
Korean MRC Question Answering

1조 김소연 김지수 안희진 정영빈

2021.11.16.

2nd MRC Project

Table of Contents

1. 프로젝트 개요
2. 프로젝트 팀 구성 및 역할
3. 프로젝트 진행 프로세스
4. 프로젝트 결과
5. 자체 평가 및 보완

목표/기대효과

한국어 지문을 보고 질문에 맞는 답을 생성하는 모델 설계

Frameworks & Libraries

 PyTorch

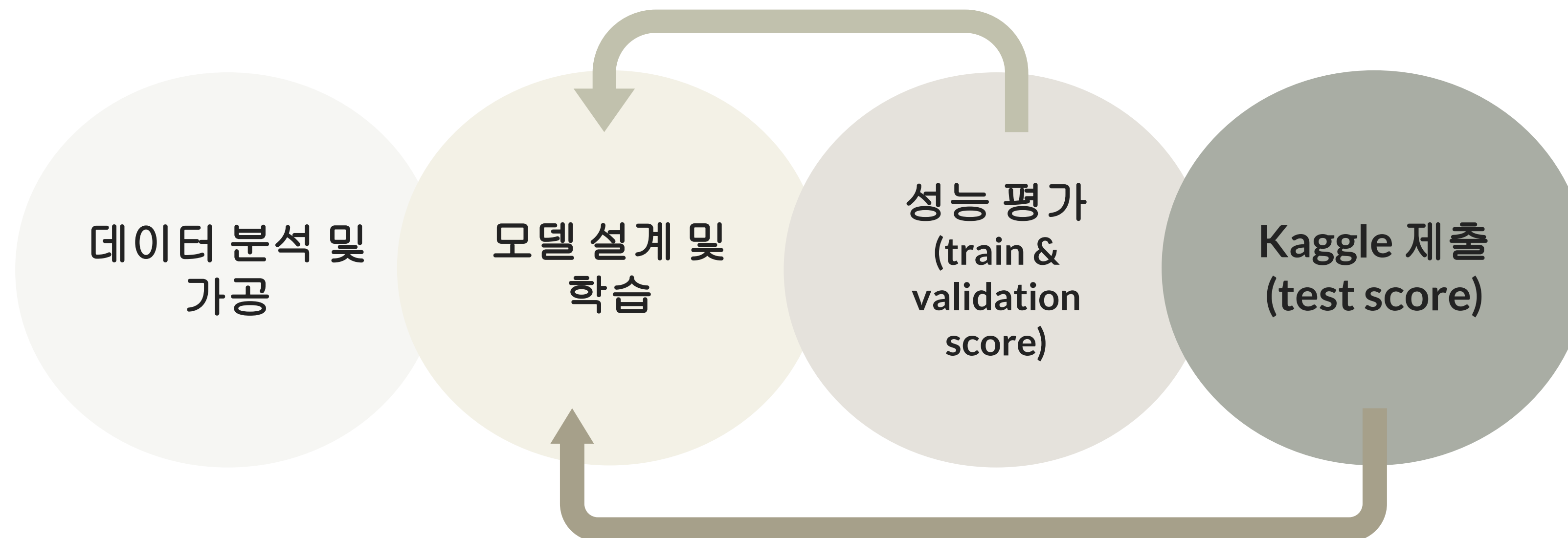


Transformers



Weights
& Biases

진행 프로세스



훈련생	역할	담당 업무
김소연	팀장	<ul style="list-style-type: none">● EDA 및 Pre-processing● Data Augmentation● Learning rate scheduler 추가● Post-processing● 발표 자료 제작
김지수	팀원	<ul style="list-style-type: none">● Metric(Levenstein, Exact match) 추가● Fine-tuning● Evaluation, Prediction 코드 작성
안희진	팀원	<ul style="list-style-type: none">● test labeling 및 Metric(Levenstein) 측정 코드작성● 자료조사 및 Aihub 자료 제공● 발표자료 제작
정영빈	팀원	<ul style="list-style-type: none">● Baseline 리뷰 및 코드정리● Post-processing● Ensemble

1. EDA 및 Preprocessing - 동일 Context, Question 에 대한 다른 Answers

```
temp_train_df[temp_train_df.duplicated(["Question"], keep=False)].sort_value("contexts")
> 10034rows
```

Contexts

"시간이동"은 미국의 방송사 ABC의 텔레비전 드라마 시리즈 로스트의 시즌 프리미어 에피소드 제목이다.

...

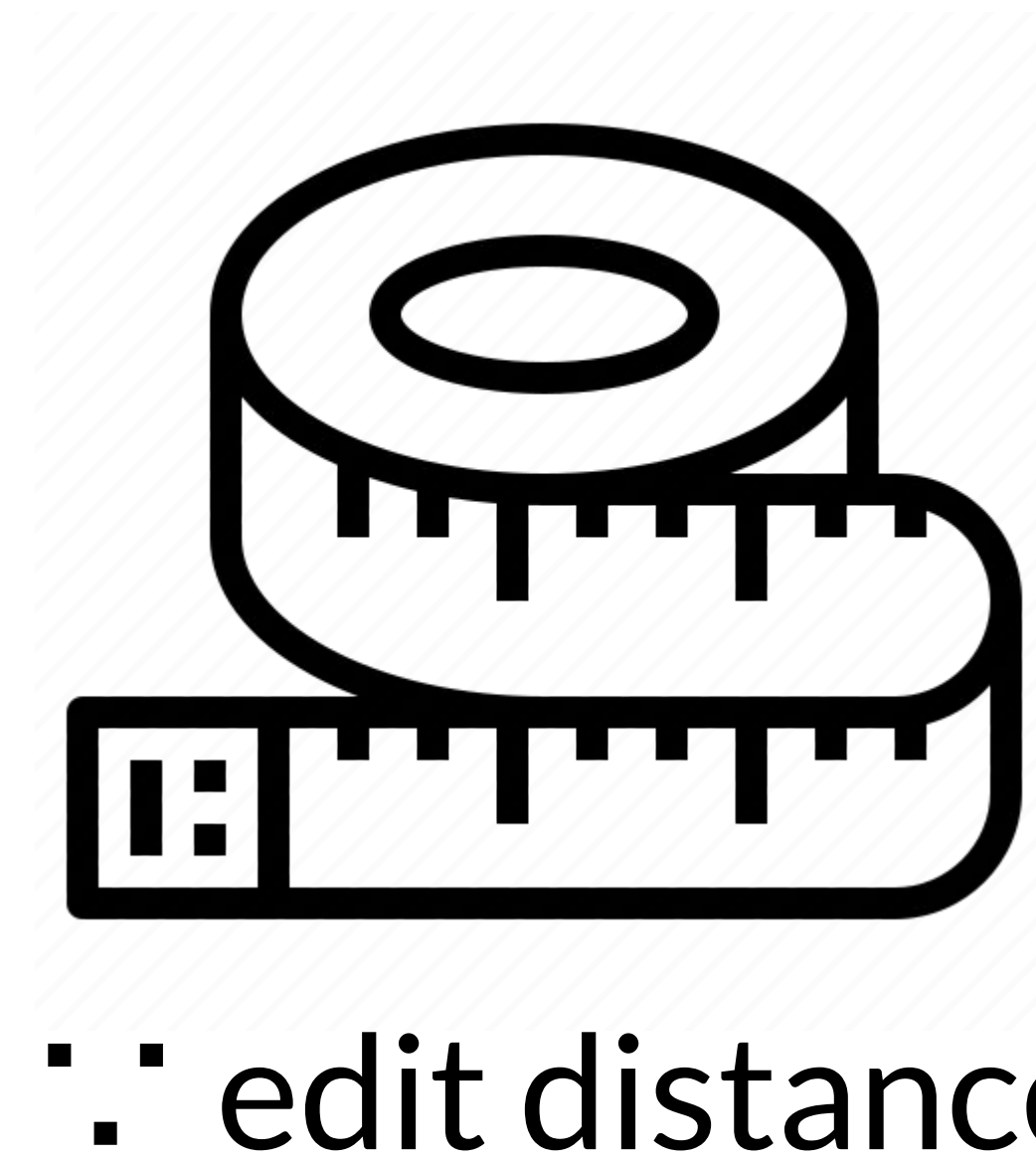
Questions

ABC에서 반영되는 '시간이동'의 장르는?

