
머신러닝 파이프라인 과제

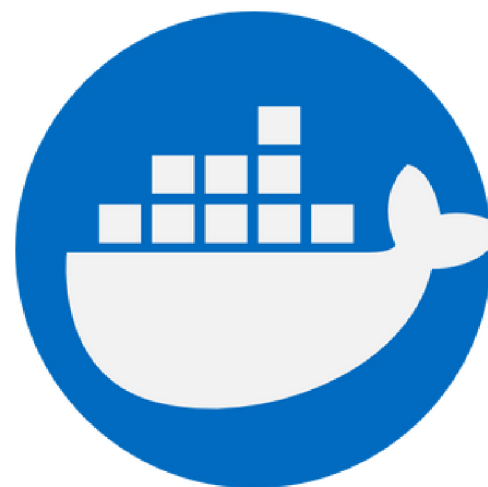
이두원



Contents

- 사용 프로그램
- DAG 파일구성
- 실행 과정
- 결과 이미지

■ 사용 프로그램



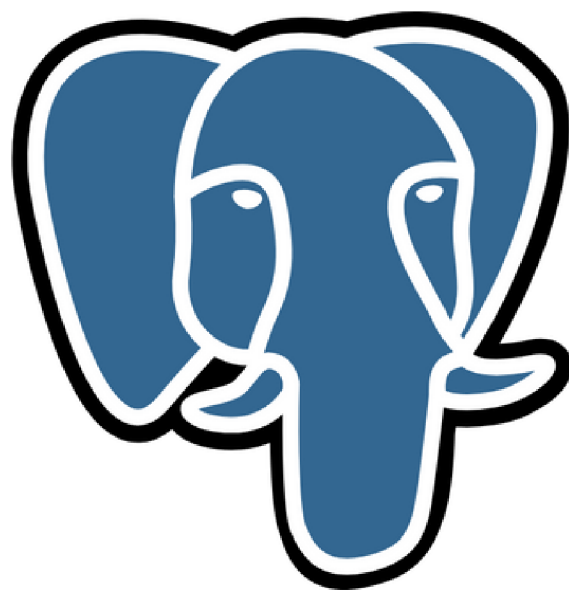
Docker

개발 환경 통일



Apache Airflow

Batch Process를 통한
데이터 전처리 자동화



Postgresql

오픈소스 DB
쓰기 작업이 빈번하고 쿼리
가 복잡한 환경(=데이터 전
처리 과정)에 적합



Amazon SageMaker

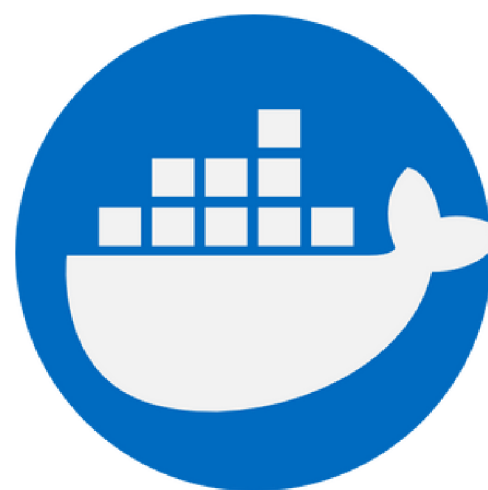


Amazon S3

SageMaker

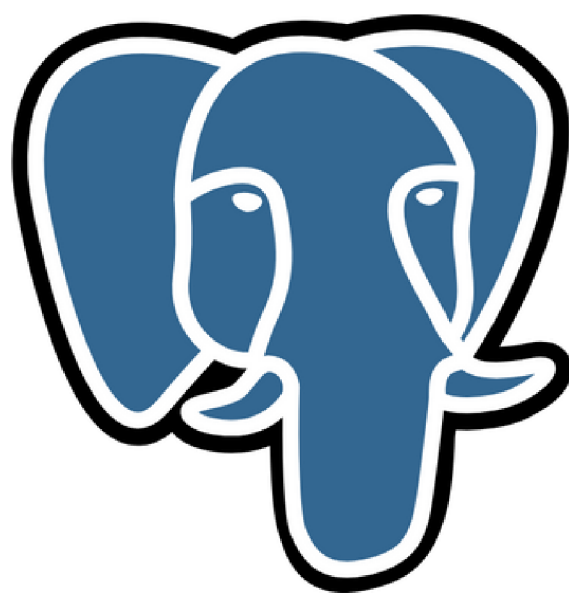
클라우드 리소스를 사용해
모델 학습 환경 구축
AWS를 사용하는 쿠버네티스팀
과의 협업 대비

