

영어 객관식 문제 풀이 서비스 개발



부산대학교

정보컴퓨터공학부

지도교수

탁성우

팀명

Re:Fresh

팀원

202055509 강형원

202055538 박규태

202055605 주연학

목차

1. 서론.....	3
- 요구조건 및 제약사항 분석.....	3
2. 본론.....	4
- 설계 상세화 및 변경 사항.....	4
3. 결론.....	12
- 모델 간 성능 검증.....	12
- 갱신된 과제 추진 계획.....	16
- 구성원 별 진척도.....	16

1.서론

- 요구조건 및 제약 사항 분석

- 요구조건

1. 영어 객관식 문제를 풀 수 있는 서비스 개발
2. 객관식 문제의 경우 주제문 찾기, 올바른 순서로 배치하기 등의 문맥 파악류의 문제풀이를 요구

- 제약 사항

1. GPT에 문제를 일임하는 것은 불가능
2. 적절한 모델을 구하여 학습을 시킨 후 사용하는 것은 가능

2.본론

- 설계 상세화 및 변경 사항

1. 최초 계획은 문제 유형 별로 모델을 두어 유형별 학습을 통해 모델 상세화

- 계획 수립 후 파인 튜닝을 위한 충분한 데이터셋을 확보하는 과정에서 저작권 등과 관련한 이유로 충분한 양의 데이터셋을 확보하는 것에 어려움을 겪음

=> 따라서, 오픈 데이터셋인 **RACE, CLOTH**를 함께 활용하는 방식을 선택

- 다만 해당 데이터셋들도 제한적인 문제 유형을 가지고 있어, 여러 유형의 문제에 대해 충분한 데이터 셋을 확보하는데 어려움이 있음.
- 즉, 데이터셋의 편향 문제가 발생함.





=> 해당 문제들을 최대한 해소하기 위해, 데이터셋을 종합한 후 해당 데이터셋들에 샘플링을 하여 데이터셋을 선정해 **training, validation, test**로 구분하여 사용

- 즉, 모델 단일화를 하여 하나의 모델에 여러 문제 유형들을 최대한 골고루 학습시켜 사용하는 방안으로 변경

1. 적절한 모델 탐색

- 모델은 llm 벤치마크 리더보드 사이트인 **llm-stats.com** 의 벤치마크 중 문맥파악 성능을 나타내는 벤치마크를 기준으로 선정

llm-stats.com ← Scroll horizontally to see all columns →

Organization ↕	Model ↕	License ↕	Parameters (B) ↕	GPQA ↕	MMLU Pro ↕
	Phi 4 Mini Reasoning	Open	3.8	52.0%	-
	Llama 3.2 3B Instruct	Open	3.2	32.8%	-
	Gemma 3 4B	Open	4	30.8%	43.6%
	Phi-3.5-mini-instruct	Open	3.8	30.4%	47.4%

- GPQA¹: 과학 분야의 박사급 전문가들이 직접 출제한, 구글 검색으로도 쉽게 답을 찾을 수 없는 고난도 객관식 문제로 구성된 벤치마크
- MMLU Pro²: 다양한 분야의 고난도 문제를 바탕으로 **LLM**이 실제로 얼마나 깊이 있는 이해와 복잡한 추론을 할 수 있는지 평가하는 벤치마크.
- GPQA, MMLU Pro 두가지 점수를 종합적으로 고려하여 5B 이하의 소형 Language Model 중 가장 괜찮다고 판단된 **Gemma 3 4B** 모델을 베이스 모델로 선택

