HAPPY:

악플 분류 및 순화 문장 재생성

NLP - 06

HAPPY | 김준휘, 류재환, 박수현, 박승현, 설유민

boostcamp ai tech

팀원 소개

김준휘

Classification model
Classification API
Data Collecting

류재환

Generation model
Generation API
Data Collecting

박수현

Classification model
Data Guideline
Data Collecting
Data Checking

박승현

Generation model
Database
Back-end
Front-end
Data Web 개발
Data Collecting

설유민

Generation model
Data Collecting
Data Checking

INDEX

1. Introduction

주제 선정 동기

기존 연구와의 비교

Simple Demo

주제 선정 동기

경제 전체

네이버, 스포츠뉴스 댓글도 중단…"선수들 고통 심각"

등록 2020.08.07 15:14 / 수정 2020.08.07 15:15

김자민 기자

"연예인 인격권 존중"...네이버, 연예뉴스 댓글 폐지

악플 탐지 '클린봇' 한계..."고통 공감이 우선"
연예뉴스→연예인-이용자 소통공간 변화 방침

등록 2020-02-19 오후 2:53:31
수정 2020-02-19 오후 3:03:36

가 가

[이데일리 한광범 기자] 카카오(035720)에 이어 네이버(035420)도 그동안 연예인들에 대한 인식공격 창구로 변질됐다는 비판을 받아온 '연예뉴스 댓글' 서비스 폐지를 결정했다. 네이버는 연예뉴스의 대대적 개편 방침도 분명히 했다.

한광범 기자

N 기자구독

2020년경 연예 및 스포츠 댓글 폐지

선수/연예인 보호 차원

주제 선정 동기

경제 전체

네이버, 스포츠뉴스 댓글도 중단…"선수들 고통 심각"

등록 2020.08.07 15:14 / 수정 2020.08.07 15:15

김자민 기자

"연예인 인격권 존중"...네이버, 연예뉴스 댓글 폐지

악플 탐지 '클린봇' 한계..."고통 공감이 우선"
연예뉴스→연예인-이용자 소통공간 변화 방침

등록 2020-02-19 오후 2:53:31
수정 2020-02-19 오후 3:03:36

가 가

[이데일리 한광범 기자] 카카오(035720)에 이어 네이버(035420)도 그동안 연예인들에 대한 인식공격 창구로 변질됐다는 비판을 받아온 '연예뉴스 댓글' 서비스 폐지를 결정했다. 네이버는 연예뉴스의 대대적 개편 방침도 분명히 했다.

한광범 기자

N 기자구독

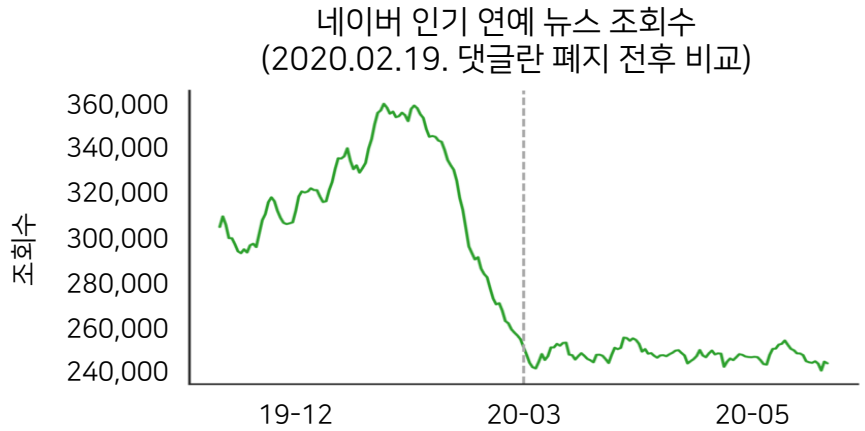
2020년경 연예 및 스포츠 댓글 폐지

선수/연예인 보호 차원

🤔 그 많던 악플은 다 사라졌을까?

주제 선정 동기

http://www.mediatoday.co.kr/news/articleView.html?idxno=306142



댓글/ 악플 타 커뮤니티로 분산, 음성화

포털 연예뉴스 조회수 20% 감소

주제 선정 동기

학술적 관점: 데이터셋 구축

Mitigating AI ethical issues: It has been repeatedly observed that large-scale language models can and often do amplify social biases embedded in text used to train them [95]. In order to disincentivize such behaviors, we proactively remove examples, from both unlabeled and labeled corpora, that reflect social biases, contain toxic content and have personally identifiable information (PII), both manually and automatically. Social biases are defined as overgeneralized judgment on certain individuals or group based on social attributes (e.g., gender, ethnicity, religion). Toxic contents include insults, sexual harassment and offensive expressions.

댓글을 이용한 다양한 연구 시,

👎 욕설, 혐오, 정치적 의견 반영되지 않음

👎 결과 편향 발생 가능

👍 혐오 표현을 제거한 의견 반영 가능

👍 결과 편향 감소

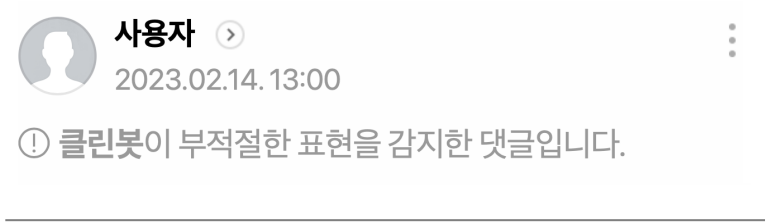
Park, Sungjoon, et al. "Klue: Korean language understanding evaluation." arXiv preprint arXiv:2105.09680 (2021).

🤔 필터링, 어떻게 하는 게 좋을까요?

기존 서비스

네이버 클린봇

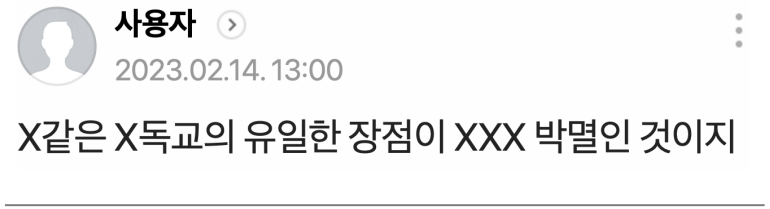
혐오 표현 필터링 기능



의견을 잘못 필터링한 경우 서비스에 치명적!

카카오 세이프봇

혐오 표현 필터링 기능 + 욕설 음표 치환 기능

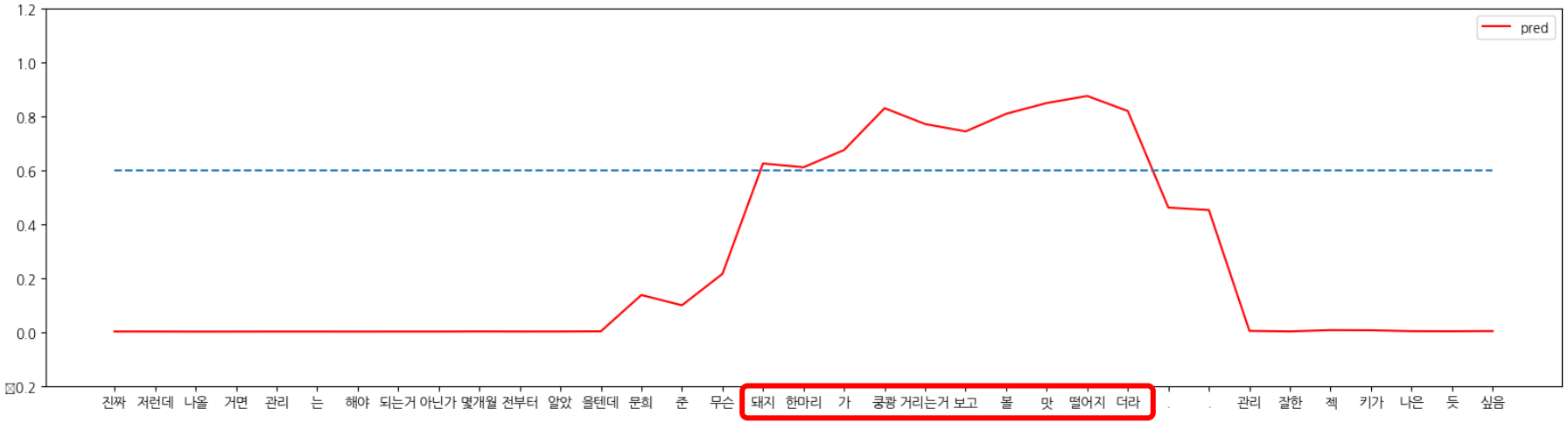


댓글에 욕이 존재하지 않지만 문장 자체가 혐오, 편향성을 드러내는 경우 필터링이 불가.

Simple Demo

▼ 작성한 댓글

진짜 저런데 나올거면 관리는 해야되는거 아닌가 몇개월전부터 알았을텐데 문희준 무슨 돼지 한마리가 쿵쾅거리는거보고 볼맛 떨어지더라.. 관리 잘한 젝키가 나은듯 싶음

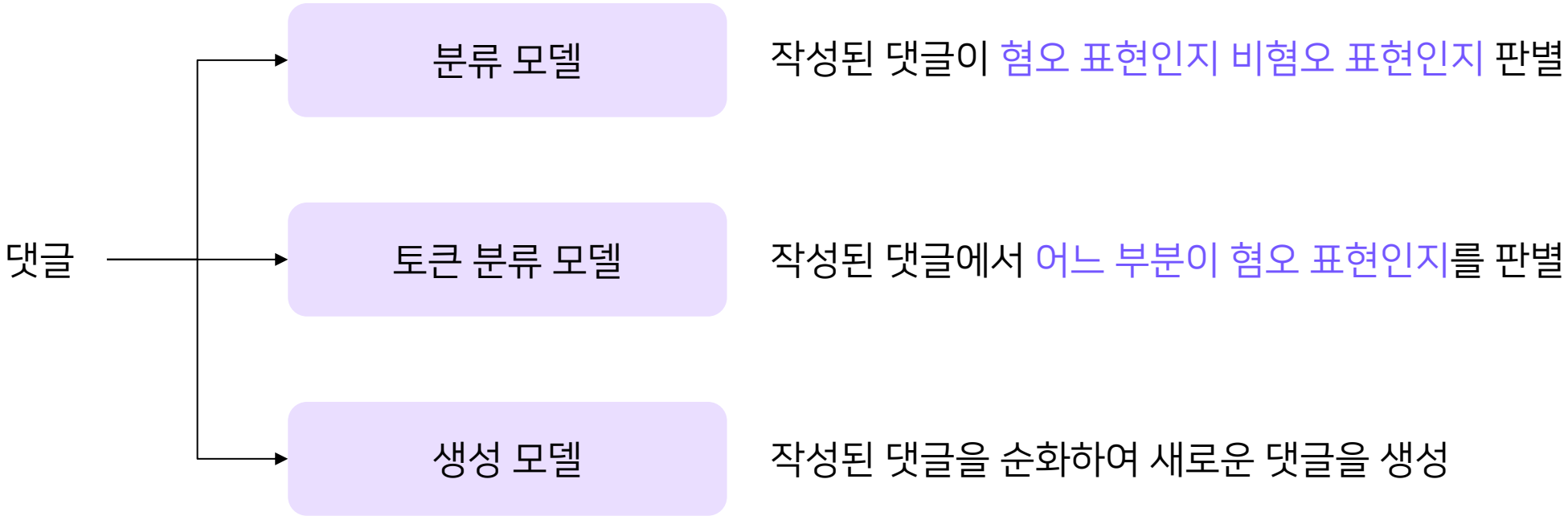


위와 같은 부분이 혐오 표현으로 분류될 여지가 있어요. 다음과 같이 순화해 보는 것은 어떨까요?