■ 사용 프로그램



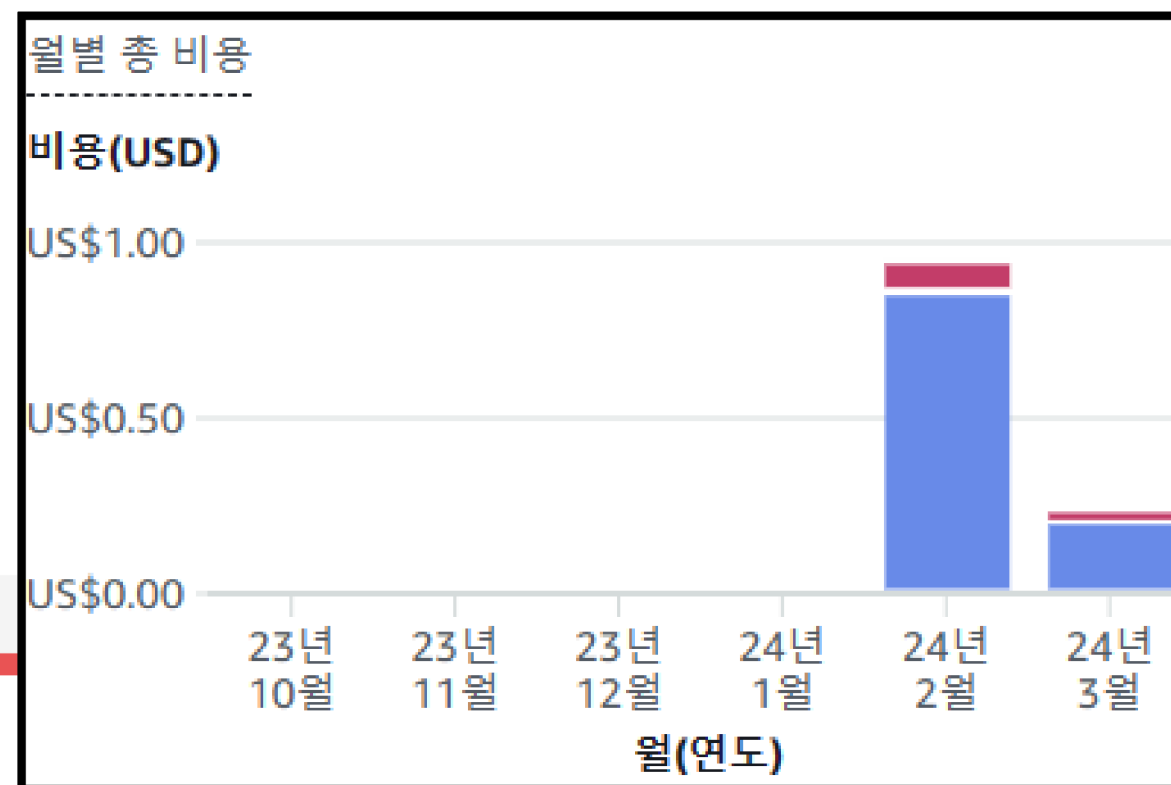
Docker

개발 환경 통일



Postgresql

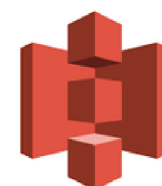
오픈소스 DB
쓰기 작업이 빈번하고 쿼리가 복잡한 환경(=데이터 전처리 과정)에 적합



