

# Open-source Large Language Model Survey

박 희 선 교수

성균관대학교 소프트웨어융합대학

2023. 05. 09

## □ Open-source Large Language Model 조사

- ChatGPT 대안 : Open-source LLMs
- Open-source LLMs : Research vs. Commercial
- Open-source LLMs
  - 연구용: LLaMA, Alpaca, GPT4All, Vicuna
  - 상업용: Dolly 2.0, Open Assistant, BLOOM, Polyglot
- Korean Open-source LLMs
  - Polyglot-ko, KoAlpaca, KoVicuna
- Conclusion : Considerations

## □ 오픈소스 LLM 모델의 이점

### 1. Data Privacy

- 데이터에 대한 컨트롤 가능

### 2. Customization

- 자신들의 데이터로 대규모 언어 모델을 훈련하고, 일부 주제에 대한 필터링 등을 허용하면서 훈련 가능

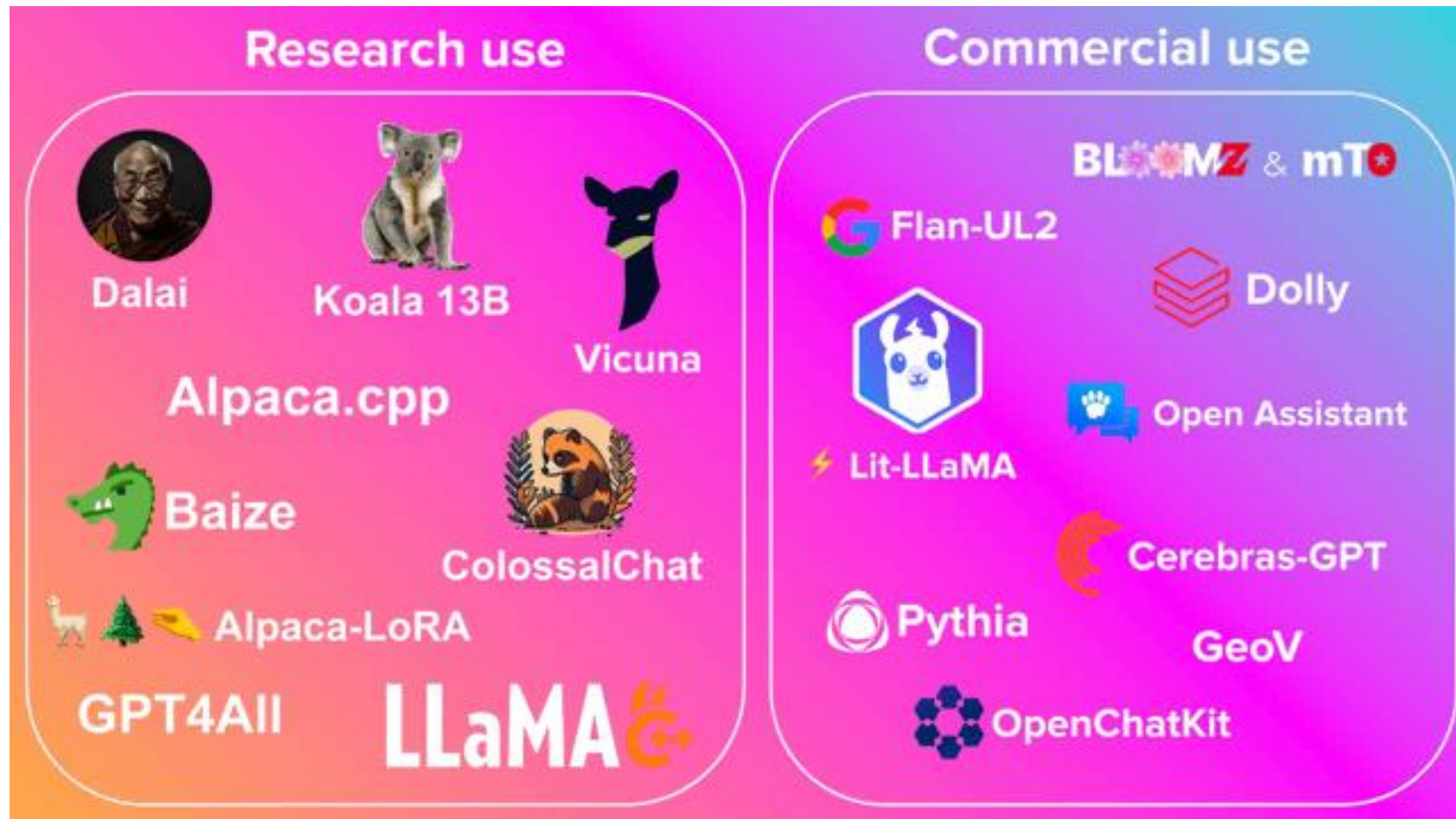
### 3. Affordability(경제성)

- 오픈소스 GPT 모델을 사용하면 값비싼 하드웨어에 대한 걱정 없이 정교한 대규모 언어 모델을 훈련할 수 있음

### 4. Democratizing AI(AI 민주화)

- 실제 문제를 해결하는 데 사용할 수 있는 추가 연구의 여지 가능성
- 오픈소스 모델은 독점 모델보다 투명하고 접근성이 뛰어나

# Open-source LLMs : Research vs. Commercial



Ref. [https://www.reddit.com/r/AllLinksandTools/comments/12kgv17/a\\_snapshot\\_of\\_todays\\_opensource\\_llms\\_space\\_every/](https://www.reddit.com/r/AllLinksandTools/comments/12kgv17/a_snapshot_of_todays_opensource_llms_space_every/)