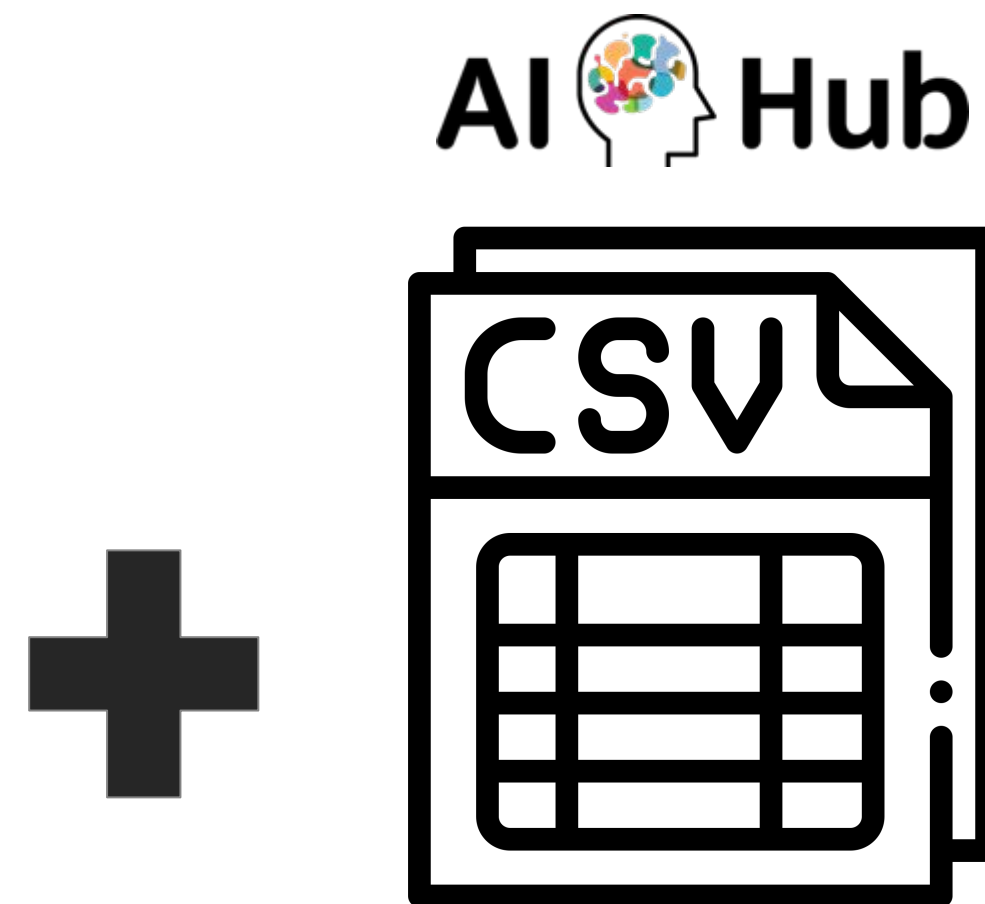
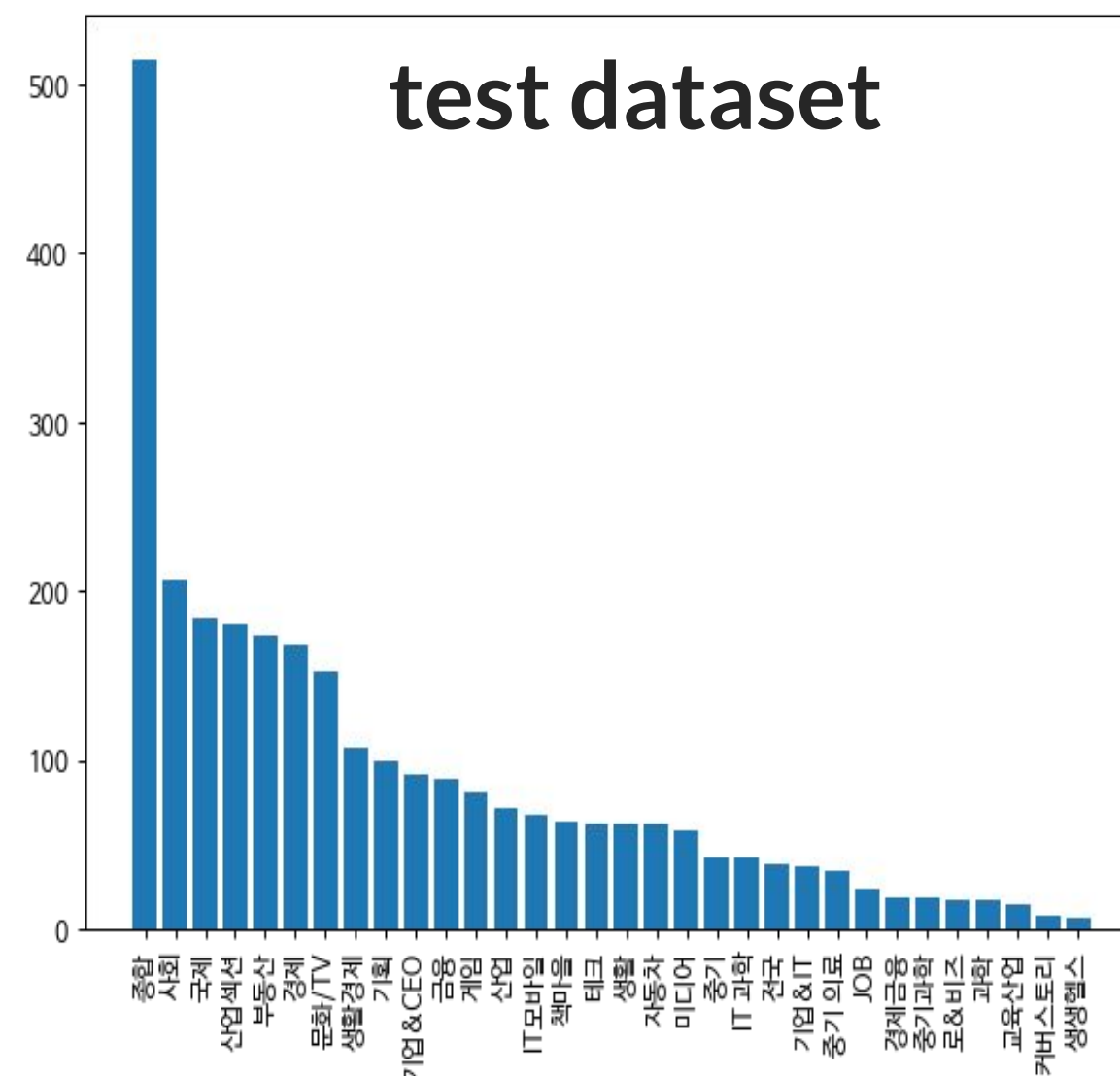
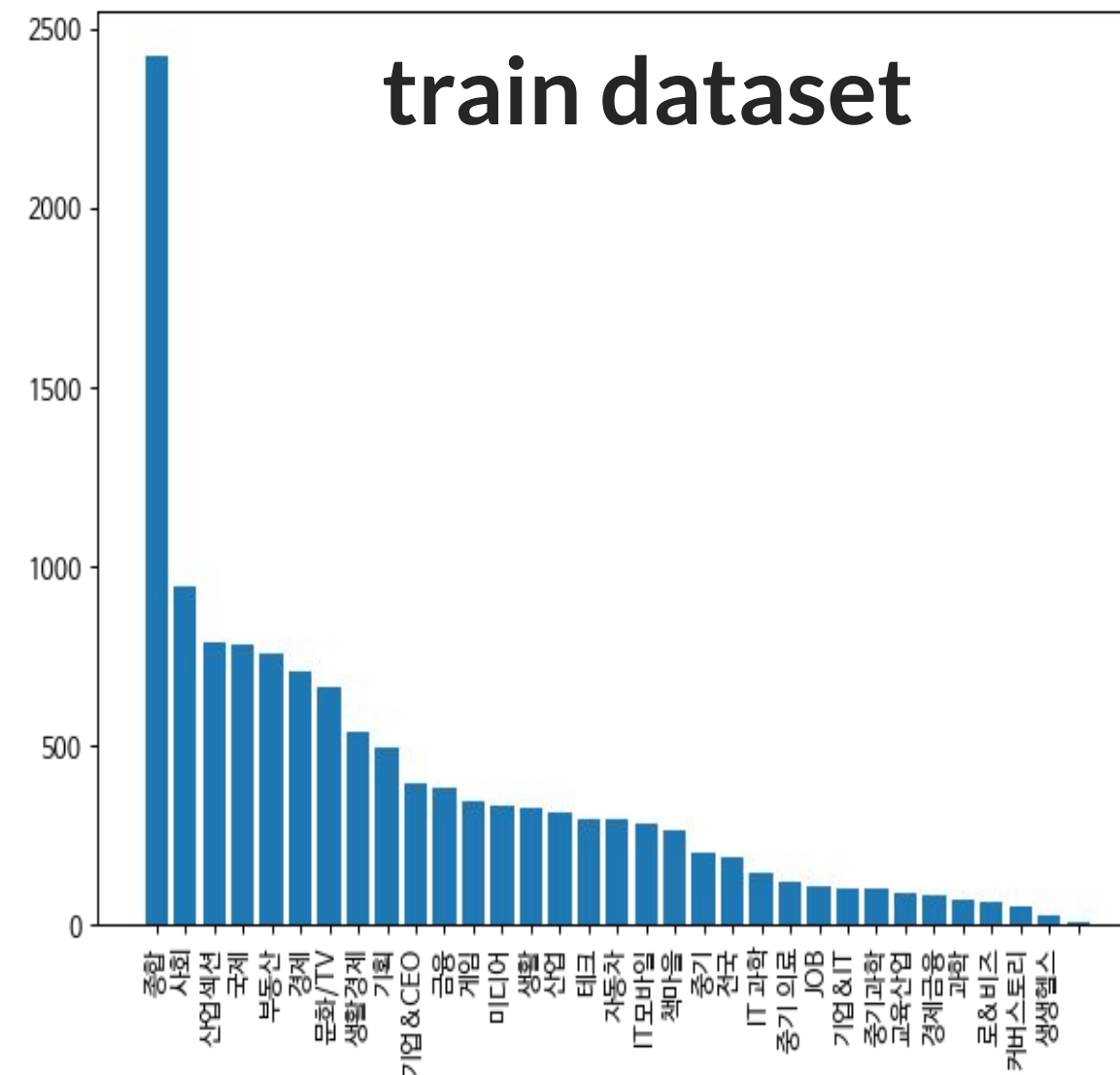
answers