W



Amazon
SageMaker

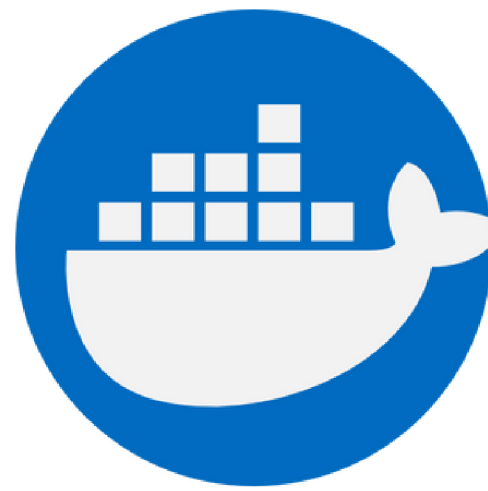


Amazon S3

SageMaker

클라우드 리소스를 사용해
모델 학습 환경 구축
AWS를 사용하는 쿠버네티스팀
과의 협업 대비

■ 사용 프로그램



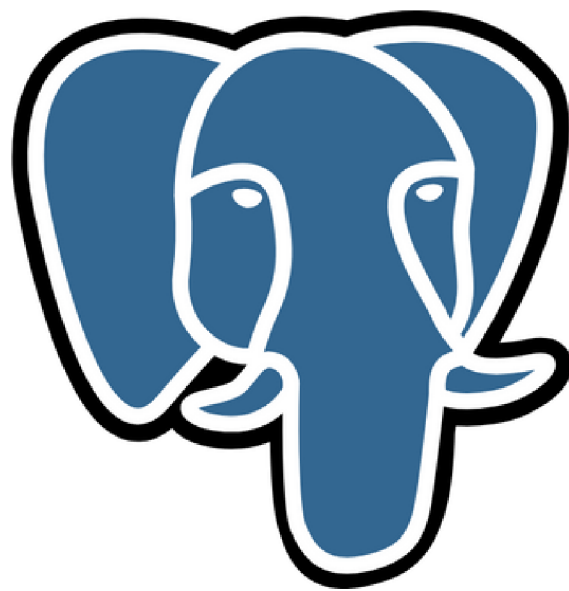
Docker

개발 환경 통일



Apache Airflow

Batch Process를 통한
데이터 전처리 자동화



Postgresql

오픈소스 DB
쓰기 작업이 빈번하고 쿼리
가 복잡한 환경(=데이터 전
처리 과정)에 적합



Google Colab

AWS 인스턴스 비용 문제로
이번 과제에서는 Colab사용



DAG 파일구성

데이터 수집

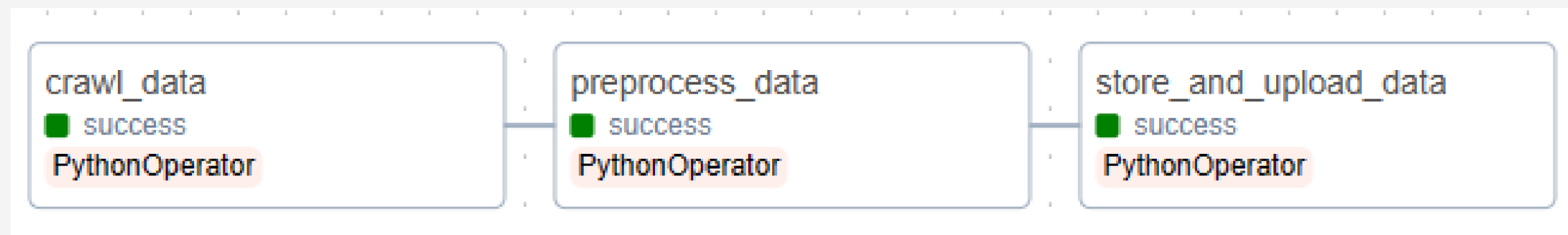
매일 0시 전날 네이버 뉴스를 크롤링

데이터 전처리

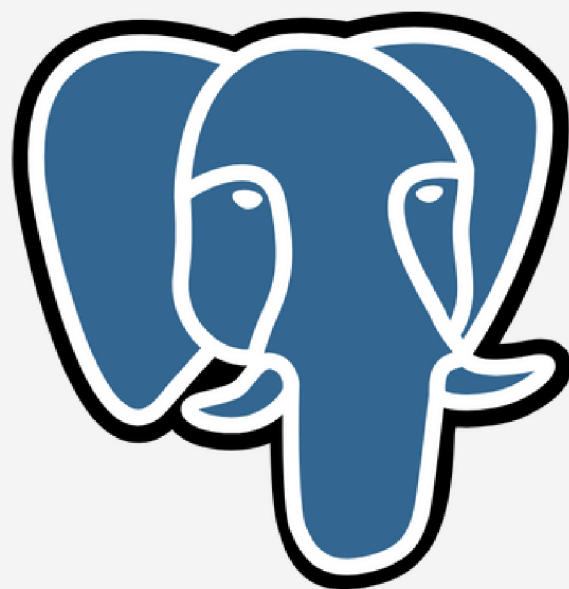
크롤링한 뉴스 전처리

저장&업로드

전처리한 뉴스를 Postgresql 에 저장
모델 학습에 사용할 열만 AWS S3 및 로컬에 업로드 및 저장
























실행 과정






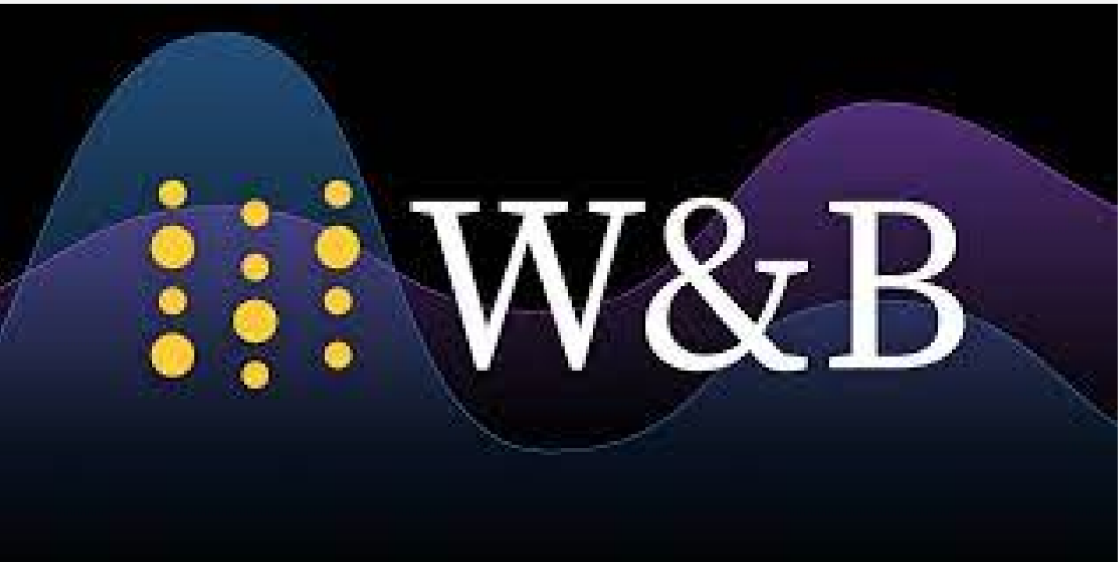
[8080:8080](#)

[5050:80](#)

pipeline		Running (4/4)	4.11%
	pgadmin4 6b814d852a2d 	dpage/pgadmin Running	0.02% 5050:80 
	postgres-1 79c5a1df0f96 	postgres:16 Running	1%
	webserver-1 5306b57c4e08 	apache/airflow: Running	0.1% 8080:8080 
	scheduler-1 07620e8e5184 	apache/airflow: Running	2.99%

DAG 		Owner 	Runs 	Schedule
<input checked="" type="checkbox"/>	data_processing_pipeline	airflow	  4  	1 day, 0:00:00
<input type="checkbox"/>	sagemaker_huggingface_finetuning	airflow	   	1 day, 0:00:00

Tables (3)	
>	 preprocessed_data
>	 raw_data
>	 training_data



Search runs .*



Name (2 visualized)