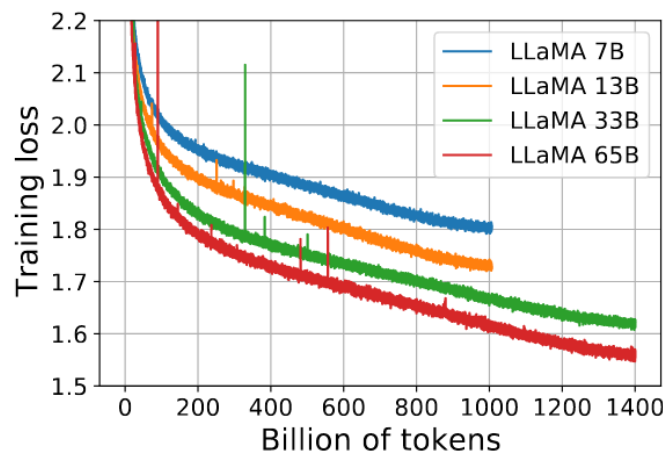
# Open-source LLMs

1. LLaMA
2. Alpaca
3. GPT4All
4. Vicuna
5. Dolly 2.0
6. Open Assistant
7. BLOOM
8. Polyglot

## □ LLaMA : Meta의 FAIR(Fundamental AI Research)팀에서 발표 (2023년 2월 24일)

- LLaMA 모델 크기: 7B, 13B, 33B, 65B(650억 파라미터)
  - LLaMA-13B 모델 성능이 GPT-3(175B) 성능을 초과 : 모델 크기 13배 작음
  - <https://github.com/facebookresearch/llama>
- LLaMA 모델은 1.4조 개의 토큰으로 훈련되었으며, 공개 데이터셋만을 사용
  - 말뭉치에 20개 언어의 데이터가 포함되어 있음
  - 토큰나이저: sentencepiece, 임베딩: Rotary Position Embedding 사용
- ✓ 현재 연구용 Pretrained Model로 가장 많이 사용되고 있음
- ✓ LLaMA는 인간의 선호도가 반영된 학습이 아님 (helpful, hones, harmless)

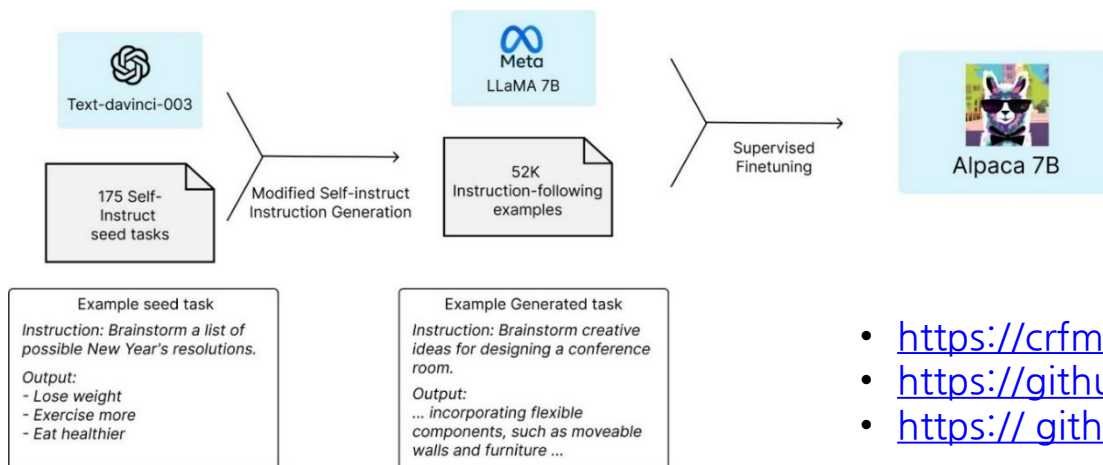
Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB



		Truthful	Truthful*Inf
GPT-3	1.3B	0.31	0.19
	6B	0.22	0.19
	175B	0.28	0.25
LLaMA	7B	0.33	0.29
	13B	0.47	0.41
	33B	0.52	0.48
	65B	0.57	0.53

## □ Alpaca : LLaMA를 Instruction-following 데이터로 학습시킨 fine-tuned 모델 (2023년 3월 발표)

- 아주 저렴한 비용으로 모델을 학습할 수 있도록, 데이터 및 모델 학습 방법을 재현 가능하도록 공개함 => 데이터 생성 비용 : \$500 (OpenAI API)
- OpenAI의 text-davinci-003으로 생성된 **52,000개 instruction-following demonstrations**을 기반으로 **LLaMA 7B 모델을 Fine-tuning한 모델**
  - Alpaca 7B 외에 13B, 30B, 65B 모델로 확장됨
- 고품질의 instruction-following 데이터 생성 방법 : self-instruct 논문 제안
  - 사람이 직접 만든 175개의 seed 데이터셋을 기반으로 데이터를 약 52,000개까지 추가 생산을 하여 학습 데이터셋으로 활용



- <https://crfm.stanford.edu/2023/03/13/alpaca.html>
- [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca)
- <https://github.com/tloen/alpaca-lora> : Fine Tuning