텔레비전 드라마

드라마 ✓

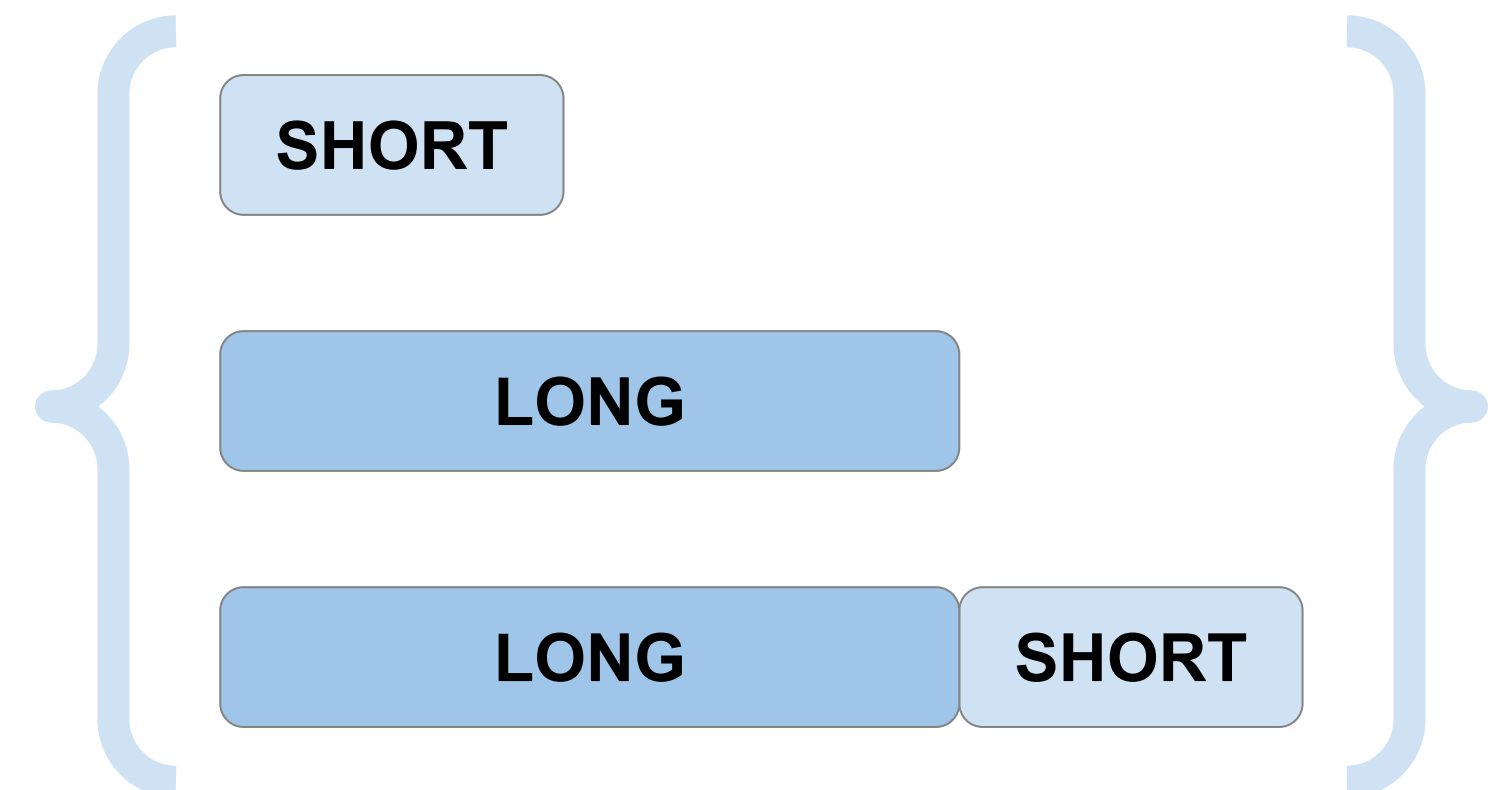
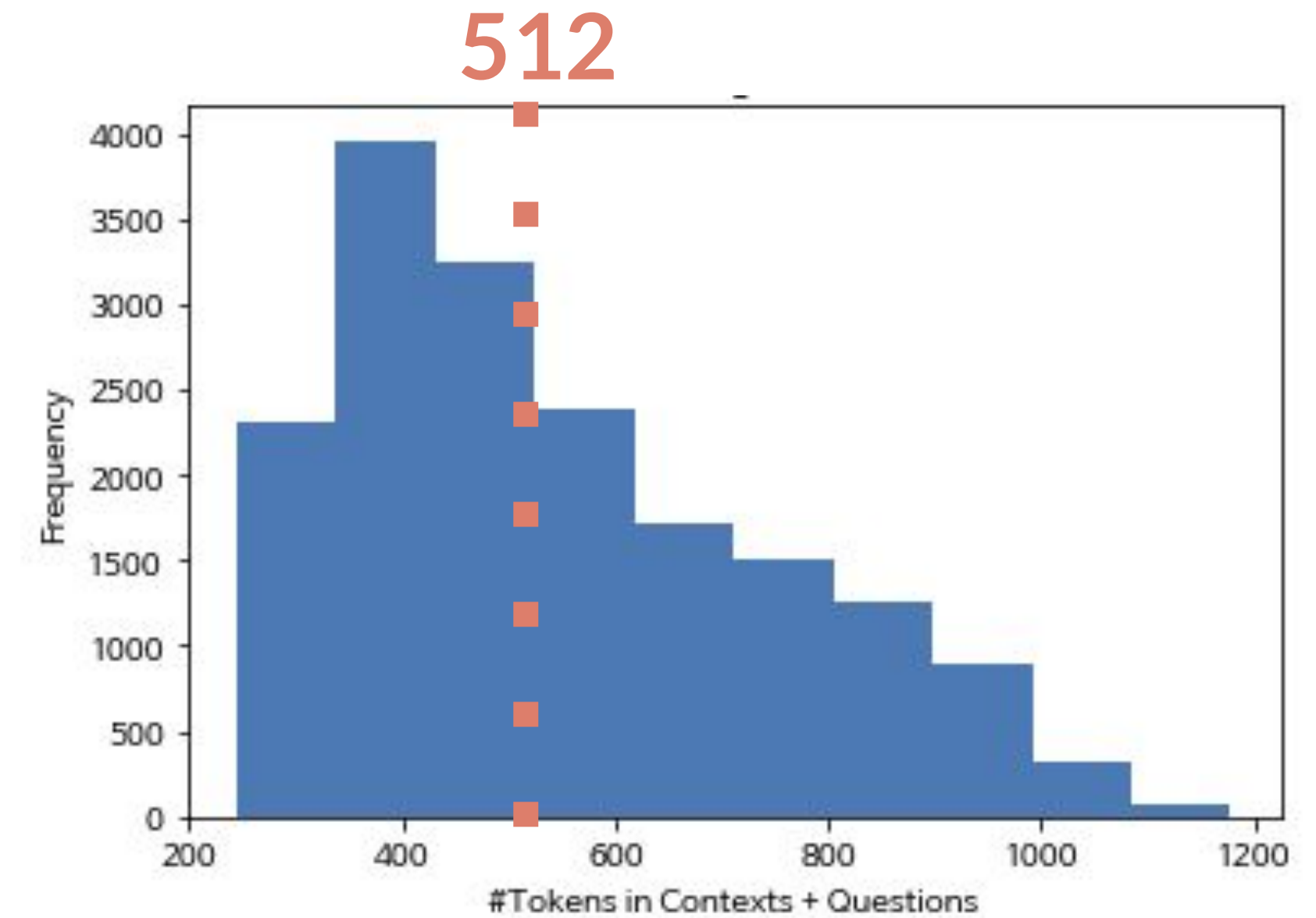