☒ polar-violet-5

☒ wandering-elevator-4

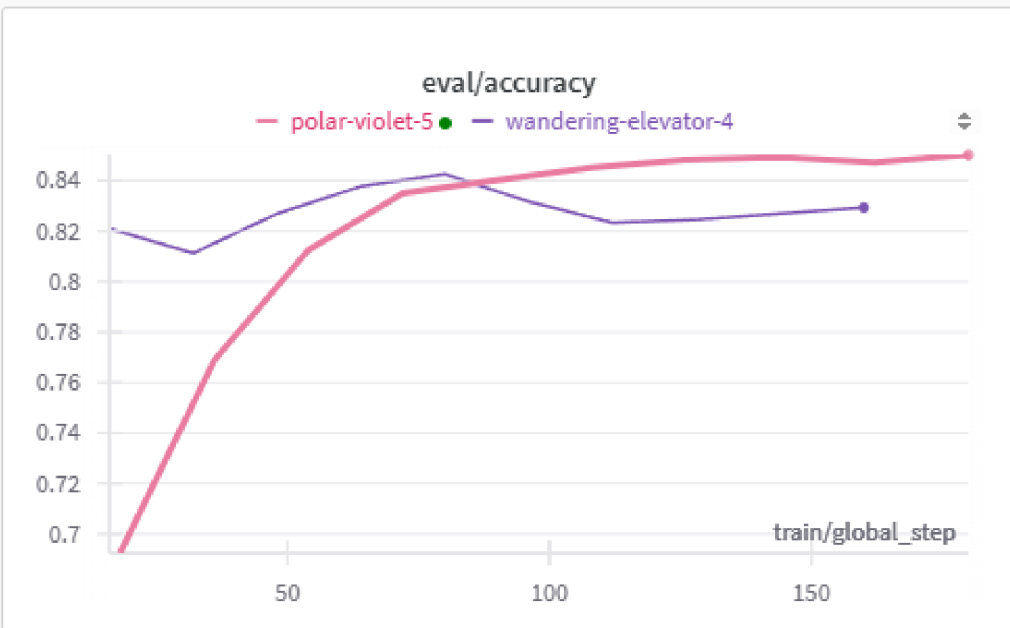
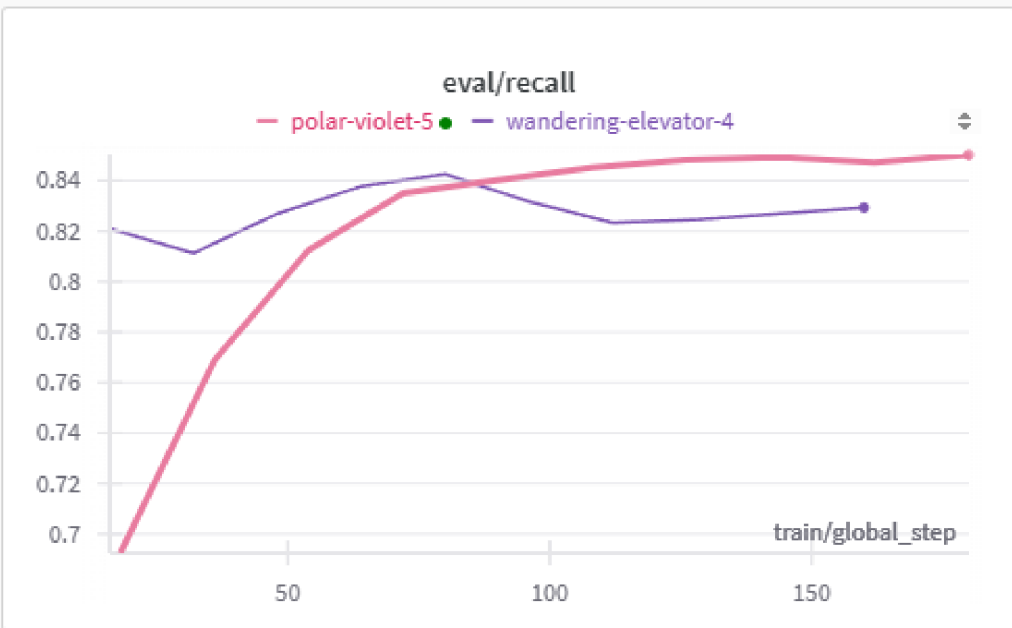
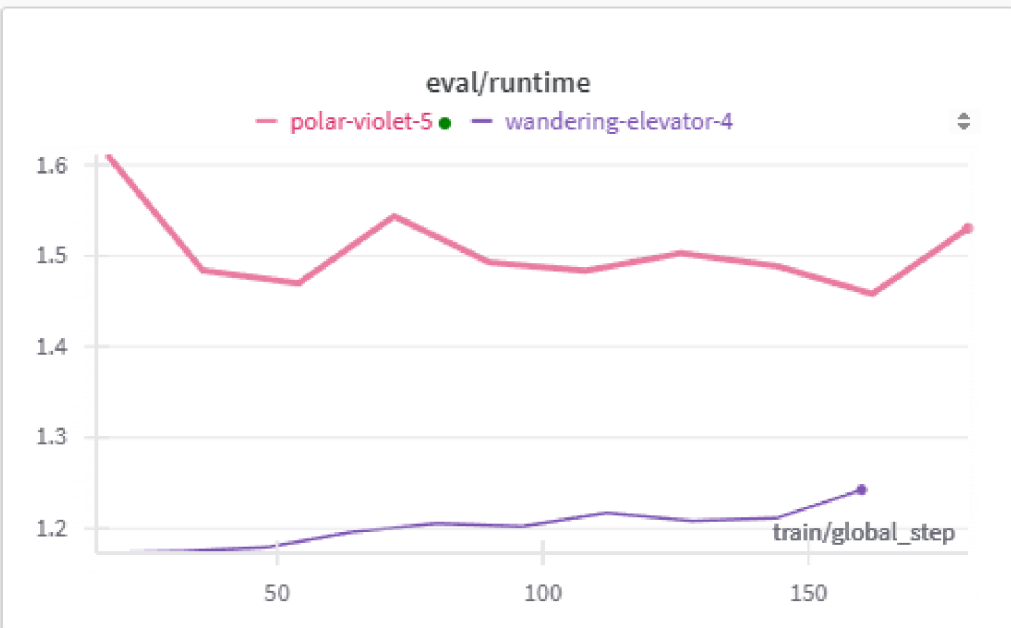
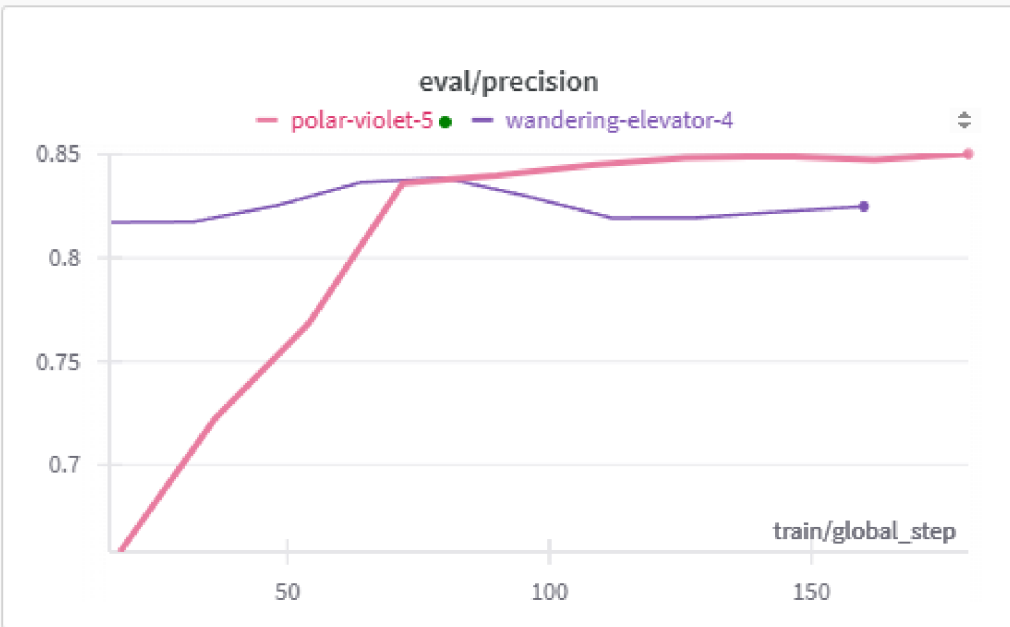
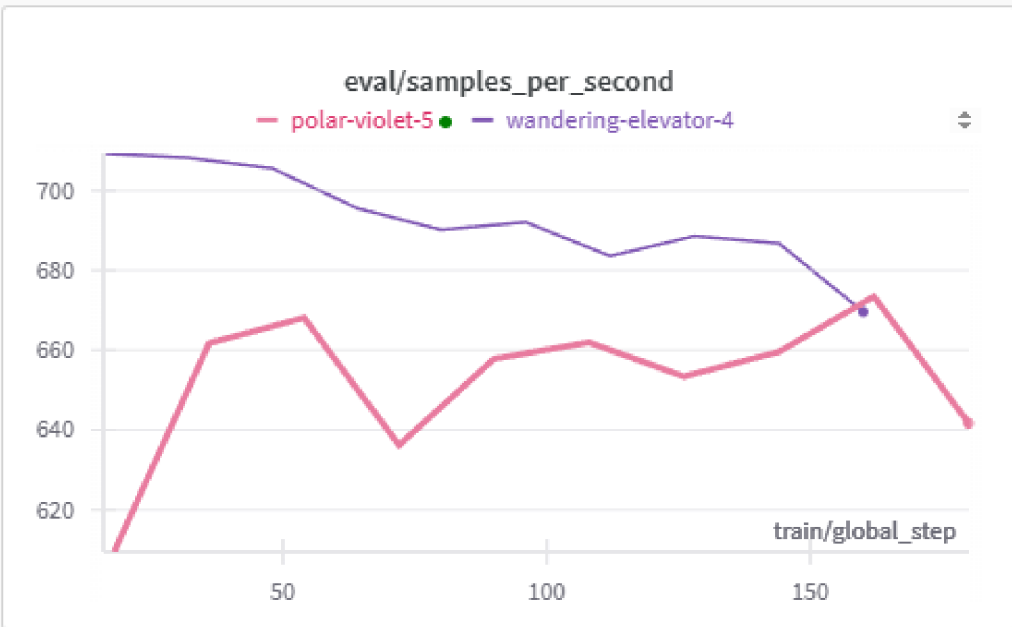
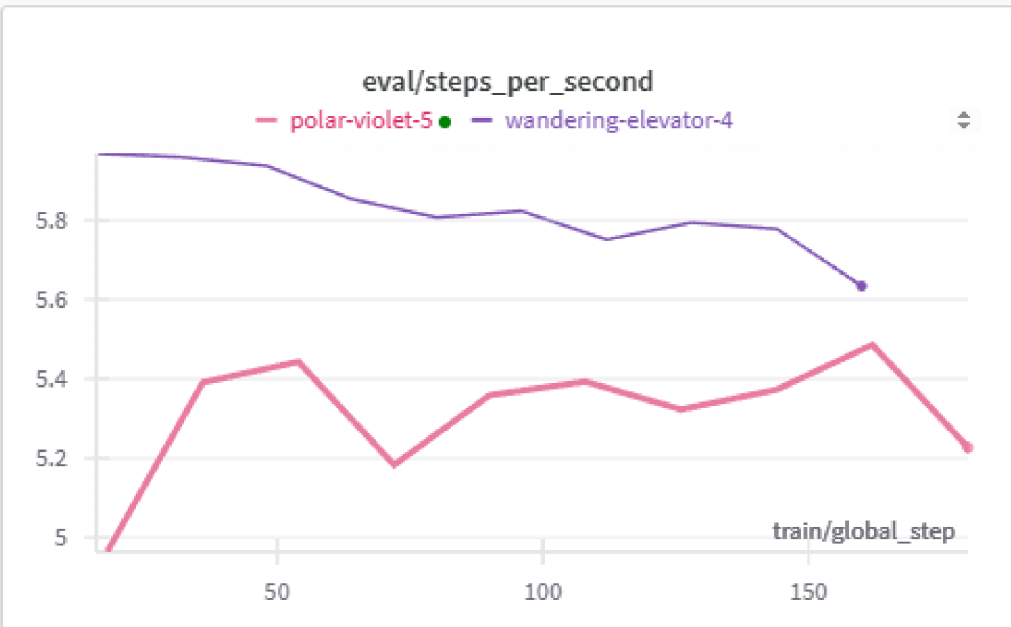
☐ trim-waterfall-3