저기에 나오려면 관리를 해야 하지 않을까? 몇 개월 전부터 알았을 텐데 문희준 보고 보기 싫어졌다. 관리 잘한 젝키가 나은 것 같다.

입력

전체 모델 구성



2. Classification

전체 모델 구성

데이터셋 선정

Recall 수치 평가

모델 선정

분류 모델

데이터셋 선정

총 5개 혐오 표현 데이터셋

APEACH

데이터 수집 목적 댓글

	비혐오	혐오	전체
Train	3,486	44,10	7,896
Valid	1,848	1,922	3,843

https://github.com/jason9693/APEACH

KOLD

2020.3 ~ 2022.3 뉴스 및 유튜브 댓글

	비혐오	혐오	전체
	20,119	20,310	40,429

https://github.com/boychaboy/KOLD

UnSmile

2019.1 ~ 2020.6 뉴스 및 커뮤니티 댓글

	비혐오	혐오	전체
Train	3,739	12,636	16,375
Valid	935	3,050	3,985

https://github.com/smilegate-ai/korean_unsmile_dataset

K-MHaS

2018.1 ~ 2020.6 뉴스 댓글

	비혐오	혐오	전체
Train	42,909	47,060	89,969
Valid	4,887	5,038	9,925

https://github.com/adlnlp/K-MHaS

BEEP!

2018.1~ 2020.2 연예 뉴스 댓글

	비혐오	혐오	전체
Train	3,486	4,410	7,896
Valid	160	311	471
Test			974

https://github.com/kocohub/korean-hate-speech

데이터셋 선정

🤔 비혐오 표현을 혐오로 잘못 분류하지 않도록 하려면?

Recall

		Training Set			
		BEEP!	K-MHaS	KOLD	UnSmile
Dev Set	APEACH	73.62	92.26	81.74	78.02
	BEEP!	93.22	100	93.51	92.38
	K-MHaS	58.20	89.84	60.53	64.08
	KOLD	77.08	90.90	82.23	74.55
	UnSmile	86.67	97.73	88.20	91.52
	평균	77.75	**94.14**	81.24	80.11

K-MHaS로 학습된 모델의 Recall 수치가 가장 높음

데이터셋 선정

🤔 비혐오 표현을 혐오로 잘못 분류하지 않도록 하려면?

Precision

		Training Set			
		BEEP!	K-MHaS	KOLD	UnSmile
Dev Set	APEACH	92.81	65.14	89.95	85.90
	BEEP!	75.24	22.18	64.95	62.37
	K-MHaS	97.35	**87.55**	95.47	94.52
	KOLD	79.37	33.71	81.84	79.17
	UnSmile	90.47	64.77	88.86	90.97
	평균	87.04	54.67	84.21	82.58

K-MHaS로 학습된 모델의 Recall 수치가 가장 높음

혐오 표현의 기준이 낮을수록, 모호한 라벨 발생 ▲

K-MHaS는 혐오 표현의 기준 매우 엄격

따라서 K-MHaS데이터셋이 가장 적합!

토큰 분류 모델은 KOLD데이터셋으로 학습.

Recall 수치 평가

🤔 실전 성능은 어떻게 평가할 수 있을까?

누가 봐도 혐오로 판별될 여지가 없는 데이터를 모아 평가

ex)

"여자가 아이를 돌보고 남자가 밖에서 돈을 벌어온다' 다 옛날 말이죠. 요즘은 다 반반씩!!"

"우리나라도 저렇게 가야지. 인구줄면 답없다"

- 데이터 편향성을 줄이기 위해 5개 데이터셋 모두에서 40개씩 총 200개 댓글 선정
- 비혐오 데이터셋의 기준은 공격적인 어투, 혐오 단어, 편견 등을 모두 포함하지 않음
- K-MHaS로 학습된 모델은 단 하나의 문장 ("사람들이 나 얼굴 못생겼으니까 얼굴 다 성형하래") 제외 모두 올바르게 예측

모델 선정

성능 관점 – F1 Score

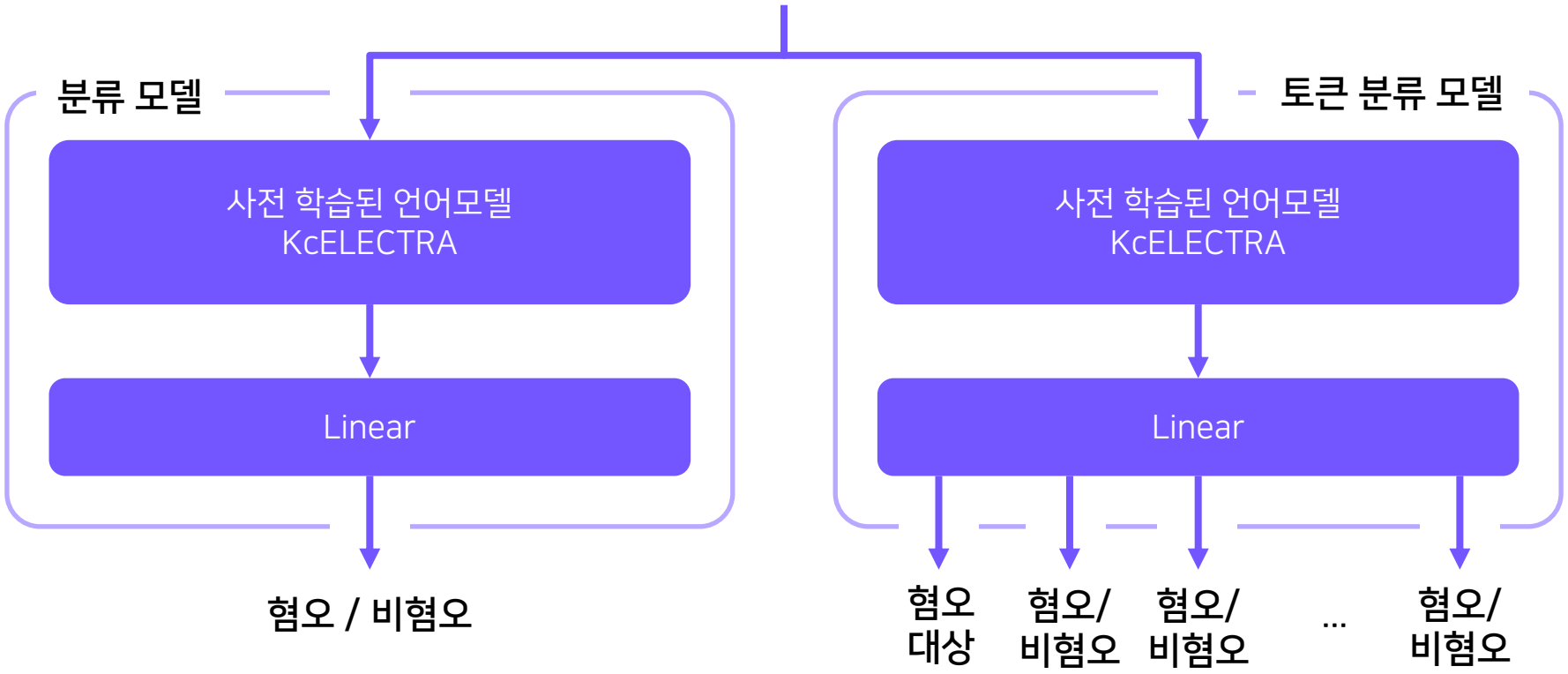
	KcELECTRA-small	KcELECTRA-base	koelectra-small	koelectra-base	CNN 기반
F1 score	87.43	90.88	79.78	87.42	83.21

추론 시간 관점 - RPS

	KcELECTRA-small	KcELECTRA-base	koelectra-small	koelectra-base	CNN 기반
RPS	319	173	319	173	765

