



# Upstage AI Lab

Natural Language Processing Seminar

2025.08.07(목)

## 목차

- 01. 팀 소개
- 02. 경진대회 수행 절차 및 방법
- 03. 분석 인사이트 및 결과
- 04. 회고

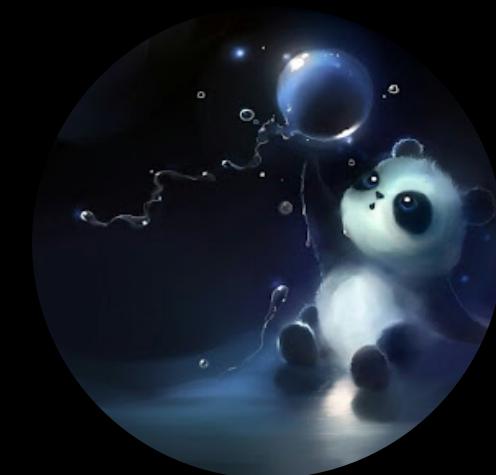
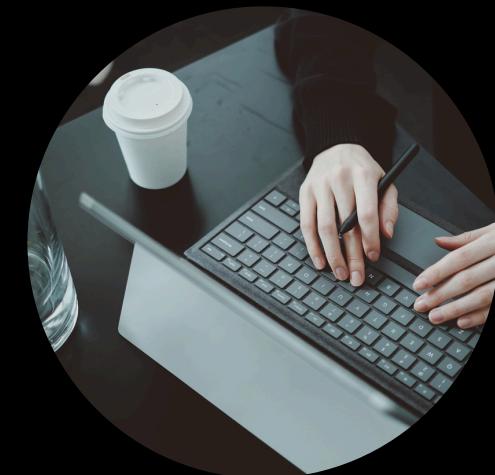
01

## 팀 소개

---

팀장/팀원 소개  
협업 방식

## [5조 - 티끌 모아 Tech] 각자의 작은 기술이 모여 힘을 발휘한다!

**팀장**

송규현

RAG / 경영정보

EDA, 베이스라인 작성,  
자료조사**팀원**

이상현

AI응용분야찾기/재료공학

Dialogue-Summary 간 상관  
데이터 특성추출**팀원**

이영준

MLOps / 컴퓨터공학

자동화 구축 후 다양한 모델,  
옵션 조합 실험**팀원**

조은별

고분자공학과  
특수 표현 전처리 및  
마스킹 토큰화 담당**팀원**

편아현

소프트웨어학과  
데이터 전처리 및  
optuna를 통한 하이퍼 파라미터 튜닝

# 경진대회 협업 방식

## Natural Language Processing [대회] Summarization

회의 및 아이디어 공유



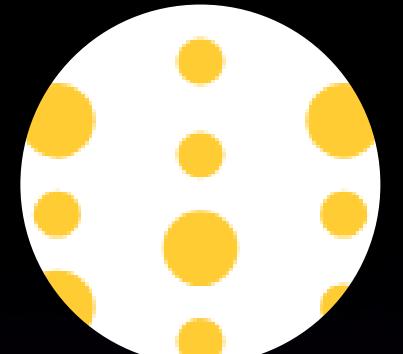
Zoom & Slack

TO-DO & 문서화



Github Issue & docs/

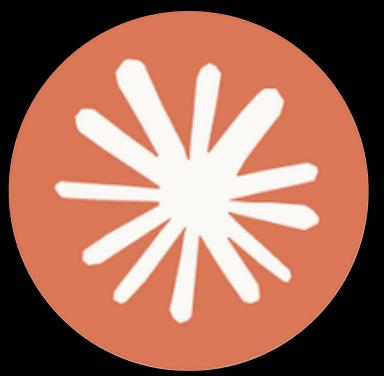
실험 정리



WandB & Notion



코드 작성



Claude Code



Gemini CLI



ChatGPT

02

## 경진대회 수행 절차 및 방법

---

목표 수립  
수행 내용 / 수행 결과

# 경진대회 목표 수립

## Natural Language Processing [대회] Summarization

주제

### Dialogue Summarization | 일상 대화 요약

학교 생활, 직장, 치료, 쇼핑, 여가, 여행 등 광범위한 일상 생활 중 하는 대화들에 대해 요약합니다.

목표

개요

#### 목표

- 대화 요약 모델 파인튜닝 및 실전 파이프라인 경험하기 [o]
- 다양한 전처리 및 데이터 정제 기법 실험하기 [o]
- 역번역(Back Translation) 기반 데이터 증강 실험하기 [o]
- ROUGE 기반의 신뢰성 있는 평가 및 검증 전략 수립 [o]
- 베이스라인 코드를 바탕으로 전처리/하이퍼파라미터/모델 구조 체계적으로 실험 [△]
- 목표 점수: ROUGE 평균 50 이상 달성 [x]

#### 소개 및 배경 설명

일상 대화문을 요약하는 모델을 개발하고, 249개의 대화문에 대해 각각 1개의 요약문을 생성하는 경진대회

#### 기간

2025.07.25 ~ 2025.08.06

# 경진대회 수행 내용

## Natural Language Processing [대회] Summarization

### 1 개발 환경 구축

- Python 3.11 (conda 가상환경)
- PyTorch 2.6.0
- transformers 4.54.0
- pytorch-lightning 2.5.2
- rouge, rouge-score (평가 metric)
- wandb (실험 관리)
- unsloth, gradio, evaluate (추론/서빙/평가)
- pandas, numpy, tqdm (데이터 처리/분석)
- kiwipiepy (형태소 분석, 한국어 전처리)