☐ rich-forest-2

☐ true-sunset-1

eval 8

Add panel





 [training_data_20240228000327.csv](#)

 [training_data_20240229012728.csv](#)

 [training_data_20240301013739.csv](#)

다중 버전 모델 예측

여러 버전의 모델 중에서 선택하여 예측을 수행합니다.



모델 버전 선택

☒ roberta0305 버전

☐ roberta0304 버전

☐ roberta0303 버전

☐ roberta0302 버전

☐ roberta0301 버전

☐ roberta0229 버전

☐ roberta0228 버전

뉴스제목

윤 대통령 지지율 41.1%...2주 연속 40%대

Clear

Submit

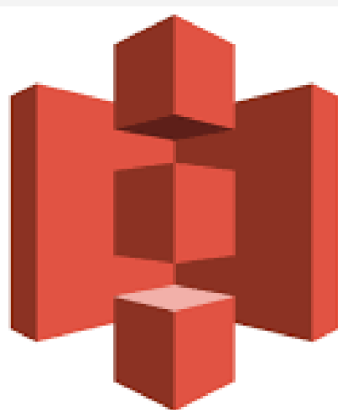
output

카테고리: 정치, 확률: 0.9518

Flag

가능한 모델 버전 목록 (Commit Hash기반)

```
model_versions = {  
    "roberta0305 버전": "b95cef9256104d6bb6912b9cc9b07822b2fb7df8",  
    "roberta0304 버전": "c5d7875611e4f849a3d228b09c68a2e2fff4cc33",  
    "roberta0303 버전": "4d5bfd262e10dc6d8edd55b9957023fa5f5186f7",  
    "roberta0302 버전": "41fd58c8e31b25b6d18e8cfa30e80aba5dca9c2b",  
    "roberta0301 버전": "4f3992e102012b68c44eb89bb7f12eb6d077fccc",  
    "roberta0229 버전": "e19197733e10c70dfbb00825481e1fd40e2f2977",  
    "roberta0228 버전": "34fe46c31f84d04b1c2da393a0026bd8dcd14b55",  
    # 필요한 만큼 추가...  
}
```



