Wrap-up report

1. 프로젝트 개요

1.1 프로젝트 개요

본 프로젝트는 수능형 국어·사회 지문 기반 객관식 문제를 해결하는 대회 과제로, 입력(지문/문항/선택지)에 대해 정답 선택지 번호를 예측하는 모델을 구축하였다. GPT·Claude 등 대형 모델 대비 상대적으로 작은 규모의 모델로도 경쟁력 있는 성능을 달성하는 것을 목표로 하였다.

EDA를 통해 4/5지선다 문항이 지문 길이와 정답 분포 등에서 서로 다른 특성을 갖는 것을 기반으로 유형 분리 전략을 적용하였다. 이후 외부 데이터 증강, 프롬프트 최적화, Qwen3-14B와 A.X-4.0 Light의 soft voting 앙상블, 그리고 마진 기반 저신뢰도 문항에 대한 Critic 재추론을 결합하여 최종 추론 파이프라인을 구성하였다.

1.2 개발 환경

Programming Language	Python 3.10.13
Experiment Tracking	Weights & Biases
Version Control	GitHub
Collaboration Tool	Slack
H/W & OS	NVIDIA Tesla V100 (32GB), Ubuntu 20.04 LTS

2. 프로젝트 팀 구성 및 역할

김영현

- 데이터 증강, 프롬프트 엔지니어링

윤준상

- 데이터 분석, 데이터 증강, 프롬프트 엔지니어링

장세현

- 데이터 분석, 전체 파이프라인 설계 및 구현, DPO 실험, 추론 로직 구현

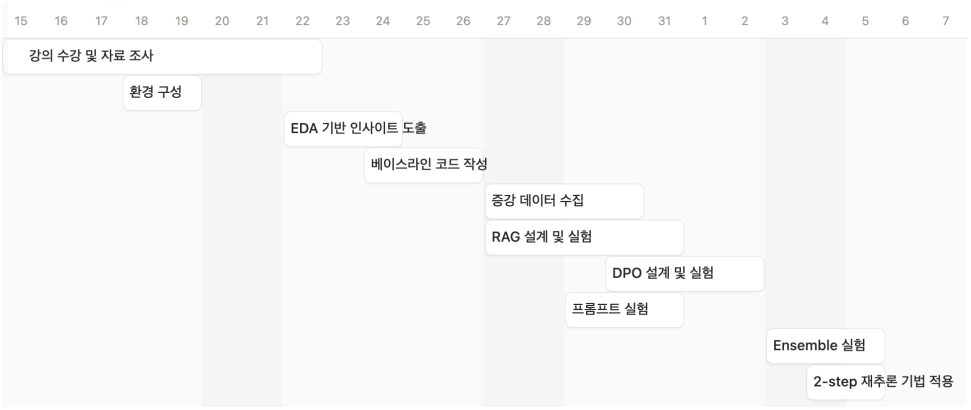
주현민

- 데이터 증강, 프롬프트 엔지니어링, EDA

한지석

- 디렉토리 구조 설계, 앙상블 구현, 추론 성능 실험

3. 프로젝트 수행 절차 및 방법



4. 프로젝트 수행 결과

4.1 데이터 분석 (EDA)

4.1.1 데이터셋 개요

본 프로젝트는 수능형 문제 해결을 위해 총 2,031개의 학습 데이터(Train)를 활용하였다. 데이터는 id, paragraph, question, choices, question_plus, answer 컬럼으로 구성되며, 국어·사회·역사 등 다양한 도메인을 포함한다. 또한 KMMLU, MMMLU, KLUE MRC 등 서로 다른 원천 데이터가 통합되어 구축된 형태로, 문항 유형과 입력 길이에서 이질성이 존재할 수 있음을 전제로 분석을 진행하였다.

- 데이터 구성의 다양성

  원본 데이터 구성과 학습 데이터를 확인한 결과, 문항에 따라 선택지 개수(4지선다/5지선다)가 혼재되어 있음을 확인하였다.

- 전처리

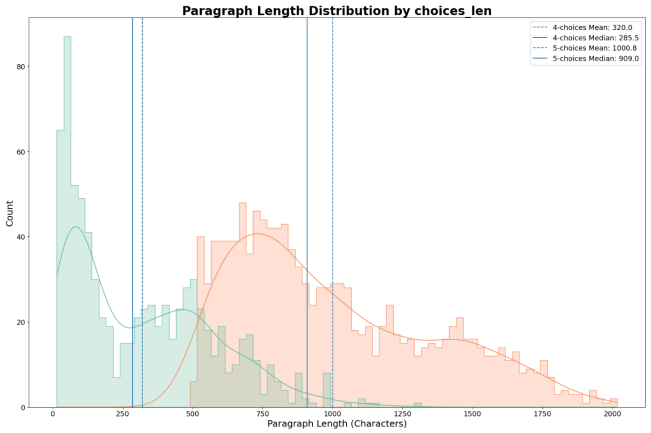
  question plus는 Train 데이터에서 2,031개 전체 행에서 결측치였으나, Test 데이터에는 해당 컬럼에 값이 존재함을 확인하였다. 따라서 학습 단계에서 해당 컬럼을 유지했다.

4.1.2 선택지 개수(4/5지 선다)에 따른 길이 분석

학습 데이터의 paragraph에 대해 문자 길이 분포를 확인한 결과, 전체 분포가 bimodal 형태로 나타났다. 선택지 개수 기준으로 분리해 확인한 결과, 4지선다는 지문 길이 중앙값이 285자(평균 320자)인 반면, 5지선다는 중앙값이 909자(평균 1000자)로 약 3배 길었다.

이는 선택지 개수가 단순 포맷 차이가 아니라 원천 데이터 및 문항 유형(지문 길이·요구 능력)의 차이를 반영하는 신호일 가능성을 시사한다. 따라서 이후 실험에서는 선택지 개수를 기준으로 문항을 구분하여 프롬프트 구

성 전략을 분리 적용하였다.