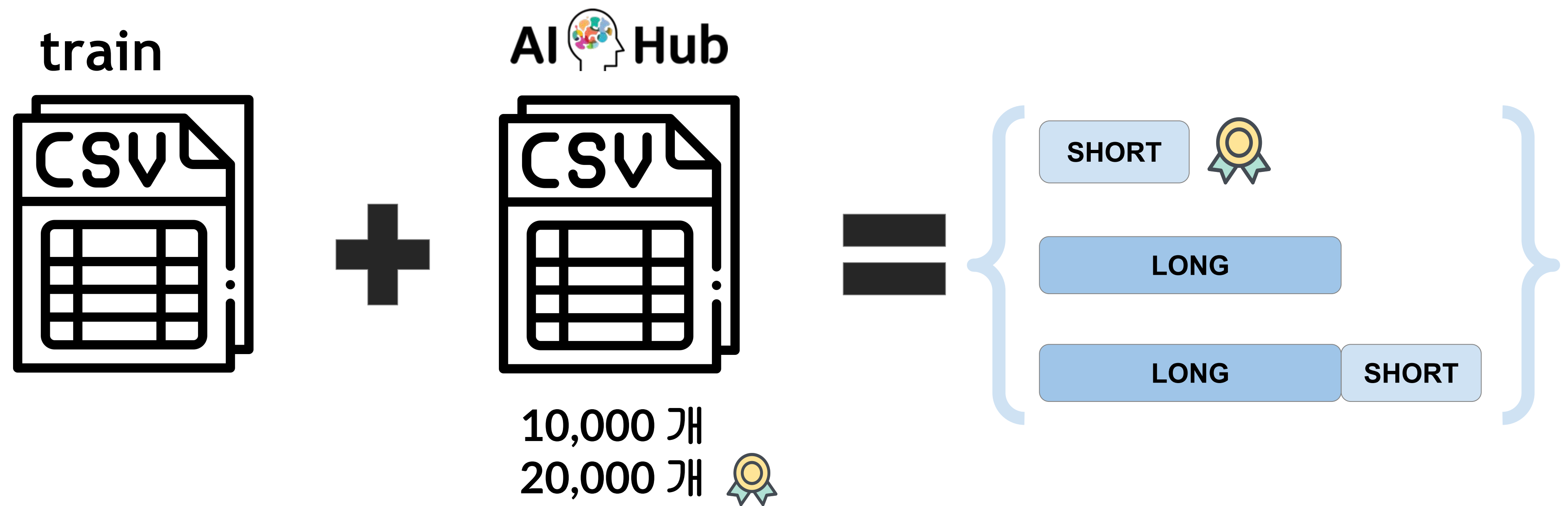
Domain Distribution



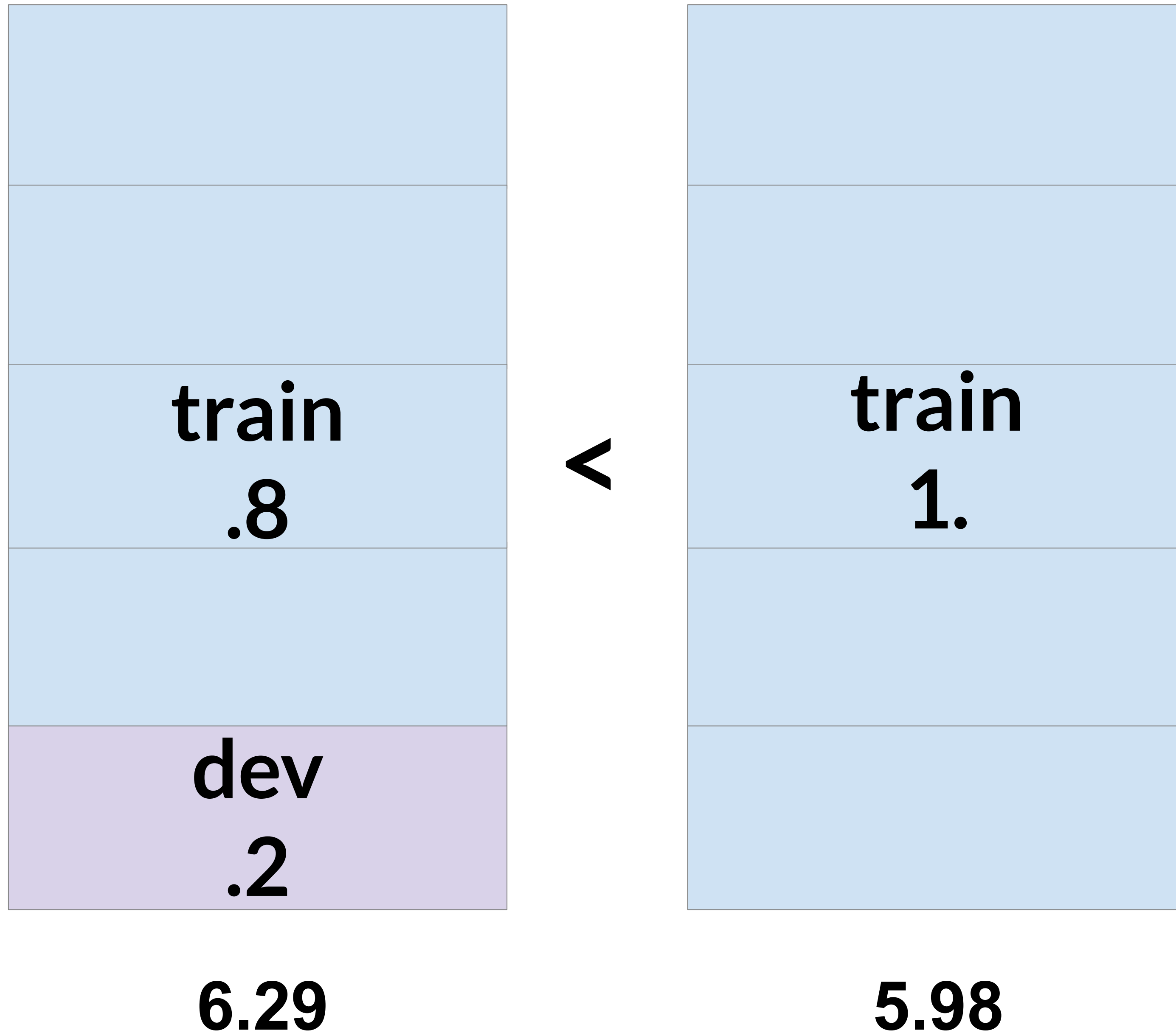
10,000 개
20,000 개

Number of Tokens

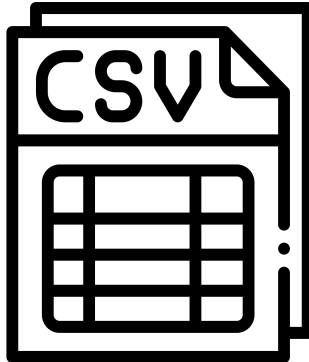
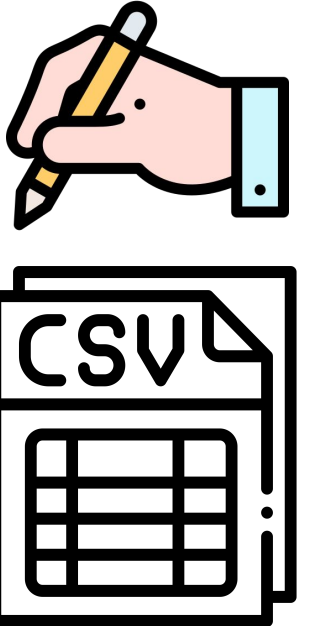





- ① Domain 불일치
- ② Domain Distribution 변화



How to evaluate?

edit_distance( , )

```
def edit_distance(data, test_data):  
    # csv파일 내 Nan값이 있으면 TypeError발생으로 인해 Nan값을 ''으로 채워준다.  
    data.fillna('', inplace=True)  
    test_data.fillna('', inplace=True)  
  
    result = []  
  
    for i in range(len(data)):  
        result.append(jamo_levenshtein(data['Predicted'][i], test_data['Predicted'][i]))  
  
    return np.mean(result) # 전체 평균으로 결과를 리턴한다.
```

 **Hugging Face**


🔍 Search models, datasets, users...

Models 16

🔍 Search Models


Edit filters

Active filters: ko, fill-mask


 klue/bert-base

📄 Fill-Mask • Updated Oct 21 • 36.5k


BERT

 klue/roberta-large

📄 Fill-Mask • Updated Oct 21 • 25.7k • ❤️ 1


 kykim/bert-kor-base

📄 Fill-Mask • Updated May 20 • 15.4k



 klue/roberta-base

📄 Fill-Mask • Updated Oct 21 • 5.79k


RoBERTa

 klue/roberta-small

📄 Fill-Mask • Updated Oct 21 • 4.68k

Metric (levenshtein)	KLUE/BERT-base	KLUE/RoBERTa-base
train.json	 5.98 vs 6.86	
train.json +Aihub.json	14.80 vs 11.22 	

Levenshtein Distance (Lower is better)



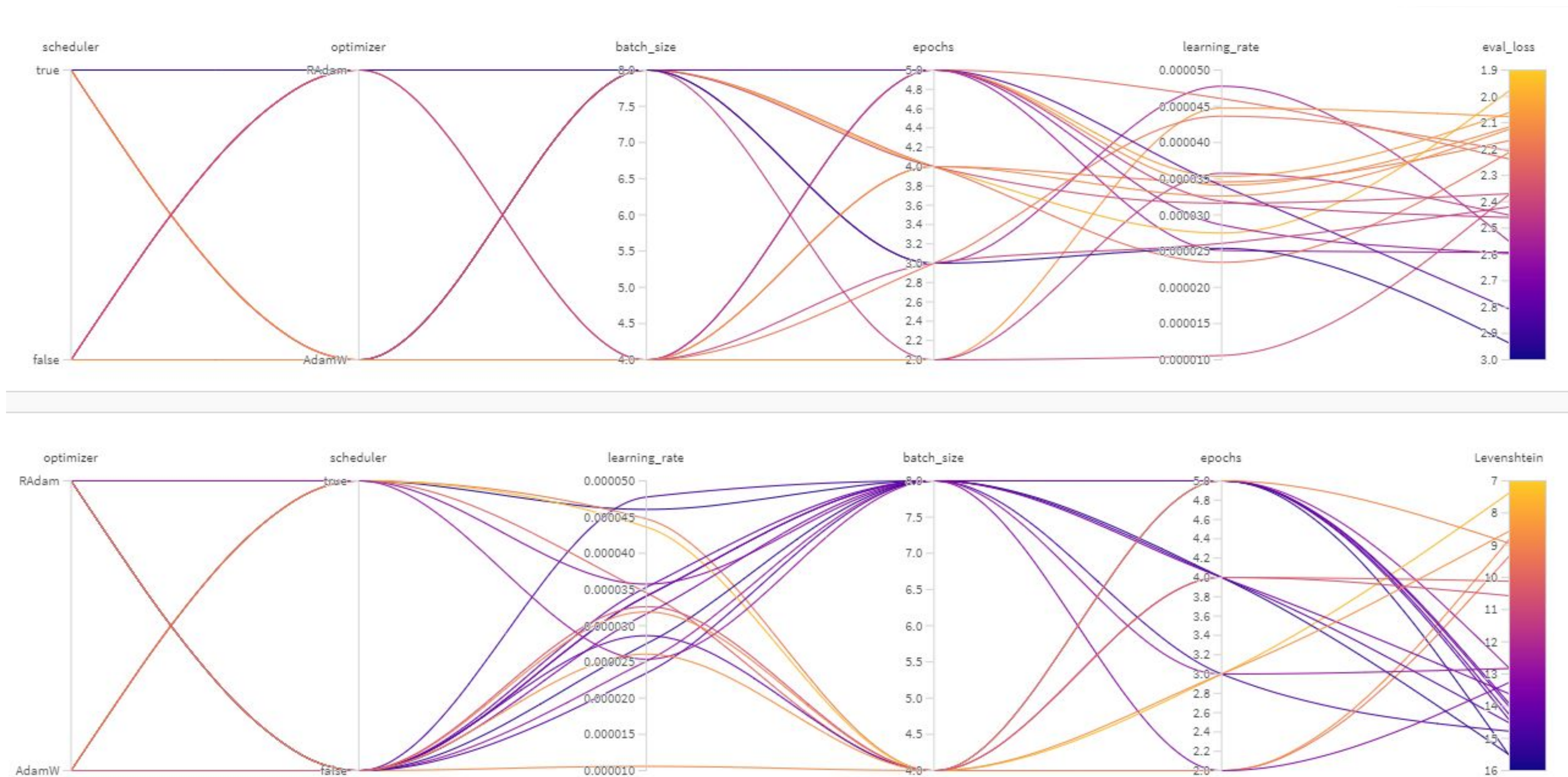
Model	Levenshtein Distance
klue/roberta-base	~14.2
klue/bert-base	~12.2

