[창의적 통합설계]

Large Language Model의 한국어 능력 평가

2023.09.01

발표자: 김형준

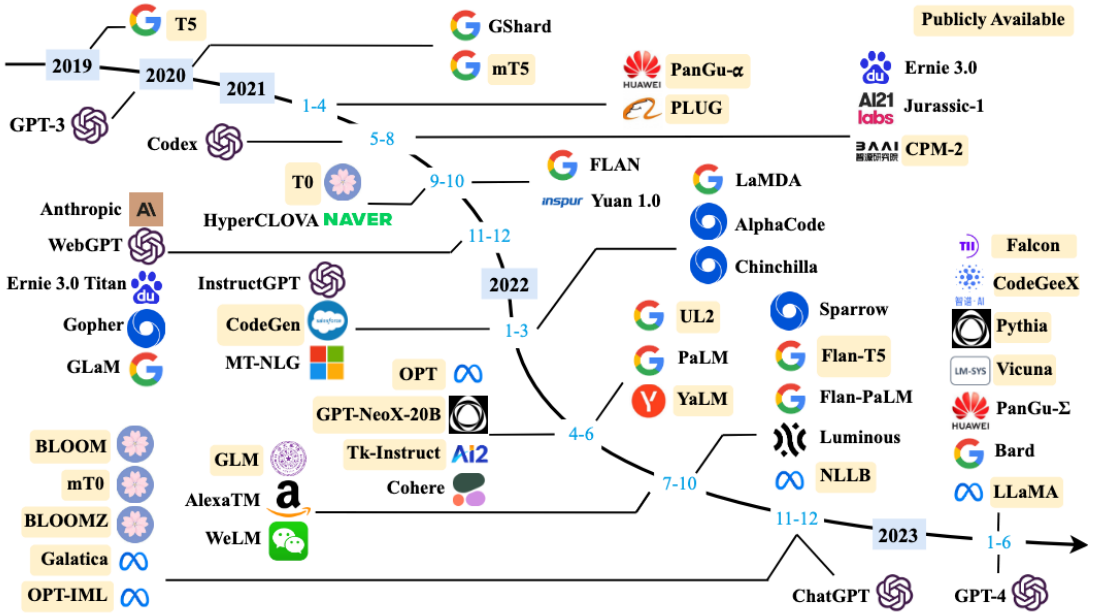
Intelligent Data Systems Lab.

- 이상구 교수님 연구실
  - 데이터베이스, 데이터 마이닝 등
  - LLM 기반 NLP 연구 활발히 진행 중.

- Recent Publications
  - CELDA: Leveraging Black-box Language Model as Enhanced Classifier without Labels. (ACL 2023)
  - Prompt-Augmented Linear Probing: Scaling Beyond The Limit of Few-shot In-Context Learners (AAAI 2023)
  - Ground-Truth Labels Matter: A Deeper Look into Input-Label Demonstrations (EMNLP 2022)
  - Enhancing Out-of-Distribution Detection in Natural Language Understanding via Implicit Layer Ensemble (EMNLP Findings 2022)