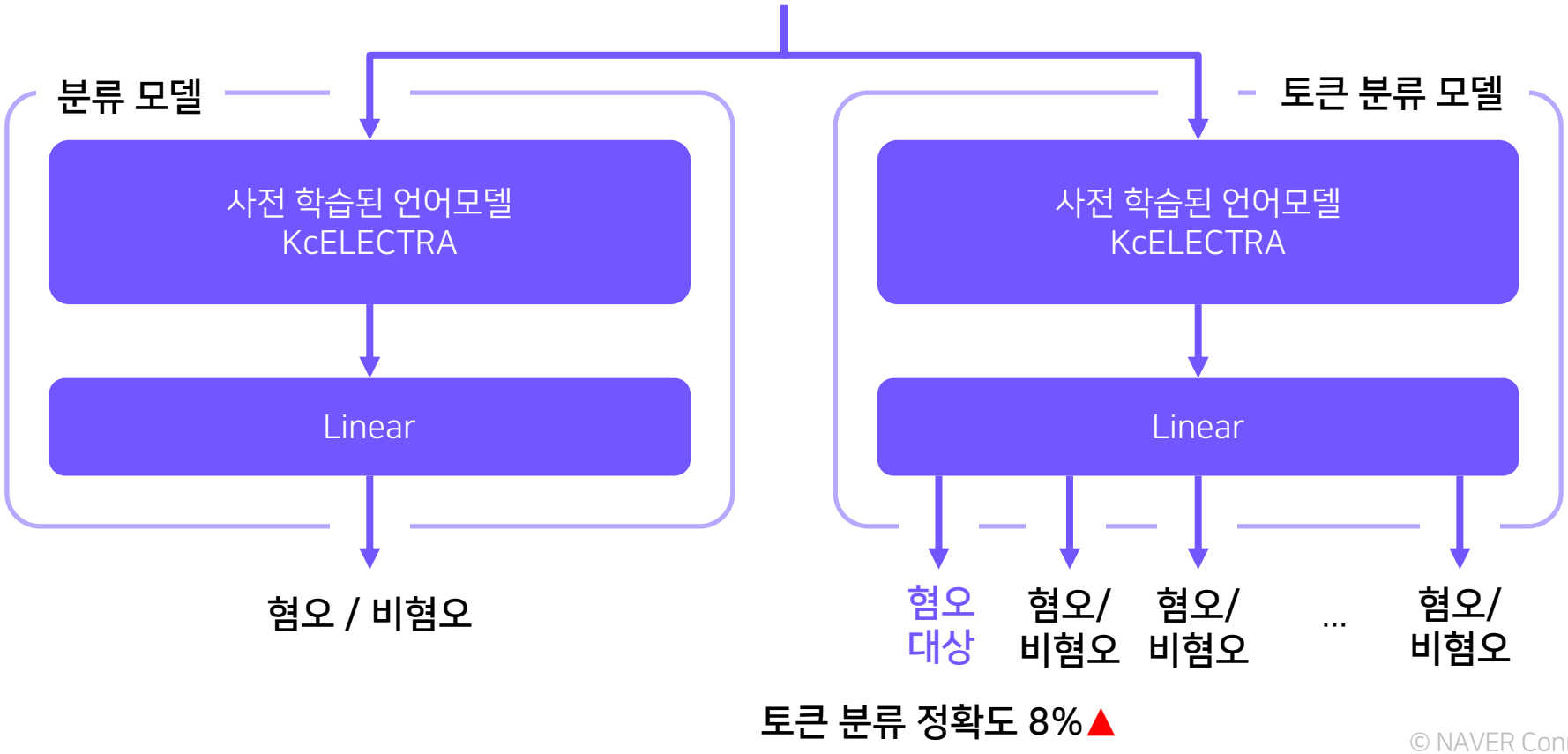
분류 모델

Sequence Classification & Token Classification



분류 모델

Sequence Classification & Token Classification



3. Data For Generation

순화 문장 생성을 위한 데이터 전략

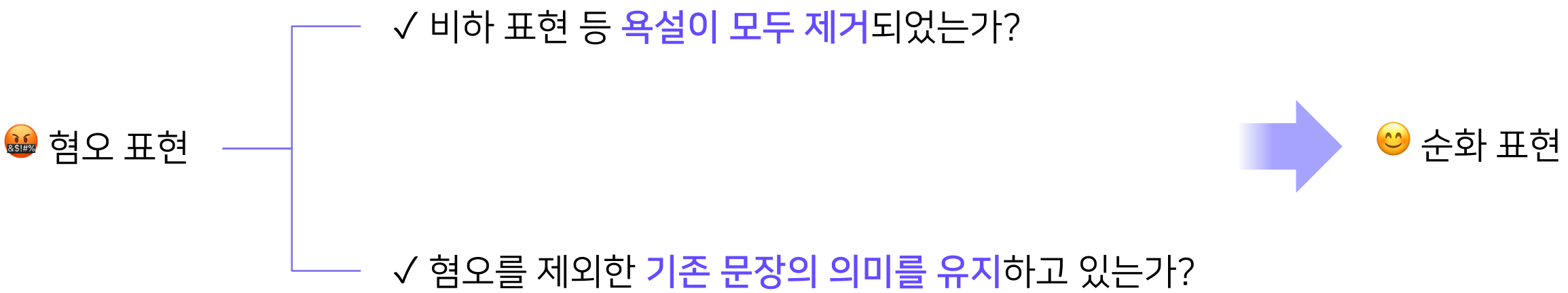
Non-parallel: Task Reformulation

Non-parallel: 유사 task dataset 활용

Parallel: 사용자 참여 대응 순화 데이터 수집

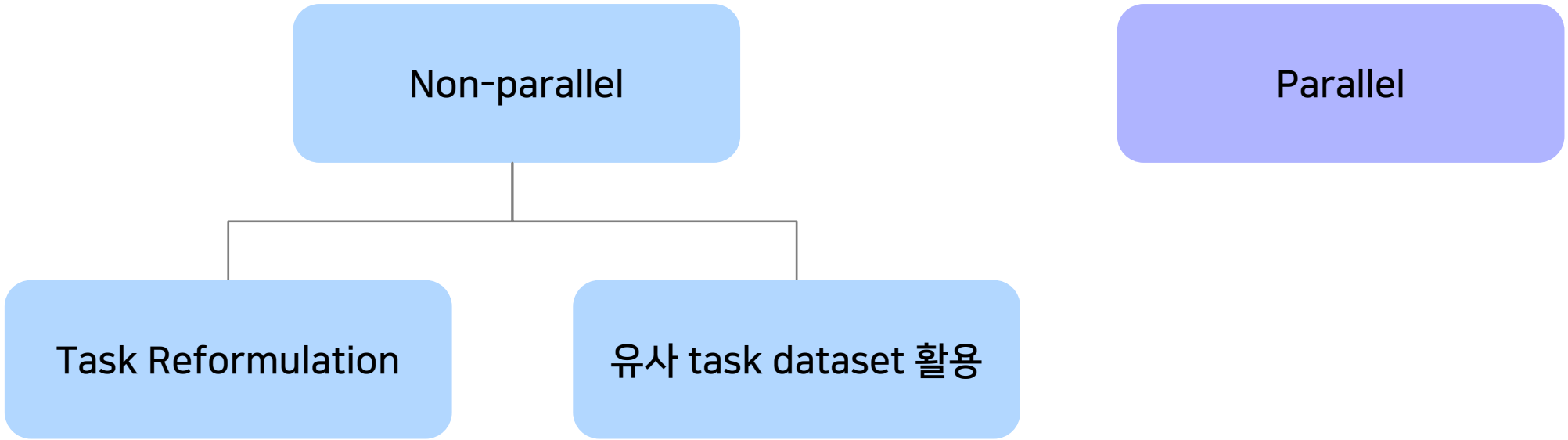
순화 문장 생성을 위한 데이터 전략

Abstract



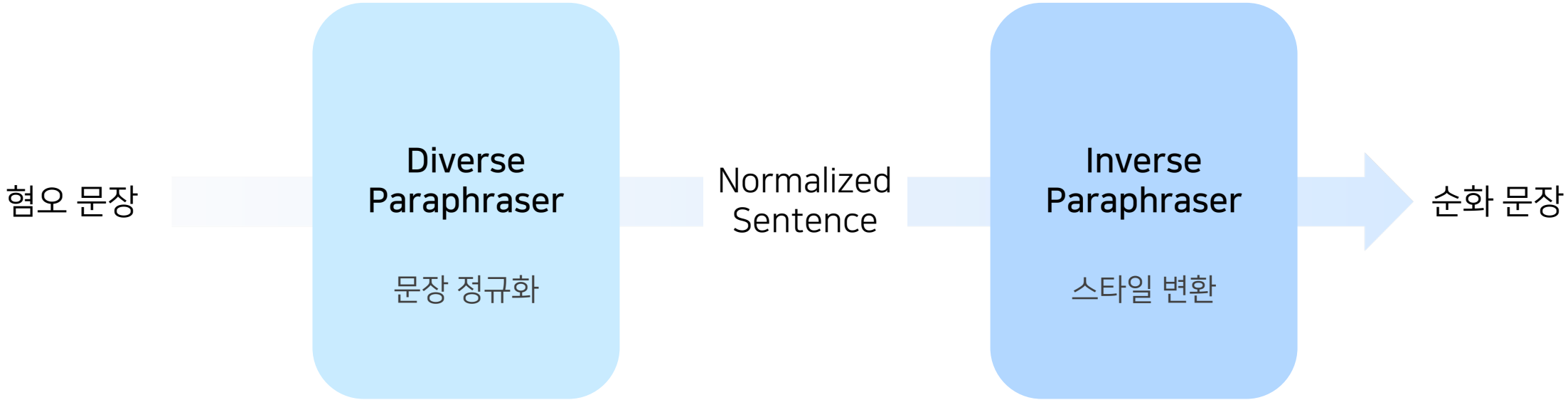
순화 문장 생성을 위한 데이터 전략

🤔 어떤 데이터를 어떻게 사용하지?



Non-parallel: Task Reformulation

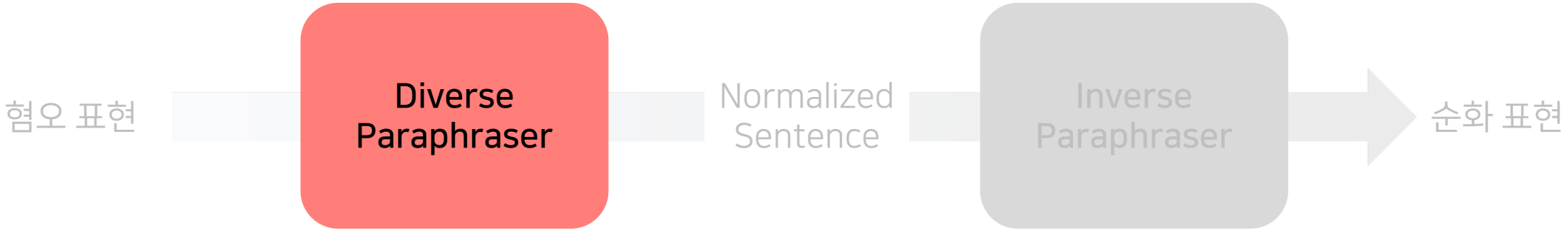
STRAP: Text Style Transfer를 2번의 paraphrasing으로 재정의