¹ <https://arxiv.org/abs/2311.12022>

² <https://arxiv.org/abs/2406.01574>

2. 데이터 탐색 및 가공

- 종류 조사
 - a. 중고등학생 수준의 영어 객관식 문제들에 대한 내용을 가지고 있는 **RACE dataset**
 - b. 영어 빈칸 문제에 대한 내용을 가지고 있는 **CLOTH dataset**
 - c. 한국 평가원에서 제작한 모의고사 / 수능 문제
 - d. 한국 검정고시 영어 문제
- HuggingFace를 통해 구할 수 있는 **ehovy/race³, AndyChiang/cloth⁴** dataset을 적절하게 가공하여 사용
- 모의고사와 수능의 경우 평가원에서 공개한 실제 문제를 읽어와 데이터셋 형태로 가공하여 사용
- 수가 적은 한국 영어 시험 문제를 고려해 결정한 최종 학습 데이터셋 크기와 비율
 - 한국 고등, 중등 검정고시, 수능, 모의고사 900문제
 - **race** 중등 문제 3500개
 - **race** 고등 문제 10500개
 - **cloth** 빈칸 추론 문제 5100개
 - 총 20000개 문제를 학습 데이터셋으로 사용
- 통일성을 위해 **cloth** 데이터와 한국 영어문제의 데이터들을 **race** 데이터셋의 형태로 가공
 - 핵심 요소는 article, question, options, answer로 각각 지문, 문제, 보기, 정답을 나타냄
 - **example_id**는 본래 문제가 몇년도에 나온 몇번문제인지 등을 기록하는 보기이나, 학습 과정에서는 사용되지 않기에 임의의 정수로 통일
 - **AndyChiang/cloth dataset**의 경우 ['distractors', 'sentence', 'answer']의 구분으로 나뉘어져있고, 각각 “답이 아닌 단어”, “빈칸이 있는 문장”, “들어갈 단어”를 의미
 - **distractor**와 **answer**를 합하여 **options**를 구성하고 **sentence**를 **article**로 한 후, 문제를 “Which word best fits in the blank?”로 하여 **race**와 형식을 통일
 - 문제의 경우 한국 문제에 대한 대응을 키우기 위해 30% 정도의 문제는 “빈칸에 들어갈 말로 적절한 것은?”, “빈칸에 들어갈 적절한 말을 고르시오.”, “다음 중 빈칸에 들어가기 가장 적절한 말은?” 중 하나로 설정
- 데이터셋은 로컬에 **HuggingFace DatasetDict** 형식의 폴더에 저장
- 추후 **local**에서 **load_from_disk** 메서드를 활용하여 불러올 수 있다.

³ <https://huggingface.co/datasets/ehovy/race>

⁴ <https://huggingface.co/datasets/AndyChiang/cloth>

가공된 dataset 예시

```
{
  'example_id': 0,

  'article': '"I planted a seed. Finally grow fruits. Today is a great
day. Pick off the star for you. Pick off the moon for you. Let it rise
for you every day. Become candles burning myself. Just light you up,
hey!... You are my little little apple. How much I love you, still no
enough."\nThis words are from the popular song You Are My Little Dear
Apple. Bae Seul-Ki acted as the leading dancer in the MV of the song.
She loves dancing. She became crazy about hip-hop when she was a school
girl.\nBai Seul-Ki was born on September 27, 1986. She is a South Korean
singer and dancer. She is 168cm tall. She loves cooking. Her favourite
food is spicy and salty. She like pink and red most. There are five
members in her family---father, mother, two younger brothers and
herself. She isn\'t married.\nAfter her father and mother broke up, she
lived with her mother and new daddy. She enjoys being alone.',

  'question': 'Bae Seul-Ki _ in the MV of the song according to the
passage.',

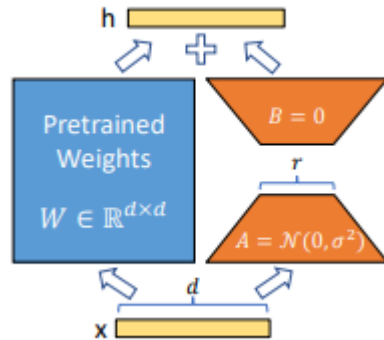
  'options': ['sang', 'danced', 'cried', 'laughed'],

  'answer': 'B'
}
```