# Stanford 'Alpaca' (2/3)

- ❑ Self Instruct 데이터 생성 예시 : Supervised Finetune Training(SFT) dataset으로 이용
  - OpenAI의 Playground(ChatGPT 사이트 등)을 통해 수집

The screenshot shows the OpenAI Playground interface. On the left, there's a 'Get started' sidebar with instructions and a 'KEEP IN MIND' section. The main area is titled 'Playground' and contains a list of 20 tasks. The first task is: '2. A GPT language model should be able to complete the instruction. For example, do not ask the assistant to create any visual or audio output. For another example, do not ask the assistant to wake you up at 5pm or set a reminder because it cannot perform any action.' The second task is: '3. The instructions should be in English.' The third task is: '4. The instructions should be 1 to 2 sentences long. Either an imperative sentence or a question is permitted.' The fourth task is: '5. You should generate an appropriate input to the instruction. The input field should contain a specific example provided for the instruction. It should involve realistic data and should not contain simple placeholders. The input should provide substantial content to make the instruction challenging but should ideally not exceed 100 words.' The fifth task is: '6. Not all instructions require input. For example, when an instruction asks about some general information, "what is the highest peak in the world", it is not necessary to provide a specific context. In this case, we simply put "<noinput>" in the input field.' The sixth task is: '7. The output should be an appropriate response to the instruction and the input. Make sure the output is less than 100 words.'

Below the tasks, there's a 'List of 20 tasks:' section. The first task is: '1. Instruction: Is there anything I can eat for a breakfast that doesn't include eggs, yet includes protein, and has roughly 700-1000 calories?'. The input is: '<noinput>'. The output is: 'Yes, you can have 1 oatmeal banana protein shake and 4 strips of bacon. The oatmeal banana protein shake may contain 1/2 cup oatmeal, 60 grams whey protein powder, 1/2 medium banana, 1tbsp flaxseed oil and 1/2 cup watter, totalling about 550 calories. The 4 strips of bacon contains about 200 calories.'

The second task is: '2. Instruction: What is the relation between the given pairs?'. The input is: 'Night : Day :: Right : Left'. The output is: 'The relation between the given pairs is that they are opposites.'

The third task is: '3. Instruction: Convert the following temperature from Fahrenheit to Celsius.'. The input is: '32°F'. The output is: '32°F is equivalent to 0°C.'

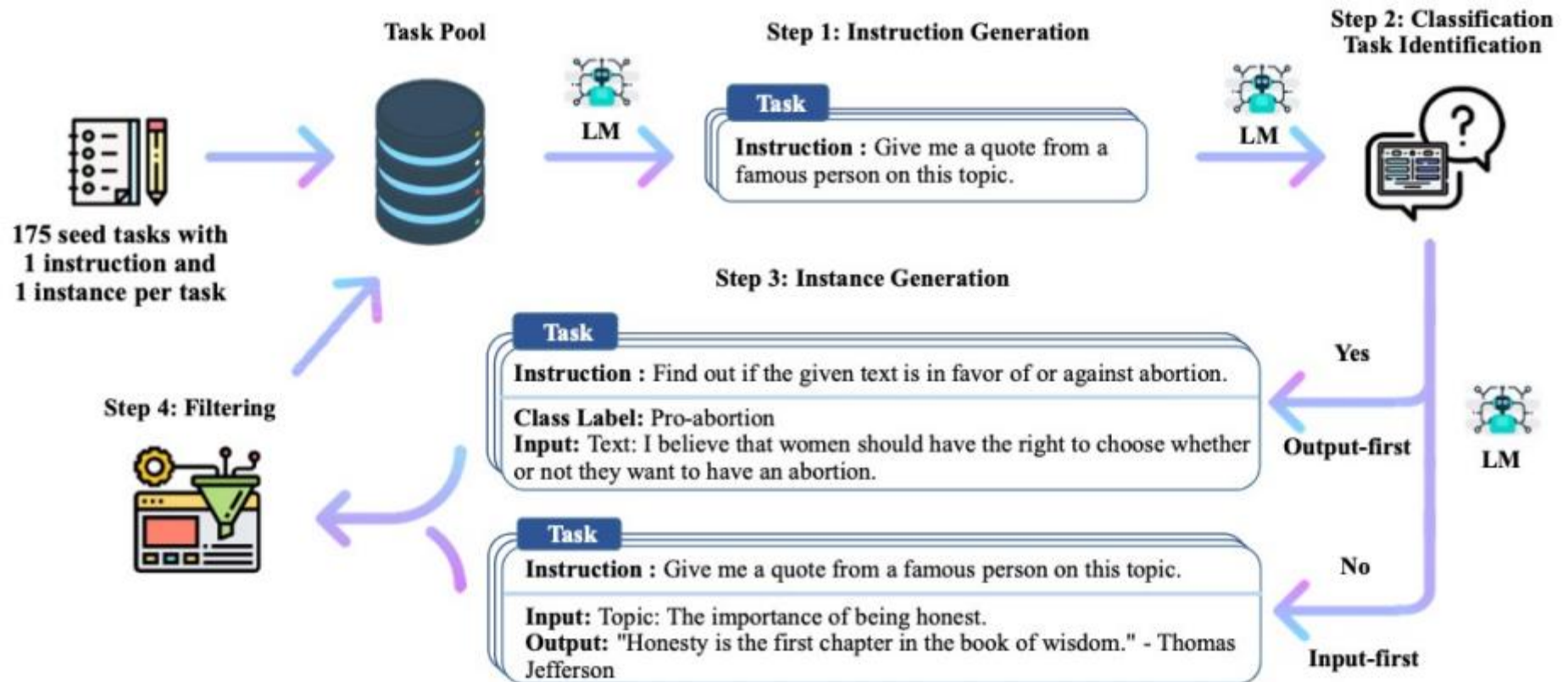
On the right side of the interface, there are settings for Mode (Complete), Model (text-davinci-003), Temperature (0.7), Maximum length (256), Stop sequences, Top P (1), Frequency penalty (0), Presence penalty (0), Best of (1), Inject start text (checked), Inject restart text (checked), and Show probabilities (Off).