🙂 존재하는 paraphrasing dataset으로, 일반화된 paraphraser 학습 및 사용 기대

Non-parallel: Task Reformulation

STRAP: Text Style Transfer를 2번의 paraphrasing으로 재정의

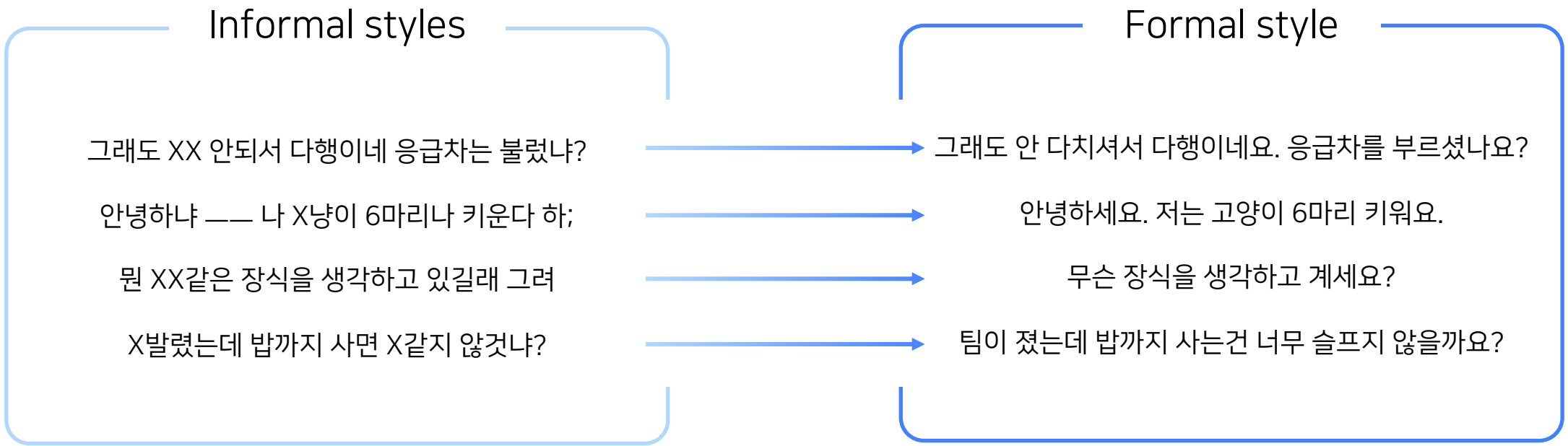


☹ 두 개의 생성 모델을 사용 ⇒ 과도한 시간, 메모리 소비

☹ Task의 특수성 ⇒ Diverse paraphraser의 정규화 성능 부족

Non-parallel: 유사 task dataset 활용

Text Style Transfer – SmileStyle Dataset 활용



욕설 및 과격한 표현이 포함된 스타일 일부가 혐오 표현과 유사 ⇒ 변환 대상으로 설정

Non-parallel: 유사 task dataset 활용

Text Style Transfer – SmileStyle Dataset 활용

XX는 대한민국에 전혀 도움이 안되는 인간들이야
이런기사는 항상 댓글통제 X재X ㅋㅋㅋ 종북XX

☹ Informal styles에서 task에 적합한 혐오 표현은 극히 일부

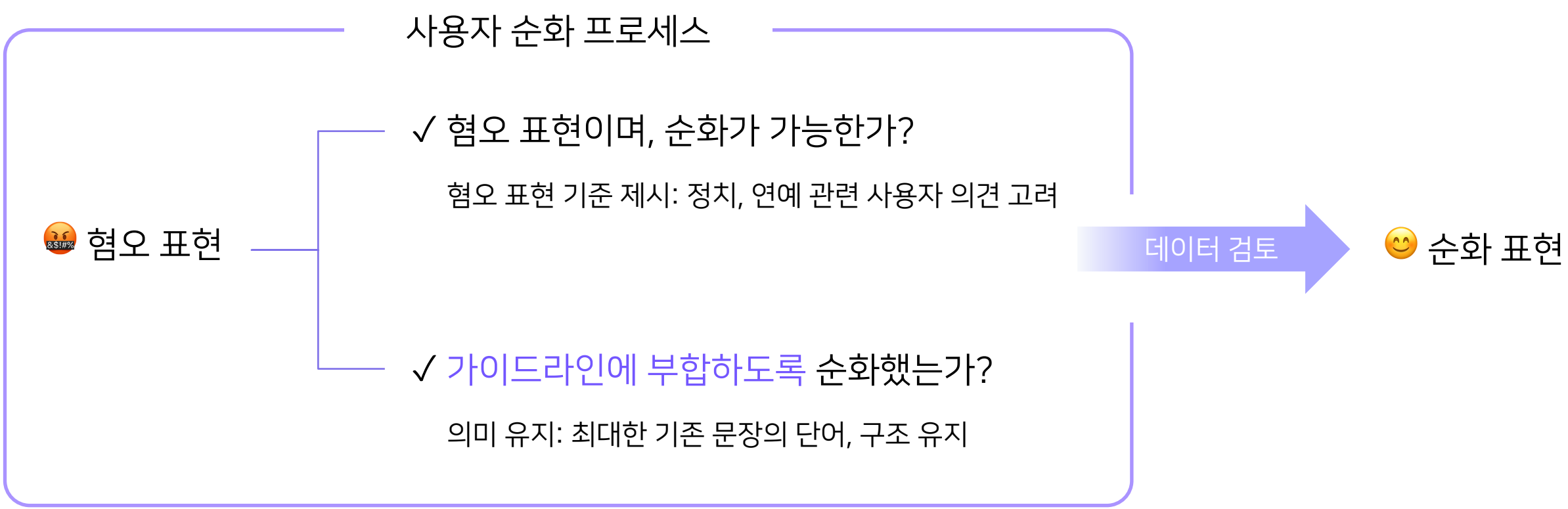
☹ 실제 댓글에서 나타나는 혐오 인물 비방, 극단적인 편견 등 를 반영하지 못함

순화 문장 생성을 위한 데이터 전략

🤔 어떤 데이터를 어떻게 사용하지?