### 2 데이터 분석

- EDA: train/dev/test 대화문 및 요약문 분석
- 대화문/요약문 길이 분포
- 특수 표현(지시어, 마스킹, 이모티콘 등) 빈도
- 중복/유사 요약문 비율
- 카테고리/주제별 분포
- 기타 전처리 필요 요소 탐색

### 3 데이터 증강

- AI Hub 대화 데이터 활용 및 백트랜슬레이션(역번역) 기반 증강 적용
- 불용어·특수문자 무작위 제거: 텍스트에서 불용어, 특수문자, 이모티콘 일부를 임의로 제거
- 지시어/마스킹 토큰 랜덤 교체: 지시어, 마스킹 토큰([#Person1#] 등)의 위치·형태를 무작위 변형
- 길이 조정 및 문장 분리/병합: 대화문 길이, 발화 단위 등을 임의로 늘리거나 줄임

### 4 모델 선택 학습 및 평가

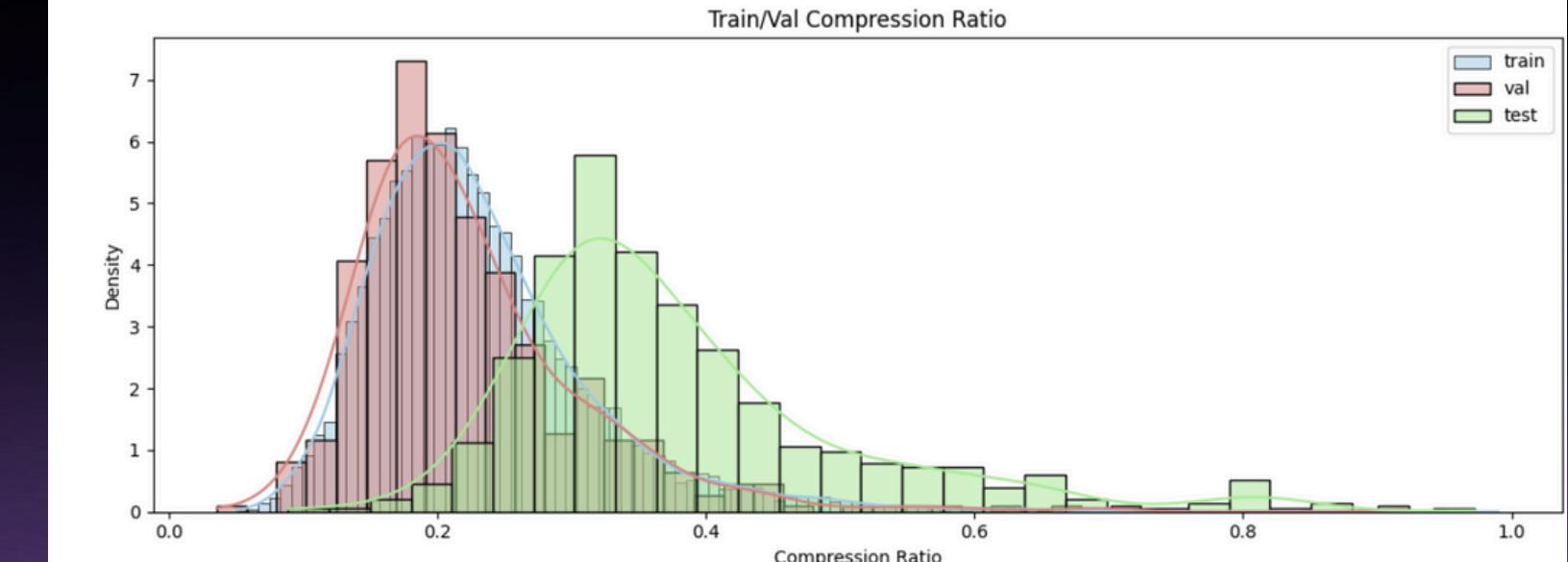
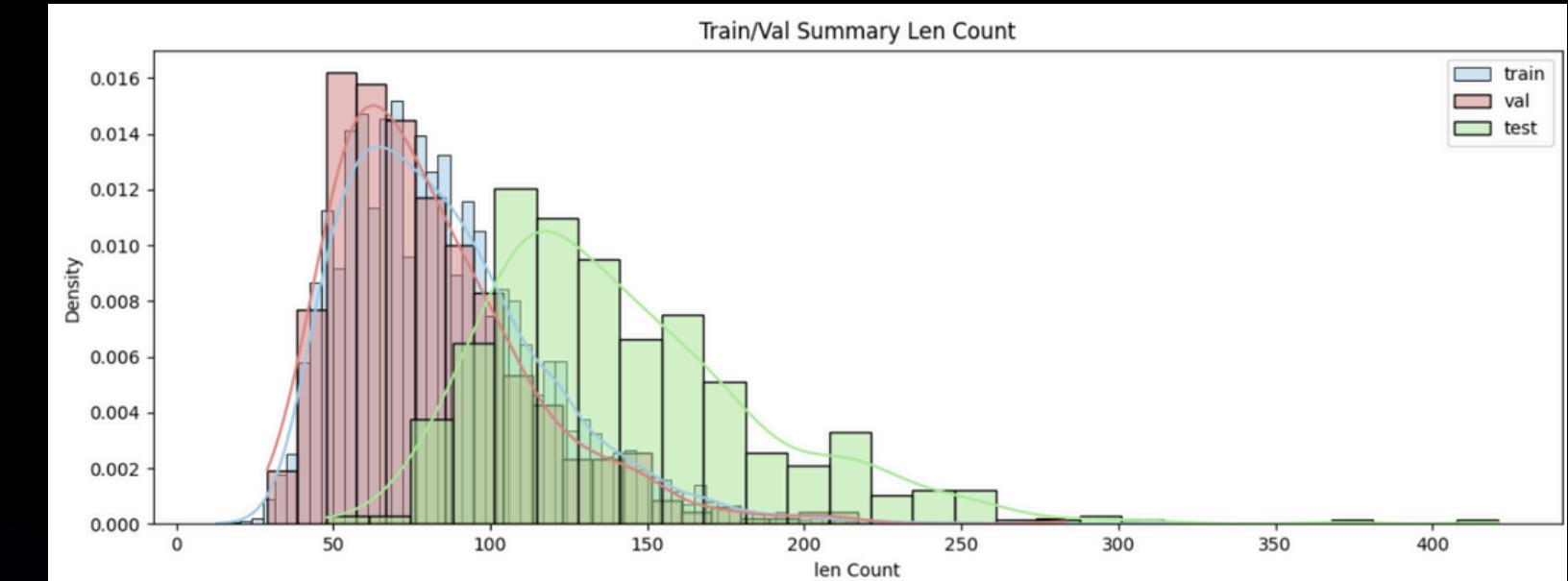
- 사전학습 요약 모델 파인튜닝: KoBART, KoT5 등 한국어 특화 사전학습 모델을 파인튜닝하여 대화 요약에 최적화
- 모델 구조 실험: 인코더 및 디코더 max length, special token 등 구조별 파라미터와 설정 실험
- 평가: ROUGE 등 자동 평가 지표를 활용한 성능 측정