4.1.3 정답 분포 분석

데이터의 선택지 개수에 따라 정답(answer) 분포를 분석한 결과, 5지선다 문항에서 정답이 1번에 크게 편향되어 있음을 확인했다. 반면, 4지선다 문항은 1~4번 정답이 상대적으로 고르게 분포하였다.

구분	정답 1	정답 2	정답 3	정답 4	정답 5
4지 선다	183	185	203	221	-
5지 선다	**765**	265	116	62	31

4.2 데이터 증강 (Data Augmentation)

학습 데이터(2,031개)의 부족한 도메인 다양성을 확보하고, 문항 유형(4/5지 선다)에 따른 모델의 추론 능력을 강화하기 위해 외부 데이터셋 4종을 선별하여 증강하였다. 외부 데이터는 지문 기반의 추론/독해 중심 문항을 기준으로 선정하였다.

4.2.1 LSAT (Law School Admission Test)

LSAT는 논리적 추론·독해 중심의 5지선다 문항이 많아, 대회 데이터의 5지선다 장문·추론형을 보완하기 위해 활용하였다. 1,827개 중 대회 5지선다 길이 분포를 기준으로 극단값을 제외하고 수작업 검수로 최종 650개를 선별하여 번역·변환 후 증강 데이터로 사용했다.

4.2.2 SAT (Scholastic Aptitude Test)

SAT는 역사(미국사/세계사) 관련 문항을 포함해 수능형 사회·역사 도메인 다양성 보완을 목적으로 선정했다. 원본 데이터는 지문/질문이 분리되지 않은 형태였고 단문 문항이 존재해 이를 제외했으며, 번역 과정에서 지문·질문 분리 전처리와 검수를 거쳐 최종 107개를 증강 데이터로 사용했다.

4.2.3 지방직 9급 공무원 한국사 시험

지방직 9급 한국사는 4지선다 객관식이며 지문/보기 기반의 추론이 포함되어 대회 데이터의 4지선다 유형 보완에 적합하다고 판단했다. 8개년 PDF(총 160문항)를 텍스트 추출 라이브러리(PyMuPDF)로 구조화한 뒤,

추출 오류(문자 깨짐/구조 훼손)를 제외하고 최종 111개를 사용했다.

4.2.4 5급 공무원 언어논리 영역 시험

5급 언어논리 문항은 장문 지문 기반의 논리적 추론을 요구하는 객관식 5지선다로, 본 대회 데이터의 5지선다 장문·추론형 문항과 유사하다고 판단하였다. 5개년 PDF(총 200문항)를 텍스트 추출 라이브러리(PyMuPDF)로 구조화한 뒤, 추출 오류(문자 깨짐/구조 훼손)를 제외하고 최종 105문항을 증강 데이터로 사용했다.

Test 데이터에서 5지선다 비중이 높게 관측되어(약 78%), 증강 구성 또한 5지선다 비율을 맞추기 위해 두 가지 후보(A/B)를 비교하였다. 최종 구성은 오프라인 검증 성능 기준으로 최종 학습에는 후보 A(LSAT 300)를 적용하여 증강 데이터 623개, 원본 Train 2,031개와 결합해 총 2,654개 규모로 학습 데이터를 구성하였다.

구성	4지 증강	5지 증강	증강 합계	5지 비율	macro-f1
증강 후보 A (LSAT 300)	218	405	623	65.0%	**0.7699**
증강 후보 B (LSAT 650)	218	755	973	77.6%	0.7681

4.3 Model 선택 및 학습 전략 (Modeling)

4.3.1 모델 선택

본 프로젝트는 소형/중형 모델 기반으로 수능형 객관식 문제를 해결하는 것을 목표로 하며, 장문 입력 처리와 출력 안정성을 고려해 Qwen3-14B를 주력 모델로 선정하였다.

4.3.2 학습 전략 비교: Curriculum Learning의 한계

초기에는 외부 증강 데이터로 기초 논리력을 학습한 후 대회 데이터를 학습하는 순차적 방식(Curriculum Learning)을 시도했으나, 두 데이터를 섞어서 학습하는 혼합 방식(Mixed Training)보다 성능이 저조했다.

이는 순차 학습 과정에서 이전 단계의 지식을 잊어버리는 망각 현상이 발생했기 때문이다. 결과적으로 서로 다른 분포의 데이터를 동시에 학습시키는 것이 과적합을 방지하고 모델의 일반화 성능을 확보하는 데 더 효과적임을 확인하였다.

그 결과 검증 macro-F1 기준으로 각각 0.7681(선학습→Train), 0.7288(혼합 학습)로 나타났으며, Curriculum Learning은 성능 저하로 최종 전략에서는 제외하였다.

4.4 프롬프트 엔지니어링 (Prompt Engineering)

EDA 결과에 기반하여 4지/5지 선다형의 문항 특성에 맞춘 이원화된 프롬프트 전략을 시도하였다. 특히 단순 독해 위주인 5지 선다형 대비, 성능 변동폭이 크고 외부 지식 활용이 필요한 4지 선다형의 성능 최적화에 집중하여 실험을 수행하였다.

4.4.1 주요 프롬프트 전략

- 페르소나 및 단계적 가이드라인

  지문만으로 해결이 어려운 문항이 존재한다고 판단하여, 지문과 모순되지 않는 범위에서 일반 지식을 활용하도록 “지식 추론 전문가” 페르소나와 단계적 추론 가이드를 제공하였다.

- 개념 확장 및 자기 검증

  오답 분석 결과, 모델이 선택지 개념을 상위 범주로 일반화하지 못해(예: ‘2차 세계대전’→‘전쟁’) 오답을 선택하는 사례가 확인되었다. 이를 보완하기 위해 선택지의 핵심 개념을 확장해 해석하도록 지시하고, 근거 추출→선택지 비교→정답 선택의 절차로 추론 과정을 구조화하였다.