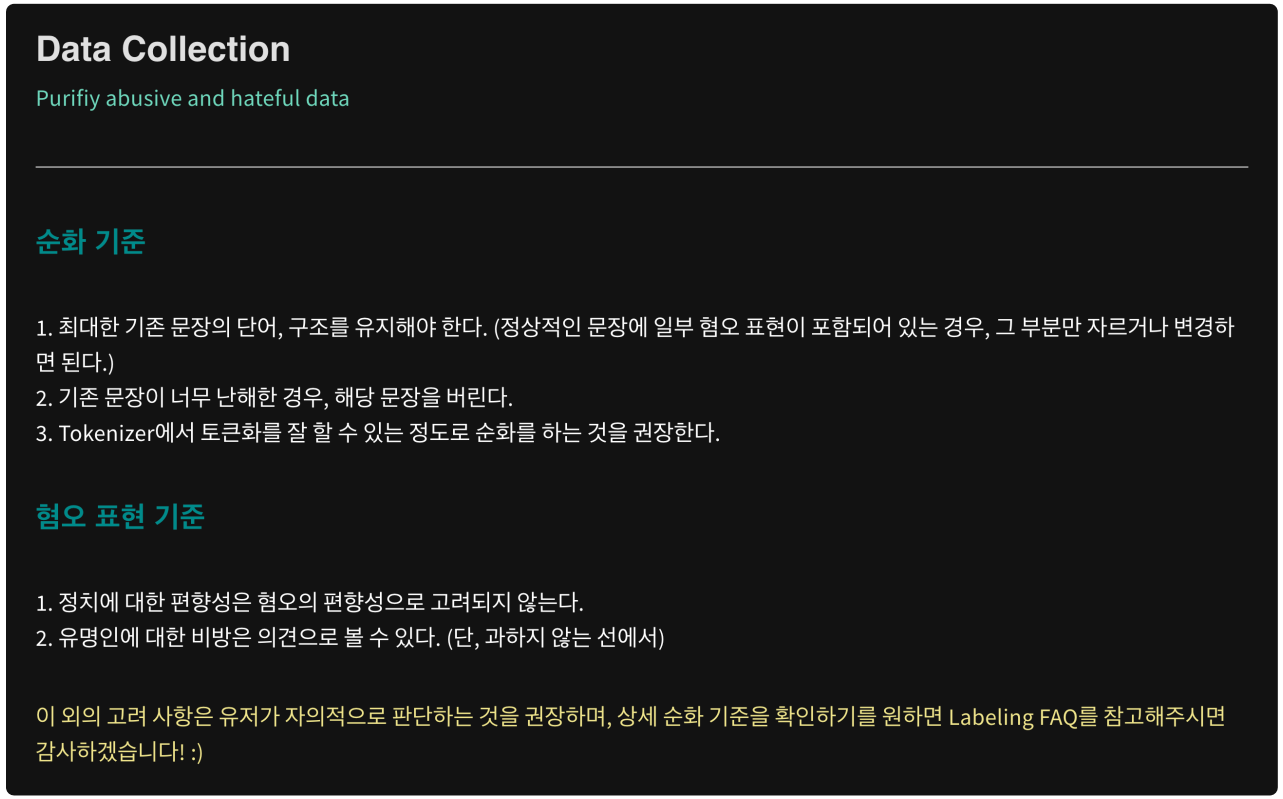
Parallel: 사용자 참여 대응 순화 데이터 수집

욕설 및 혐오 표현에 대한 사용자 순화 결과 수집



Parallel: 사용자 참여 대응 순화 데이터 수집

사용자 순화 데이터 수집

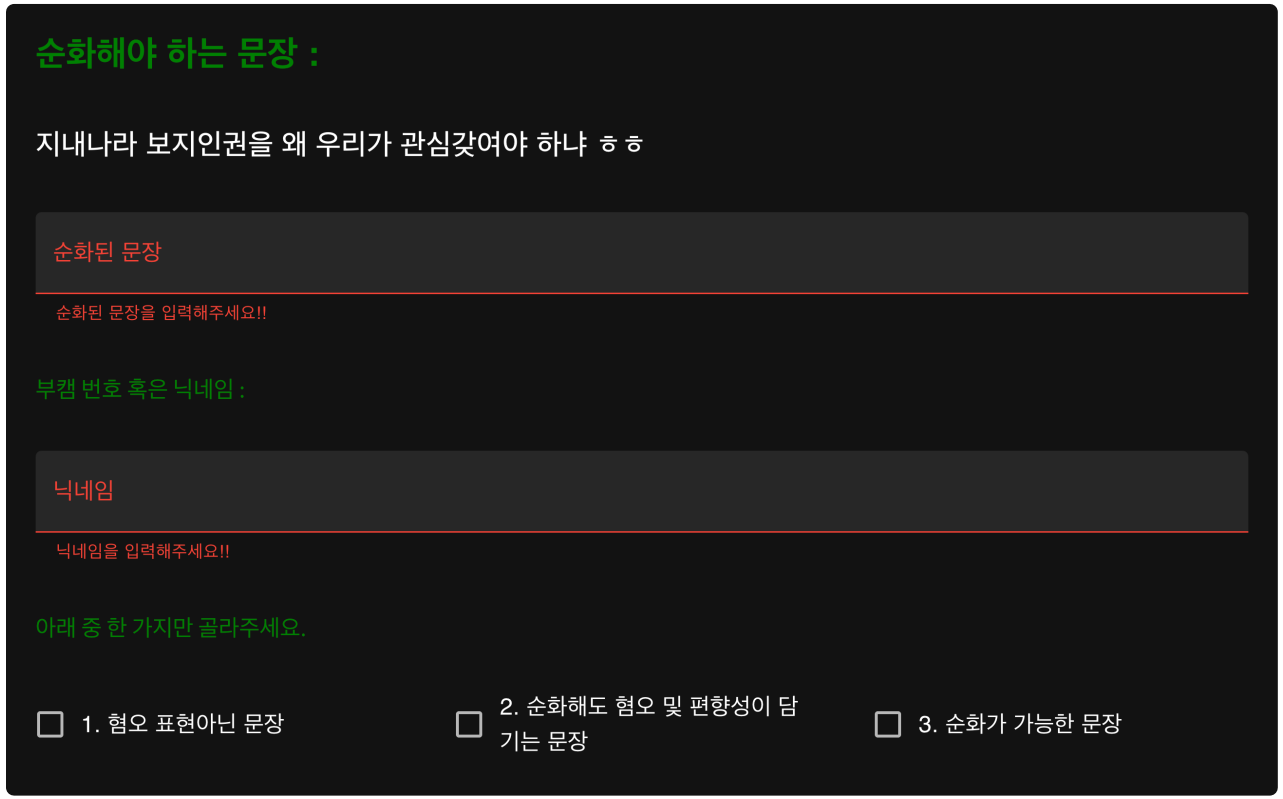


순화 및 혐오 표현 핵심 기준 제시

FAQ 기준 상세 설명, 순화 예시 제시

Parallel: 사용자 참여 대응 순화 데이터 수집

사용자 순화 데이터 수집

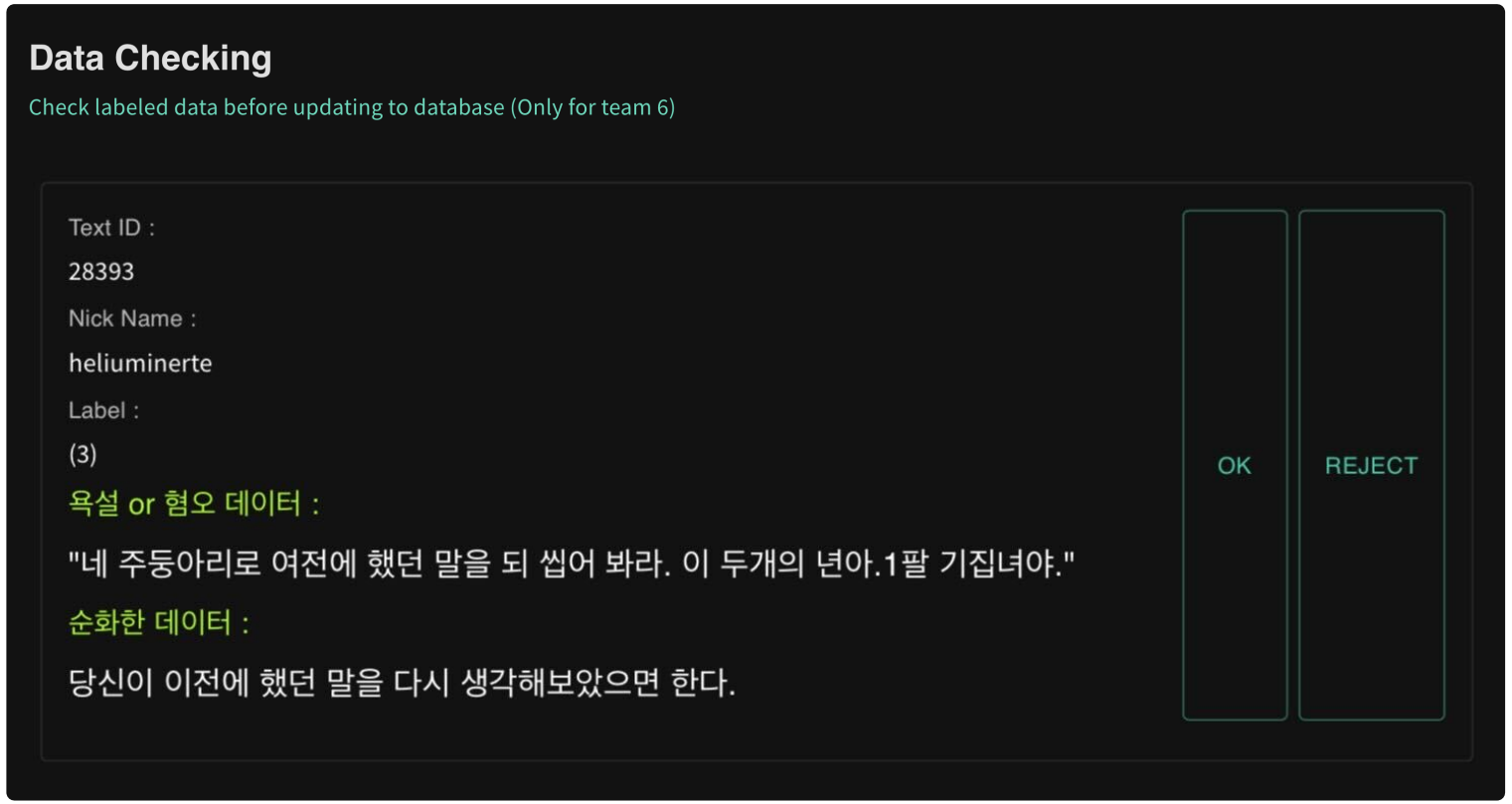


순화 대상 혐오 표현 제시

해당 문장 순화 결과 작성 및
순화 여부 분류 선택

Parallel: 사용자 참여 대응 순화 데이터 수집

사용자 순화 결과 검토



사용자 순화 결과 검토 및
반영 여부 결정

Parallel: 사용자 참여 대응 순화 데이터 수집

데이터셋 구축 결과

👏 총 3,133개의 혐오 문장 – 순화 문장 쌍 수집 👏

참여자 22명

수집된 순화 문장 5,256개

Parallel: 사용자 참여 대응 순화 데이터 수집

데이터셋 구축 결과

단순 욕설 순화

자기 생각대로 하는거지 꼰대마냥 이래라저래라 존나 강요하네	자기 생각대로 하는 건데 자신의 생각을 너무 강요를 한다.
IT는 쫏도 모르면서 틀딱 새끼가 나대네. ㅋ	IT에 대해 잘 알고 의견을 공유했으면 좋겠다.

비하 / 비꼬는 표현 순화

나라도 믿을만한 재단에 계속 기부할거같은데? 선행해도 ㅈㄹ이네?	나라도 같은 재단 믿을만한 재단에 계속 기부할 거 같은데? 선행에 과도하게 참견하지 말자.
할게없어서 감옥소재로 드라마를찍냐 다음엔 범죄 저지른 연예인들 나오는 감빵 예능도 만들어라 ㅅㅂ	감옥 소재의 드라마는 적절하지 않다고 생각한다

4. Generation

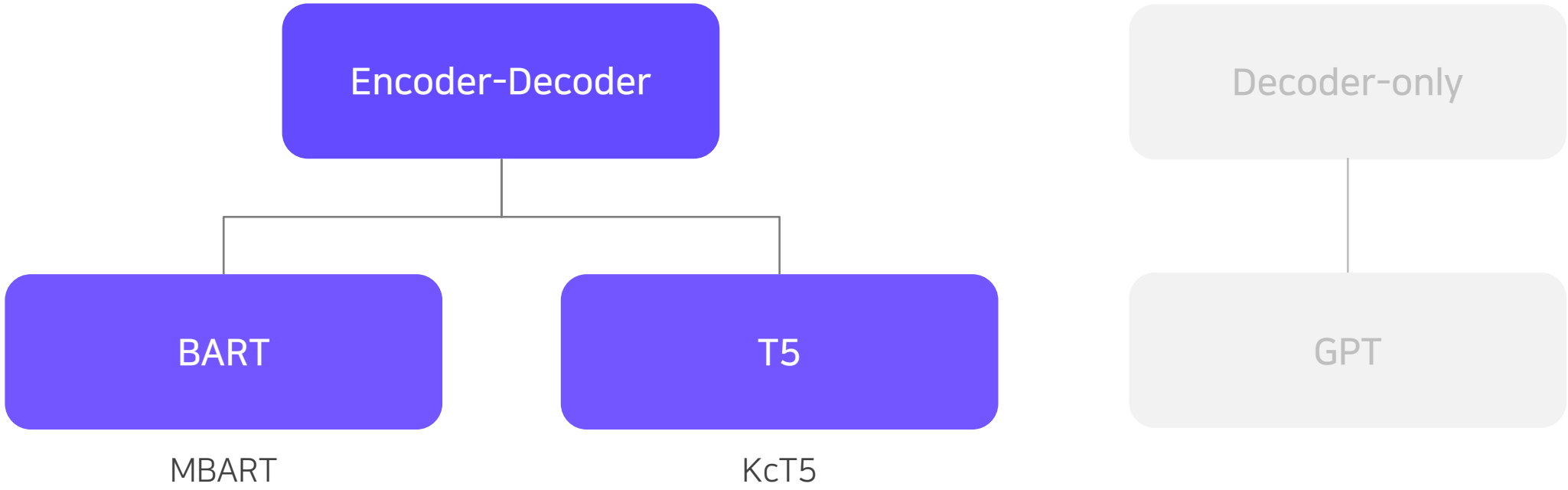
Model Selection

Generation Methods

Results

Model Selection

🤔 어떤 모델을 어떻게 사용하지?



Generation Methods

순화 성능 평가 지표 설정

🤔 어떤 문장이 순화가 잘 된 문장일까?

Cross-Entropy Loss

예시 순화 문장과 일치하는가?