- ✓ 좋은 SFT Dataset의 조건
  - 다양한 표현으로 작성된 Prompt
  - Prompt에 대해 적절히 작성된 응답
  - GPT-3 API만을 이용하여 모든 작업 수행



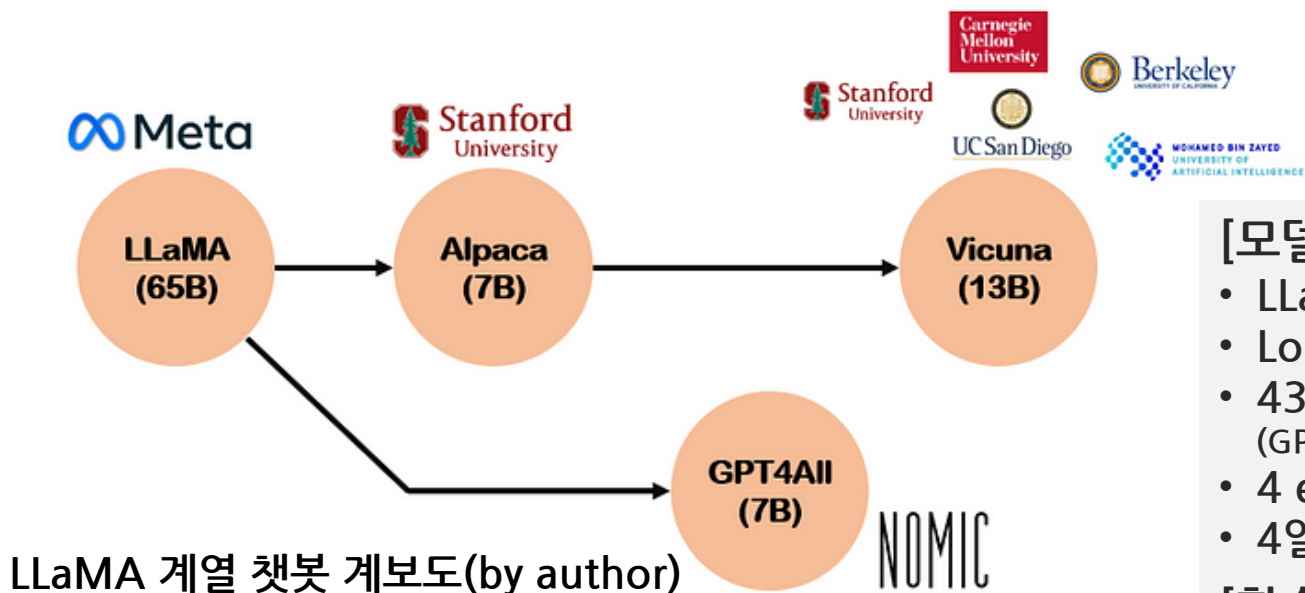
# Stanford 'Alpaca' (3/3)

## □ Self Instruct



## □ GPT4All : Nomic AI팀은 LLaMA-7B를 fine-tuned 모델 공개 (2023년 3월)

- ChatGPT의 경량화된 모델 : 3GB ~ 8GB file
- A free-to-use, locally running, privacy-aware chatbot
  - Memory 요구사항: 8GB RAM의 CPU에서 실행 가능
- Github에 공개되자마자 2주만에 24.4k개의 star(2023/4/8 기준)를 얻을 만큼 큰 인기를 끌고 있음 (두 달 전에 발표된 LLaMA의 github star는 16.9k)
  - <https://github.com/nomic-ai/gpt4all>, Demo : <https://gpt4all.io/index.html>



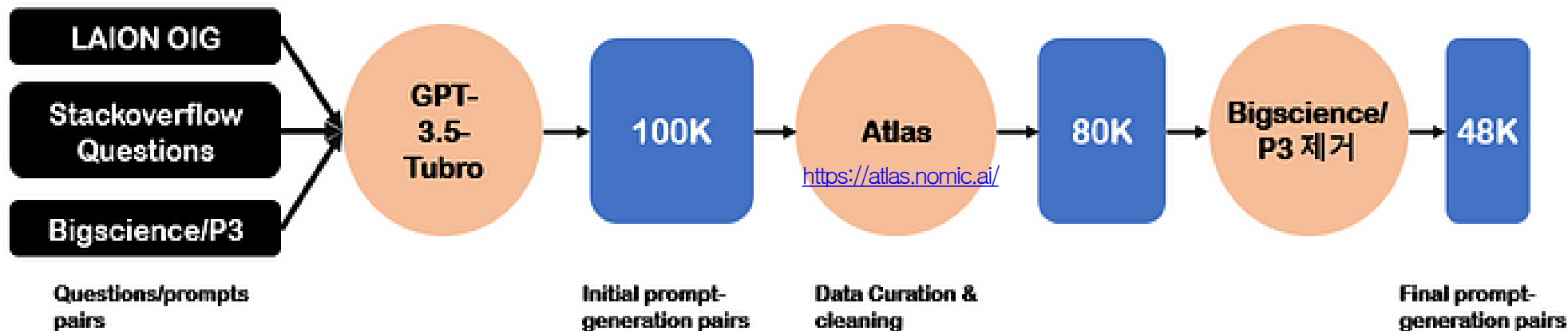
LLaMA 계열 챗봇 계보도(by author)