WandB Sweep
Configuration

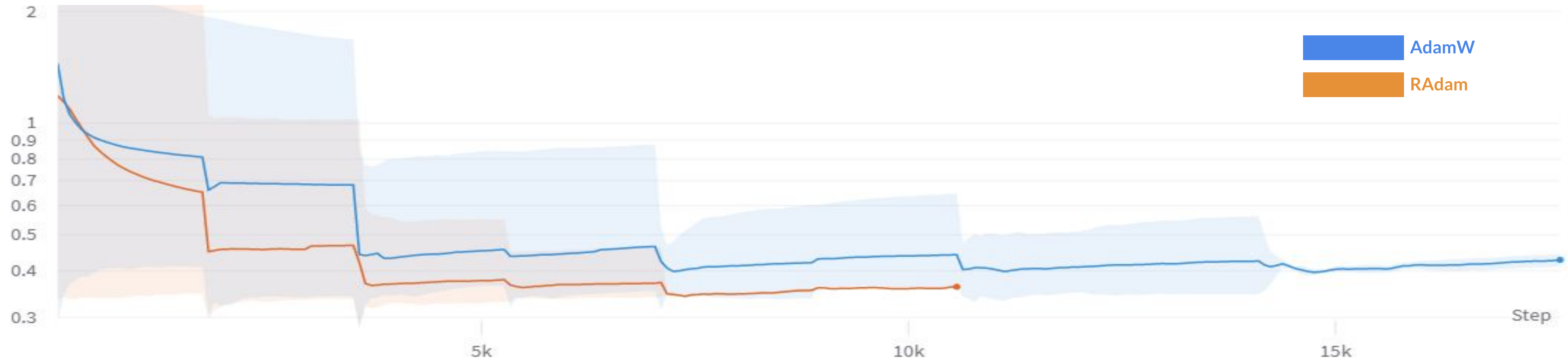
```

method: random
name: bert-base-2
parameters:
  batch_size:
    values:
      - 4
      - 8
  epochs:
    values:
      - 2
      - 3
      - 4
  learning_rate:
    distribution: uniform
    max: 5e-05
    min: 1e-05
  optimizer:
    values:
      - RAdam
      - AdamW
  scheduler:
    values:
      - true
      - false

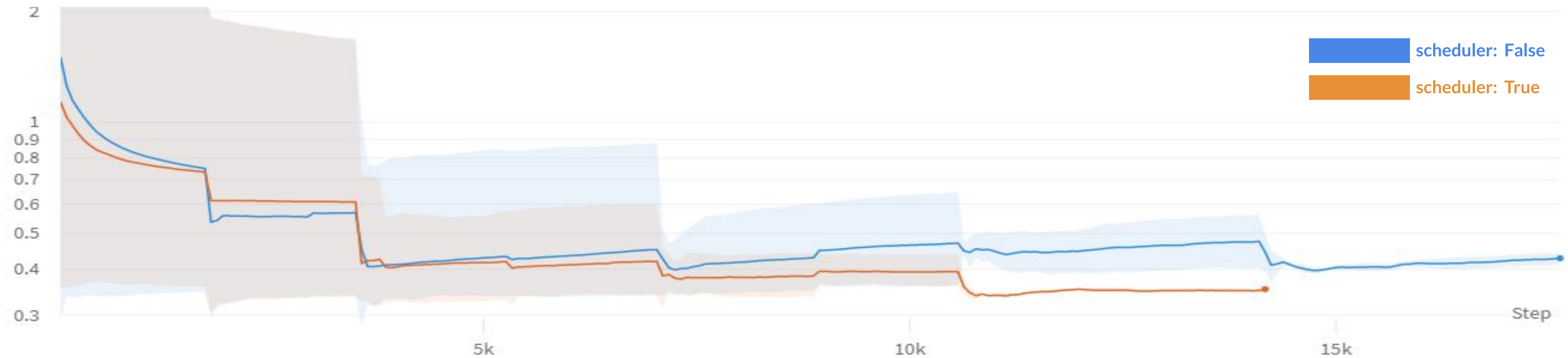
```