Amazon S3



performance/

폴더



model_performance_20240228000327.json



model_performance_20240229012728.json



model_performance_20240301013739.json



model_performance_20240302220857.json



model_performance_20240303223216.json



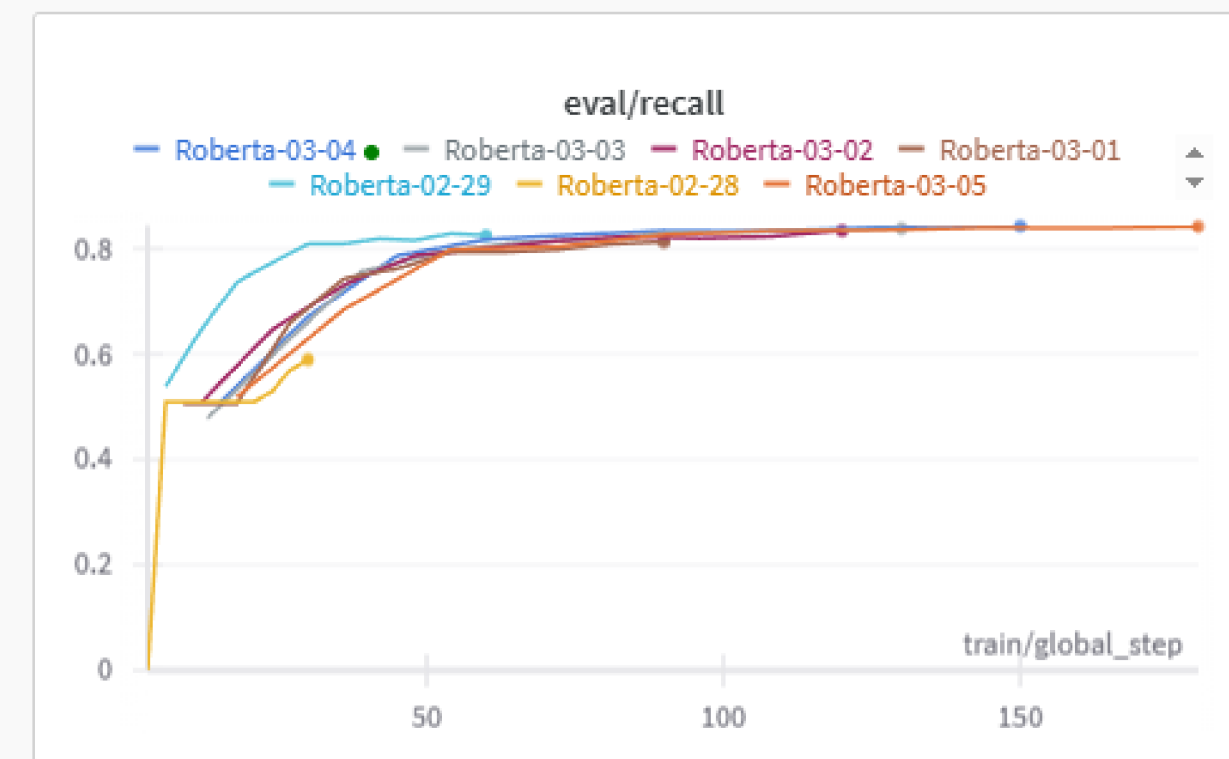