  또한 최종 답안 출력 전 자기검토 단계를 추가하여 근거-정답 불일치 및 논리적 오류를 스스로 점검·수정하도록 설계하였다.

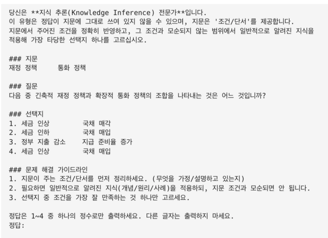
- Few-shot CoT

  제한된 Train 데이터의 효율적 사용과 모델에 입력하는 명령 구체화를 위해 문제 해결을 위한 추론 과정 예시를 few shot 프롬프트로 추가하였다.

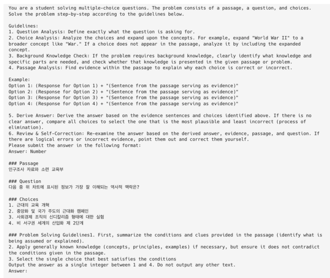
  모델에 추론 예시를 프롬프트에 추가해 추상적인 추론 가이드라인이 아닌 예시의 명확한 추론과정을 따도록 설계하였다.

- 예시

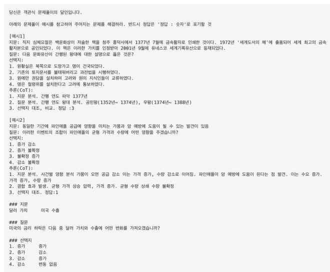
  페르소나 및 단계적 가이드라인 | 개념 확장 및 자기 검증 | Few-shot CoT



```
당신은 **지식 추론(Knowledge Inference) 전문가**입니다.
이 유형은 정답이 지문에 그대로 쓰여 있지 않을 수 있으며, 지문은 '조건/단서'를 제공합니다.
지문에서 주어진 조건을 정확히 반영하고, 그 조건과 모순되지 않는 범위에서 일반적으로 알려진 지식을
적용해 가장 타당한 선택지 하나를 고르십시오.

### 지문
재정 정책      통화 정책

### 질문
다음 중 긴축적 재정 정책과 확장적 통화 정책의 조합을 나타내는 것은 어느 것입니까?

### 선택지
1. 세금 인상         국채 매각
2. 세금 인하         국채 매입
3. 정부 지출 감소     지급 준비율 증가
4. 세금 인상         국채 매입

### 문제 해결 가이드라인
1. 지문이 주는 조건/단서를 먼저 정리하세요. (무엇을 가정/설명하고 있는지)
2. 필요하면 일반적으로 알려진 지식(개념/원리/사례)을 적용하되, 지문 조건과 모순되면 안 됩니다.
3. 선택지 중 조건을 가장 잘 만족하는 것 하나만 고르세요.

정답은 1~4 중 하나의 정수로만 출력하세요. 다른 글자는 출력하지 마세요.
정답:
```



```
You are a student solving multiple-choice questions. The problem consists of a passage, a question, and choices.
Solve the problem step-by-step according to the guidelines below.

Guidelines:
1. Question Analysis: Define exactly what the question is asking for.
2. Choice Analysis: Analyze the choices and expand upon the concepts. For example, expand "World War II" to a
broader concept like "War." If a choice does not appear in the passage, analyze it by including the expanded
concept.
3. Background Knowledge Check: If the problem requires background knowledge, clearly identify what knowledge and
specific parts are needed, and check whether that knowledge is presented in the given passage or problem.
4. Passage Analysis: Find evidence within the passage to explain why each choice is correct or incorrect.

Example:
Option 1: (Response for Option 1) + "(Sentence from the passage serving as evidence)"
Option 2: (Response for Option 2) + "(Sentence from the passage serving as evidence)"
Option 3: (Response for Option 3) + "(Sentence from the passage serving as evidence)"
Option 4: (Response for Option 4) + "(Sentence from the passage serving as evidence)"

5. Derive Answer: Derive the answer based on the evidence sentences and choices identified above. If there is no
clear answer, compare all choices to select the one that is the most plausible and least incorrect (process of
elimination).
6. Review & Self-Correction: Re-examine the answer based on the derived answer, evidence, passage, and question. If
there are logical errors or incorrect evidence, point them out and correct them yourself.
Please submit the answer in the following format:
Answer: Number

### Passage
인구조사 자료와 소련 교육부

### Question
다음 중 위 차트에 표시된 정보가 가장 잘 이해되는 역사적 맥락은?

### Choices
1. 근대의 교육 개혁
2. 중앙화 및 국가 주도의 근대화 캠페인
3. 사회경제 조직의 신디칼리즘 형태에 대한 실험
4. 비 서구권 세계의 산업화 제 2단계

### Problem Solving Guidelines1. First, summarize the conditions and clues provided in the passage (identify what is
being assumed or explained).
2. Apply generally known knowledge (concepts, principles, examples) if necessary, but ensure it does not contradict
the conditions given in the passage.
3. Select the single choice that best satisfies the conditions
Output the answer as a single integer between 1 and 4. Do not output any other text.
Answer:
```



```
당신은 객관식 문제풀이의 달인입니다.

아래의 문제풀이 예시를 참고하여 주어지는 문제를 해결하라. 반드시 정답은 '정답 : 숫자'로 표기할 것

[예시1]
지문: 직지 심체요절은 백운화상이 저술한 책을 청주 흥덕사에서 1377년 7월에 금속활자로 인쇄한 것이다. 1972년 '세계도서의 해'에 출품되어 세계 최고의 금속
활자본으로 공인되었다. 이 책은 이러한 가치를 인정받아 2001년 9월에 유네스코 세계기록유산으로 등재되었다.
질문: 다음 문화유산이 간행된 왕대에 대한 설명으로 옳은 것은?
선택지:
1. 원황실은 북쪽으로 도망가고 명이 건국되었다.
2. 기존의 토지문서를 불태워버리고 과전법을 시행하였다.
3. 원에만 권당을 설치하여 고려와 원의 지식인들이 교류하였다.
4. 명은 철령위를 설치한다고 고려에 통보하였다.
추론(CoT):
1. 지문 분석. 간행 연도 파악 1377년
2. 질문 분석. 간행 연도 왕대 분석. 공민왕(1352년~ 1374년), 우왕(1374년~ 1388년)
3. 선택지 대조, 비교. 정답 :3

[예시2]
지문: 동일한 기간에 파인애플 공급에 영향을 미치는 가뭄과 암 예방에 도움이 될 수 있는 발견이 있음
질문: 이러한 이벤트의 조합이 파인애플의 균형 가격과 수량에 어떤 영향을 주겠습니까?
선택지:
1. 증가 감소
2. 증가 불확정
3. 불확정 증가
4. 감소 불확정
추론(CoT):
1. 지문 분석. 사건별 영향 분석 가뭄이 오면 공급 감소 이는 가격 증가, 수량 감소로 이어짐. 파인애플이 암 예방에 도움이 된다는 점 발견. 이는 수요 증가.
가격 증가, 수량 증가
2. 결합 효과 발생. 균형 가격 상승 압력, 가격 증가. 균형 수량 상쇄 수량 불확정
3. 선택지 대조. 정답:1

### 지문
달러 가치      미국 수출

### 질문
미국의 금리 하락은 다음 중 달러 가치와 수출에 어떤 변화를 가져오겠습니까?

### 선택지
1. 증가      증가
2. 증가      감소
3. 감소      증가
4. 감소      변동 없음
```