Optimizer에 따른 loss 평가

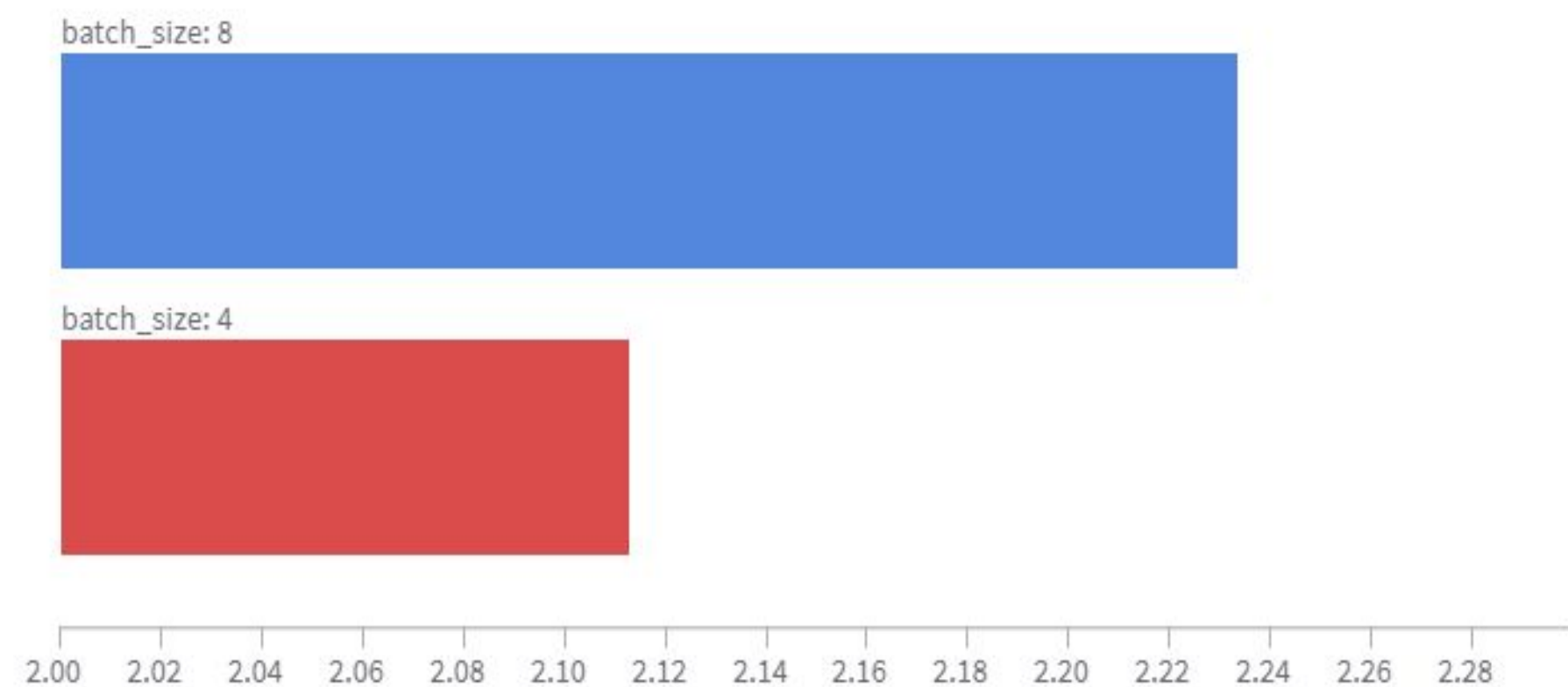


lr Scheduler 유무에 따른 loss 평가

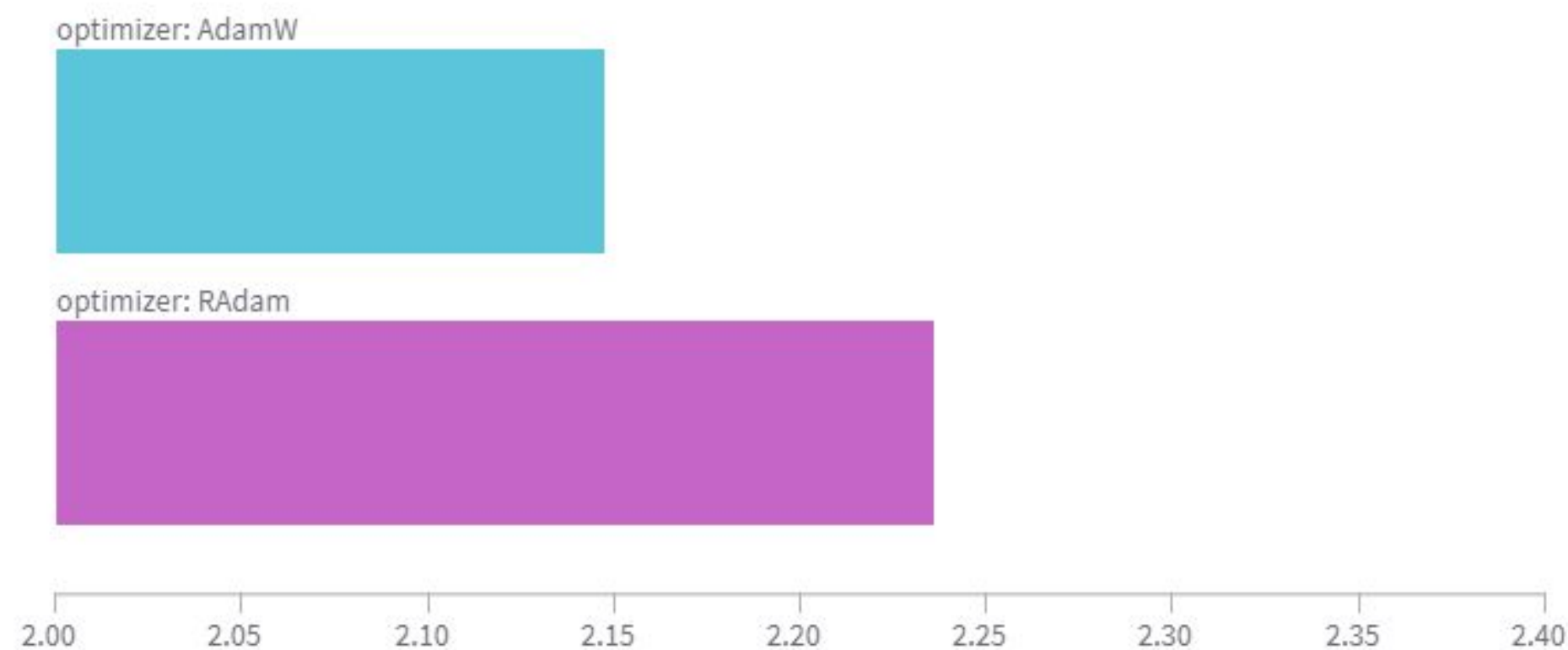