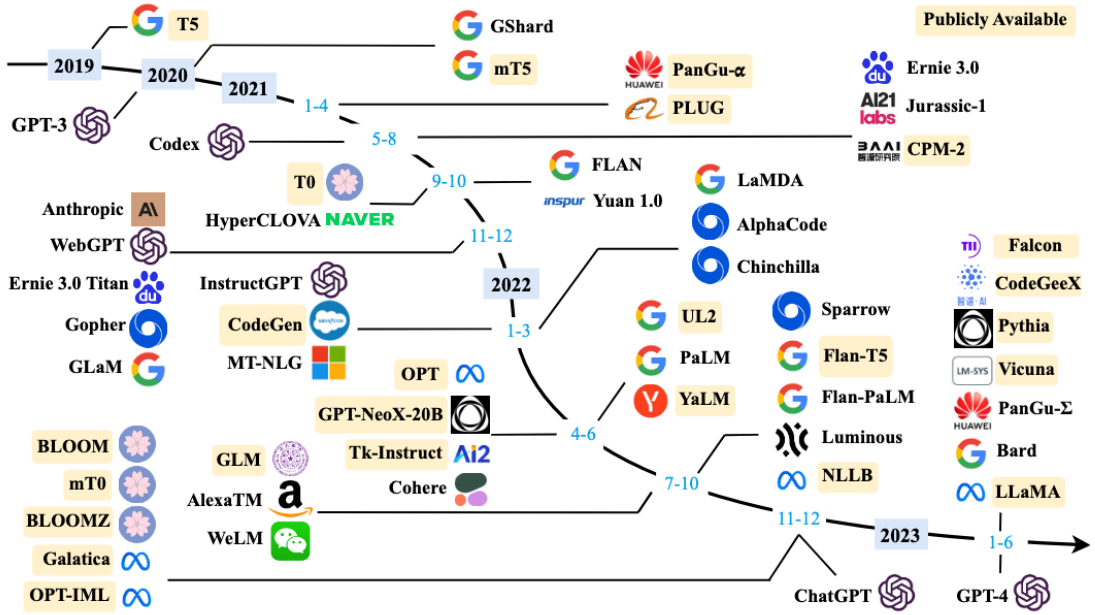
Project Background

- Large Language Model (LLM)
  - 인간의 **언어를 이해하고 생성**하도록 훈련된 모델.
  - 사전에 대규모 언어 데이터를 학습, 문장 구조나 문법, 의미 등을 이해하고 생성할 수 있음.
    - 대화 응답 생성, 수학 문제 해결, 번역, 코드 생성 등 다양한 능력을 보유하고 있음.
  - 대표적으로, ChatGPT, GPT-4, HyperCLOVA X 등이 있음.



Project Background

- Large Language Model (LLM)
  - LLM이 보유한 능력을 평가하고자 하는 다양한 노력들이 있었음.
  - 하지만, 이러한 모델들의 **한국어 능력**에 대한 직접적인 평가는 부족한 상태.



Project Details

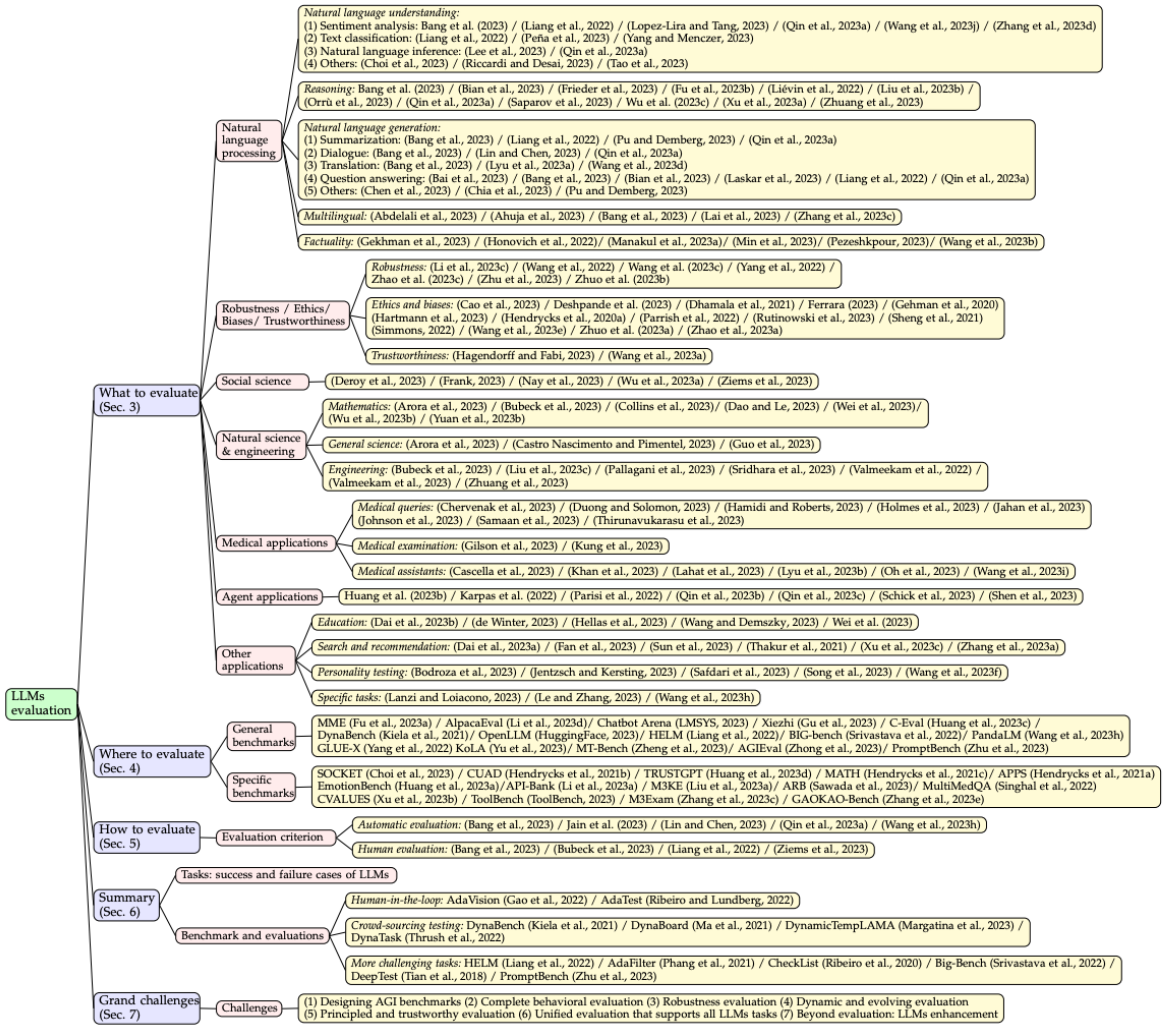
- 프로젝트 개요
  - 현존하는 **LLM의 한국어 이해 정도**를 파악하기 위한 **한국어 능력 평가 지표** 정의 및 적용
    - 접근 가능한 5~10개의 LLM 선정