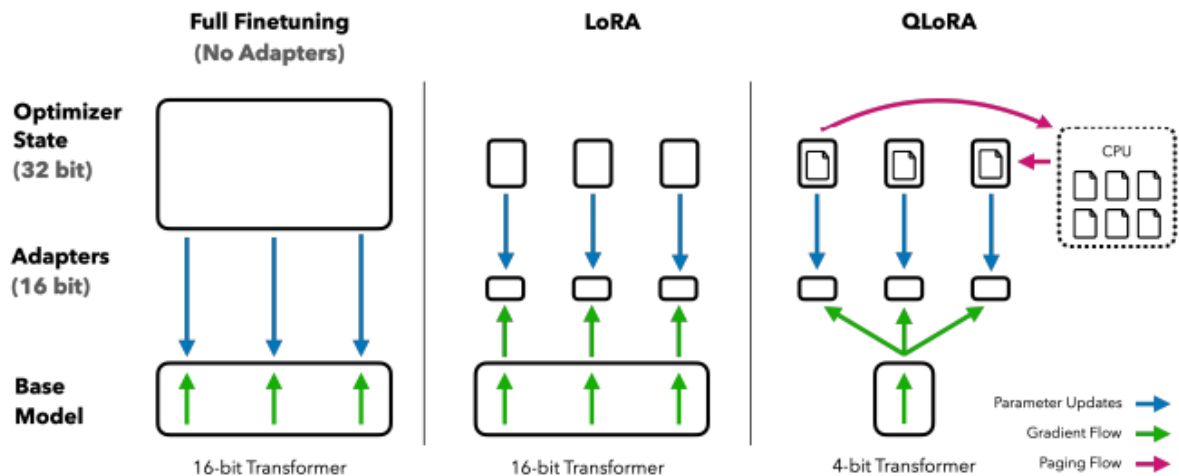
3. 모델 파인 튜닝

- 파인 튜닝 기법

- **Low-Rank Adaptation** 파인 튜닝: llm의 파인튜닝을 효율적으로 수행하기 위한 기법으로 전체 사전학습 가중치를 학습하는 **full fine-tuning**과 달리 저랭크 행렬 쌍 학습한다. **LoRA**는 **full fine-tuning**에 비해 최대 3배까지 하드웨어 성능 요구를 낮출 수 있다⁵.



- **QLoRA**: LoRA에 양자화 기법을 적용시켜 메모리 사용을 줄이는 기법이다. 베이스 모델의 가중치를 그대로 사용하는 **LoRA**와 달리 베이스 모델의 가중치를 4비트 **NF4**로 양자화해 메모리 요구량을 줄인다⁶.



⁵ <https://arxiv.org/pdf/2106.09685>

⁶ <https://arxiv.org/pdf/2305.14314>

- 데이터 셋 가공
 - 데이터를 학습시키기 위해 다음과 같은 형식의 프롬프트로 가공

<bos><start_of_turn>user

You are a helpful AI assistant. Please answer the user's questions correctly. Look for the evidence in the text when answering. Underlined replaced with highlights. Example: **Was Underline**

[QUESTION]

Are Jim and Ann in the same school? ---- _ .

[PASSAGE]

This is a teacher's family. The father's name is Lake Smith. He's forty - four. The mother's name is Kate Green. She's forty - two. The Smiths have a son, Jim, and a daughter, Ann. Jim is fourteen, and Ann is twelve. The son looks like his father, and the daughter looks like her mother. They are all in No.4 Middle School. But the Smiths are teachers; the son and daughter are students.