### [모델 학습 방법]

- LLaMA-7B를 fine-tuning
- LoRA 적용
- 43만개의 prompt-generation pairs 사용 (GPT-Turbo-3.5 이용)
- 4 epochs
- 4일간의 학습

[학습 비용 : 총 \$1,400]

- 수집된 데이터, 데이터 수집 프로시저, 훈련 코드, 최종 모델 가중치 등을 모두 공개 ([https://s3.amazonaws.com/static.nomic.ai/gpt4all/2023\\_GPT4All\\_Technical\\_Report.pdf](https://s3.amazonaws.com/static.nomic.ai/gpt4all/2023_GPT4All_Technical_Report.pdf))
- 데이터셋 준비 과정: LLaMA 기반으로 43만개 명령셋(Prompt-response쌍)으로 훈련
  - GPT-3.5-Turbo로 생성한 초기 100k의 prompt-response 쌍에 data curation과 cleaning를 시행하여 80만개의 고품질의 prompt-generation 쌍 구축
    - Prompt-generation 쌍 데이터셋 수집 : [LAION OIG\(Open Instruction Generalist\)](#) 데이터셋, 스택오버플로우의 코딩 질문, [Big-Science/P3\(Public Pool of Prompts\)](#)
    - Stanford 'Alpaca' 등을 참고하고, 데이터를 ATLAS에 올려서 큐레이션 및 클리닝



GPT4All의 데이터셋 준비 과정 (by author)

## □ UC Berkeley, CMU, Stanford, UC San Diego 연구원 팀이 만든 오픈소스 챗봇 모델 Vicuna-13B 공개 (2023년 3월)

- Vicuna-13B : [ShareGPT.com](https://sharegpt.com)에서 수집한 사용자 공유 대화(7만 건)로 LLaMA-13B를 Fine-tuning
  - 예비 평가에서 **OpenAI ChatGPT 품질의 92% 달성 (훈련비용 : \$300)**
    - LLaMA, Stanford Alpaca의 13B 모델을 능가.
  - <https://lmsys.org/blog/2023-03-30-vicuna/>
- <https://github.com/lm-sys/FastChat>

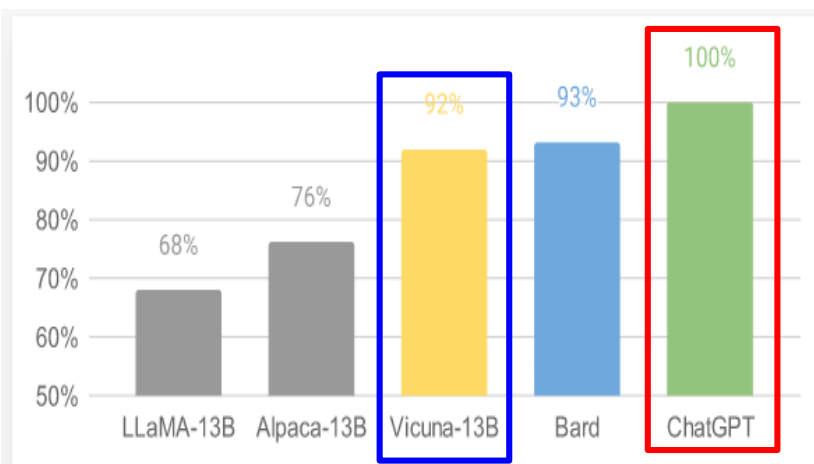
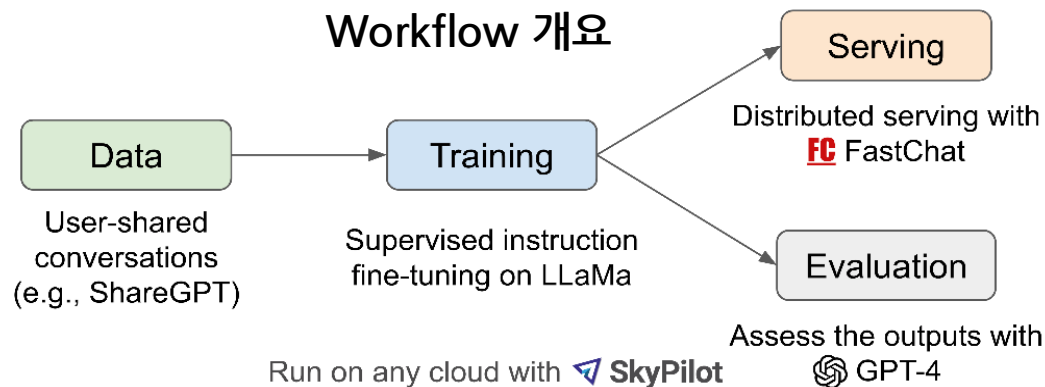


