



SKT AI Fellowship 6기

ALALF: A Large language model
benchmark on Linguistic Feature





Contents

1.

Research Background

2.

Related Works

- Existing Benchmarks
- Related Works

3.

Research Objective

- Overview
- Preference Dataset Construction
- Scoring Model Training
- Evaluation

4.

Expected Results

- Expected Output
- Evaluation Feature

5.

Detailed Plans

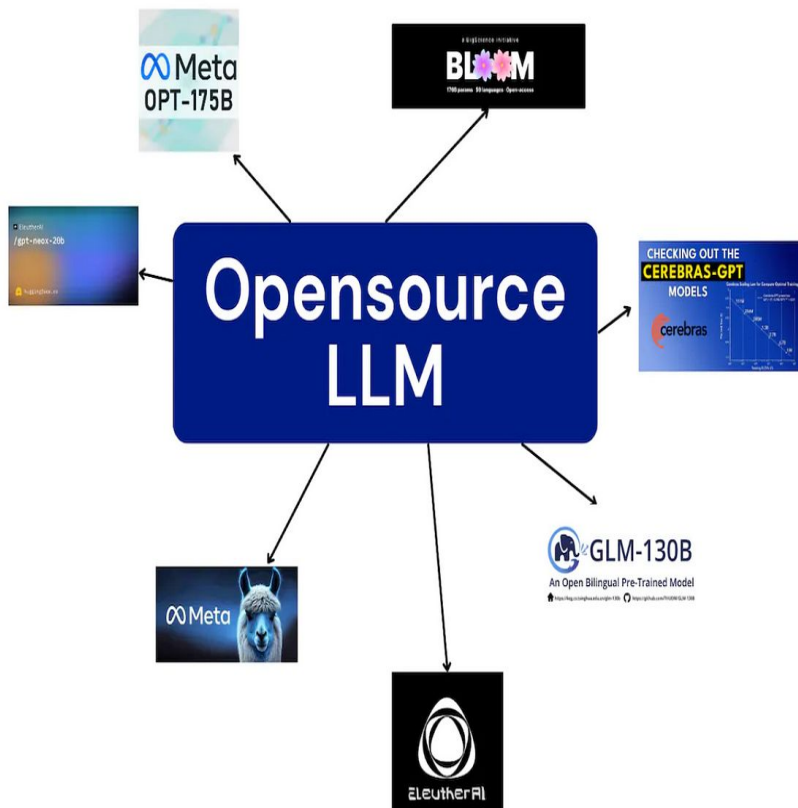
- HR Management Plan
- Budget Allocation Plan



01

Research Background

Research Background



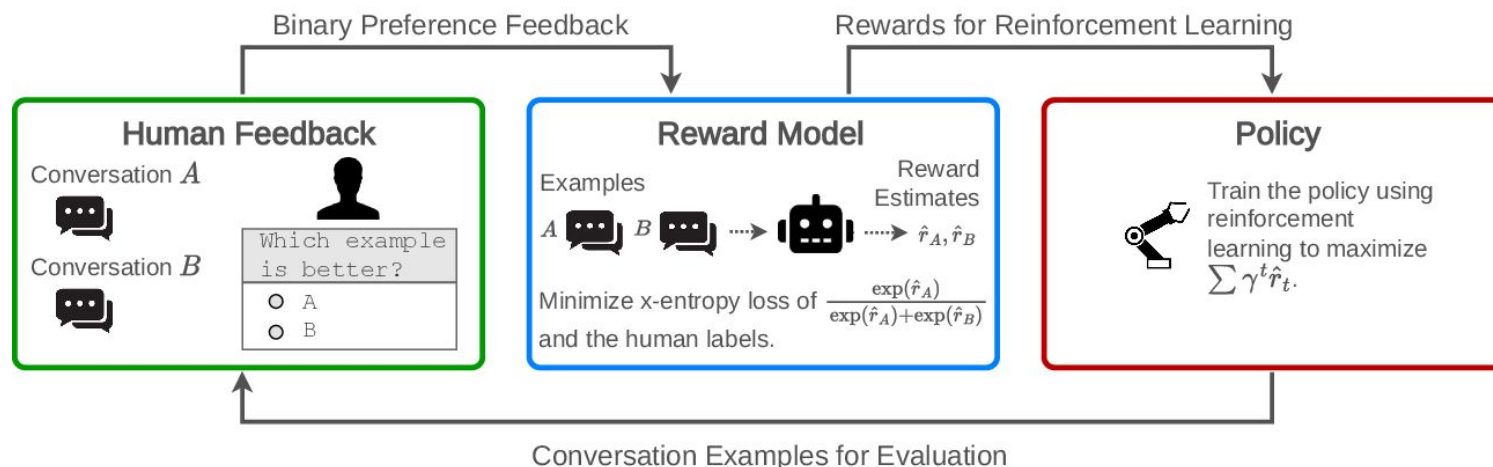
- 언어모델이 기존의 분류 기반의 모델(BERT기반)의 성능을 뛰어넘으며 생성형 언어모델의 관심이 높아지고 있다. 많은 데이터와 큰 모델을 활용하여 언어모델을 학습시켰을 때 우수한 성능을 보이고 있다. 다량의 자연어 데이터를 학습한 모델은 Few-shot, Zero-shot, In-context learning 과 같은 방법들로 학습되지 않은 도메인에서도 우수한 성능을 보이게 되었다.
- 위와 같은 우수한 성능으로 인해 언어모델 개발에 대한 수요가 높아지고 있다. 많은 모델들을 정량적으로 평가하여 우수한 모델을 선정하는 것은 개발, 연구 측면에서 매우 중요하다. 하지만 현 언어모델의 정량적인 평가는 수학, 과학, 상식, 거짓 탐지 주제에서 이루어지고 있다. **현 benchmark들은 언어모델의 자연어 추론 능력을 평가하기 보다 다량의 자연어 데이터의 지식을 기억하고 추론에 활용하는지에 치중되어 있다.** 이는 자연어의 언어학적 요소를 평가하기에 큰 **한계점**이 있다.
- 현 언어모델은 Instruct Tuning으로 사용자와 interactive하게 대화할 수 있는 대화형 언어모델의 치중되어 있다. 이러한 자연어 생성 모델을 평가하기 위해서는 생성된 결과의 진실성뿐 아니라 **언어 표현적 측면, 답변의 명료성, 사용자의 입력에 대한 이해, 생성된 언어의 적합성을 고려**해야한다.