# 경진대회 수행 내용

## Natural Language Processing [대회] Summarization

### [데이터 분석] - Train & Test & Val Summary Length EDA

- 대화문 길이:
  - train/val/test 모두 100~600자(또는 단어)에 주로 분포, test 세트가 다소 더 긴 대화도 존재
  - 최장 대화문은 2000자 이상도 일부 포함됨
- 요약문 길이:
  - train/val은 40~100자에 집중
  - test는 80~150자 구간에 더 많이 분포, 최대 400자 이상인 케이스도 일부 존재
  - test 데이터의 요약문이 전반적으로 더 길
    - 모델 파라미터(encoder/decoder max\_length) 설정 시  
train/val 기준만 사용하면 test 데이터의 긴 대화/요약 일부가 잘릴 위험  
충분히 넉넉한 max\_length(encoder 512 / 1024, decoder 200으로 설정  
=> 정보 손실 방지

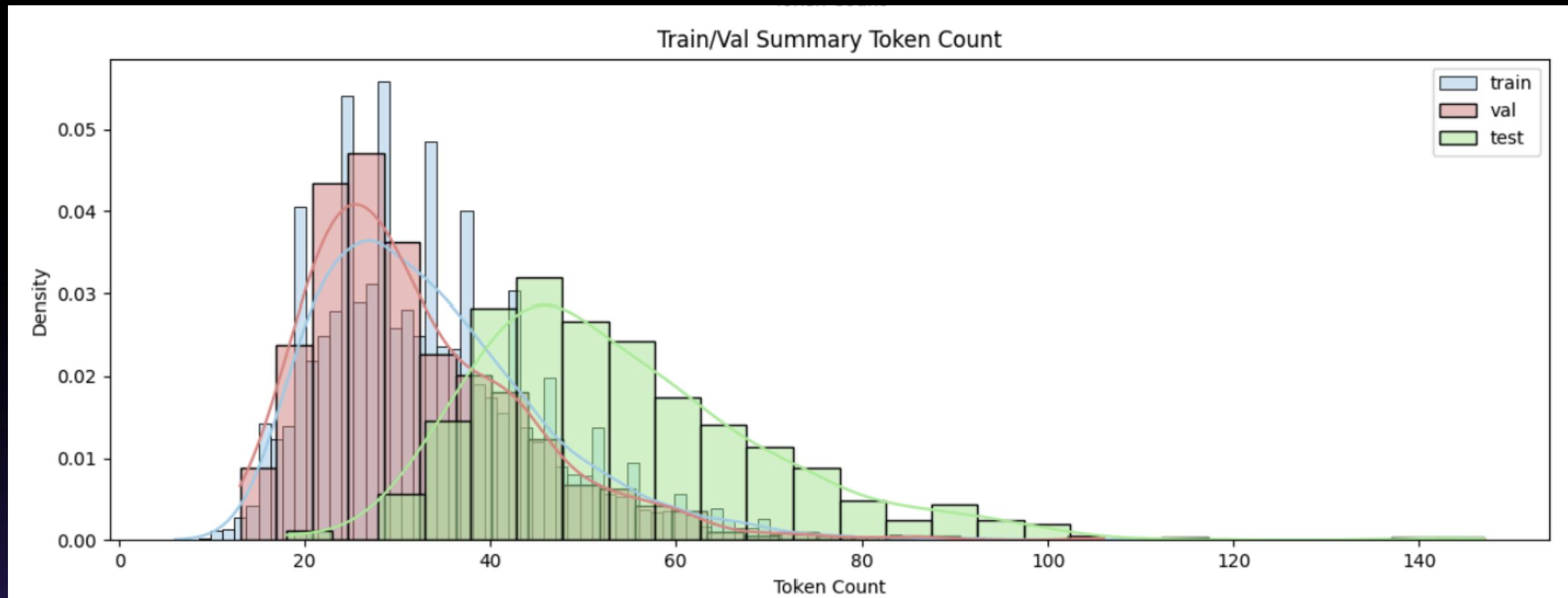


# 경진대회 수행 내용

## Natural Language Processing [대회] Summarization

### [데이터 분석] - Train & Test & Token Count EDA

- [요약문 토큰 수 분포(Train/Val/Test)]
- train/val 데이터의 요약문은 주로 20~40 토큰에 분포
- test 데이터는 더 긴 요약문이 많으며, 40~60 토큰에 집중, 100개 이상 토큰도 일부 존재