4.4.2 결과

전략 (Strategy)	Accuracy	Macro F1
1. 페르소나 + 가이드라인	0.700	0.557
2. 개념 확장 + 검증 (Review)	**0.760**	**0.612**
3. Few-shot CoT	0.700	0.562

3가지 방법 모두 기본 베이스라인보다 높은 성능을 보였으나 그 중 개념 확장 및 검증이 높은 성능을 보였다. 확실한 추론 과정과 자기 검증을 줌으로써 모델이 추론 과정에서 발생할 수 있는 논리적 오류를 스스로 교정하고 정답의 신뢰도를 높이는데 기여한 것으로 분석된다.

4.5 앙상블

단일 모델이 가질 수 있는 편향 및 추론 오류를 완화하고 예측 안정성을 확보하기 위해 모델 앙상블을 적용하였다. 최종 예측은 두 모델의 단일 성능 차이를 반영하여 Soft Voting을 사용했으며, 가중치는 0.6(Qwen3-

14B) : 0.4(A.X-4.0 Light)로 설정하였다.

4.5.1 모델 선정

수능형 문제는 복합적인 논리 추론 능력과 한국어 특유의 문맥 이해력이 동시에 요구된다. 이에 고성능 모델과 한국어 특화 모델을 결합하는 전략을 수립하였다.

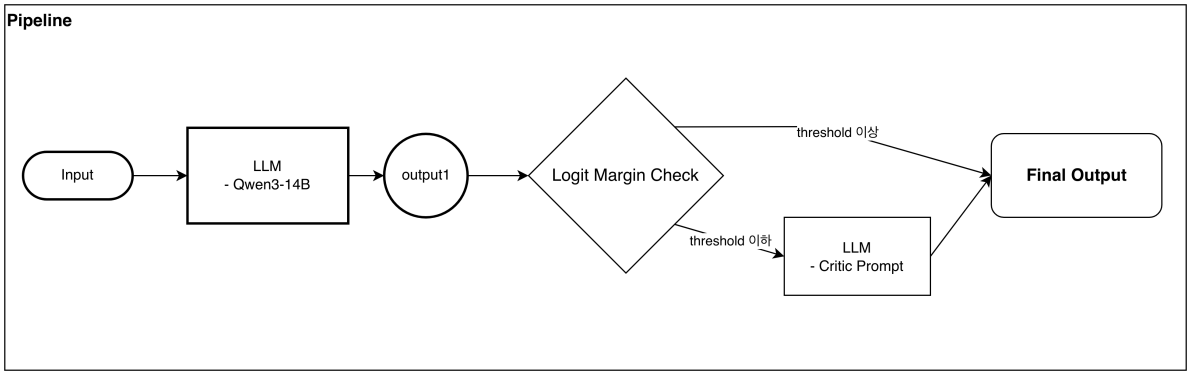
- **Qwen3-14B**: 비교적 강한 논리적 추론 및 일반화 성능을 기대할 수 있어 메인 모델로 선정하였다.
- **A.X-4.0 Light:** 한국어 도메인에 특화된 모델로, 오답 분석 결과 타 모델이 놓치는 한국어 맥락/뉘앙스 기반 문항에서 정답을 도출하는 경향이 확인되어 보완 모델로 선정하였다.

4.5.2 단일 모델과의 비교

오프라인 검증 성능 기준으로 Soft Voting 앙상블이 단일 모델 대비 성능이 가장 높음을 알 수 있다.

모델	macro-F1
Ensemble (0.6 Qwen3-14B / 0.4 A.X-4.0 Light)	**0.7861**
Qwen3-14B (실험명/버전)	0.7621
A.X-4.0 Light	0.7591

4.6 2단계 추론 (2-step Inference)



4.6.1 저신뢰도 문항 정의

모델의 확신(Confidence)을 측정하기 위해, 확률값(Probability) 대신 비선형 압축 전 단계인 로짓(Logit) 값을 활용하였다. 구체적으로 1순위와 2순위 선택지 간의 로짓 격차(Margin)를 신뢰도 지표로 정의하였다.