그림 1. GPT-4\*에서 평가한 상대적 응답 품질

Model Name	LLaMA	Alpaca	Vicuna	Bard/ChatGPT
Dataset	Publicly available datasets (1T token)	Self-instruct from davinci-003 API (52K samples)	User-shared conversations (70K samples)	N/A
Training code	N/A	Available	Available	N/A
Evaluation metrics	Academic benchmark	Author evaluation	GPT-4 assessment	Mixed
Training cost (7B)	82K GPU-hours	\$500 (data) + \$100 (training)	\$140 (training)	N/A
Training cost (13B)	135K GPU-hours	N/A	\$300 (training)	N/A

## □ Vicuna-13B 훈련 비용: \$300

- Stanford ‘Alpaca’의 Training Script를 개선하여 다단계 대화와 긴 시퀀스를 잘 처리하도록 개선
  - **Memory optimizations** : Gradient checkpointing, flash attention 적용
    - 긴 대화를 더 잘 이해할 수 있도록 Alpaca의 최대 컨텍스트 길이를 4배(512에서 2048) 확장하면서, GPU 메모리 요구 사항이 크게 증가하여 메모리 최적화
  - **Multi-round conversations**
    - multi-turn 대화를 위해 training loss 조정, fine-tune 시 챗봇의 출력만을 고려해 loss 계산
  - **Cost reduction via Spot Instance**
    - SkyPilot의 spot instance 도입을 통해 학습 비용 절감



## □ 비영리 AI 연구단체인 EleutherAI에서 [pythia](#) 모델 기반 12B 모델 출시 (2023년 4월)

- Pre-trained Model : [pythia](#) 모델
- Databricks 직원들이 생성한 지침에 따라 15,000개의 Instruction-following 데이터 세트에서 모델을 Fine-tuning
  - Databricks 직원 5,000명이 참여: 게임화된 프로세스에서 예제를 수집
  - Instruction-following records 데이터 세트 : [databricks-dolly-15k.jsonl · databricks/databricks-dolly-15k at main \(huggingface.co\)](#)

## □ Pythia-12B: <https://huggingface.co/EleutherAI/pythia-12b>

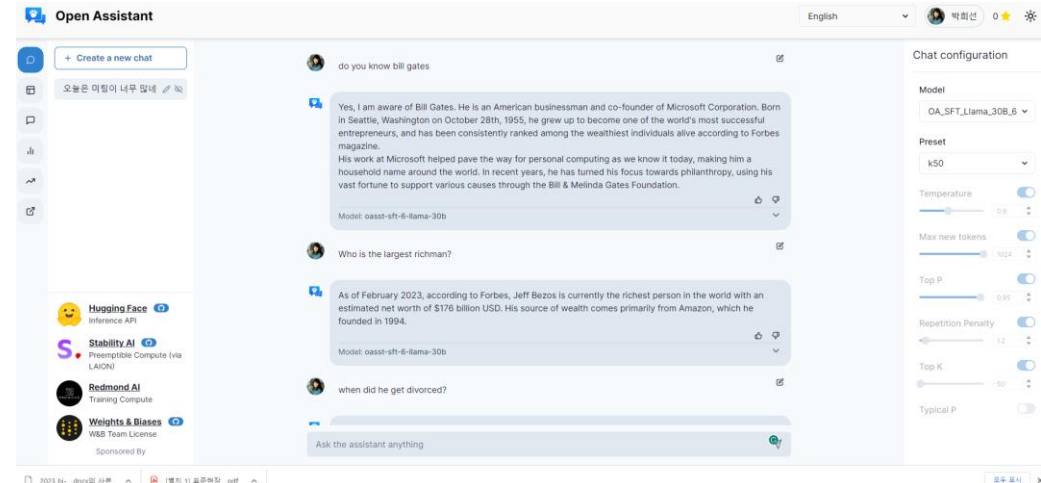
- EleutherAI에서 해석 가능성 연구를 용이하게 하기 위해 개발한 LLM
  - [Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling](#)
  - 모델은 [GPT-NeoX](#) 라이브러리를 사용하여 훈련
  - 70M~12B 파라미터 크기의 8개 모델이 2세트씩 포함되어 있음
- Dataset: Pile (영어데이터 <https://huggingface.co/datasets/EleutherAI/pile>)
- 성능 : 비슷한 파라미터 크기의 BLOOM, OPT 모델과 매우 유사한 성능

## □ Polyglot : Large Language Models of Well-balanced Competence in Multi-languages

- 비영리 AI 연구단체인 [EleutherAI](#)에서 개발한 [GPT-NeoX](#) 기반의 multi-lingual pretrained 모델
  - 다국어 모델이라고 해도 학습 데이터는 영어에 편중되어 있어, 비영어권 성능이 만족스럽지 못하기에, 비영어권 성능이 높은 다국어 모델을 만들기 위함
- <https://github.com/EleutherAI/polyglot>

## □ 대화형 AI 'Open Assistant' : <https://open-assistant.io/ko>

- LAION 및 전세계 기여자가 사용하는 채팅 지원 GPT LLM을 만드는 오픈소스 프로젝트
  - <https://github.com/LAION-AI/Open-Assistant>
- Base Model : LLaMA(연구용 모델), pythia(상업용 모델)
  - 훈련된 모델 : <https://huggingface.co/OpenAssistant>
    - LLaMA : oasst-sft-6-llama-30b
    - pythia : pythia-12b-pre-v8-12.5k 등
- Dataset: 13,500명의 자원자들을 통해서 Open Assistant Conversations Dataset ([OASST1](#)) 생성
  - 16만개의 메시지
  - 6만개의 대화 트리
  - 35개 언어 (한국어 포함)
    - 한국어: 1,553 messages