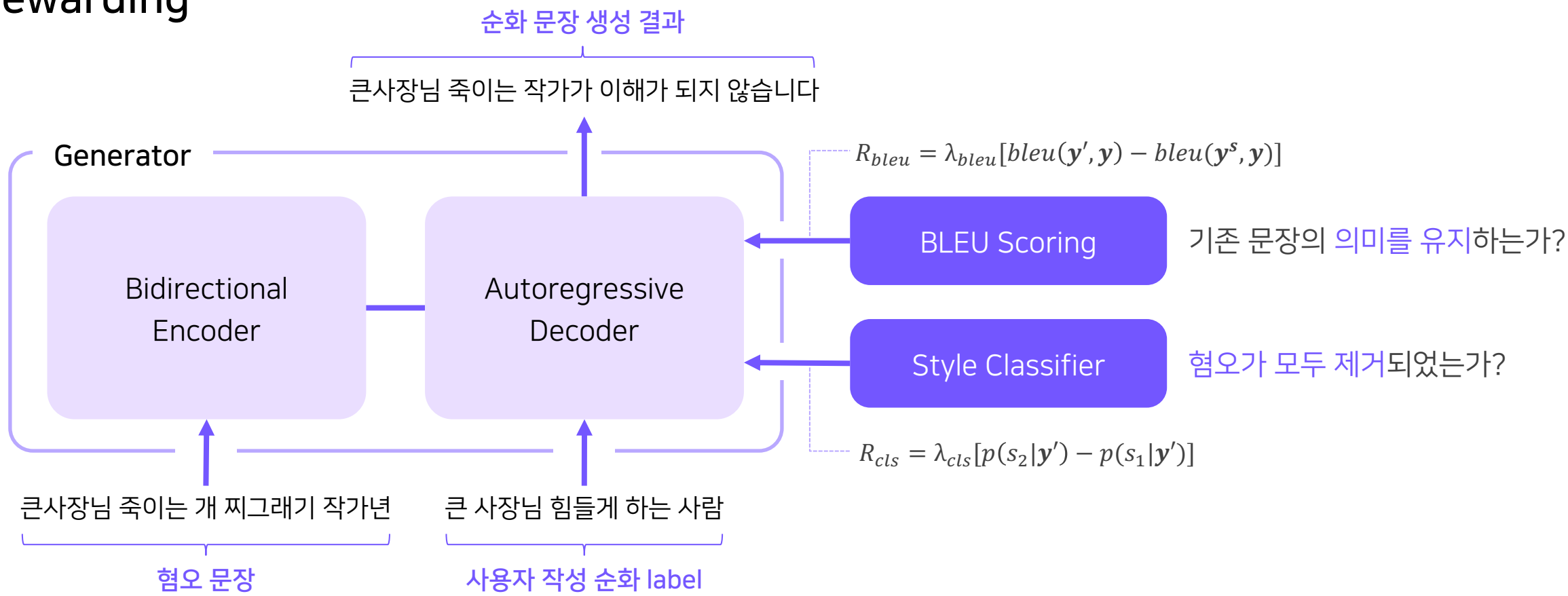
Human Evaluation

Classifier

혐오가 모두 제거되었는가?

Generation Methods

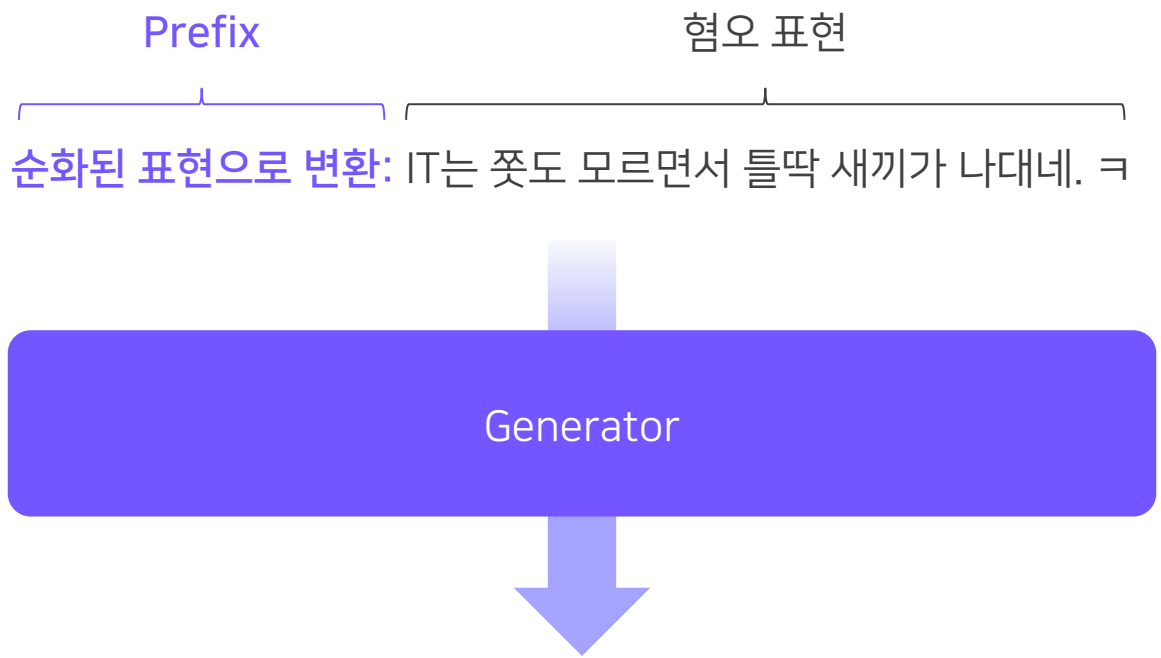
Rewarding



Lai, Huiyuan, Antonio Toral, and Malvina Nissim. "Thank you BART! rewarding pre-trained models improves formality style transfer." arXiv preprint arXiv:2105.06947 (2021).

Generation Methods

Prompt-based Learning



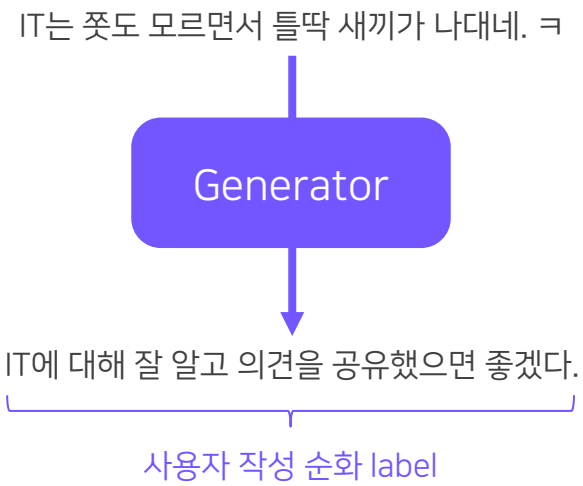
Ouyang, Long, et al. "Training language models to follow instructions with human feedback." arXiv preprint arXiv:2203.02155 (2022).

Generation Methods

Instructions with Human Feedback

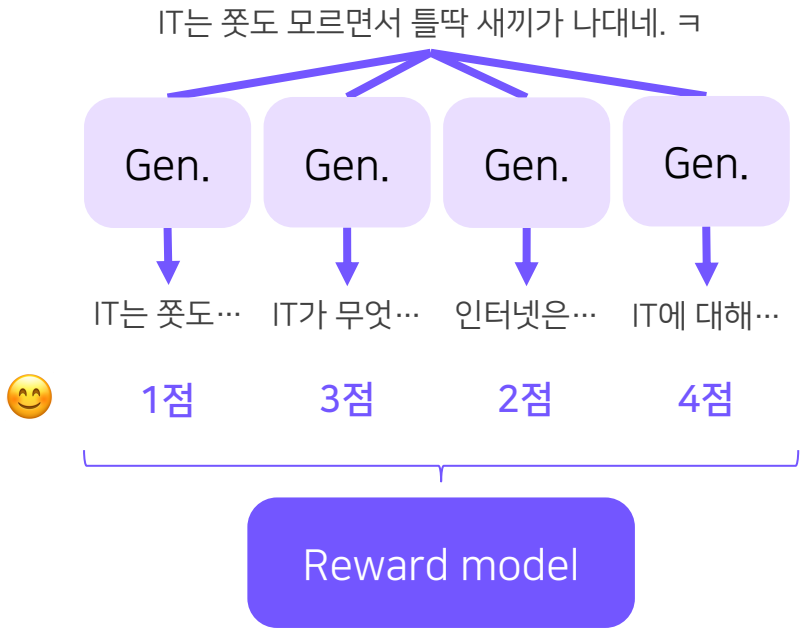
Supervised Fine-Tuning

Parallel data로 generator fine-tuning



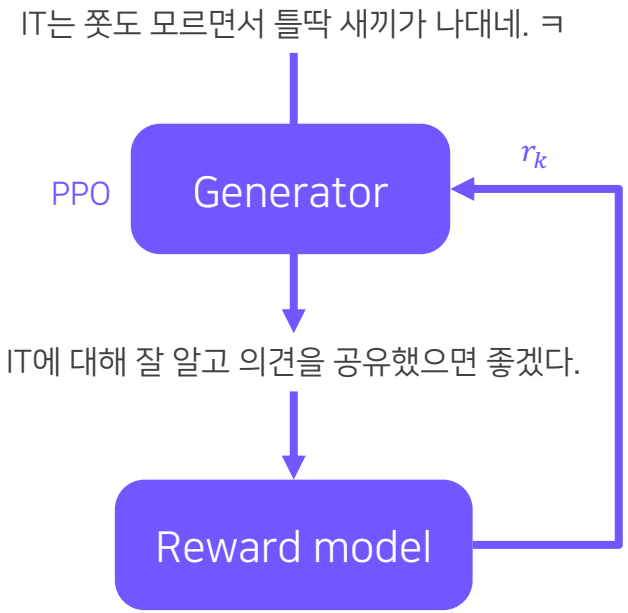
Reward Model 구축

여러 모델의 생성 결과 human ranking
→ Reward model 학습



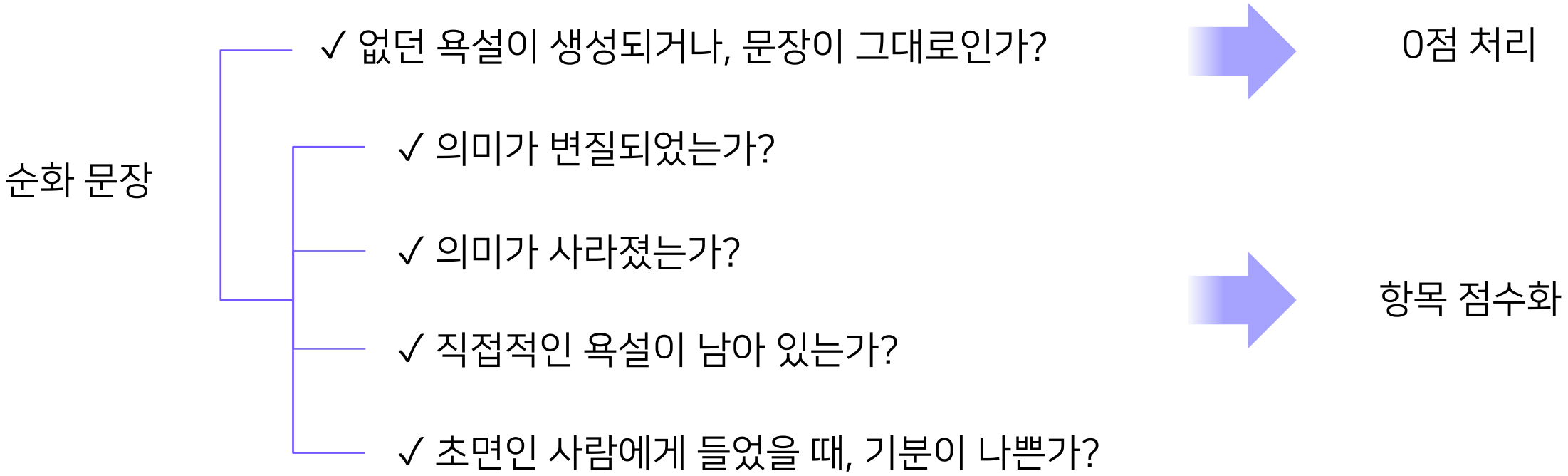
Reinforcement Learning

Reward를 이용해 policy update



Results

User Evaluation



Results

User Evaluation

	Raw 학습	Reward	Reward + Prompt	Instruct + Prompt
평가 평균 점수	59	64	**66**	63
가장 높은 평가 비율	12%	26%	**38%**	24%
가장 낮은 평가 비율	48%	20%	**10%**	22%
평가 점수 표준편차	0.87	0.9	**0.84**	0.86