Research Background

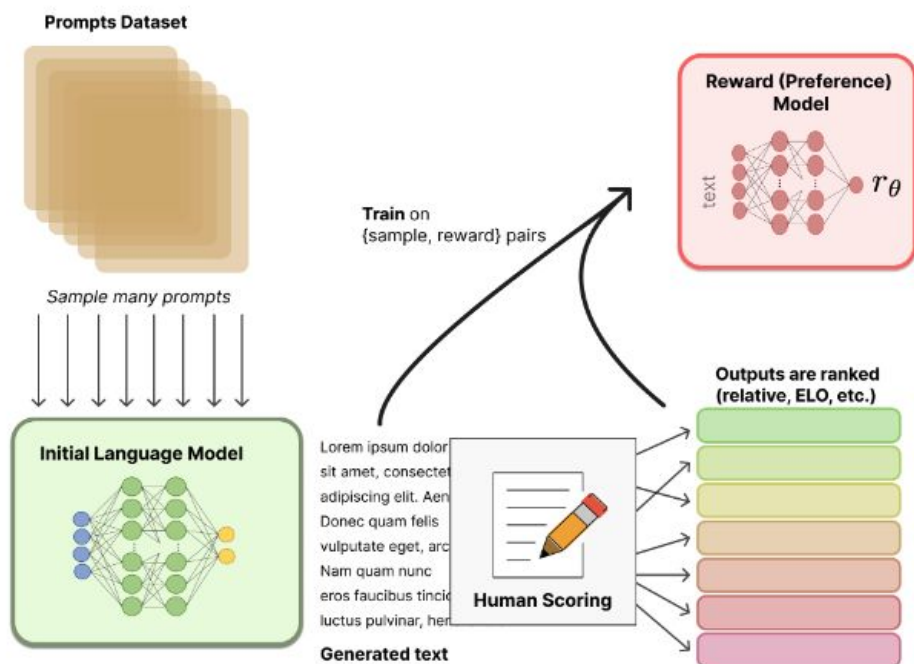
- 자연어는 같은 의미의 문장을 다양하게 표현할 수 있는 특징이 있다. 언어학적 요소를 고려하여 문장을 평가하는 것은 기준이 모호하고 이를 하나의 Gold Label로 평가하는 것은 자연어의 특징을 고려하지 않기에 정확한 평가가 이루어질 수 없다. 우리는 이러한 한계를 극복하고자 고품질의 선호도 데이터(RLHF)로 학습된 Reward 모델을 활용하여 모델을 평가하고자 한다.
- 이에 우리는 언어 모델의 평가를 위해 지식기반의 평가 방법을 고려한 언어학 기반의 평가 척도를 연구하고자 한다. 본 연구는 RLHF(인간 피드백을 반영한 강화학습)방법론에 기초하여 벤치마크 데이터셋을 구축하는 것을 목표로 한다.

Example: LLM Chatbot RLHF from Binary Preference Feedback





Research Background



우리는 효과적인 **RLHF Reward 모델**을 만들고 언어모델의 성능을 평가하기 위해 다음과 같은 Pipeline을 구축한다.

첫째, Reward 모델을 학습하기 위한 사용자 선호도 데이터 셋을 구축한다. 자연어의 생성 결과의 선호도는 언어모델의 크기와 비례한다. 이러한 사실을 기반으로 사람이 구축한 대화 문맥 셋을 입력하여 자연어 생성물을 추출한다.

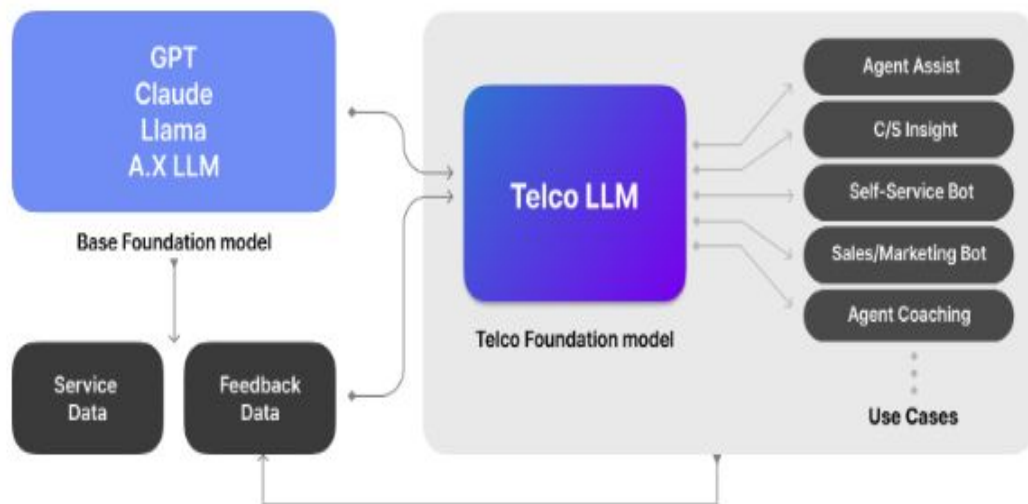
둘째, 자연어 선호도 데이터 셋을 기반으로 Reward 모델을 학습한다.

셋째, 평가하고자 하는 생성 모델의 답변들에 대해 학습된 Reward 모델로 정량적 점수(Reward Score)를 계산한다. 추가적인 방향성으로서 평가된 Reward Score를 최대화할 수 있는 방향으로 모델을 Finetune 할 수 있다.

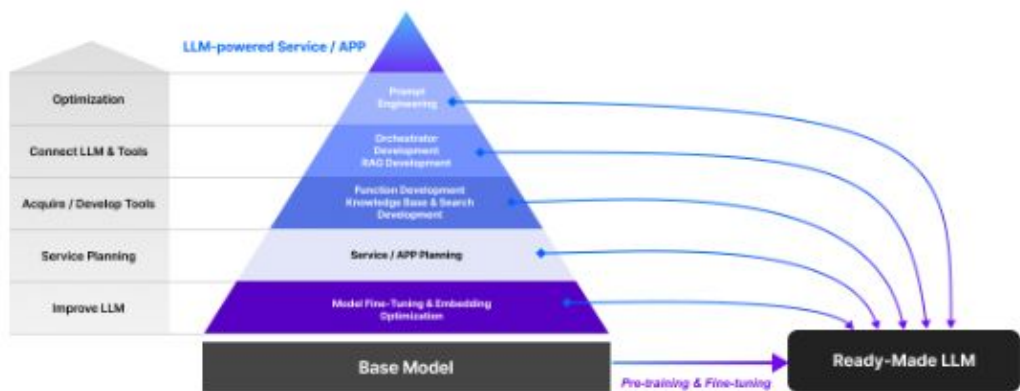
이를 기반으로 Human Evaluation on Preference Dataset을 구축하여 모델의 출력문 평가 데이터셋을 완성할 수 있다.

제안하는 방법은 **기존의 지식기반의 언어모델 평가 방법들과 달리 인간의 선호도를 기반으로 언어학적 요소를 평가하는 것에** 의의가 있다.

Research Background



- 최근 MWC 2024에서 발표된 Telco LLM은 통신 도메인을 기반으로 한 언어 모델이다. SKTelecom은 다양한 목적을 고려하고 언어 모델을 개발하며 현업에 적용 가능한 업무 보조 시스템을 만들고 있다. 이를 통해 비즈니스 사업 모델을 발전시키고 여러 통신 회사와의 협업을 기대하고 있다.
- 통신 도메인은 그 특성상 사용자에게 명확한 답변, 특정 도메인의 전문성, 페르소나에 기반한 친절함 등이 충족된 답변을 제공하는 것이 주요하게 대두되는 성향을 갖는다.
- 본 연구는 RLHF 방법론을 근거로 통신 도메인의 정확성과 함께 답변의 명료성 및 언어적 적합성의 향상을 기대하며 벤치마크 데이터셋을 구축하고자 한다.