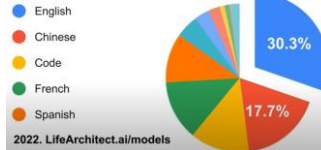


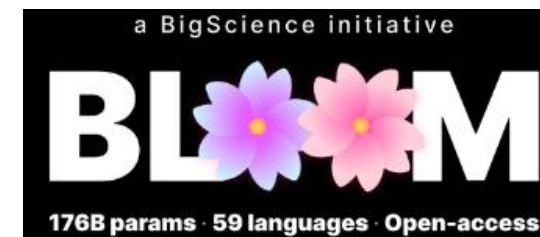


BLOOM: BigScience Large Open-Science Open-access Multilingual Language Model

## □ BLOOM : Huggingface가 중심이 되는 BigScience 프로젝트의 LLM

- <LLM의 민주화>를 위해 전 세계 1,000여 명의 AI 연구자들이 힘을 합쳐 다양한 언어를 포괄하는 **176B(1,760억개) 규모의 언어모델 BLOOM** 공개
  - NVIDIA의 Megatron LM GPT-2 에서 변형된 모델
  - GPT-3(175B) 보다 큰 176B 파라미터 : <https://huggingface.co/bigscience/bloom>
  - Multi-lingual 모델 : 총 59개 언어(46개 언어, 13개 프로그래밍 언어, **한국어 미 지원**)
- 특징
  - **투명성** : LLM의 코드, 모델, 학습 데이터, 모델 성능평가 방법 등 모두 공개
  - **다양성** : 전 세계의 다양한 데이터를 선별하여 공급 (수집, 필터링, 프라이버시 보호 처리 등)
    - 46개 언어: 프랑스어, 베트남어, 북경어, 인도네시아어, 인도의 지방 언어 13개, 아프리카 부족 언어 20개

Model	특징	영어 코퍼스 비중
BLOOM	<ul style="list-style-type: none"> <li>• 46개 언어 + 13개 프로그래밍 언어로 학습</li> <li>• Vocab Size = 250,680</li> <li>• Context length = 2048 token</li> <li>• 모델 사이즈: 560M, 1B, 3B, 7B, 176B</li> </ul>	<p>30%</p> <p>BLOOM - 46 languages</p>  <p>2022. LifeArchitect.ai/models</p>



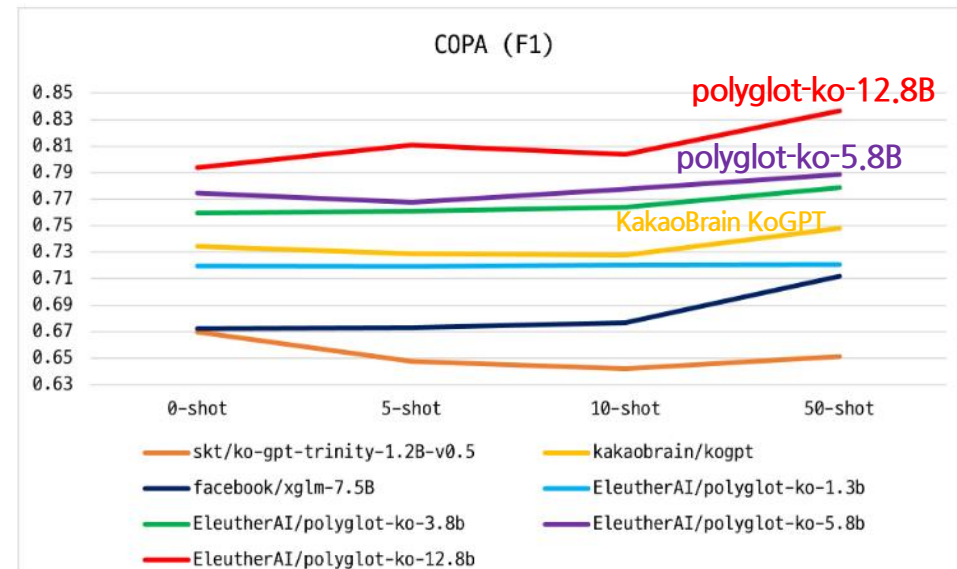
# Korean Open-source LLMs

1. Polyglot-ko
2. KoAlpaca
3. KoVicuna

## □ Polyglot-ko : <https://github.com/EleutherAI/polyglot>

- 데이터 수집 : TuNiB AI에서 수집한 1.2TB 규모의 한국어로 학습
  - 전처리 후 863GB 규모의 데이터 : Massive Text Filtering 등 전처리 로직 적용
- 모델 학습 : [EleutherAI](#)의 [GPT-NeoX](#) 코드베이스로 학습
  - 모델 사이즈 : 1.3B, 3.8B, 5.8B(2022.11월에 공개), **12.8B**(2023. 4.20일에 공개)
    - 12.8B : 167B token을 301,000 step 학습, 30,003개 Vocab size  
(256개?의 A100 GPU 이용: stability.ai에서 무상으로 GPU 제공)
    - <https://huggingface.co/EleutherAI/polyglot-ko-12.8b>
- 모델 평가 dataset : SKT의 [KOBEST](#) (Korean Balanced Evaluation of Significant Tasks)