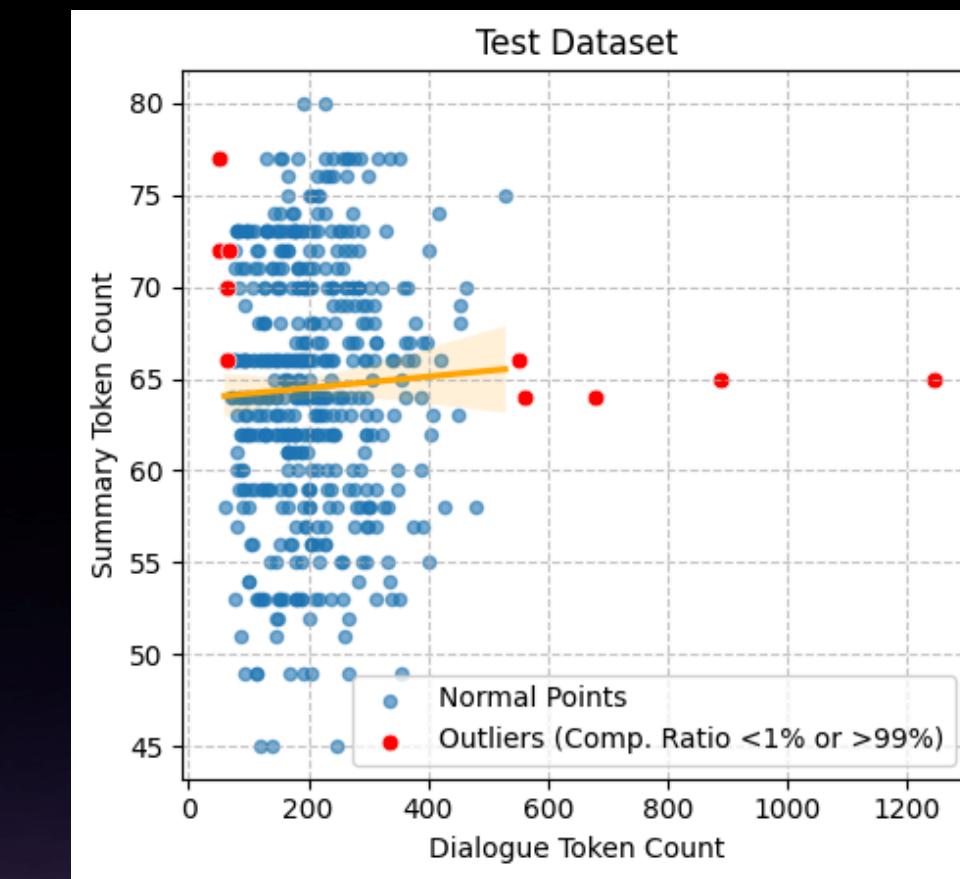
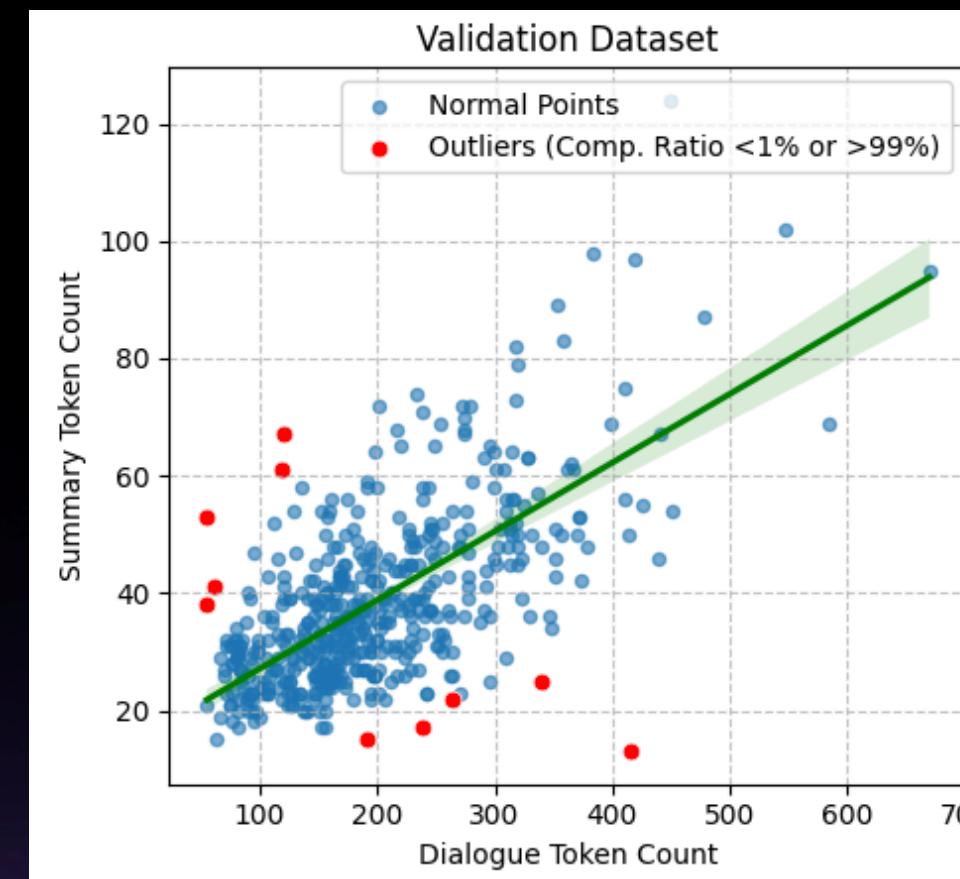
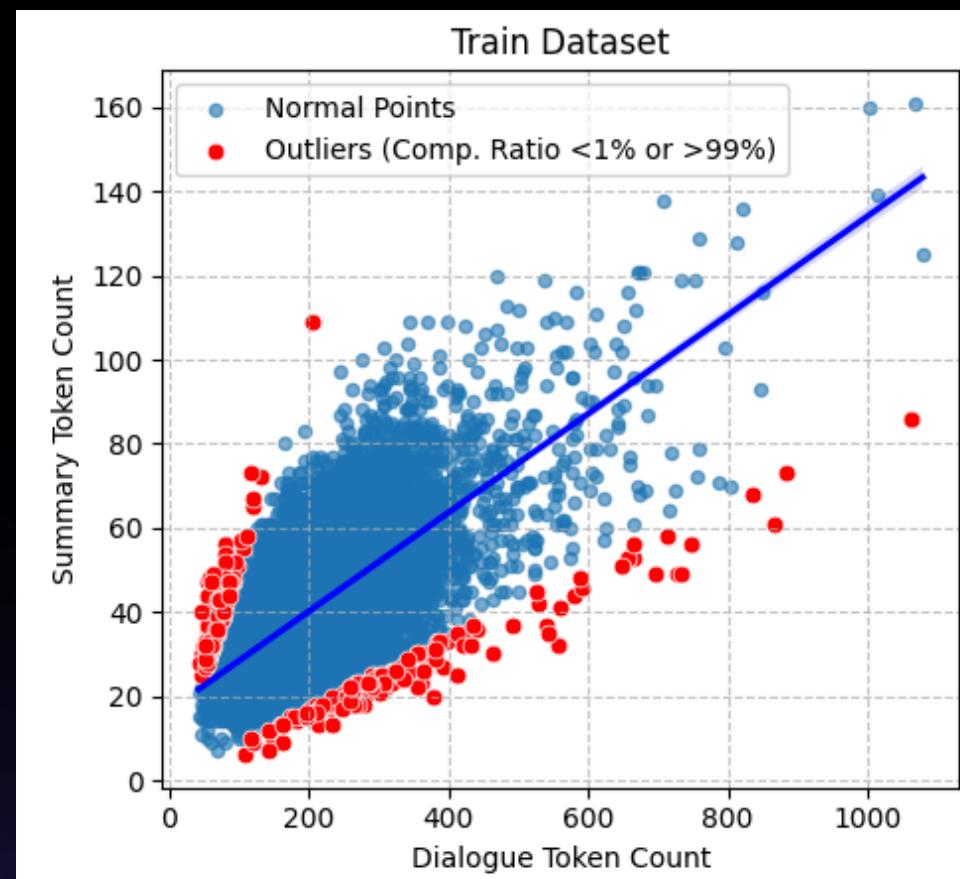