[OPTIONS]

- A. They are not at school
- B. They are in different schools
- C. Yes, they are
- D. No, they aren't

[ANSWER]<end_of_turn>

<start_of_turn>model

C<end_of_turn>

-

- 초기 하이퍼 파라미터 설정
 - google ai studio의 gemma 3 파인튜닝 튜토리얼⁷을 따르나, **LoRA alpha**의 경우 적은 데이터셋에 대해 모델이 좀 더 새로운 데이터를 받아들이도록 16에서 32로 상향 조정
 - **learning rate**는 QLoRA 학습에서 **2e-4**가 제안되나⁸ 낮은 값에서 높은 값까지 비교를 위해 2e-5로 조정
- LoRA r = 16
- LoRA alpha = 32
- LoRA dropout = 0.05
- batch_size = 2
- gradient_accumulation_steps = 4
- Learning Rate = 2e-5
- Warmup ratio = 0.03

⁷ https://ai.google.dev/gemma/docs/core/huggingface_text_finetune_glora?hl=ko

⁸ <https://arxiv.org/pdf/2305.14314>

4. 모델 검증

- 테스트 데이터 셋을 사용, 각각의 모델에서 추론 정답과 각 선택지 별 예상 확률을 기억
- epoch를 다르게 학습한 모델을 비교
- 각 모델에 대해, F-1 점수를 구함
F-1 점수 계산에는, `sklearn.metrics` 라이브러리의 `f1_score` 함수를 사용

```
# model_results 딕셔너리의 각 모델에 대해 F1 점수를 계산합니다.  
for model_name in model_results.keys():  
    # one-hot 인코딩된 정답 리스트를 레이블 인코딩으로 변환합니다.  
    y_true = np.argmax(answer_list[model_name], axis=1)  
  
    # 확률 리스트에서 가장 높은 확률의 인덱스를 예측값으로 사용합니다.  
    y_pred = np.argmax(probs_list[model_name], axis=1)  
  
    # 가중 F1 점수를 계산합니다.  
    f1_weighted = f1_score(y_true, y_pred, average='weighted')  
  
    # 계산된 F1 점수를 딕셔너리에 저장합니다.  
    f1_scores[model_name] = f1_weighted
```