- 세부 목표
  1. LLM의 한국어 능력 평가 기준 정의
  2. 한국어 능력 평가를 위한 실험 설계
  3. LLM 평가 및 분석

Project Details

1. LLM의 한국어 능력 **평가 기준** 정의
   - 어떤 요소를 평가할 것인가?

   - Natural Language Understanding (NLU)
     - Sentiment analysis, text classification, ...
   - Natual Language Generation (NLG)
     - Summarization, Dialogue Generation, ...
   - Mathematics, Medical Knowledge, General Science, ...



A Survey on Evaluation of Large Language Models (Chang et al., arxiv 2023)

Project Details

2. 한국어 능력 평가를 위한 **실험 설계**

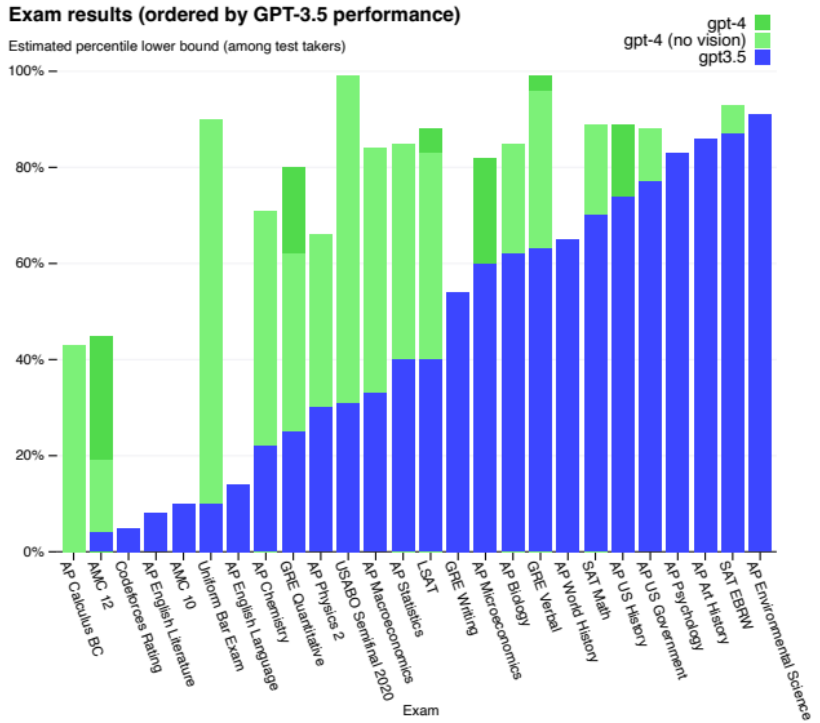
- 평가 기준을 확인하기 위해 어떤 실험을 할 것인가?
  - 기존에 존재하는 Dataset / Benchmark 사용.
    - MMLU, BIG-bench (math, linguistics, common-sense, ..), Chatbot Arena, MT-bench (evaluating dialogue generation), ...
  - 필요에 따라 적합한 데이터셋 구축.