# 경진대회 수행 내용

## Natural Language Processing [대회] Summarization

### [데이터 분석] - Dialogue와 Summary의 길이 상관관계 EDA

- train, val은 상관계수 0.66, 0.64로 유의미한 관계가 있고, test는 유의미한 관계가 없음



# 경진대회 수행 내용

## Natural Language Processing [대회] Summarization

### [데이터 전처리]

- **지시표현 보완:**
  - 대화문 내 “그 사람”, “이것” 등 지시어를 직전 발화자 정보로 치환하여 맥락 해소
- **텍스트 정제:**
  - \\n, <br> 등 줄바꿈 표현을 모두 \n으로 통일
  - 특이 케이스 'ㅎ ㅎ'는 문맥에 맞게 '나도 행복해.'로 대체
  - 자음/모음 약어(ㅋㅋ, ㅇㅋ, ㅜㅜ 등) 제거
  - 중복 줄바꿈 및 중복 공백 제거
  - 기타 불필요한 태그, 특수문자, 이모티콘 등 정리
- **프롬프트 추가:**
  - 토픽 등 메타정보 프롬프트(#Topic#, #Dialogue#) 자동 삽입 (special token 연동)
  - 학습/테스트셋 변환:
  - 필요한 컬럼만 추출, BART/KoBART 입력 포맷(bos\_token, eos\_token)에 맞게 변환
  - 인코더/디코더 입력, 디코더 라벨 구성

```

# 텍스트 클린 함수
def clean_text(text: str) -> str:
    if not isinstance(text, str):
        return ""

    # 줄바꿈 표현 통일
    text = text.replace("\n", "\n").replace("<br>", "\n").replace("</s>", "\n")

    ### 특이 케이스 : train.csv에는 'ㅎㅎ'가 오직 1개 존재한다. 그런데 이것이 #Person2#: ㅎㅎ 라서 빈문자열로 대체된다.
    # 문맥과 summary에 맞춰 '나도 행복해.'로 바꾼다.
    text = text.replace("ㅎㅎ", "나도 행복해.")

    # 자소만 있는 단어 제거 (예: ㅋㅋ, ㅇㅋ, ㅜㅜ) > 이모티콘
    text = re.sub(r"\b[ㄱ-ㅎㅏ-ㅣ]{2,}\b", "", text)

    # 중복 줄바꿈 제거
    text = re.sub(r"\n+", "\n", text)

    # 중복 공백 제거
    text = re.sub(r"\t+", ' ', text)

    def add_instructions(row:pd.Series) -> pd.Series:
        """지시어 프롬프트 추가.

        :param str dialogue: _description_
        :return str: _description_
        """

        topic = str(row['topic']).strip()
        dialogue = row['dialogue']
        dialogue = f"#Topic#{topic}\n#Dialogue#{dialogue}"
        row['dialogue'] = dialogue
        ##Topic#, '#Dialogue#', '#Summary#', '#SEP#'

        return row