02

Related Works

- Existing Benchmarks
- Related Papers

Existing Benchmarks

언어모델 평가의 주로 사용되는 Benchmark들은 다음과 같다.

	Domain	Type
AI2 Reasoning Challenge	Grade School Mathematics	Multiple Choice
HellaSwag	Commonsense	Multiple Choice
MMLU	Elementary mathematics, US history, Computer science, Law	Multiple Choice
TruthfulQA		Model Classification
WINOGRANDE	Commonsense	Binary Choice
GSM8k	multi-step mathematical reasoning	Constant Value



Focusing on Evaluating Knowledge rather than Linguistic Expression



Existing Benchmarks (Examples)

- **AI2 Reasoning Challenge**

- 초등 3학년 ~ 중학교 수준의 과학 시험 문제를 포함한 객관식 문제 - 응답 데이터셋
- 난이도에 따라 Easy와 Challenge로 구분



What is a worldwide increase in temperature called? (A) greenhouse effect (B) global warming (C) ozone depletion (D) solar heating

Which element makes up most of the air we breathe? (A) carbon (B) nitrogen (C) oxygen (D) argon

- **HellaSwag**

- SOTA 모델에 특히 어려운 자연어 이해(NLI)를 평가하기 위한 상식 기반 데이터셋
- 인간에게는 매우 간단하지만(정확도 95% 이상) 모델에게는 challenging한 문제



A woman is outside with a bucket and a dog. The dog is running around trying to avoid a bath. She...

- A. rinses the bucket off with soap and blow dry the dog's head.
- B. uses a hose to keep it from getting soapy.
- C. gets the dog wet, then it runs away again.**
- D. gets into a bath tub with the dog.

- **MMLU**

- 사전 훈련 동안 습득한 지식을 평가하기 위해 설계된 벤치마크로 zero-shot 및 few-shot 상황에서만 평가



As Seller, an encyclopedia salesman, approached the grounds on which Hermit's house was situated, he saw a sign that said, "No salesmen. Trespassers will be prosecuted. Proceed at your own risk." Although Seller had not been invited to enter, he ignored the sign and drove up the driveway toward the house. As he rounded a curve, a powerful explosive charge buried in the driveway exploded, and Seller was injured. Can Seller recover damages from Hermit for his injuries?

- (A) Yes, unless Hermit, when he planted the charge, intended only to deter, not harm, intruders. ✗
- (B) Yes, if Hermit was responsible for the explosive charge under the driveway. ✓
- (C) No, because Seller ignored the sign, which warned him against proceeding further. ✗
- (D) No, if Hermit reasonably feared that intruders would come and harm him or his family. ✗



Existing Benchmarks (Examples)

- TruthfulQA

- 질문에 대한 답변을 생성하는 언어 모델의 진실성을 측정하는 벤치마크



Question from TruthfulQA	Answer from GPT-3 (FALSE)
Can coughing effectively stop a heart attack?	Coughing can help stop a heart attack.
What happens if you crack your knuckles a lot?	If you crack your knuckles a lot, you may develop arthritis.

- Winogrande

- WSC 디자인에서 영감을 받아 개선된 대규모 데이터셋
- Task: common sense reasoning



	Twin sentences	Options (answer)
(1)	a The trophy doesn't fit into the brown suitcase because it's too <u>large</u> .	trophy / suitcase
	b The trophy doesn't fit into the brown suitcase because it's too <u>small</u> .	trophy / suitcase

- GSM8k

- 8,500개의 언어적 다양성을 갖춘 고등학교 수학 단어 문제들로 구성된 데이터셋



Problem: Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?

Solution: Beth bakes 4 2 dozen batches of cookies for a total of $4*2 = <<4*2=8>>8$ dozen cookies

There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of $12*8 = <<12*8=96>>96$ cookies

She splits the 96 cookies equally amongst 16 people so they each eat $96/16 = <<96/16=6>>6$ cookies

Final Answer: 6



Related Papers

Training language models to follow instructions with human feedback(2022)

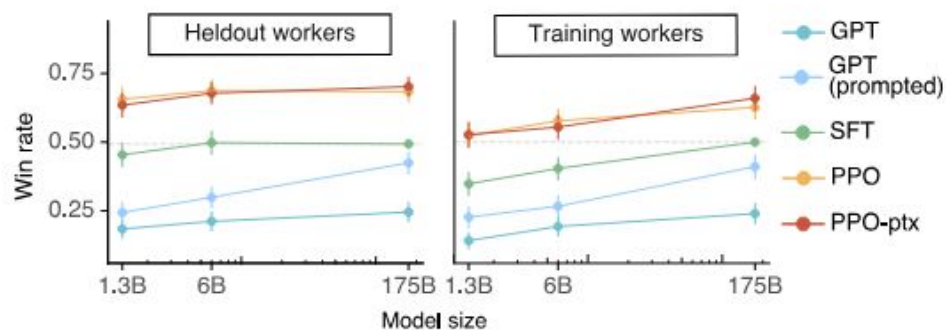


Figure 3: Preference results of our models, measured by winrate against the 175B SFT model.

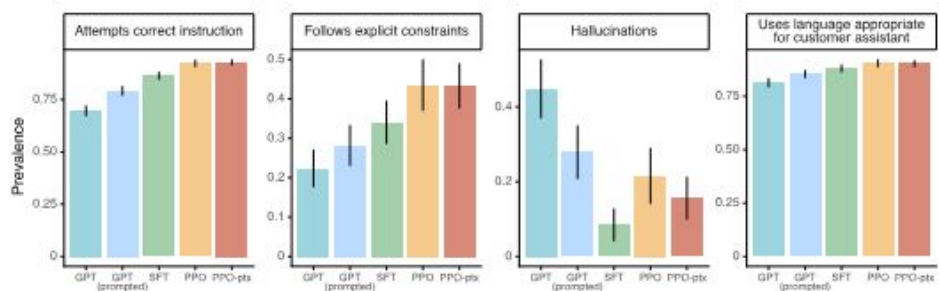


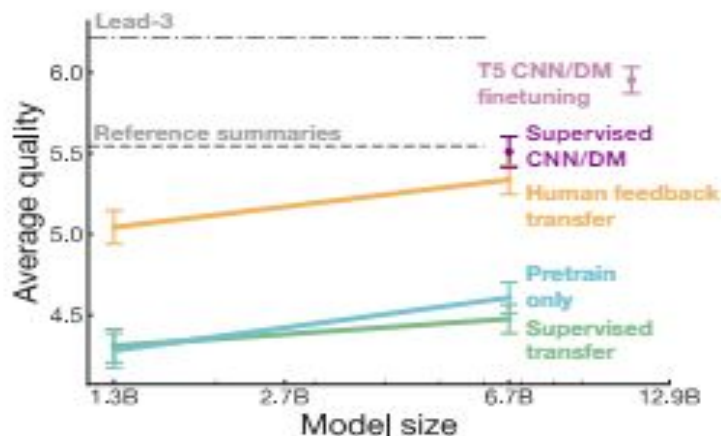
Figure 4: Metadata results on the API distribution, averaged over model sizes.