① Loss

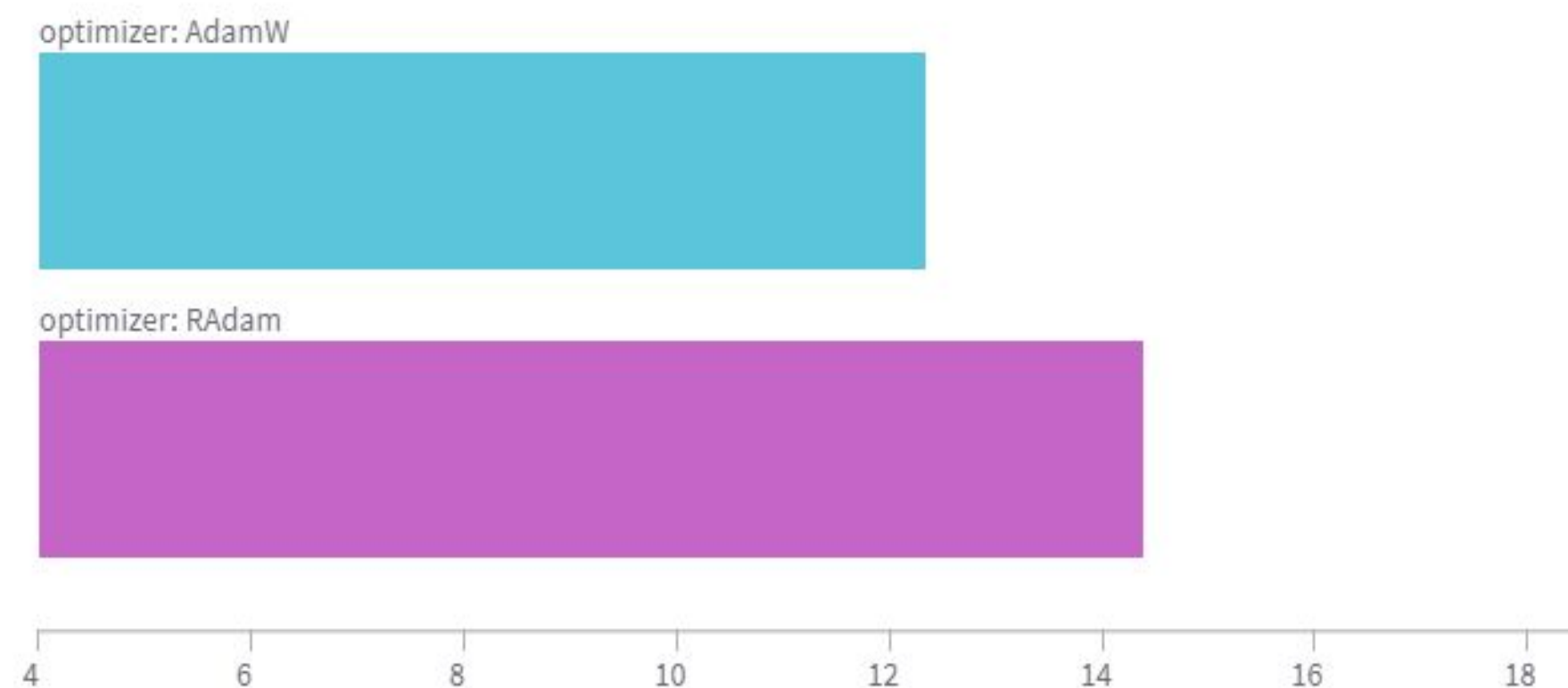
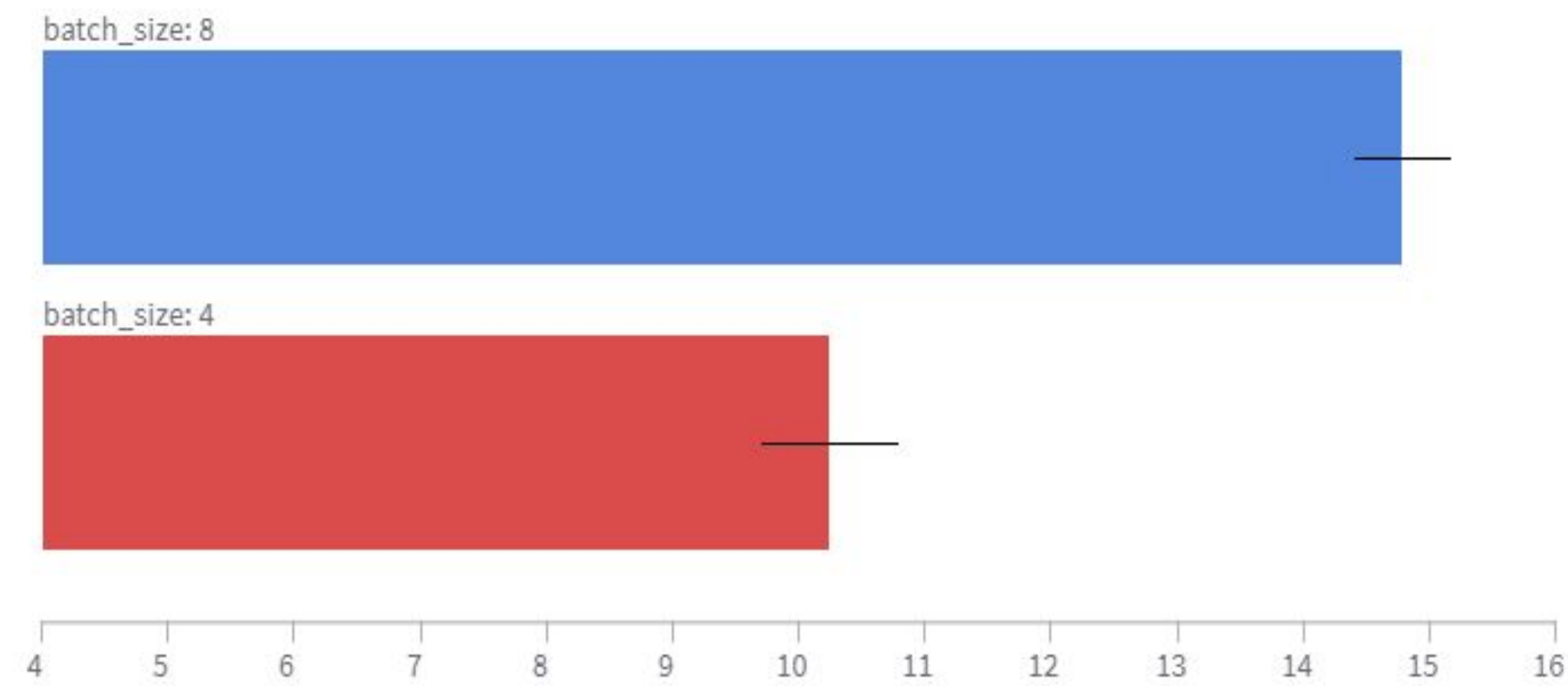
batch size