```

# 경진대회 수행 내용

## Natural Language Processing [대회] Summarization

### [데이터 증강]

- solar api를 사용하여, 기존의 학습 데이터셋을 한 → 영 → 일 → 한 back translation을 통해 2배로 증강

#### 기존 데이터

train\_136,"#Person1#: 진짜야?  
#Person2#: 그런 것 같아. 안나가 이제 아이가 네 명인데 또 임신했대.  
#Person1#: 와, 아이가 그렇게 많은데도 정말 멋있어 보여. 임신 중인데도 프라다 옷을 입고  
있잖아!  
#Person2#: 그게 바로 안나지. 스타일리시한 이탈리아 여성의 이미지를 잘 유지하고 있네.  
저기 온다.  
#Person1#: 빛이 나오고 있어. 임신해서 그런가 봐.  
#Person2#: 응, 아니면 비싼 이탈리아식 피부관리 덕분일지도 몰라.",#Person1#와  
#Person2#는 임신했지만 여전히 패셔너블한 안나에 대해 이야기한다.,임신과 패션

#### back translation

train\_136,"#Person1#: 그게 정말이에요?  
#Person2#: 그런 것 같아. Anna는 지금 아이가 4명이고, 또 임신 중이야.  
#Person1#: 대단하네, 그렇게 많은 아이가 있어도 멋진 모습이야. 임신 중에 Prada를  
입다니!  
#Person2#: 그게 바로 Anna야. 그녀는 정말 스타일리시한 Italian 여성 이미지를 유  
지하고 있어. 저기 그녀가 오네.  
#Person1#: 빛나네요. 임신 때문이겠죠.  
#Person2#: 그렇겠지, 아니면 그녀의 고급 Italian 스킨케어 덕분일 수도 있겠  
네.",#Person1#와 #Person2#는 임신했지만 여전히 패셔너블한 안나에 대해 이야기  
한다.,임신과 패션

# 경진대회 수행 내용

## Natural Language Processing [대회] Summarization

### [모델링] KoBART 기반 대화 요약 모델 학습

- 모델/토크나이저 로딩
  - HuggingFace BartForConditionalGeneration, AutoTokenizer 활용
  - config에서 지정한 사전학습 모델명(digit82/kobart-summarization 등) 사용
  -
- 특수토큰(Special Token) 처리
  - #Person1#, #PhoneNumber# 등 대화문 마스킹 토큰을 tokenizer에 추가 등록
  - 모델 임베딩 크기를 special token 수에 맞춰 재조정
- 파인튜닝
  - 전처리된 대화문과 요약문을 모델 입력/라벨로 변환
  - 전체 파라미터 full-finetuning

```
special_tokens_dict={'additional_special_tokens':config['tokenizer']['special_tokens']}
tokenizer.add_special_tokens(special_tokens_dict)

generate_model.resize_token_embeddings(len(tokenizer)) # 사전에 special token을 추가했으므로 재구성
generate_model.to(device)
print(generate_model.config)
```

# 경진대회 수행 내용

## Natural Language Processing [대회] Summarization

### [추론] – test셋 요약문 생성 및 제출

- 데이터 전처리 및 토크나이즈  
test셋 대화문 전처리(줄바꿈, 클린, 프롬프트 삽입 등)  
tokenizer로 max\_length 등 맞춰 입력 토크나이즈
- 요약문 생성  
DataLoader로 batch inference  
모델 generate 함수로 beam search, max\_length, early\_stopping 등  
파라미터 적용  
각 샘플별 생성 결과를 tokenizer로 decode
- 후처리 및 제출 파일  
스페셜 토큰, 중복 공백 등 불필요 문자 제거  
결과를 DataFrame에 저장, 제출용 CSV 파일 생성

```
with torch.no_grad():
    for item in tqdm(dataloader):
        text_ids.extend(item['ID'])
        generated_ids = generate_model.generate(input_ids=item['input_ids'].to(device),
                                                no_repeat_ngram_size=config['inference']['no_repeat_ngram_size'],
                                                early_stopping=config['inference']['early_stopping'],
                                                max_length=config['inference']['generate_max_length'],
                                                num_beams=config['inference']['num_beams'],
                                                length_penalty=config['inference']['length_penalty'],
                                                )
    for ids in generated_ids:
        result = tokenizer.decode(ids)
        summary.append(result)
```

# 경진대회 수행 결과

## Natural Language Processing [대회] Summarization

### [대회 성적]

- 중간 평가에서는 9팀 중 3위를 기록했지만, 최종 평가에서는 5위 순위 하락

Leaderboard [mid]		Leaderboard [final]	
Rank	Team Name	Team Member	
The leaderboard provided during the competition is a result of scoring using part of the evaluation dataset.			
			Last update: 2025.08.07 11:15:54
rouge1	rouge2	rougeL	final_result
My Rank 3	NLP 5조	AI 은별	0.5758 0.3824 0.4922 48.3465 85 1d
1	NLP_7조 ♂	j	0.5999 0.4165 0.5337 51.6701 18 21h
2	NLP_1조 ♀	준석 국현 승현	0.5827 0.3986 0.5095 49.6957 56 18h
3	NLP 5조 ♂	AI 은별	0.5758 0.3824 0.4922 48.3465 85 1d

