더 높일 수도 있음
  - 분류 모델에 있어서도 더 정확한 분류를 기대할 수 있을 듯함