optimizer



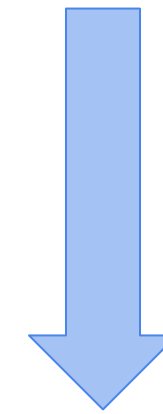
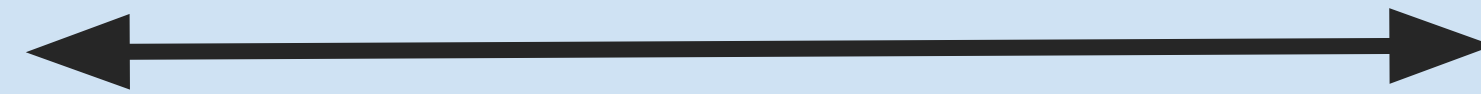
② Levenshtein



∴ batch size 작을 때
AdamW 사용했을 때

answers

축에서 영양물의 밸런스를 유지하기 위해



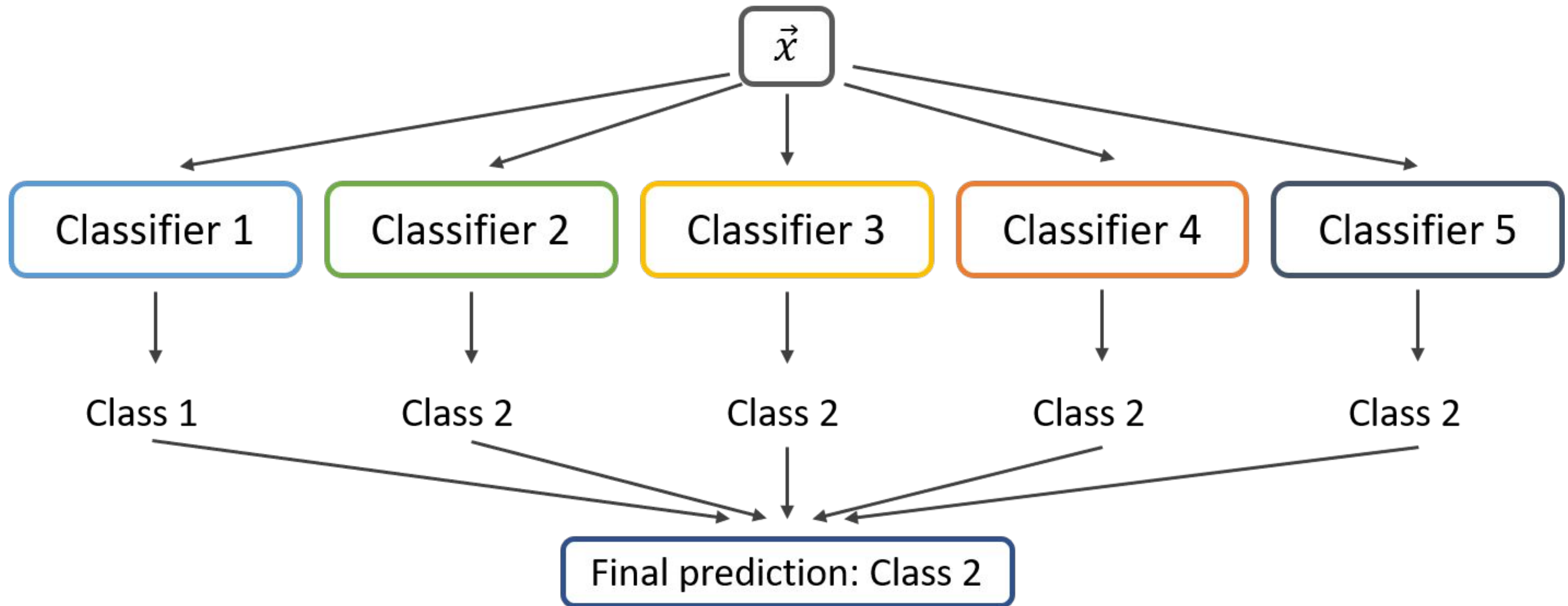
뒤에서부터 12글자 답변길이 자르기

answers

밸런스를 유지하기 위해

성능 향상 (▲)

6.7 → 2.9



최고 성능 달성!

2.68933



1조

Team Name	Score
1조	2.59334

Pre-processing	Data Augmentation	Model	Others
<p>한계</p> <p>긴 문장들을 별도의 전처리 없이 사용</p> <p>개선 방안</p> <p>모델의 max_token_length에 맞게 효과적인 input을 만들어주는 방법 고려</p>	<p>한계</p> <p>문장의 길이만을 고려한 Random Sampling 방식</p> <p>개선 방안</p> <p>Domain Distribution을 고려한 data augmentation</p>	<p>한계</p> <p>2가지의 extraction-based model 만을 이용</p> <p>개선 방안</p> <p>KoELECTRA 또는 T5 이용</p>	<p>한계</p> <p>한 조합의 train/ validation set 으로만 평가 -> 모든 train set 을 이용하지 못함</p> <p>개선 방안</p> <p>k-fold validation 이용</p>

Reference

<https://arxiv.org/pdf/1810.04805.pdf>

<https://huggingface.co/models?language=ko&sort=downloads&search=bert>

<https://huggingface.co/klue/bert-base>

<https://github.com/ainize-team/klue-mrc-workspace/blob/master/klue-bert-base-mrc.ipynb>

<https://towardsdatascience.com/question-answering-with-a-fine-tuned-bert-bc4dafd45626>

Thank you

Contacts

mynameissoyeonkim@gmail.com

kuotientdev@gmail.com

annigin46@gmail.com

ybok2005@gmail.com

Github

<https://github.com/SYKflyingintheSKY/Question-Answering>

<http://shinsimko.net>