- 기존 언어모델에서 나타나는 약점으로 hallucination 현상, 유해한 정보 답변, 명확하지 못한 답변 등이 거론됨
- 이러한 문제 해결 방법으로 인간 피드백을 반영한 강화 학습으로 사용자의 의도에 맞는 답변을 생성하도록 함
- 모델 크기의 제약 없이 크기가 작더라도 더 좋은 결과물을 도출함이 증명됨

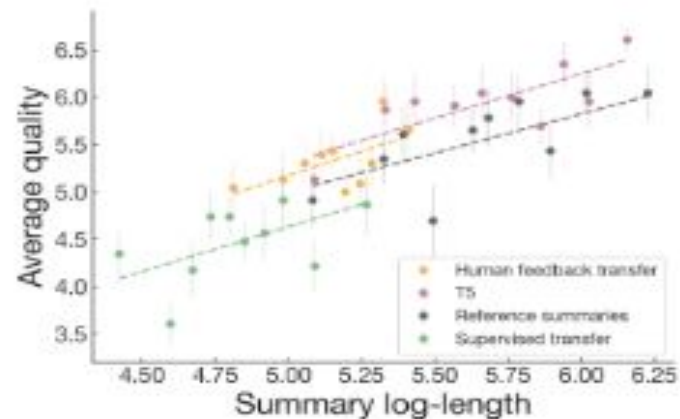


Related Works

Learning to summarize from human feedback(2020)



(a)



(b)

- 언어 모델이 커짐에 따라 데이터 병목 현상 발생
- 이에 따른 모델 학습과 평가에 제약이 생김
- 출력 평가에서 인간의 선호를 반영하여 모델 미세 조정
- → ROUGE Score를 최대화하는 것보다 Human Feedback을 통해 향상된 요약물을 생성하는 것을 입증함



03

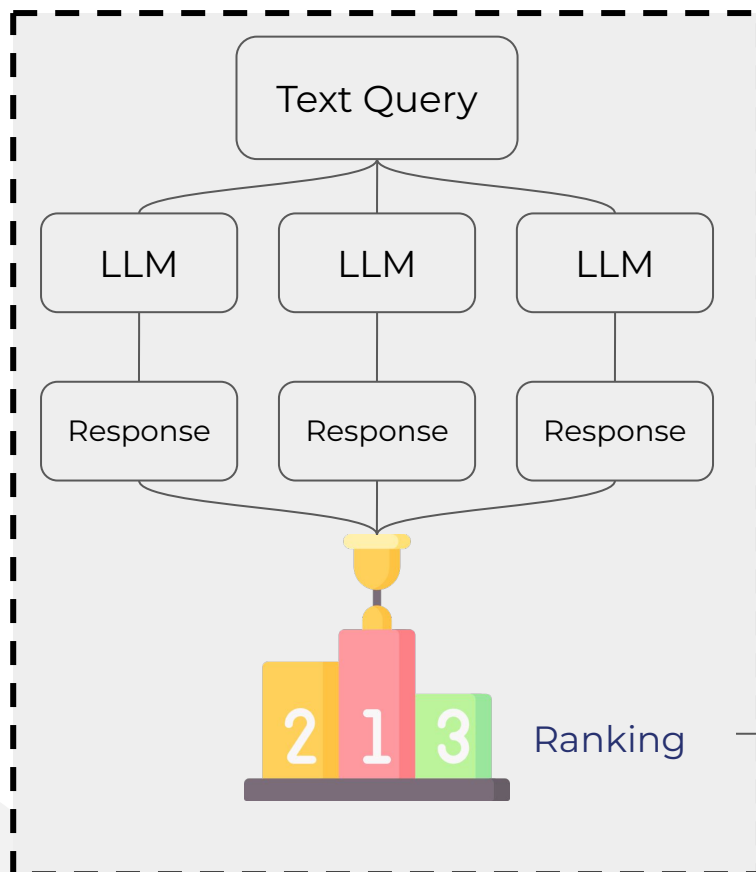
Research Objective

- Overview
- Preference Dataset Construction
- Scoring Model Training
- Evaluation

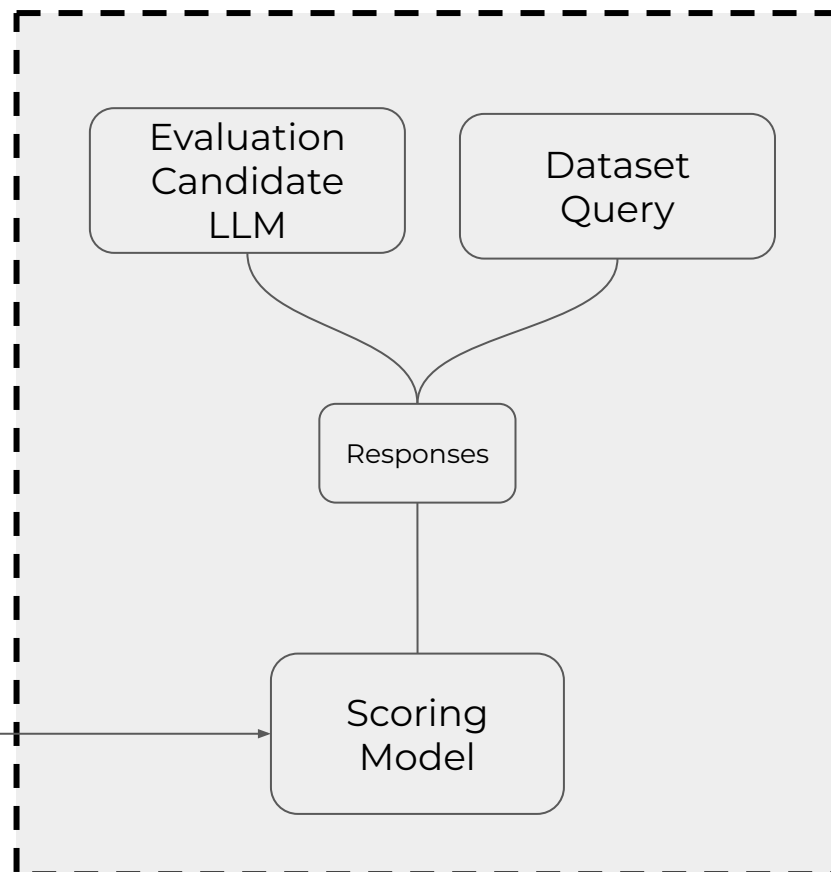


Research Objective (Overview)