4.6.2 재추론 전략

마진이 작은 문항에만 재추론을 적용하기 위해, 마진 기준 하위 25%를 **저신뢰도 문항**으로 분류하였다. 이후 해당 문항에 한해 Critic(반증형 검증자) 프롬프트로 2차 추론을 수행하고, 나머지 문항은 1차 예측을 그대로 사용하였다.

또한 재추론 대상 비율은 15/25/35%로 비교 실험하였다. 15%는 교정되는 문항 수가 제한적이었고, 35%는 25%와 유사한 수준의 변경만 발생해 임계 비율을 크게 잡는다고 해서 답안 변경(교정)이 선형적으로 증가하지 않음을 확인하였다. 이에 효율과 효과의 균형 측면에서 25%를 최종 기준으로 사용하였다.

4.6.3 Critic 프롬프트 설계

저신뢰도 문항을 해결하기 위해, 모델을 반증형 검증자(Critic)로 정의하는 특화 프롬프트를 설계하였다. 이는 모델이 자신의 1차 추론 결과를 의도적으로 의심하게 함으로써 확증 편향을 억제하고, 지문에 명시된 텍스트 근거만을 바탕으로 오답 가능성을 배제하여, 추론의 객관성과 최종 답안의 신뢰도를 확보하도록 설계했다.

Critic 프롬프트

```
당신은 객관적인 '반증형 검증자(critic)'입니다.
주어진 지문만을 근거로, 1차 예측 답안이 틀렸는지 먼저 검증하세요.

규칙:
1. 외부 지식 사용 금지. 지문에 명시된 내용만 근거로 판단.
2. 먼저 1차 예측 답안이 "틀릴 수밖에 없는" 반증 근거가 있는지 찾기.
3. 반증 근거가 명확하면 정답을 교정(다른 선택지 번호로 변경).
4. 반증 근거가 불충분하면 1차 예측 답안을 유지.
5. 최종 출력은 반드시 선택지 번호 하나(1~4)만. 다른 글자 출력 금지.

### 문제 해결 가이드라인
1. 지문을 끝까지 읽고 핵심 정보를 다시 정리하세요.
2. 질문이 요구하는 정보(수치/인물/원인/결과/요지 등)를 정확히 확인하세요.
3. 1차 예측 답안이 지문과 **모순되는 부분이 있는지** 먼저 검토하세요.
4. 각 선택지를 지문의 근거와 다시 1:1로 대조하세요.
5. 1차 예측 답안을 **반증할 수 있는 명확한 근거가 있으면 교정**하고,
   그렇지 않으면 1차 예측 답안을 유지하세요.
```

5. 자체 평가 의견

5.1 전체 프로젝트에 대한 평가

최종적으로 리더보드에서 다음과 같은 성과를 달성하였다.

구분	Macro-F1
Private	0.7010
Public	0.7861

본 프로젝트는 소형 모델로 수능형 문제 해결을 목표로 하여

특히 EDA(탐색적 데이터 분석) 결과를 바탕으로 전체 파이프라인을 설계하고, 명확한 역할 분담(R&R)을 통해 체계적으로 업무를 수행한 점은 긍정적인 성과로 생각한다.

5.2 한계점 및 아쉬운점

5.2.1 검증 데이터셋의 신뢰성 및 대표성 부족

학습 데이터(Train)의 10%인 약 204개 문항만을 단순 랜덤 분할하여 검증 데이터로 활용하였다. 그러나 해당 검증 세트에서는 학습 중 지표가 과대하게 상승하는 반면, 실제 제출 점수와의 괴리가 크게 나타나 검증 세트가 일반화 성능을 충분히 대변하지 못하는 문제가 발생했다.

5.2.2 RAG 성능 미확보

외부 지식이 필수적인 문항 해결을 위해 RAG를 도입했으나, 다음과 같은 기술적 난관으로 인해 성능 향상으로 이어지지 못했다.

- **Query 최적화 실패:** 질문(Query)을 어떤 형태로 변환하여 검색 엔진에 입력해야 최적의 문서가 추출되는지에 대한 실험과 튜닝이 부족하였다.
- **Multi-hop 정보 통합 부재:** 수능형 문제는 단일 지식이 아닌 여러 문맥의 정보를 종합해야 풀 수 있다. 그러나 현재 시스템은 파편화된 문서들을 단순히 나열하는 수준에 그쳐, 복합적인 근거(Evidence)를 유기적으로 연결하여 정답을 도출하는 데 실패하였다.

5.2.3 DPO 적용의 한계

오답을 줄이기 위해 DPO를 도입하고자 했으나, DPO 데이터셋(chosen/rejected)을 구성하는 과정에서 선호 응답의 텍스트 품질 차이를 충분히 반영하지 못했다. 특히 chosen/rejected를 단순히 정답 여부에 기반해 구성하면서, 모델이 학습해야 할 선호 신호가 약해져 결과적으로 성능 향상이 제한적이었다.

5.2.4 팀 운영 및 실험 공유

프로젝트 초기 수립한 실험 설계 및 공유 체계가 후반부의 일정 압박으로 인해 무너지는 현상이 발생하였다. 기록과 공유가 지연됨에 따라 실험 결과가 분산되고, 팀원 간 불필요한 중복 실험이 발생하는 비효율을 겪었다.

5.3 향후 개선 방향

5.3.1 소통 프로세스 개선

프로젝트 후반부로 갈수록 실험 결과 공유가 늦어지며 중복 실험 비용이 증가하였다. 이를 개선하기 위해 데일리 스크럼 기반의 짧은 공유 템플릿(목표–변경점–결과)을 사용하고, 실험 페이지에 실험을 최소 단위로 즉시 기록하도록 할 예정이다. 또한 아이디어는 완성도가 낮더라도 우선 공유하고, 팀 단위로 발전시키는 방식으로 의사결정 속도를 높이려고 한다.