- F-1 점수를 `matplotlib` 라이브러리를 사용해 그래프로 나타냄

3. 결론

- 모델 간 성능 검증

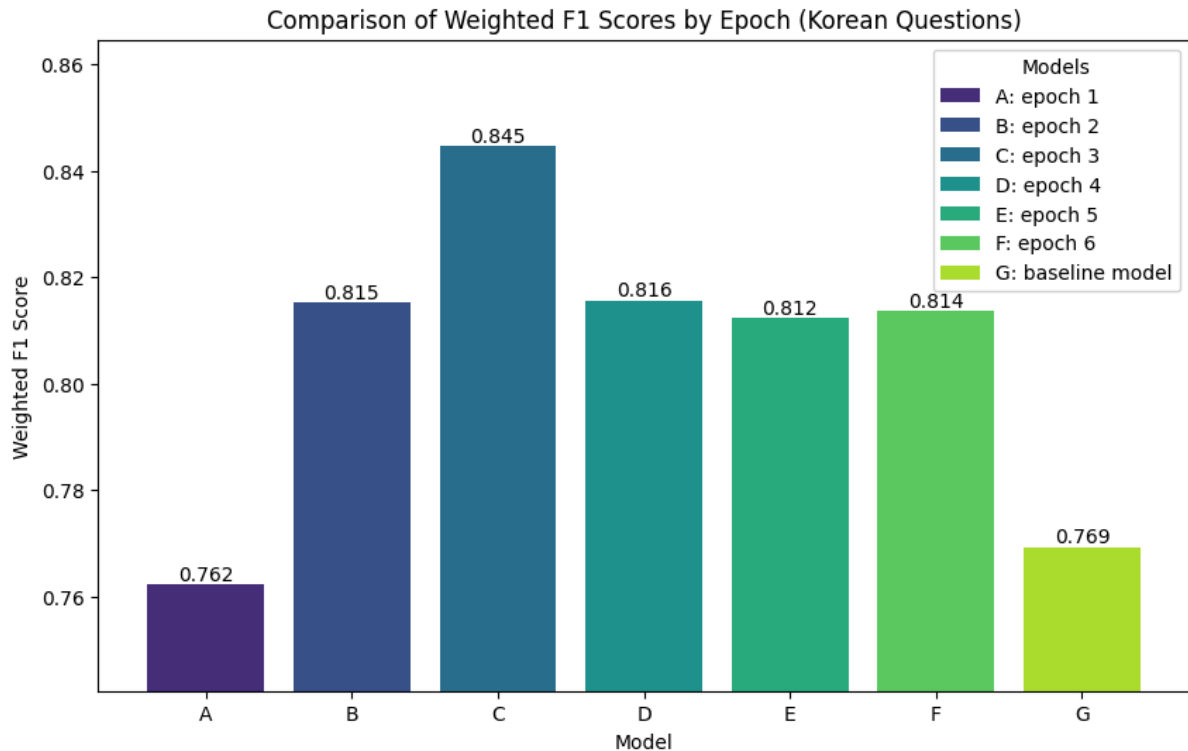
측정 데이터셋

2가지 유형에 대해 테스트를 진행

	Full Dataset	Korean Questions
수능 + 검정고시	20	100
race 중학생용 문제	78	-
race 고등학생용 문제	233	-
cloth 문제	113	-
총합계	444	100