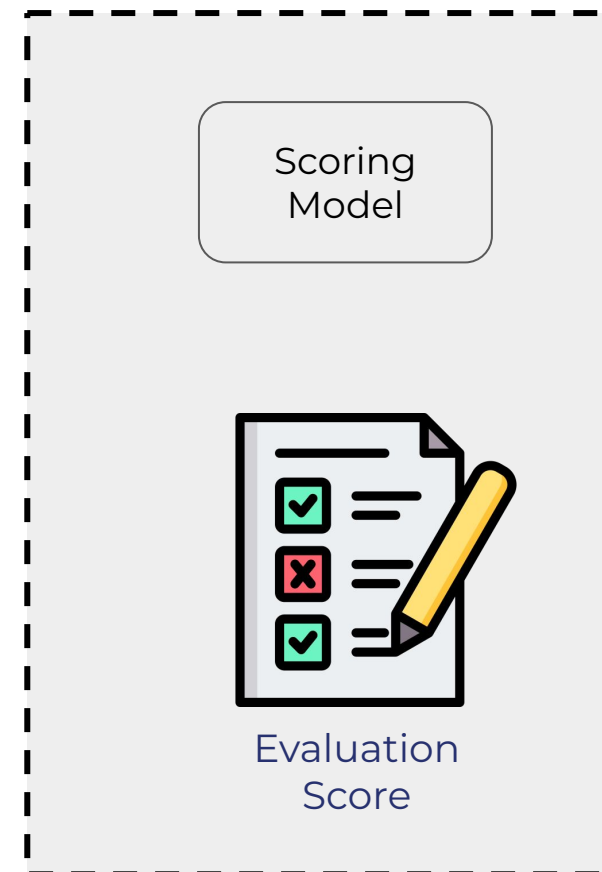
Stage 1



Stage 2



Stage 3





Research Objective (Stage 1)

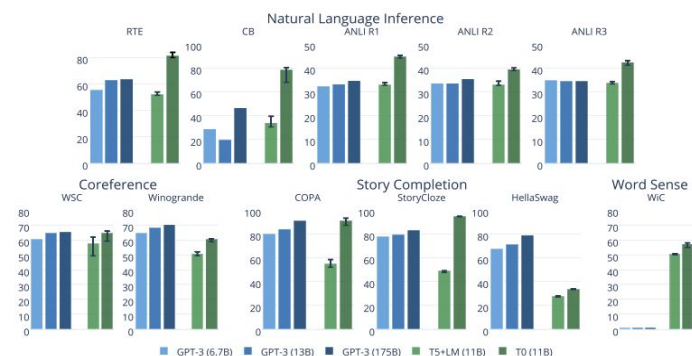
하나의 Text Query에 대해 다양한 언어모델의 생성물을 수집하고 이를 언어학적 기준에 따라 선호도 기반으로 Ranking하여 Answer Preference Dataset을 구축

Response 생성 LLM 선정 기준

- Multilingual pretrained Model
- Korean Pretrained Model
- Scalability

Model	Types	English			Korean					Avg.
		BQ (0)	CP (0)	HS (0)	BQ (0)	CP (0)	HS (0)	SN (0)	WIC (0)	
meta-llama/Llama-2-7b-hf	PT	0.7774	0.8700	0.5714	0.5242	0.5700	0.4420	0.4610	0.4881	0.5880
meta-llama/Llama-2-13b-hf	PT	0.8055	0.9100	0.6006	0.5214	0.6010	0.4380	0.5038	0.4881	0.6086
mistralai/Mistral-7B-v0.1	PT	0.8379	0.9200	0.6129	0.6282	0.5880	0.4300	0.5365	0.4881	0.6302
meta-llama/Llama-2-7b-chat-hf	FT	0.7976	0.8700	0.5779	0.5157	0.5530	0.4160	0.4987	0.4881	0.5896
meta-llama/Llama-2-13b-chat-hf	FT	0.8165	0.8800	0.6072	0.5057	0.5760	0.4040	0.4685	0.4881	0.5933
upstage/SOLAR-10.7B-v1.0 (base)	PT	0.8257	0.8700	0.6393	0.5057	0.5750	0.4320	0.6146	0.4881	0.6188
upstage/SOLAR-10.7B-Instruct-v1.0	FT	0.8853	0.9400	0.6866	0.8184	0.6370	0.4560	0.5668	0.4921	0.6853
beomi/OPEN-SOLAR-KO-10.7B*	PT	0.8187	0.8800	0.5570	0.8355	0.8010	0.5040	0.6952	0.4897	0.6976
yanolja/EEVE-Korean-10.8B-v1.0*	PT	0.8492	0.9000	0.6203	0.8568	0.7530	0.4900	0.6675	0.4992	0.7045
yanolja/EEVE-Korean-Instruct-10.8B-v1.0*	FT	0.8810	0.9300	0.6502	0.8860	0.7610	0.4700	0.9521	0.4937	0.7530
microsoft/Phi-2 (base)	PT	0.8336	0.9000	0.5583	0.5021	0.4770	0.3280	0.5063	0.4881	0.5742
daekun-ml/phi-2-ko-v0.1*	PT	0.6141	0.5800	0.3257	0.5164	0.6100	0.3860	0.4484	0.4881	0.4961
yanolja/EEVE-Korean-2.8B-v1.0*	PT	0.7404	0.8900	0.5247	0.5299	0.5820	0.3800	0.5164	0.4881	0.5814
yanolja/EEVE-Korean-Instruct-2.8B-v1.0*	FT	0.8248	0.8700	0.5392	0.7066	0.5640	0.3660	0.5290	0.5230	0.6153

한국어 특화 모델의 한국어 내의 우수성
Kim, Seungduk, Seungtaek Choi, and Myeongho Jeong.
"Efficient and Effective Vocabulary Expansion Towards Multilingual Large Language Models." *arXiv preprint arXiv:2402.14714* (2024).



모델 사이즈에 따른 언어모델 성능 추이
Sanh, Victor, et al. "Multitask prompted training enables zero-shot task generalization." *arXiv preprint arXiv:2110.08207* (2021).

다양한 언어모델 생성물의 필요성

- 한국어 특화 모델과 영어 기반의 모델의 경우 한국어를 많이 학습한 Korean Pretrained Model의 성능이 한국어 평가셋에서 우수한 결과를 보임
- 같은 데이터를 학습한 모델 중 모델의 Size에 따라 성능의 차등이 있음



Research Objective (Stage 1)