Leaderboard [mid]		Leaderboard [final]						
Rank	Team Name	Team Member	rouge1	rouge2	rougeL	final_result	Entries	Final
My Rank 5	NLP 5조	AI 은별	0.5535	0.3487	0.4654	45.5898	85	1d
1	NLP_7조 ♂	j	0.6030	0.4083	0.5258	51.2398	18	21h
2	NLP_1조 ♀	준석 국현 승현	0.5582	0.3613	0.4769	46.5497	56	18h
3	NLP_2조 ♂	승민 상원	0.5527	0.3498	0.4709	45.7824	60	17h
4	NLP_8조 ♂	정민 J	0.5548	0.3457	0.4692	45.6567	58	1d
5	NLP 5조 ♂	AI 은별	0.5535	0.3487	0.4654	45.5898	85	1d

03

## 분석 인사이트 및 결과

---

문제 및 인사이트 도출  
해결 방법 및 결과

# 경진대회 인사이트 공유

## Natural Language Processing [대회] Summarization

### [데이터 전처리]

- 데이터에 \\n이나 <br>이 있는 부분과 ㅎ ㅎ 확인
- 발언을 중단시키는 ...이 많음
- 대괄호([])로 감싸진 문장들이 있는데, 대부분은 그 문장을 해당 발화자의 상대방 화자의  
발화로 인식하여 분리 후 새로운 발화 줄로 할당해야 하는 모습이 보임



- 전처리 후 : 성능 점수 하락



- 인사이트 : 데이터 중 전처리 해야할 부분들이 보이지만,  
전처리를 할수록 성능이 떨어짐

# 경진대회 인사이트 공유

## Natural Language Processing [대회] Summarization

### [데이터 증강]

- back translation을 통해 훈련 데이터셋을 2배로 증강 후 점수 1점 상승



- 증강 전

rouge1 (Final)	rouge2 (Final)	rougeL (Final)	final_result (Final)
0.5686	0.3703	0.4794	47.2777
0.5478	0.3373	0.4536	44.6236

- 인사이트 : 데이터 증강의 효과가 큼

- 증강 후

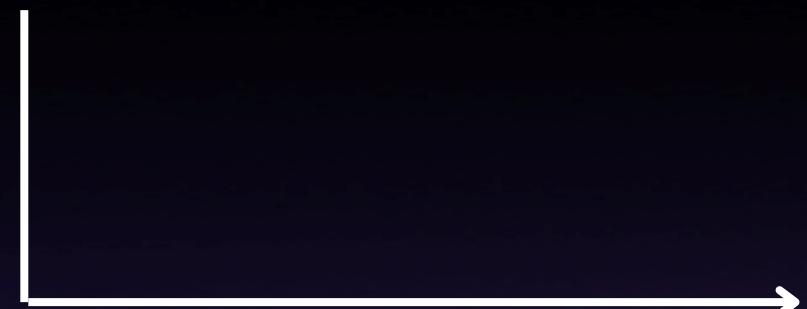
rouge1 (Final)	rouge2 (Final)	rougeL (Final)	final_result (Final)
0.5758	0.3824	0.4922	48.3465
0.5535	0.3487	0.4654	45.5898

# 경진대회 인사이트 공유

## Natural Language Processing [대회] Summarization

### [모델링 및 실험 분석]

- Encoder max\_length
  - 1026보다 512로 설정 시 ROUGE 점수 더 높게 나옴
  - 핵심: 대화문 앞부분 중심의 정보만 활용하는 것이 요약 성능에 유리
  -
- Decoder max\_length
  - 200으로 설정했을 때 정보 손실 없이 긴 요약문도 충분히 생성 가능
  - 너무 짧으면 누락, 더 길면 불필요한 문장 증가 → 200이 최적
  -
- Beam Search
  - num\_beams=2에서 성능 및 속도 모두 가장 안정적
  - 3 이상에서는 오히려 점수 하락/속도 저하 발생



- 인사이트
  - 불필요하게 배치 크게, patience 길게, decoder 길이 짧게, beam depth 줄이기  
→ 실성능에서는 큰 개선 없음
  - 오히려 데이터셋 및 task 특성에 맞는 적정/보수적 하이퍼파라미터가 더 우수

04

## 회고

---

우리 팀의 목표 달성도  
느낀점 및 향후 계획

# 경진대회 회고

## Natural Language Processing [대회] Summarization

Point 1

### 우리 팀의 처음 목표에서 어디까지 도달했는가

목표했던 바를 어느 정도 이뤄냈기에 80%정도 도달했다고 생각한다.  
데이터 증강, 모델 실험 등 최대한 많은 시도를 했다.

Point 2

### 우리 팀이 잘했던 점

다음 회의 때까지의 각자의 역할을 잘 분배해서 진행했고, TODO 리스트를 작성하여 공유해놓고 해당 결과를 기록해놓았다.  
서로 의견 충돌을 하며 더 좋은 아이디어를 추출해낼 수 있었고, 이를 바탕으로 멘토링 시간에 양질의 정보를 얻었고 많은 시도를 할 수 있었다.

