HAE-RAE Bench
Evaluation of Korean Knowledge in Large Language Models

김송성, 김수완, 김정우, 김휘서, 손규진, 염제원, 이재철, 이한울, 정지휴

May. 2023

Sponsored By.

ONELINE AI

소개: 대규모 한글 Instruction-Tuning 데이터셋 제작 프로젝트 “해례”

“해례 (HAE-RAE)”는 영어로 수집된 데이터셋을 번역하는 것에서 나아가 한국어의 고유한 특성을 반영한 Instruction Dataset을 구축하는 오픈소스 프로젝트입니다.

- Status Quo.
  - 2023.2 월 LLaMA 모델의 공개를 바탕으로 오픈소스 커뮤니티는 Standford Alpaca, Vicuna, Koala, Dolly, ChatGLM 등 무수히 많은 “Instruction Model” 을 개발하였습니다.
  - 이에 맞추어 국내에서도 EleutherAI팀의 Polyglot-Ko 모델을 바탕으로 하는 여러 한국어 Instruction Model이 공개되었습니다.
    - “Beomi” 님의 KoAlpaca_V1.1 과 “이창원”님의 ChangGPT 프로젝트에서 네이버 지식인과 AiHUB 데이터셋 등 영문 데이터셋을 번역하는 것에서 나아가 한국어의 고유한 특성과 지식을 더욱 잘 반영하기 위한 시도가 있었으며 저희 해례 프로젝트는 이를 더 발전 시켜보고자하는 취지에서 시작되었습니다.
- Progress.
  - 2023.5 월 킥오프 이후 여러 방법론을 활용해 한국어 Instruction Model을 학습하기 위한 데이터를 제작 및 생성 중에 있습니다.
  - 더불어, 언어 모델의 “한국어 지식”을 평가할 수 있는 HAE-RAE Bench 벤치마크도 제작 중에 있으며 그 중 일부를 프로젝트 홍보차 먼저 소개하게 되었습니다.

HAE-RAE BENCH

HAE-RAE Bench는 한국어 어휘, 독해, 문법과 지식, 총 4가지 영역에 걸쳐 언어 모델의 능력을 평가하는 벤치마크 입니다.

- HAE-RAE Bench의 구성

Category	SubCategory	Sample Size	Source
Vocabulary	Rare Words	402	우리말 겨루기
	Loan Words	166	NIKL
	Standard Nomenclature	150	NIKL
Knowledge	History	188	
	Culture	제작 진행중	
Reading Comprehension	Reading Comprehension	445	KLAT
Grammatical Error Detection	Grammatical Error Detection	450	NIKL

- 모든 문제는 사지선다 와 오지선다 문항으로 구성되어 있으며, 2023년 5월 11일 기준 Rare Words, Loan Words, Standard Nomenclature, History, Reading Comprehension 카테고리의 초기버전이 완성되어 벤치마킹이 진행되고 있습니다.
- 모든 벤치마킹은 3-shot 설정에서 진행됩니다.

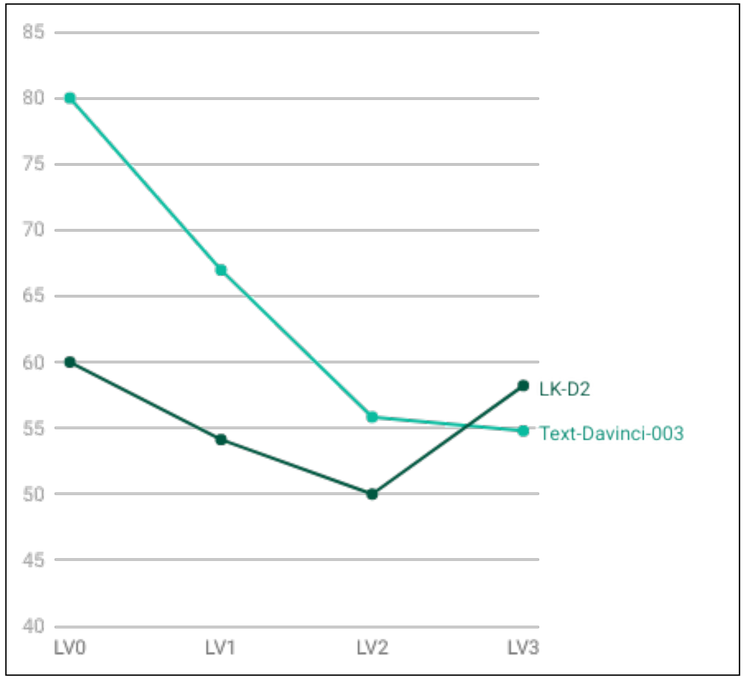
GPT-3.5 와 HyperClova 중 한국어를 더 잘하는 모델은?

HyperClova가 독해를 제외한 영역에서 GPT-3.5를 앞서는 모습을 보여줌

- HyperClova (LK-D2) 와 GPT-3.5 (Text-Davinci-003) 벤치마크 결과 비교

	Rare Words	Loan Words	Standard Nomenclature	History	Reading Comprehension
LK-D2	**82.34%**	**83.13%**	**78%**	**81.62%**	54.50%
Text-Davinci-003	62.18%	62.27%	58.66%	30.27%	**60.13%**

- 그러나, Reading Comprehension 데이터셋의 난이도가 올라감에 따라 성능이 linear 하게 감소하는 GPT-3.5와 달리 HyperClova 모델은 일정한 성능을 보여줌.



- HyperClova가 저난이도 문제(LV0~LV1) 에서는 GPT-3.5를 underperform 하는 모습을 보이지만, 문제 난이도 상승에 따른 성능 감소폭은 HyperClova가 더 작으며 가장 고난이도 문제 (LV3)에서는 오히려 향상된 성능을 보입니다.
- 원인을 알아보기 위해 현재 데이터셋을 분석하는 과정 중에 있습니다.
- Polyglot-Ko, KoGPT-Ryan 등과 같은 한국어 언어 모델과 mT5, XGLM 과 같은 multilingual 언어 모델에도 벤치마킹을 진행할 예정입니다.