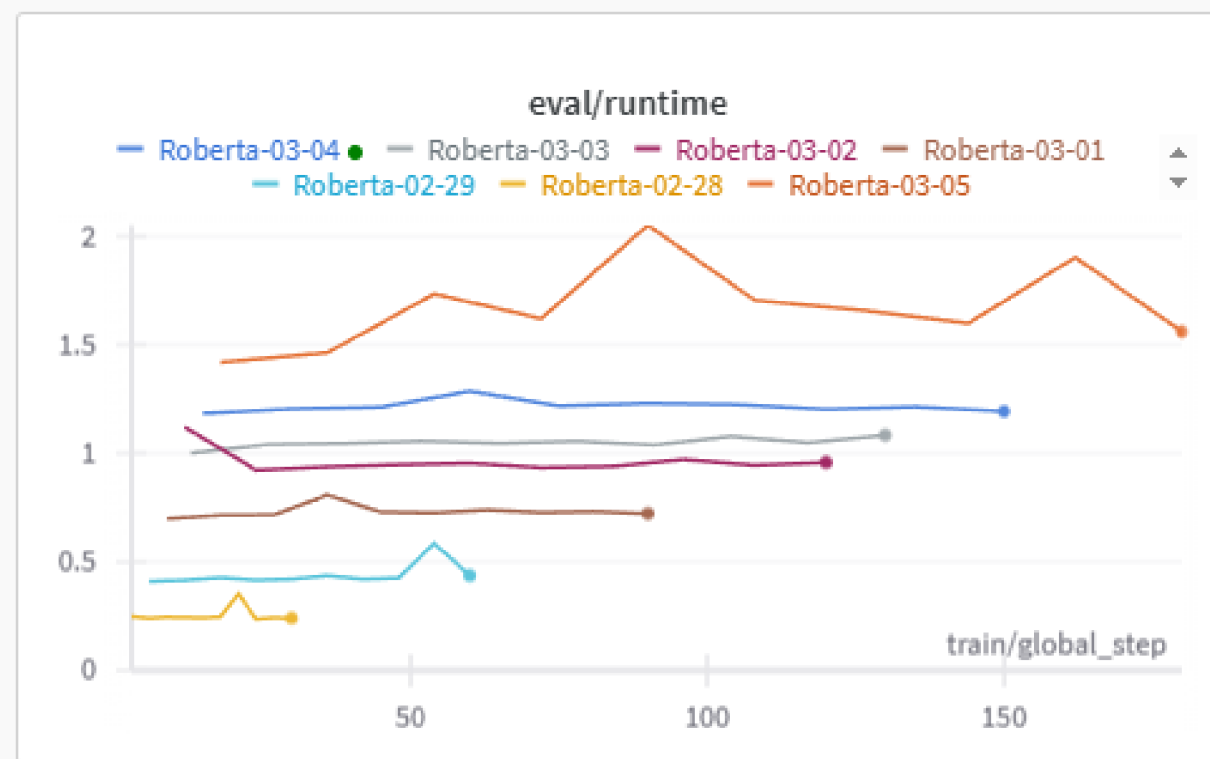
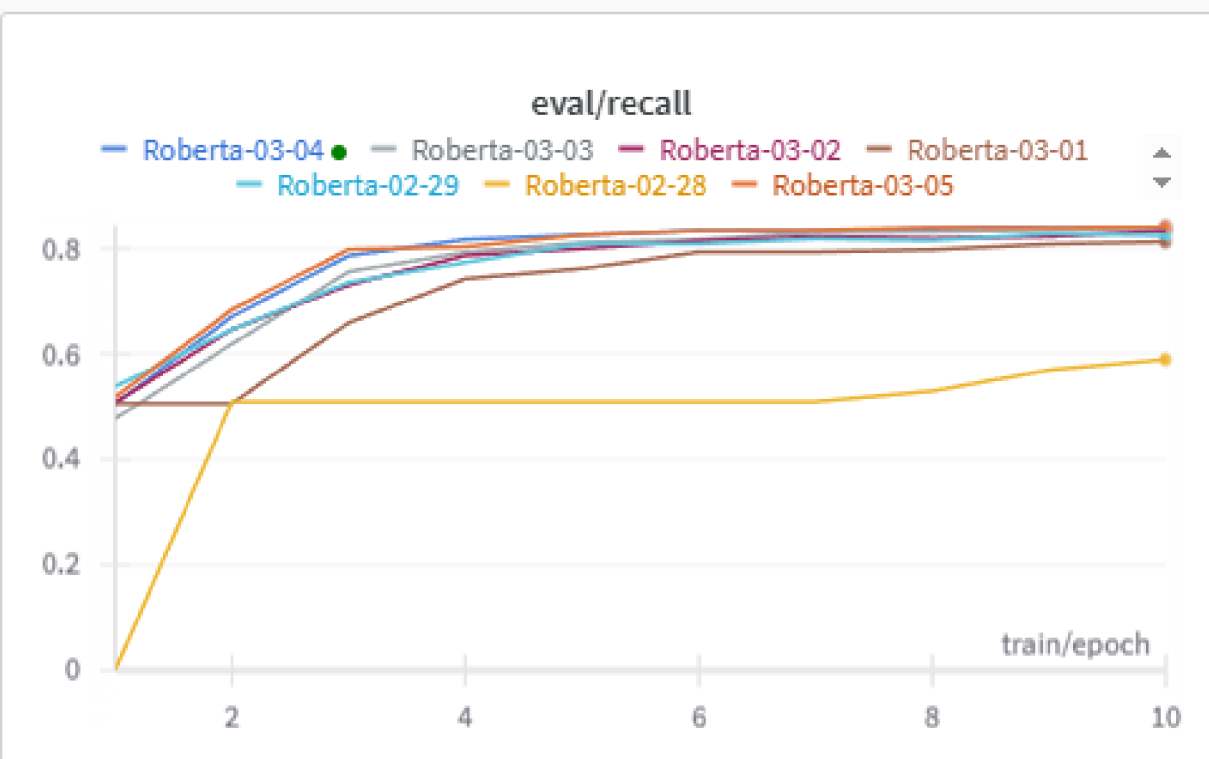
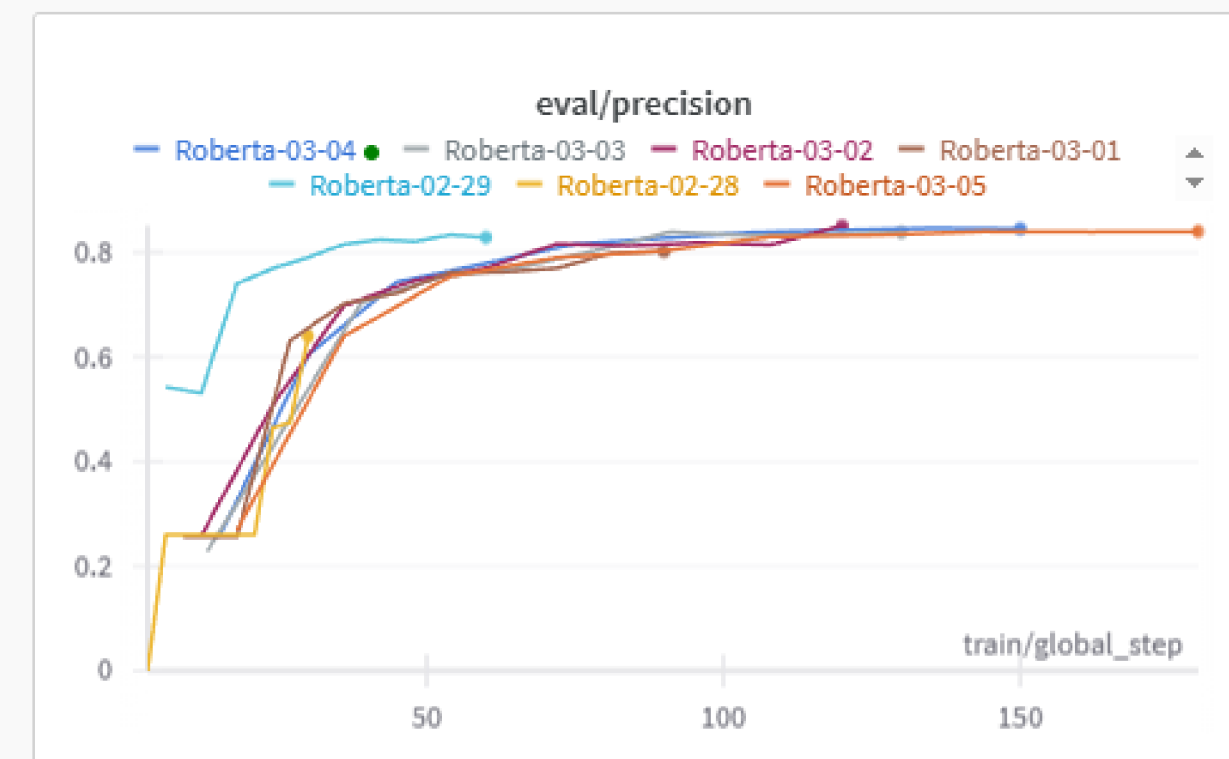
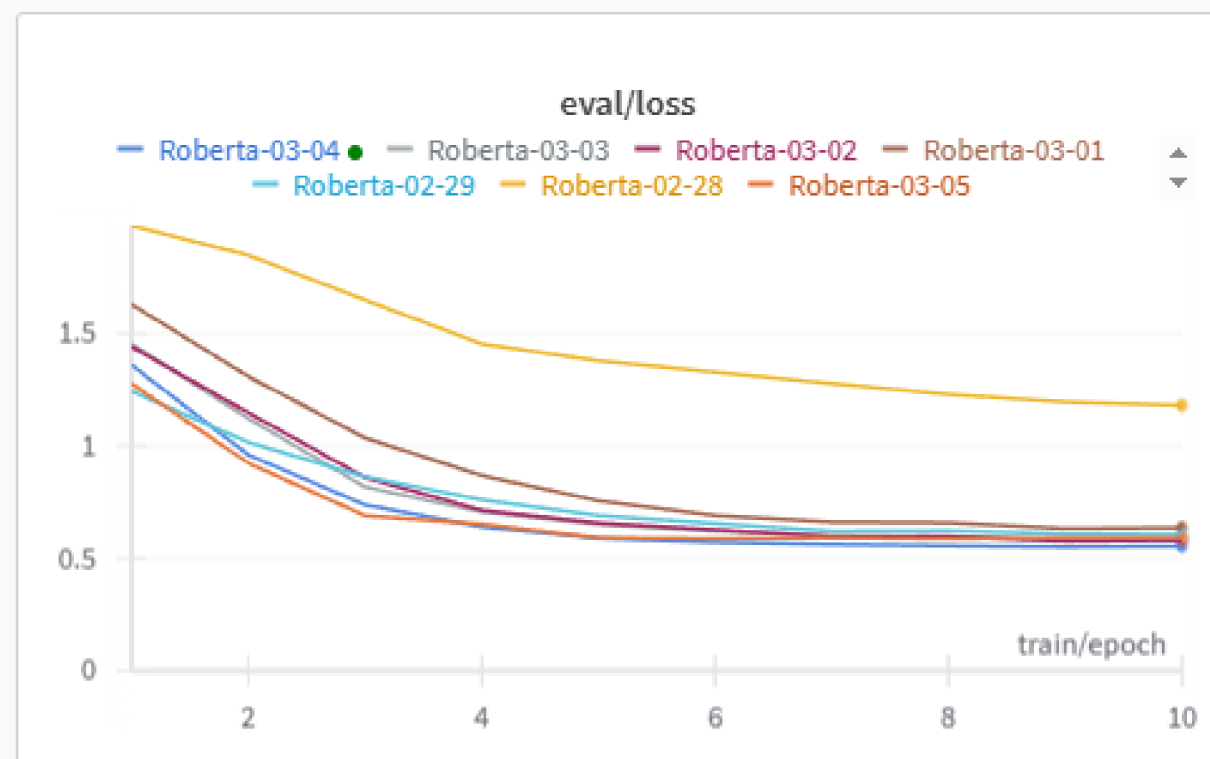
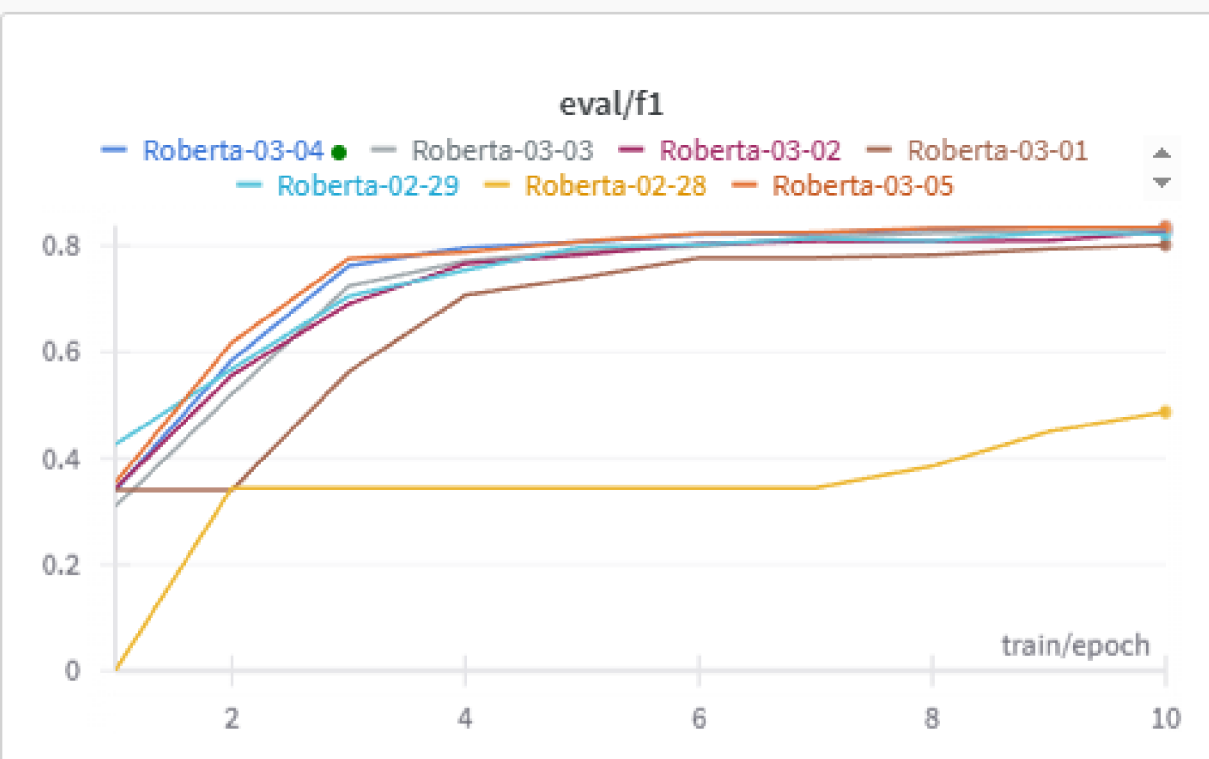
model_performance_20240304224328.json