😊 Reward + Prompt 모델이 프로젝트에 적합하다고 생각되어 최종 순화 모델로 선정

5. Architecture

Overall Architecture

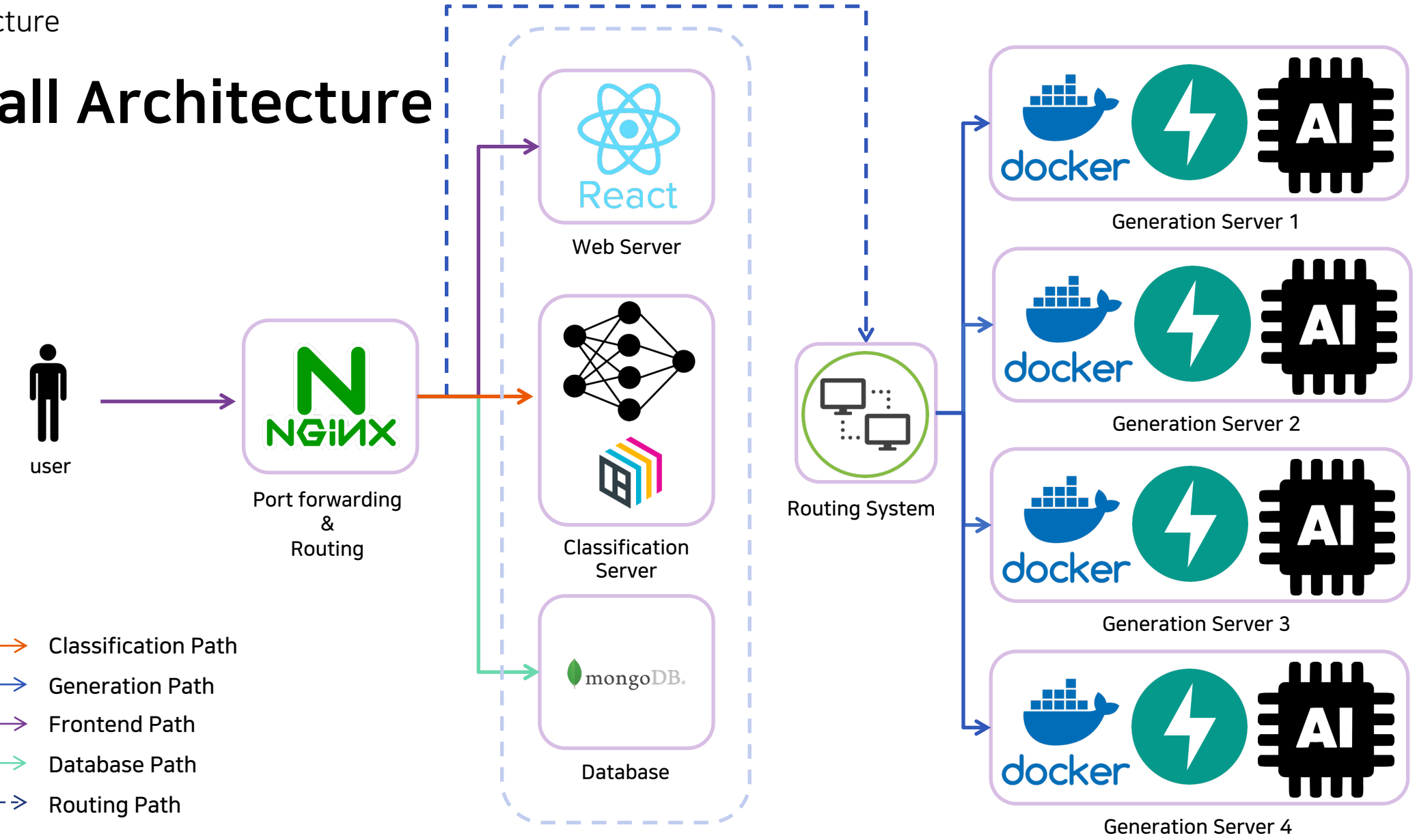
Stress Test (Classification)

Stress Test (Generation)

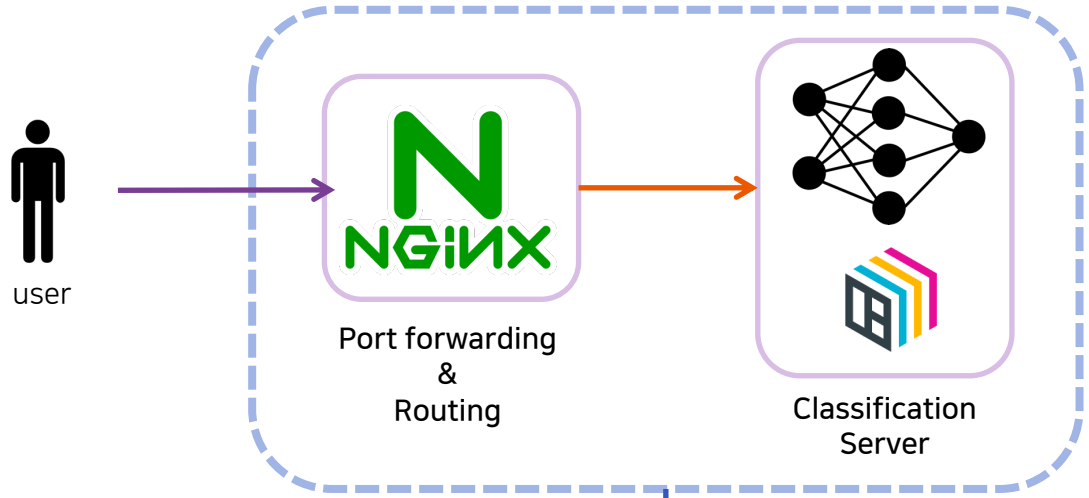
Server

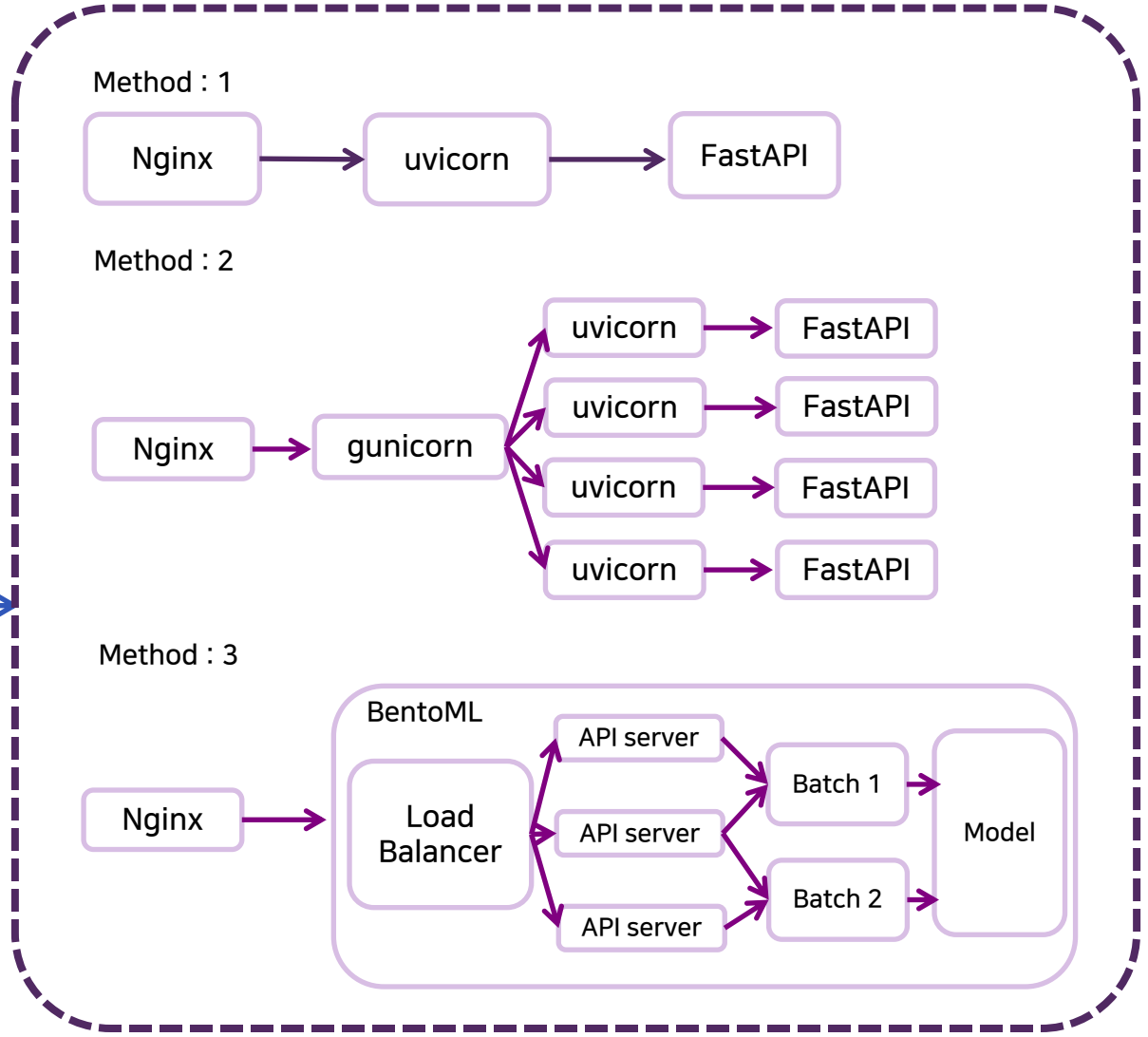
Real Time Test

Overall Architecture



Stress Test (Classification)