5.3.2 코드 리뷰 및 피드백 보완

진행된 실험 및 리팩토링 과정에서 팀원 간 상호 검토가 부족하여 코드 이해도의 편차가 발생하였다. 향후에는 이를 방지하기 위해 PR을 의무화하여 코드 리뷰를 강제하고, 단순 리뷰로 해결되지 않는 복잡한 로직은 별도의 '리뷰 이슈 티켓'을 발행하여 팀원 전체가 해당 코드를 완전히 이해하고 넘어가는 문화를 정착시키고자 한다.

6. 개인회고

김영현

- 나는 내 학습목표 달성을 위해 무엇을 어떻게 했는가?

  프롬프트 엔지니어링, 데이터 증강 등을 통해 모델 성능을 개선하고자 하였다.
- 전과 비교해서, 내가 새롭게 시도한 변화는 무엇이고, 어떤 효과가 있었는가?

피어세션 시간에 진행한 기본 개념 스터디와 딥러닝 책을 통해서 모호했던 개념을 많이 잡은 것 같다. 전보다 의사소통함에 있어서 수월했으나 여전히 많이 부족한 것 같다.

- 나는 어떤 방식으로 모델을 개선했는가?

  프롬프트, 데이터 증강

- 내가 한 행동의 결과로 어떤 지점을 달성하고, 어떤 깨달음을 얻었는가?

  4지선다 프롬프트를 통해서 약간의 성능 향상을 달성하였으나, 5지선다 프롬프트에서는 유의미한 성능 향상을 하지 못하였다. 개인적으로 모호한 파트라는 생각이 들었고 흥미로운 분야같다. 또한 이번 데이터 증강을 통해, 그 전에 진행하였던 전처리 없는 데이터 증강이 얼마나 모델에 악영향을 미쳤을지에 대한 회고를 하게 되었다.

- 마주한 한계는 무엇이며, 아쉬웠던 점은 무엇인가?

  기초 지식이나 근거를 찾아 실험을 하는 부분이 많이 부족하다고 느꼈다.

- 한계/교훈을 바탕으로 다음 프로젝트에서 시도해볼 것은 무엇인가?

  근거있는 실험, 부족한 분야에 대해 즉각적인 소통을 통한 피드백

윤준상

- 나는 내 학습목표 달성을 위해 무엇을 어떻게 했는가?

  Train 데이터셋 내부 구조를 확인하고 초기에 RAG를 활용하려는 방향성을 위해 4지선다 유형들을 유형별로 분류하는데 집중하였고, 프롬프트의 중요한 요소들을 알아보며 메타 프롬프팅을 활용하는 방식을 통해 프롬프트 엔지니어링을 하였고, 공무원 시험 관련하여 데이터를 증강하였다.

- 전과 비교해서, 내가 새롭게 시도한 변화는 무엇이고, 어떤 효과가 있었는가?

  이전에는 프로젝트 경험이 없어 체계적이지 못했으나 그러지 않기 위해 노력하였고, 되도록 한 작업의 문서화를 하는 것에 집중하였다. 프로젝트 이해도를 올려 이전에 비해 참여도를 올리는 것을 목표로 하였다.

- 나는 어떤 방식으로 모델을 개선했는가?

  프롬프트 엔지니어링(메타 프롬프팅 활용), 데이터 증강, 데이터 EDA 및 분류

- 내가 한 행동의 결과로 어떤 지점을 달성하고, 어떤 깨달음을 얻었는가?

  RAG를 활용하는 방향성을 초기에 잡고 유형별 분류를 하였으나 생각 이상으로 모델의 성능이 좋았고, Vector DB를 구축하는데에 있어 난항을 겪어 포기를 하였다. 프롬프트 엔지니어링 과정에서 페르소나를 주입하거나 영어/한국어 중 사용을 어떻게 할지 여부 또는 few shot을 프롬프트에 활용하느냐 여부를 테스트해보는 과정에서 성능 향상을 확인할 수 있었다.

- 마주한 한계는 무엇이며, 아쉬웠던 점은 무엇인가?

  RAG의 Vector DB를 구축을 해냈다면 조금 더 유의미한 실험 결과를 만들어낼 수 있었지 않았을까 하는 아쉬움이 남았다.

  베이스라인 원리 이해까지는 하였으나 코드 모듈화나 그런 부분에 있어서 이해에 난항을 겪었다.

- 한계/교훈을 바탕으로 다음 프로젝트에서 시도해볼 것은 무엇인가?

  코드 모듈화에 관해 자세히 알아볼 것이고, 프로젝트 실험 방향성을 명확히 이해하고 task를 수행하고자 할 것이다.

장세현

- 나는 내 학습목표 달성을 위해 무엇을 어떻게 했는가?

  이번 프로젝트에서의 개인 목표는 수능형 문항 특성을 근거로 분석하고, 그 근거를 바탕으로 실험을 설계·구현하는 역량을 올리는 것이었다.

  이를 위해 먼저 데이터셋 구조와 분포를 EDA로 정리하고(4/5지선다 혼재, 지문 길이 bimodal, 정답 분포 편향), 원천 데이터(KMMLU/MMMLU/KLUE-MRC)의 특성을 직접 확인하며 문제 유형을 이해했다. 이후 분석 결과를 기준으로 파이프라인을 설계하고, DPO·RAG 등 고도화 기법과 함께 앙상블 및 2-step 재추론 로직을 구현하며 성능 개선을 시도했다.

- 전과 비교해서, 내가 새롭게 시도한 변화는 무엇이고, 어떤 효과가 있었는가?

  이전에는 아이디어가 떠오르면 빠르게 구현하고 결과로 판단하는 방식에 가까웠다. 이번에는 “왜 가능하다고 생각하는가 / 근거는 무엇인가 / 어떤 가정이 필요한가 / 관련 방법론은 무엇인가”를 계속 질문하며 실험 전 논리를 먼저 구조화했다.