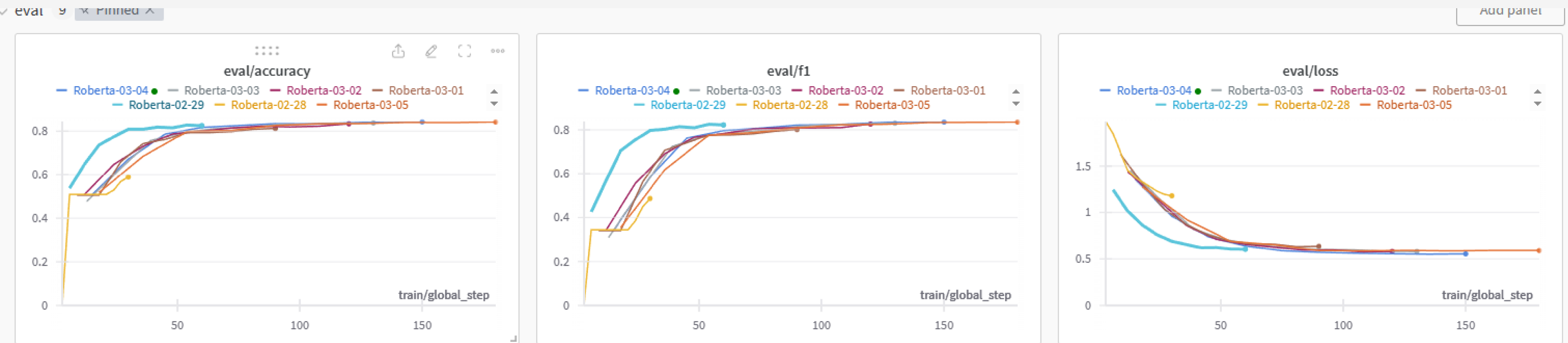
model_performance_20240305112649.json

```
{"eval_loss": 0.5738615989685059,  
  "eval_accuracy": 0.8503054989816701,  
  "eval_f1": 0.8458629470566328,  
  "eval_precision": 0.8504772259285096,  
  "eval_recall": 0.8503054989816701,  
  "eval_runtime": 1.4576,  
  "eval_samples_per_second": 673.712,  
  "eval_steps_per_second": 5.488,  
  "epoch": 10.0}
```

결과 이미지



결과 이미지



End 