- 결과를 어떻게 평가할 것인가?
  - Automatic Evaluation, Human Evaluation, ...

Project Details

3. LLM **평가 및 분석**

- LLM의 한국어 능력을 다방면으로 평가할 수 있도록 실험 및 분석 진행.



Exam	GPT-4	GPT-4 (no vision)	GPT-3.5
Uniform Bar Exam (MBE+MEE+MPT)	298 / 400 (~90th)	298 / 400 (~90th)	213 / 400 (~10th)
LSAT	163 (~88th)	161 (~83rd)	149 (~40th)
SAT Evidence-Based Reading & Writing	710 / 800 (~93rd)	710 / 800 (~93rd)	670 / 800 (~87th)
SAT Math	700 / 800 (~89th)	690 / 800 (~89th)	590 / 800 (~70th)
Graduate Record Examination (GRE) Quantitative	163 / 170 (~80th)	157 / 170 (~62nd)	147 / 170 (~25th)
Graduate Record Examination (GRE) Verbal	169 / 170 (~99th)	165 / 170 (~96th)	154 / 170 (~63rd)
Graduate Record Examination (GRE) Writing	4 / 6 (~54th)	4 / 6 (~54th)	4 / 6 (~54th)
USABO Semifinal Exam 2020	87 / 150 (99th - 100th)	87 / 150 (99th - 100th)	43 / 150 (31st - 33rd)
USNCO Local Section Exam 2022	36 / 60	38 / 60	24 / 60
Medical Knowledge Self-Assessment Program	75 %	75 %	53 %
Codeforces Rating	392 (below 5th)	392 (below 5th)	260 (below 5th)
AP Art History	5 (86th - 100th)	5 (86th - 100th)	5 (86th - 100th)
AP Biology	5 (85th - 100th)	5 (85th - 100th)	4 (62nd - 85th)
AP Calculus BC	4 (43rd - 59th)	4 (43rd - 59th)	1 (0th - 7th)
AP Chemistry	4 (71st - 88th)	4 (71st - 88th)	2 (22nd - 46th)
AP English Language and Composition	2 (14th - 44th)	2 (14th - 44th)	2 (14th - 44th)
AP English Literature and Composition	2 (8th - 22nd)	2 (8th - 22nd)	2 (8th - 22nd)
AP Environmental Science	5 (91st - 100th)	5 (91st - 100th)	5 (91st - 100th)
AP Macroeconomics	5 (84th - 100th)	5 (84th - 100th)	2 (33rd - 48th)
AP Microeconomics	5 (82nd - 100th)	4 (60th - 82nd)	4 (60th - 82nd)
AP Physics 2	4 (66th - 84th)	4 (66th - 84th)	3 (30th - 66th)
AP Psychology	5 (83rd - 100th)	5 (83rd - 100th)	5 (83rd - 100th)
AP Statistics	5 (85th - 100th)	5 (85th - 100th)	3 (40th - 63rd)
AP US Government	5 (88th - 100th)	5 (88th - 100th)	4 (77th - 88th)
AP US History	5 (89th - 100th)	4 (74th - 89th)	4 (74th - 89th)
AP World History	4 (65th - 87th)	4 (65th - 87th)	4 (65th - 87th)
AMC 10^3	30 / 150 (6th - 12th)	36 / 150 (10th - 19th)	36 / 150 (10th - 19th)
AMC 12^3	60 / 150 (45th - 66th)	48 / 150 (19th - 40th)	30 / 150 (4th - 8th)
Introductory Sommelier (theory knowledge)	92 %	92 %	80 %
Certified Sommelier (theory knowledge)	86 %	86 %	58 %
Advanced Sommelier (theory knowledge)	77 %	77 %	46 %
Leetcode (easy)	31 / 41	31 / 41	12 / 41
Leetcode (medium)	21 / 80	21 / 80	8 / 80
Leetcode (hard)	3 / 45	3 / 45	0 / 45

GPT-4 Technical Report (OpenAI, 2023)

Contact

- 김형준
  - 전화: 010-5105-8199
  - e-mail : heyjoonkim@europa.snu.ac.kr

- 고영록
  - 전화: 010-4164-9756
  - e-mail: yrko1@europa.snu.ac.kr

Questions?