  그 결과, 실험 도중 방향이 자주 바뀌거나 비교 조건이 흐려지는 일이 줄었고, 동일한 실험을 반복해 비용을 낭비하는 경우도 줄었다. 즉, 단순 구현 속도보다 실험의 재현성과 비교 가능성을 우선하는 방식으로 작업 습관을 개선했다.

- 나는 어떤 방식으로 모델을 개선했는가?

  철저하게 데이터 주도적(Data-Driven) 방식으로 모델 개선에 접근했다. 단순히 최신 기법(SOTA)을 적용하는 것이 아니라, 원천 데이터를 하나하나 직접 확인하며 문항 유형별 특성과 오답 패턴을 먼저 분석한 뒤 솔루션을 도출했다.

  예를 들어, 4지선다와 5지선다의 지문 길이 분포와 추론 방식이 다름을 확인한 후 유형별 분리 전략을 취했고, 이를 바탕으로 외부 데이터 증강 구성과 프롬프트 최적화 방향을 결정했다.

  이후 DPO, RAG, 2단계 추론 로직 등을 구현하며 성능 개선을 시도했다. 특히 모델이 헷갈리는 문항을 로짓 마진(Logit Margin)으로 식별하여, 저신뢰도 문항에 한해 Critic 프롬프트로 재추론하는 방식을 설계했다. 이를 통해 정확도와 효율성을 동시에 고려한 추론 파이프라인을 구축했다.

  무엇보다 코드를 작성하기 전에 논리적 타당성을 검증하는 과정을 거침으로써, 중간에 로직을 전면 수정하는 리스크를 줄이고 체계적으로 성능을 향상시킬 수 있었다.

- 내가 한 행동의 결과로 어떤 지점을 달성하고, 어떤 깨달음을 얻었는가?

  이번 프로젝트를 통해 최신 기법을 단순히 적용하는 것보다, 데이터 분석을 바탕으로 문제를 명확히 정의하는 것이 성능 향상의 핵심임을 체감했다.

  무엇보다 "왜?"를 끝까지 파고드는 사고 과정의 중요성을 깨달았다. 이전에는 떠오른 아이디어를 대충 정리하고 바로 구현했지만, 이번에는 "왜 효과적일까?", "근거는 무엇인가?", "어떤 가정이 필요한가?"를 반복적으로 질문하며 논리를 구조화했다. 그 결과 구현 후 방향을 전면 수정하는 일이 줄었고, 실험의 목적과 비교 기준이 명확해졌다. 깊이 있게 고민한 후 구현하는 것이 빠르게 구현하는 것보다 훨씬 효율적이며, 이전의 방식이 얼마나 허술했는지 뼈저리게 느꼈다.

- 마주한 한계는 무엇이며, 아쉬웠던 점은 무엇인가?

코드 구현 능력의 한계 를 체감했다. 구현하고 싶은 아이디어는 많았으나, 실제 코드로 옮기는 과정에서 어려움을 겪었다. 특히 Jupyter Notebook 중심의 작업 습관으로 인해 .py 모듈 구조로 전환할 때 import 오류와 의존성 문제 해결에 예상보다 많은 시간이 소요되었다.

코드 가독성과 컨벤션측면에서도 부족함을 느꼈다. 이전에는 긴 주석으로 설명하면 충분하다고 생각했으나, 팀 협업 과정에서 설명 없이도 의도가 전달되는 코드 작성이 더 중요함을 깨달았다.

라이브러리 버전 이슈가 발생했을 때 대응 능력이 부족했다. 특정 함수가 deprecated되면서 대체 방법을 찾아야 했는데, 라이브러리를 표면적으로만 이해하고 있어 여러 부분을 수정하는 데 어려움을 겪었다.

- 한계/교훈을 바탕으로 다음 프로젝트에서 시도해볼 것은 무엇인가?

  적극적인 질문과 협업을 실천하려 한다. 혼자 고민하는 시간을 줄이고, "나 혼자 고민 → 팀 단위 논의 → 더 잘하는 사람에게 질문"의 순서로 문제해결 경로를 구조화할 것이다. 계속 고민한다고 해서 답이 나오는 것이 아닐 수 있다는 점을 인식하고, 빠르게 도움을 요청하는 습관을 정착시키려 한다.

  Python 모듈화 및 버전 관리 역량을 강화하기 위해, 자주 사용하는 라이브러리는 핵심 구조와 주요 함수의 동작 원리를 깊이 있게 이해하여, 버전 이슈나 deprecated 문제에 유연하게 대응할 수 있는 기반을 마련하려 한다.

주현민

- 나는 내 학습목표 달성을 위해 무엇을 어떻게 했는가?
  EDA결과 제공된 학습데이터가 수능 국어, 철학, 역사 문제를 풀기에 충분하지 않다는 판단을 했다. 그 판단의 결과로 공무원 데이터를 수집하고, 학습데이터 구조로 재가공하여 데이터를 증강했다.
  또 프로젝트의 목표인 '수능 문제풀이'에 대해 고민하여 프롬프트 엔지니어링을 했다.
- 전과 비교해서, 내가 새롭게 시도한 변화는 무엇이고, 어떤 효과가 있었는가?
  프로젝트 막바지에 모든 일상을 프로젝트에 할애했다.
  다른 팀원, 다른 캠퍼들보다 코드 구현 능력이 부족하고, 지식이 부족하기에 더 많은 시간을 할애하는게 당연한 일이었지만 뒤늦게 그렇게 했다.
  프로젝트에 있어 스스로에게 떳떳할 수는 없지만, 이전 프로젝트에 비해 조금 더 어깨를 필 수 있었다.
  팀적으로는 더 원활한 소통을 도모하고자 피어세션 뿐만 아니라 모든 시간에 줌 회의에 참여했다.
  가능한 계속 얼굴을 보고 각자 역할에 충실하며 모르는 것, 애매한것, 궁금한 것이 생기면 바로바로 소통을 할 수 있었던 것이 좋았다.
- 나는 어떤 방식으로 모델을 개선했는가?
  모델의 성능을 위해서 데이터를 수집했다. 또 모델의 개선을 위해 프롬프트를 고민하고, 실험했다.
- 내가 한 행동의 결과로 어떤 지점을 달성하고, 어떤 깨달음을 얻었는가?
  내가 수집한 데이터로 학습 데이터를 증강했을 때 성능이 소폭 상승했다.
  위에서 언급한 데이터 외에 프로젝트 막바지에 추가로 수집한 데이터가 있었지만, 학습 데이터로 바로 사용할 수 있을만큼 전처리를 하지 못했기에 모델에 활용하지 못했다.
  협업할 때 데이터를 Ready to use로 전처리하고 전달하는 것이 시간을 아끼고 팀원들을 배려하는 것, 팀에게 도움이 되는 것이라는걸 다시금 깨달았다.