Point 3

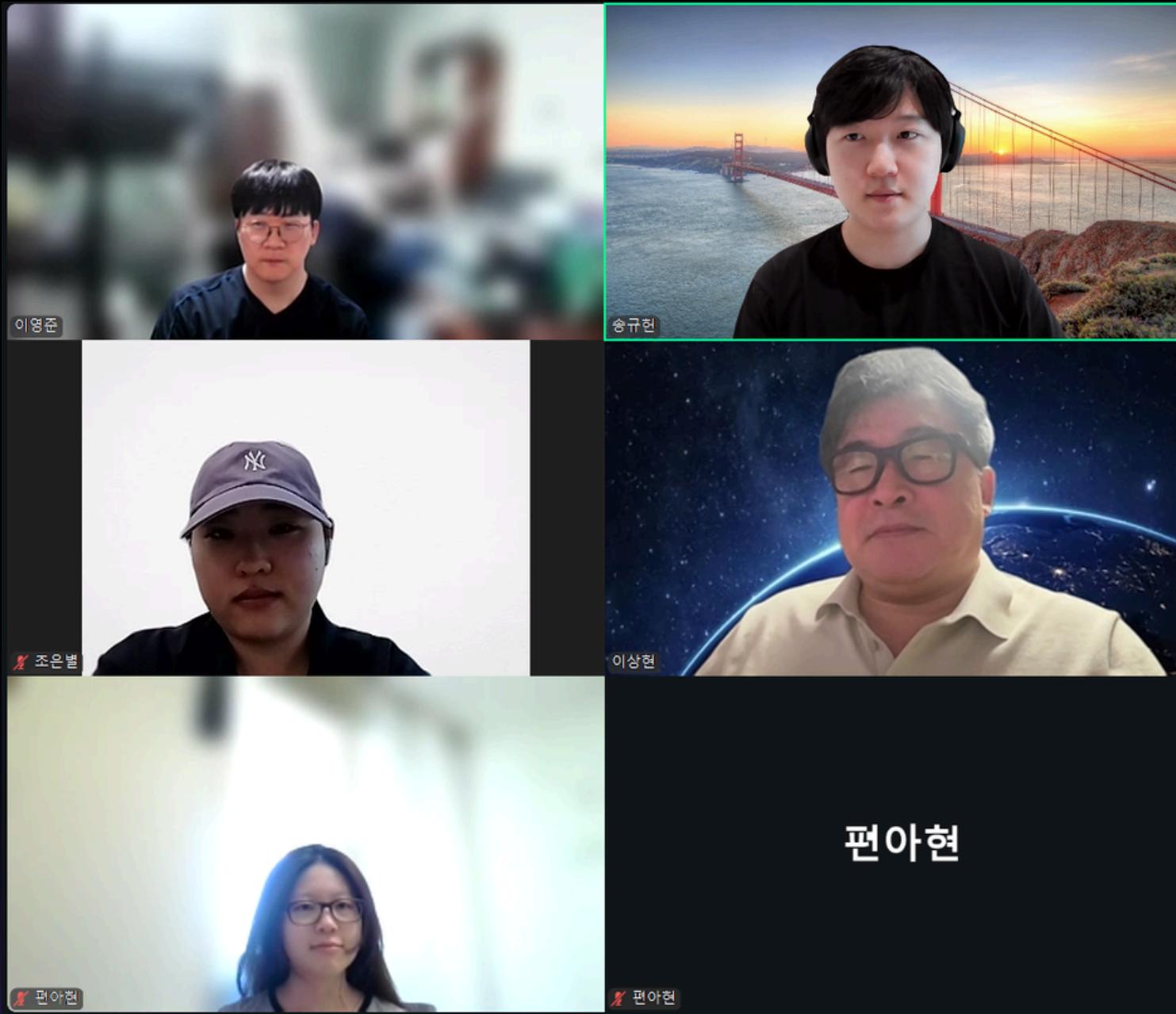
### 협업하면서 아쉬웠던 점

마지막 날에 시간이 더 많았다면 점수를 더 올릴 수 있었지 않을까 싶다.

- 향후 계획 : 다음에도 이런 대회를 참여한다면 이번 인사이트를 기반으로 더 높은 점수를 노려보고 싶다.

# 경진대회 진행 소감

## Natural Language Processing [대회] Summarization



송규현

CV 경진대회와 마찬가지로 코드를 개선하는 데 시간을 많이 소요해 요약 task 논문 조사를 많이 하지 못한 것이 아쉽다. DialogSum 논문에 기반해 계획은 잘 수립했지만, 또 코드 때문에 계획대로 하지 못했던 것 같다.

이상현

요약이라는 Task를 딥러닝으로 해결하는 기법을 강의로 듣고 AI의 탄생의 적용이라는 주제를 깊이 고민할 수 있었다. 2인 대화록에서 요약을 끌어내는 Task란 인간만이 가진 함축적 사고력의 결과를 담아야 하는 주제라 심도있는 문제라는 걸 배우게 되었는데, 이 회고록을 작성하는 순간까지는 아직 그 해결방법을 알지 못해, 발표대회가 궁금하다.

이영준

자동화를 구축하여 다양한 실험을 하려고 하였습니다. 두 대형 모델을 사용해서 실험을 하려고 며칠을 고생하면서 했지만, 끝내 Claude MCP가 해결 못하는 부분이 있어서 해당 작업은 포기하고, 조에서 공동으로 사용하는 것으로 하게 되었습니다. 오랜시간 공들인 것이 허사가 된 부분이 있어서 아쉽습니다.

조은별

Solar API로 프롬프트 기반 Back Translation을 직접 실험해보면서, 프롬프트의 작은 차이만으로도 결과 품질이 크게 달라진다는 점을 실감했다. 모델이 다양한 대화와 요약 데이터를 학습해 자동으로 요약을 생성하는 과정에서, 정확하고 구조화된 프롬프트가 성능에 매우 중요하다는 점을 확인했다. 여러 실험을 통해 데이터 다양성, 프롬프트 설계, 하이퍼파라미터 튜닝이 성능 개선의 핵심임을 다시 한 번 느꼈다.

편아현

전처리를 통해 더 좋은 결과를 기대했지만, 오히려 성능이 떨어져서 아쉬웠습니다. 사람이 보기에 더 자연스럽도록 전처리한 게 잘못됐던 것인지, 성능 저하의 원인이 아직도 의문입니다. 우리 팀을 포함한 다른 팀들도 멘토링 이후에 갑자기 치고 올라왔는데, 확실히 멘토링에서 인사이트를 많이 얻어가는 것 같습니다. 잠깐이었지만 1등을 경험해볼 수 있어 좋았습니다.

Life-Changing Education

감사합니다.

---