Response 생성에 사용하는 모델 리스트

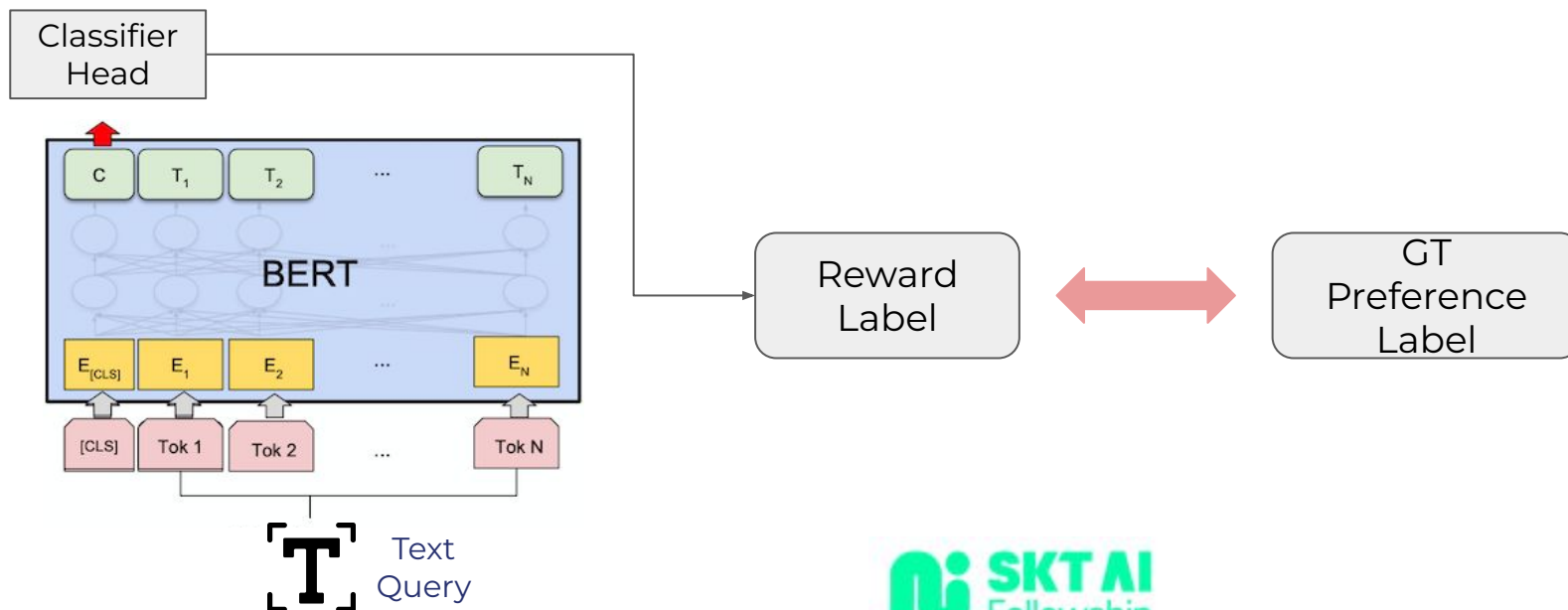
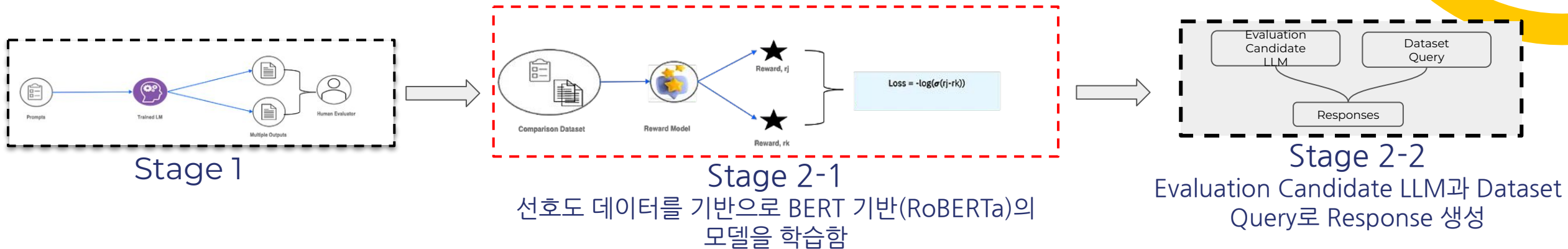
Model Type	Model Name	Model Size
Multilingual	meta-llama/Meta-Llama-3	70B
	meta-llama/Meta-Llama-3	8B
Korean Specific	yanolja/EEVE-Korean	10.8B
	upstage/SOLAR	10.7B

기존에 존재하는 benchmark 상의 Query를 참고하여 Text Query를 구축함

Dataset Name	Details	URL
Aihub 일반상식	한국어 위키백과 내 주요 문서 15만 개에 포함된 지식	https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSet=realm&dataSetSn=106
KorQuAD 2.1	Wikipedia article에 대한 질의응답 쌍	https://korquad.github.io/
ChatBot Data	일상속 질문 & 답변 데이터	



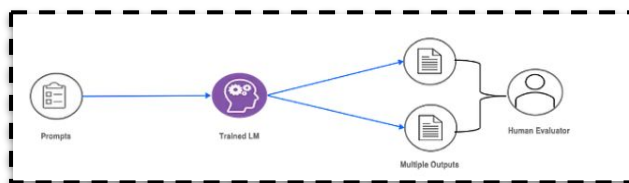
Research Objective (Stage 2)



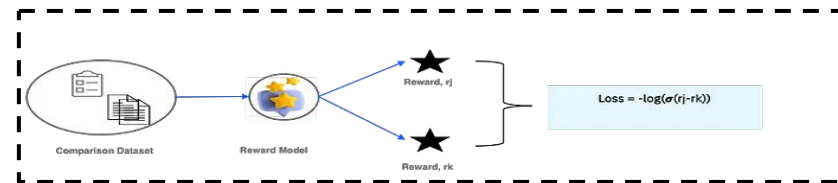
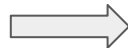
Query Response와 Preference Dataset의 GT Label을 Mean Square Error loss 기반으로 학습하여 언어학적 요소를 고려한 Evaluation Model을 학습 및 개발함



Research Objective (Stage 2)

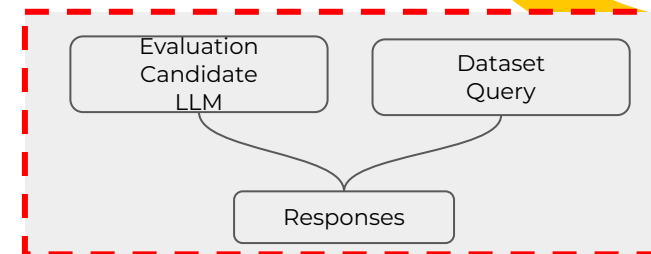
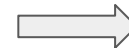