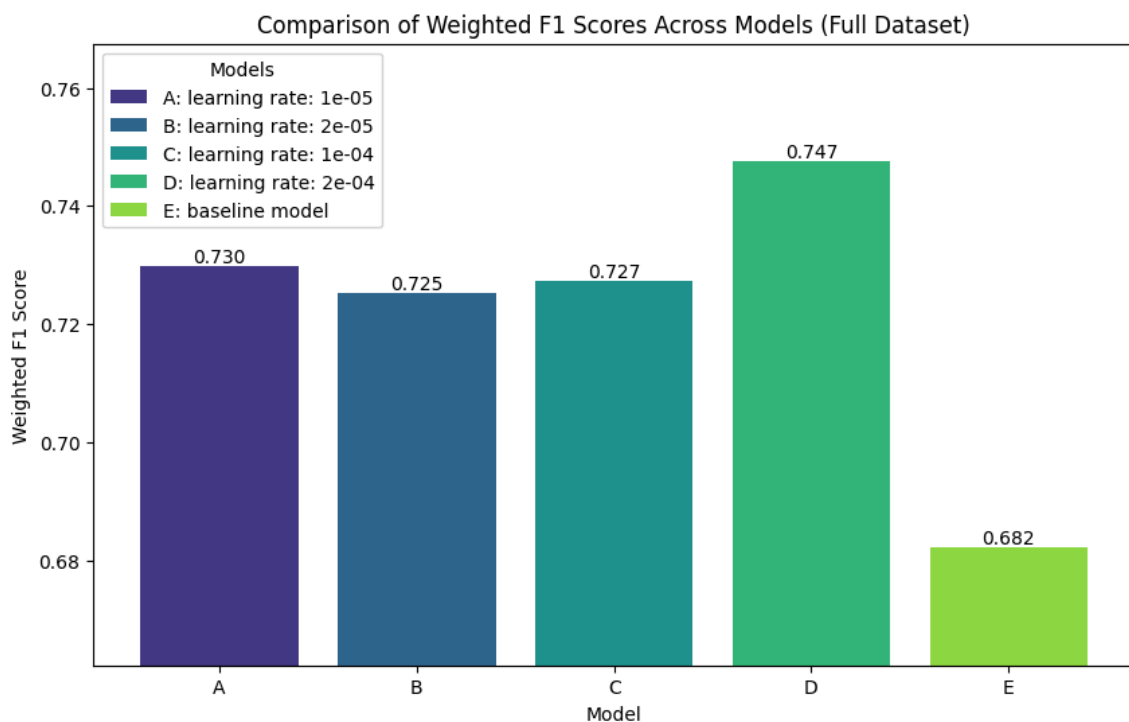
1. 에포크 증가에 따른 모델 f1 score 비교

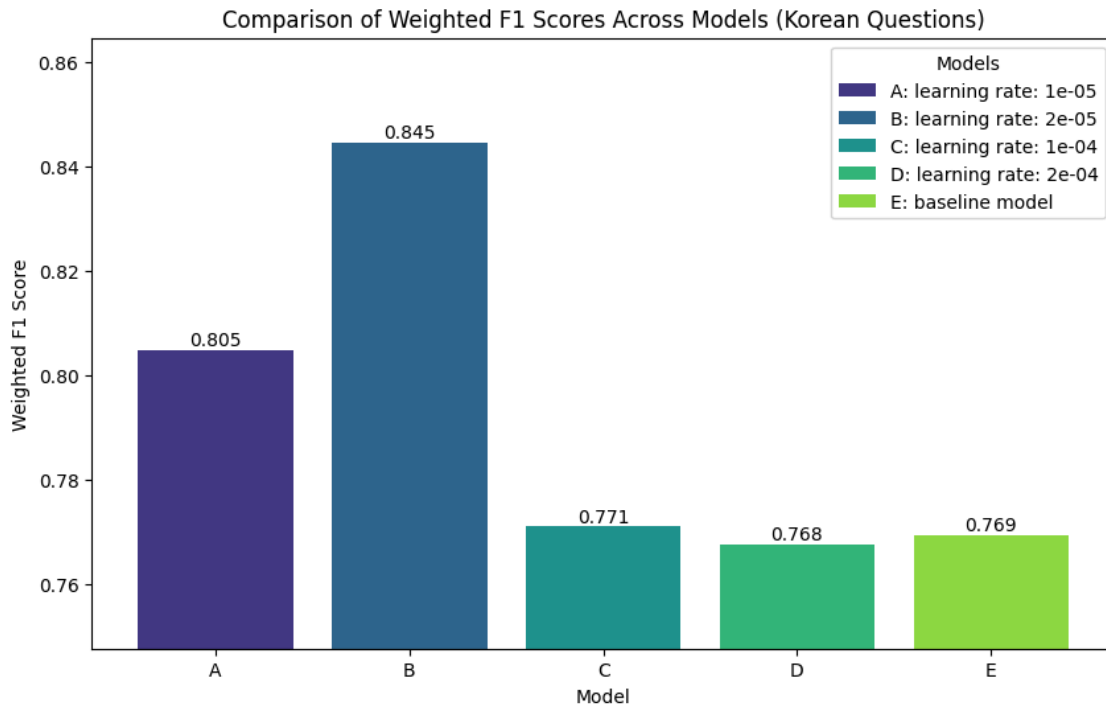




- epoch가 커지면 커질수록 모델의 F1 score도 높아지는 경향을 보이지만, 2~3 이후로는 그렇게 큰 점수 상승의 폭을 보이지는 않는다. epoch가 증가할 때마다 학습 시간이 증가하는 것에 비해 유의미한 성능 향상이 보이지 않았기 때문에, 추후 모델 개선 시 epoch 횟수를 3으로 두고 계속해서 모델 개선을 진행할 예정이다.