- 마주한 한계는 무엇이며, 아쉬웠던 점은 무엇인가?
  아이디어를 구상하고, 팀원들을 설득하는 것이 아쉬웠다.
  이전 프로젝트에서 근거있는 사고, 아이디어를 하고자 반성했지만. 여전히 내 아이디어들은 근거가 충분하지 못했다.
  근거가 있어도, 코드로 구현하고 설득하기에 실력이 부족했다.
  '눈이 오면 겨울이다' 라는 말처럼 당연한 말이지만, 여전히 코드 구현 능력이 한계고, 아쉬웠다.
- 한계/교훈을 바탕으로 다음 프로젝트에서 시도해볼 것은 무엇인가?
  최종 프로젝트는 컴페티션이 아닌 만큼, 팀원들을 충분히 설득히킬 만한 아이디어들을 말 그대로 던지고 하고 싶은 걸 다 해볼 것이다. 그러기 위해서 부단히 공부하고 노력할 것이다.

한지석

- 나는 내 학습목표 달성을 위해 무엇을 어떻게 했는가?

  특정 기술을 적용하기 전, 사전 실험을 통해 실제 성능 향상 여부를 검증하였다.

  스크립트 방식의 모듈 구조에 익숙해지기 위해 디렉토리 구조를 설계했다.
- 전과 비교해서, 내가 새롭게 시도한 변화는 무엇이고, 어떤 효과가 있었는가?

  이전 프로젝트에서는 좋다고 하는 기술들을 무작정 사용하려고 했으나, 이번 프로젝트에서는 실제로 유의미한 결과를 내는지 확인하고 없다면 과감히 사용하지 않았다.

  전체 코드의 디렉토리 구조를 설계했다. 작업 시간은 많이 소요되었으나, 코드 전반에 대한 이해도가 높아졌고 구조적인 모듈화 방향에 대해서도 깊이 고민할 수 있었다.
- 나는 어떤 방식으로 모델을 개선했는가?

  논리 추론 성능이 검증된 Qwen3-8B 모델과 한국어 특화 모델인 A.X 4.0 Light를 결합한 Soft Voting 앙상블을 통해 모델의 최종 성능을 개선하였다.
- 내가 한 행동의 결과로 어떤 지점을 달성하고, 어떤 깨달음을 얻었는가?

  문맥 파악과 고도의 언어적 추론 능력을 요구하는 수능형 문제 풀이에서 기본적인 추론 능력을 가진 비교적 큰 모델과 비교적 작은 한국어 특화 모델의 앙상블은 성능향상을 가져온다는 것을 깨달았다.

  RAG 활용을 위해 Retriever 구현을 시도했으나, 수능 문제 풀이에 실질적인 도움이 되는 문서를 추출하는 데 어려움을 겪어 유의미한 성능 변화를 이끌어내지 못하는 경험을 했다.
- 마주한 한계는 무엇이며, 아쉬웠던 점은 무엇인가?

  생성형 AI의 도움 없이 스스로 코드를 작성하려 노력했으나, 아직은 직접 구현할 수 있는 범위가 넓지 않다는 점을 확인했다. 스스로 코드를 작성하는 역량의 부족함을 실감한 계기가 되었다.

  빠른 모델 학습과 추론 결과 확인에 치중한 나머지, 완벽히 검증되지 않은 코드로 실행 시간을 낭비한 경우가 있었다.

  시간적 여유 부족으로 충분한 자료조사를 수행하지 못해, 더 다양한 아이디어를 가설로 연결하지 못한 점이 아쉬움으로 남는다.
- 한계/교훈을 바탕으로 다음 프로젝트에서 시도해볼 것은 무엇인가?

  다음 프로젝트에서는 시간이 더 걸리더라도 AI 의존도를 낮추고 스스로 코드를 작성하며, 디버깅을 고려한 견고한 코드를 짜는 연습을 병행해야겠다.

많은 아이디어를 내지 못한 이유는 자료조사를 많이 하지 못해서 였던 것이라 판단된다. 부족했던 아이디어를 보완하기 위해, 할 일이 많더라도 반드시 자료조사 시간을 확보해야겠다. 이를 통해 전체 구조 내에서 유효한 가설을 세우고 실험을 설계하는 데 더 많은 시간을 투자할 계획이다.

프로젝트의 시작 단계에서 데이터의 흐름과 형태를 구체적으로 파악하는 것이 얼마나 중요한지 깊이 체감했다. 다음 프로젝트에서는 이를 최우선 순위로 두고, 구조를 완전히 이해한 뒤 설계에 착수해야겠다.

7. 부록

- LSAT
  - https://huggingface.co/datasets/tasksource/lsat-rc
- SAT
  - https://huggingface.co/datasets/UniqueData/sat-questions-and-answers-for-llm?not-for-all-audiences=true
- 9급 공무원 한국사 영역 시험
  - https://www.gosi.kr/uat/uia/gosiMain.do
- 5급 공무원 언어논리 영역 시험
  - https://www.gosi.kr/uat/uia/gosiMain.do