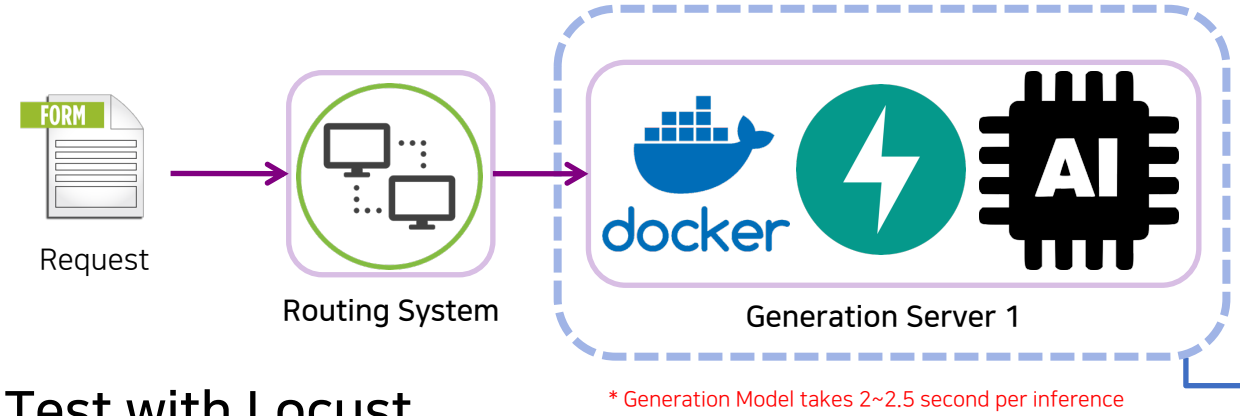




Test with Locust

	FastAPI + uvicorn	FastAPI + uvicorn	FastAPI + gunicorn	FastAPI + gunicorn	BentoML
Num workers	1	1	1	1	1
Num model	1	4	4	1	1
RPS	71	71	73	75	210

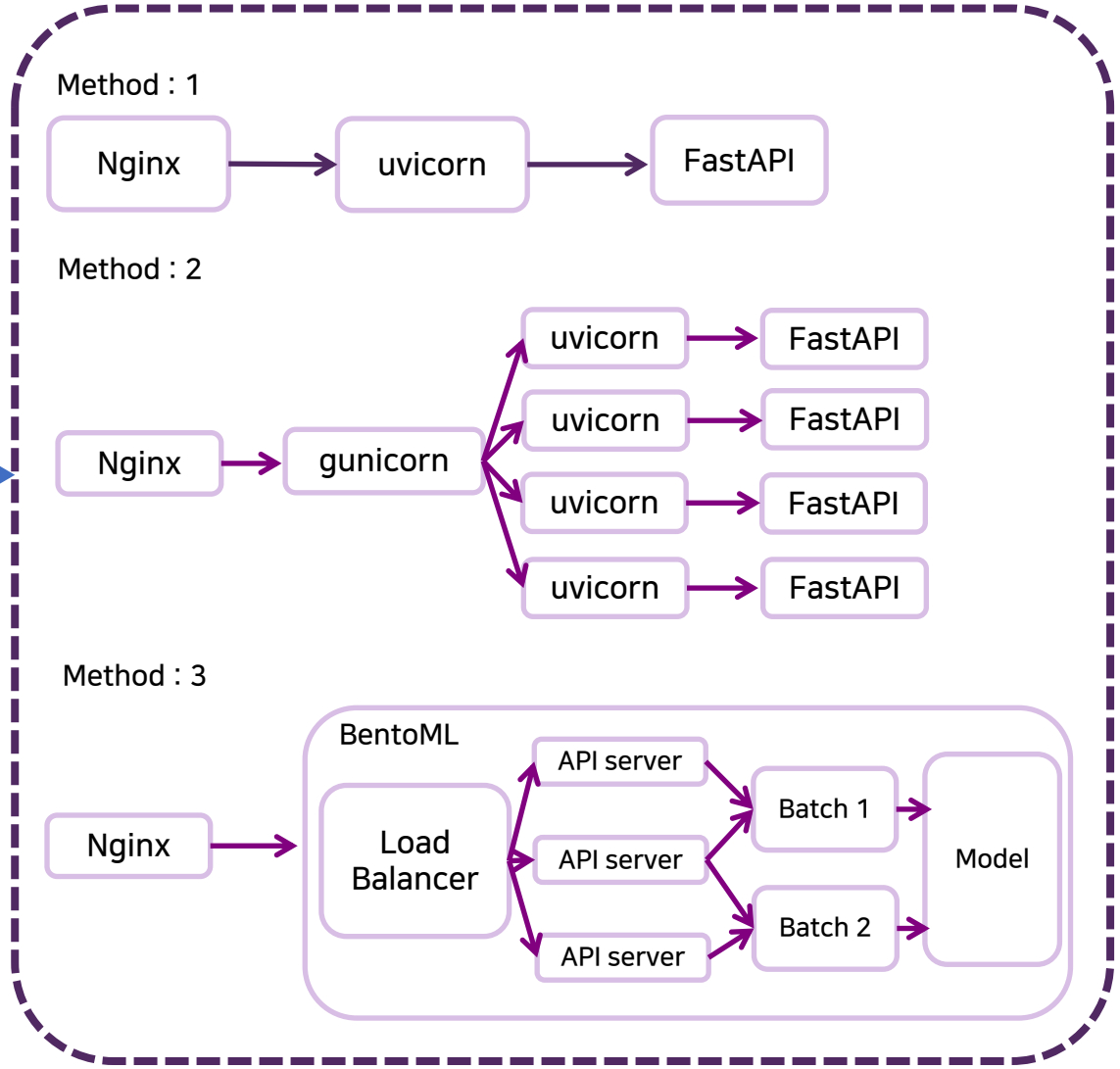
Stress Test (Generation)



* Generation Model takes 2~2.5 second per inference

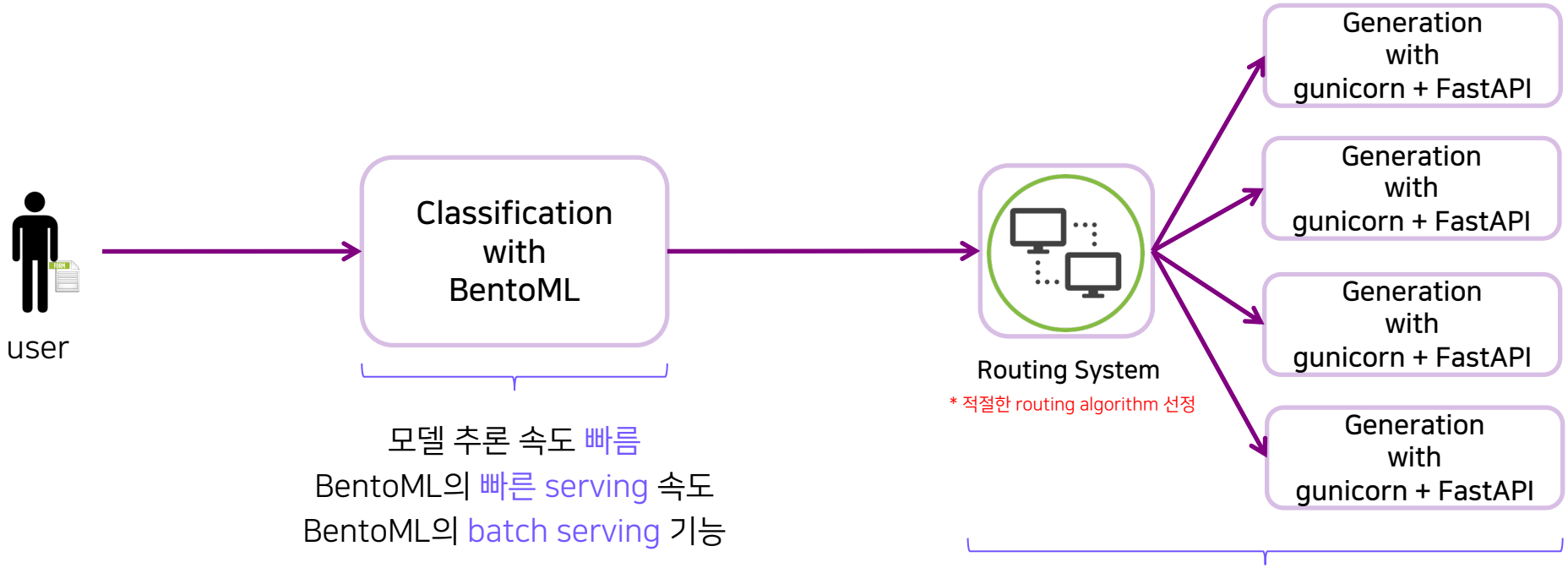
Test with Locust

	Uvicorn	Uvicorn	Uvicorn	Gunicorn	Gunicorn	BentoML	BentoML
Num workers	1	1	1	2	3	initial	initial
Model per worker	1	2	1	1	1	1	1
Num server	1	1	2	2	4	2	4
RPS	2.1	2	4.2	8.3	24	4.6	9.4
95th percentile response time(ms)	4700	6100	2400	2500	1400	2600	880



Server

🤔 어떤 구조가 서비스에 제일 적합할까?



모델 추론 속도 빠름
BentoML의 빠른 serving 속도
BentoML의 batch serving 기능

모델 추론 속도 느림
여러 서버 동시 운영 필요
Batch serving 시 대기 시간 발생
비혐오 표현인 경우 generation 수행하지 않음

Real Time Test

Single User Test

Multiple User Test

Scenario-based Test

	Single User Test	Multiple User Test	Multiple User Test	Scenario based Test	Scenario based Test	Scenario based Test
Num Users	1	3	5	8	10	15
Request per 1 Second	x	x	x	24	40	60
Score (1~5)	4.1	3.7	3.6	x	x	x

👍 플랫폼 사용 문제 없음

👎 단시간에 많은 request 발생 시 서버 부하

👎 Generation 최적화 필요

End of Document

Thank You.