2. Learning Rate 별 모델 f1 score 비교



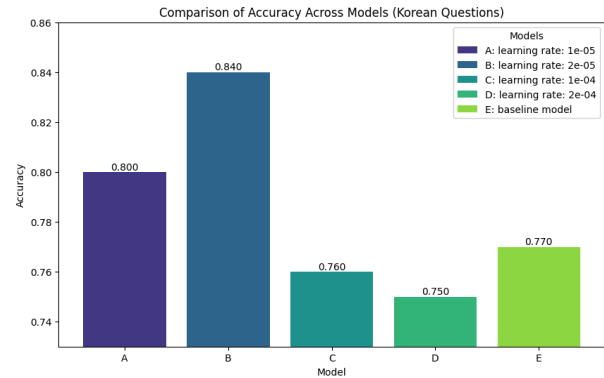
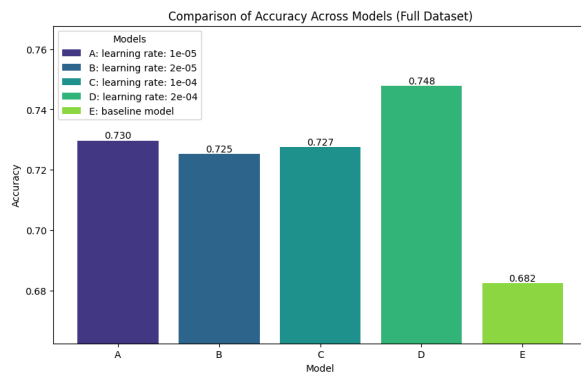
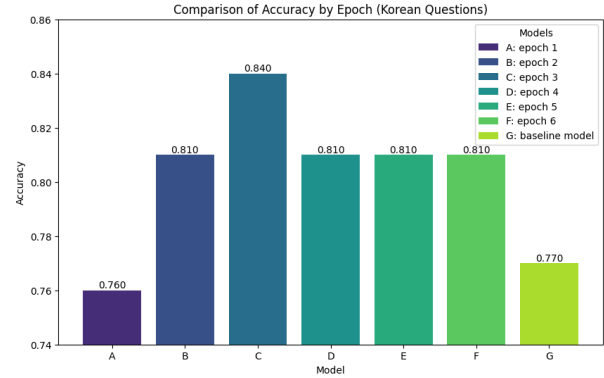
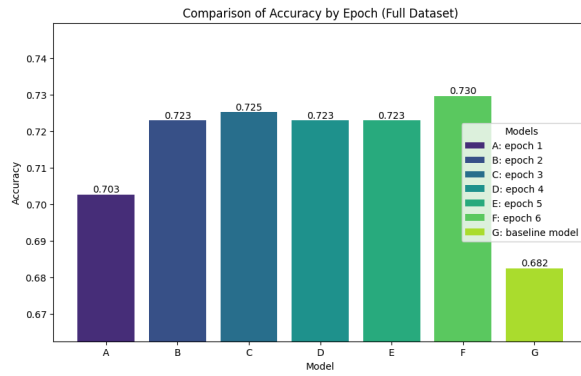


learning rate의 비율을 다르게 하여 각 데이터셋에 대해 f1 score를 평가한 그래프이다. full dataset의 경우 추천 학습률⁹인 2e-04에서 가장 좋은 점수를 기록하였으며, korean questions의 경우 2e-05에서 가장 좋은 점수를 기록하였다. 따라서 추후 모델 개선 시 2e-05 ~ 2e-04 사이의 학습률 사이에서 조정하여 모델을 개선할 계획이다.

⁹ <https://arxiv.org/pdf/2305.14314>

3. model 간 정확도 비교

- 정확도를 “모델이 생성한 답이 문제의 답과 동일한 비율”로 정의하였다.
- 정확도의 경우 **F1-score**와 거의 비슷하게 나오는 것을 볼 수 있다.
- 즉, 학습을 한 모델은 학습을 안 한 모델에 비해 정답을 더 잘 맞힌다고 할 수 있다.



- 갱신된 과제 추진 계획
- 모델에 대한 보완
 - 계속해서 파라미터 및 데이터셋을 조정하여 최종 제출 기간 내에 만들 수 있는 최적의 모델 학습
- 모델을 활용한 웹 서비스 개발
 - 모델을 서비스에 직접 탑재하거나, 모델을 포함하는 이미지를 서버에 올린 후 통신하는 방식 등 “개발한 모델을 활용하여 문제를 풀이할 수 있는 서비스” 개발
- 구성원별 진척도

강형원

- 데이터셋 정제 및 병합 (CLOTH / RACE / 수능 / 모의고사 / 검정고시)
- 웹 서비스 서버 개발 예정

박규태

- 모델 테스트 코드 작성
- 웹 서비스 클라이언트 개발 예정

주연학

- 모델 파인튜닝 코드 작성, 프롬프트 작성
- 모델 학습 진행
- 모델 추가 학습 및 보완 예정