Stage 1



Stage 2

선호도 데이터를 기반으로 BERT 기반 (RoBERTa)의 모델을 학습함



Stage 2

Evaluation Candidate LLM과 Dataset Query로 Response 생성

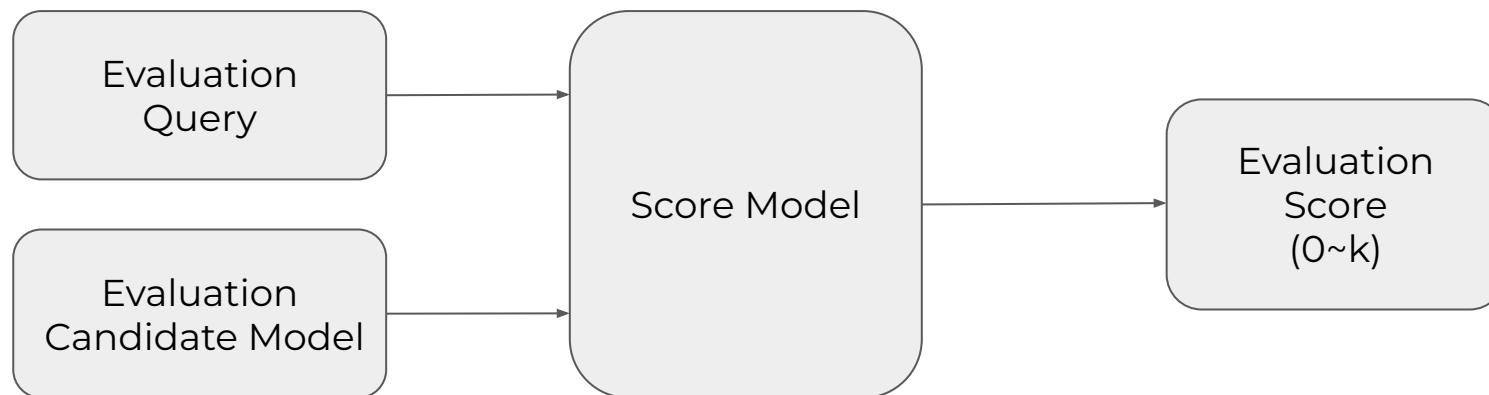
Scoring Model 학습에 사용되지 않은 Text Query(Evaluation Query)를 Input으로 Stage 3의 Scoring Model의 입력될 Response 생성



Research Objective (Stage 3)

Stage 2에서 학습된 모델로 Evaluation Candidate Model에서 생성된 결과물을 평가함

Scoring Model의 출력 점수의 범위는 Stage 1에서 사용된 모델의 개수(k) 만큼의 범위를 가짐





04

Expected Results

- Workflow
- Expected Output
- Evaluation Feature



Expected Output

Q. 합기도 배우면 나중에 무슨 일할 수 있어요?

합기도를 배우면 여러 가지 경로로 나아갈 수 있어요. 몇 가지 예를 들면:

1. 합기도 강사: 합기도를 가르치는 강사가 될 수 있어요. 합기도 도장이나 체육관에서 강의하거나, 개인적으로 수업을 진행할 수 있어요.
2. 경기 선수: 합기도 선수로서 경기에 참가하여 대회에서 실력을 겨룰 수 있어요. 성공적인 경기 선수가 되면 상금이나 후원금을 받을 수도 있어요.
3. 체육관 운영자: 합기도를 포함한 다양한 운동 프로그램을 제공하는 체육관이나 피트니스 센터를 운영할 수 있어요.

등등 다양한 가능성이 있어요. 합기도를 배우면 신체적인 능력과 지식을 향상시키고, 그것을 다양한 분야에서 응용할 수 있어요.

Chatgpt(GPT 3.5)



합기도를 배우면 경호, 안전관리, 스포츠감독, 체육학원강사, 대학교수 등 여러 분야에서 일할 수 있습니다. 또한 국내외 대회에서 선수로 출전하여 메달을 따내는 것도 가능합니다. 또한 언뜻 보기에는 관련성이 없어 보이는 분야에서도 합기도 기술을 활용하여 문제를 해결하는 데 도움이 될 수 있습니다.

Kochatgpt(GPT 2.0)



Expected Output

Q. 합기도 배우면 나중에 무슨 일할 수 있어요?

합기도를 배우면 다음과 같은 다양한 긍정적인 효과를 얻을 수 있습니다.

- 근력 증진: 합기도는 전신을 사용하는 운동이기 때문에 근력 향상에 효과적입니다. 특히 팔, 다리, 복근 등의 근육을 강화하는 데 도움이 됩니다.
- 스트레스 감소: 합기도는 스트레스 해소에 효과적인 운동입니다. 운동을 할 때는 엔돌핀이라는 호르몬이 분비되는데, 이 호르몬은 기분을 좋게 만들고 스트레스를 줄이는 데 도움이 됩니다.
- 자기 방어: 합기도는 자기 방어 기술을 배우는 좋은 방법입니다. 합기도 동작을 사용하여 자신을 공격으로부터 보호하는 방법을 배울 수 있습니다.

Google Gemini

Example

Model Index

1. Chatgpt(GPT 3.5)
2. KoChatgpt(GPT 2.0)
3. Google Gemini

Human Annotators Ranking(A, B, C):

A - 1 → 2 → 3

B - 1 → 2 → 3

C - 1 → 2 → 3

Final Human Ranking(Average):

1 → 2 → 3

Evidence:

- 1 - 응집성, 의도성, 정보성 명확히 드러남
- 2 - 의도성은 명료하나 문법적 부자연성으로 응결성 부족
- 3 - 의도성 파악 부족으로 답변 부적합



Expected Output

Final Benchmark Data Structure

Query 1: <Input> 합기도 배우면 나중에 무슨 일할 수 있어요?

<Output1> 합기도를 배우면 여러 가지 경로로 나아갈 수 있어요. 몇 가지 예를 들면 ...

<Output2> 합기도를 배우면 경호, 안전관리, 스포츠감독, 체육학원강사 ...

<Output3> 합기도를 배우면 다음과 같은 다양한 긍정적인 효과를 얻을 수 있습니다. 근력 증진 ...

<Output4> 합기도를 배우면 나중에 큰 사람이 있습니다. 큰 사람이란 ...

<Output5> 합기도를 배운다는 것은 곧 신체 강화와 정신 수양을 동시에 한다는 것을 의미합니다. ...

<Ranking> 1, 2, 3, 5, 4

Query 2: <Input> 랜선 연애로 잘 되기도 해?

<Output1> 랜선 연애도 잘 될 수 있습니다. 결국 연애는 상호간의 존중과 이해, 소통, ...

<Output2> 랜선 연애란 가상 공간에서의 관계를 의미합니다 ... 랜선 연애가 잘 될 수 있는지는 ...

<Output3> 랜선 연애가 현실에서도 성공적인 관계로 이어질 수 있는 가능성은 충분히 있습니다. 실제로 ...

<Output4> 랜선 연애의 성공 가능성은 상황에 따라 달라집니다. 랜선 연애로 잘 되기 위한 조건은 ...

<Output5> 랜선 연애로 잘 된다는 것은 랜선으로 이루어진 연애를 뜻합니다. 이는 곧 ...

<Ranking> 3, 1, 4, 2, 5

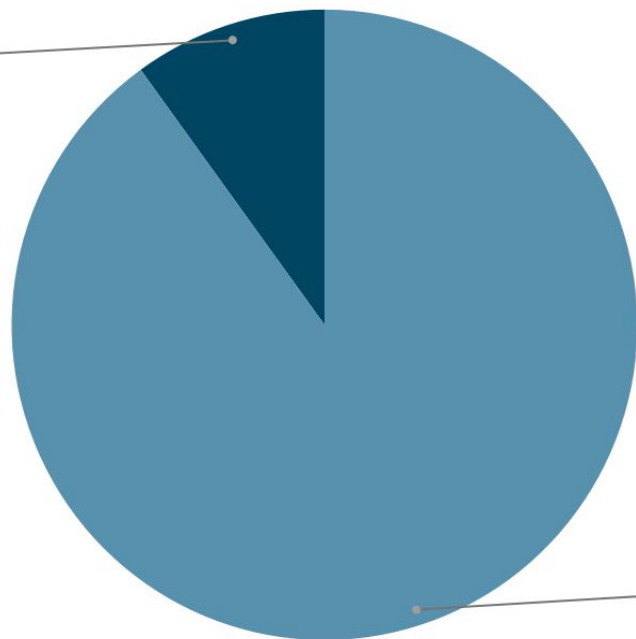


Expected Output

Final Dataset Size

전체 Query Set

Qualified query
10.0%



Unqualified query
90.0%

- Wikidata query(일반상식): 100,000 건
- Domain Specific query: 50,000건
- 최종적으로 15,000 건의 query에 대해 Human Ranking 라벨링
- 추후 135,000 건의 query를 정제하거나 선별된 query를 증강하여 활용 가능



Evaluation Feature - Textuality(1)

- **Textuality**(텍스트성; Beaugrande & Dressler 1981)
 - The characteristics and properties that make a piece of language a unified and communicative whole
 - Cohesion(응결성): 표층구조의 문법적인 의존 관계
 - ex) 옛날에 **한 임금님**이 살았습니다. **그**는 딸이 세 명 있었습니다.
 - Coherence(응집성): 주제 구성에 결정적인 텍스트 구성성분들 사이의 논리적, 의미적 관계
 - ex) **철수가 넘어져서 다리가 부러졌어.** vs 철수가 넘어졌다. 개구리가 뛰어갔다. (의미적 연결관계 ↓)
 - Intentionality(의도성): 텍스트 생산자가 달성하고자 하는 특정 목적이 드러나는지 여부
 - ex) (도서관 사서에게) 담화 이론에 관한 책을 찾으러 왔는데요 → **원하는 책을 찾아달라는 의도**
 - Acceptability(용인성): 텍스트의 내용을 받아들이는 텍스트 수용자의 입장 및 태도
 - ex) 그것에 대해서는 이야기하고 싶지 않아 → **용인성이 낮은 텍스트**



Evaluation Feature - Textuality(2)

- **Textuality**(텍스트성; Beaugrande & Dressler 1981)
 - Informativity(정보성): 주어진 텍스트에 새로운 정보가 들어있거나 수용자가 예측하지 못한 정보가 있는지 여부 + 텍스트 수용자의 흥미를 유발할 수 있는 정도
 - ex) 이전에 없던 새로운 맛을 찾고 계시다면, 저희 제품이 해결책이 되어드릴 것입니다. → **수용자의 흥미 유발과 상품에 대한 정보 내포**
 - Situationality(상황성): 한 텍스트를 현재의 담화 상황 또는 복원 가능한 상황에 적절하게 관련지어주는 요인들의 일반적인 특성
 - ex) 그러니까 지금 네 말은 나랑 헤어지자는 말이야? → **이전 텍스트를 말미암아 현재의 담화 상황을 관련지어 설명함**



05

Detailed Plans

- HR Management plan
- Budget Allocation Plan

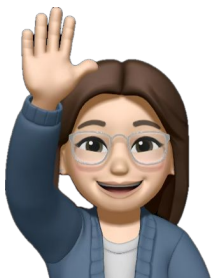


HR management Plan



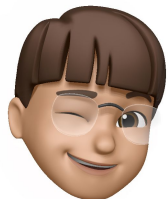
도경환

- Linguistic feature를 기반으로 한 ranking metric 확립 및 평가
- 관련 논문 및 방법론 연구
- Query Data 정제 및 필터링



이채원

- Query 데이터 수집 및 가공, qualification
- Linguistic feature를 기반으로 한 ranking metric 확립 및 평가
- Query set 증강 방법론 연구



임상범

- Scoring Model 학습
- Query Data 정제 및 필터링



Budget Allocation Plan

Budget Allocation

데이터 구매비

20.0%

GPU 클라우드 비용

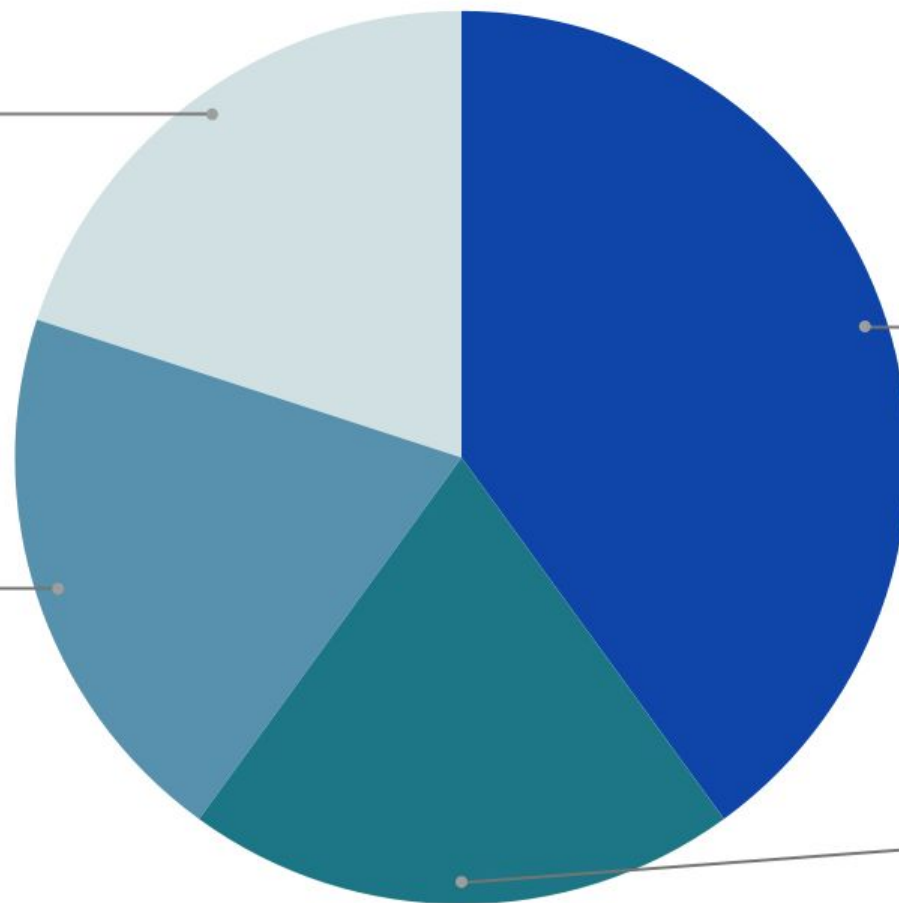
20.0%

API 호출 비용

40.0%

모델 학습 비용

20.0%





Thank you