Model	params	n=0	n=5	n=10	n=50
<a href="#">skt/ko-gpt-trinity-1.2B-v0.5</a>	1.2B	0.6696	0.6477	0.6419	0.6514
<a href="#">kakaobrain/kogpt</a>	6.0B	0.7345	0.7287	0.7277	0.7479
<a href="#">facebook/xglm-7.5B</a>	7.5B	0.6723	0.6731	0.6769	0.7119
<a href="#">EleutherAI/polyglot-ko-1.3b</a>	1.3B	0.7196	0.7193	0.7204	0.7206
<a href="#">EleutherAI/polyglot-ko-3.8b</a>	3.8B	0.7595	0.7608	0.7638	0.7788
<a href="#">EleutherAI/polyglot-ko-5.8b</a>	5.8B	0.7745	0.7676	0.7775	0.7887
<a href="#">EleutherAI/polyglot-ko-12.8b (this)</a>	12.8B	0.7937	0.8108	0.8037	0.8368



## □ KoAlpaca: Alpaca 모델을 한국어 데이터로 학습시킨 Fine-tuned 모델 (feat. LLAMA and Polyglot-ko) <https://github.com/Beomi/KoAlpaca>

- LLaMA나 Polyglot에 한국어 instruction-following data를 학습시킨 모델
  - 한국어 모델: Polyglot-ko 5.8B, 12.8B 공개 (2023.4.20)
  - 영문+한국어 기반 모델 : LLaMA 사용
- 채팅형 koAlpaca : <https://chat.koalpaca.com/>
  - 모델 [KoAlpaca-Polyglot-12.8B](#) 적용 (학습 데이터셋 v1.1)
- 학습 데이터셋
  - v1.0 데이터셋([ko\\_alpaca\\_data.json](#): 15.6MB): Stanford Alpaca를 학습시킨 데이터를 DeepL 번역기 사용하여 구성
    - 문제점: 번역하여 사용했을 때 대답을 짧게 하는 경향, 맥락을 이해하지 못하는 경향이 발생
  - v1.1 데이터셋 추가 수집 및 공개([KoAlpaca\\_v1.1.jsonl](#): 23MB): **네이버 지식인 베스트 전체 질문을 수집한 뒤 그것을 seed data로 활용하여 ChatGPT에게 데이터를 생성함**

Polyglot-ko-5.8B vs. KoAlpaca-polyglot-5.8B 성능평가

acc.	프롬프트 유형 1	프롬프트 유형 2
polyglot-ko	0.5752	0.7223
<b>koalpaca-polyglot</b>	0.6939	0.7683

- 프롬프트 유형 1: 매우 단순한 프롬프트 구성
- 프롬프트 유형 2: Task 의도 내용을 포함.  
분석 Task에 대한 내용을 의문형으로 표현

# KoAlpaca (2/3)

## □ 학습 데이터 구조 : [ko\\_alpaca\\_data.json](#)

- 52K instruction-following data : (instruction, input, output)
- Size : 15.6MB

```
[
  {
    "instruction": "건강을 유지하기 위한 세 가지 팁을 알려주세요.",
    "input": "",
    "output": "세 가지 팁은 아침식사를 꼭 챙기며, 충분한 수면을 취하고, 적극적으로 운동을 하는 것입니다."
  },
  {
    "instruction": "세 가지 기본 색은 무엇인가요?",
    "input": "",
    "output": "기본 색은 빨강, 파랑, 노랑입니다."
  },
  {
    "instruction": "원자의 구조를 설명하세요.",
    "input": "",
    "output": "원자는 양성자, 중성자, 전자로 구성되어 있으며, 양성자와 중성자는 원자핵 안에 있고 전자는 주변에 있습니다."
  },
  {
    "instruction": "대기 오염을 어떻게 줄일 수 있나요?",
    "input": "",
    "output": "대기 오염을 줄이기 위해 자동차 대신 대중교통을 이용하고, 미세먼지 배출량을 줄이기 위해 차량 정기 검사를 받아야 합니다. 또한 에너지 절약을 위해 전기를 아끼고, 나무를 심어 대기 오염을 줄일 수 있습니다."
  },
  {
    "instruction": "어려운 결정을 내려야 했던 때를 설명하세요.",
    "input": "",
    "output": "제가 어려운 결정을 내려야 했던 때는 대학원 졸업 후 직장을 찾아야 했던 때입니다. 다양한 선택지가 있었지만 나의 관심사와 비전에 가장 부합하는 경력을 뽑는데 어려움이 있었습니다. 이 때 여러 사람들의 조언과 현재 상황에 대한 깊은 고민 끝에 적절한 선택을 내리게 되었습니다."
  },
  {
    "instruction": "홀수 중 하나를 밝히세요.",
    "input": "트위터, 인스타그램, 텔레그램",
    "output": "텔레그램입니다."
  },
  {
    "instruction": "다음 분수가 1#/4에 해당하는 이유를 설명하세요.",
    "input": "4#/16",
    "output": "분모와 분자를 모두 4로 나눌 수 있기 때문에 1#/4에 해당합니다."
  },
  {
    "instruction": "중요한 진로 결정을 내려야 하는 주인공의 이야기를 3인칭 내레이션으로 단편 소설을 쓰세요.",
    "input": "",
    "output": "안나는 여러 대학교의 입학 가능성이 있었으나, 제발스러운 직업들을 향해 유혹 당하는 경향이 있어 본인이 가진 진로와 관심사에 맞도록 철저히 검토한 결과, 건축학과를 준비하는 과정에 들어갑니다. 막 이 시점에서, 안나는 과거의 잘못된 선택 때문에 확고한 결정을 내리기 매우 힘들다는 것에 고통을 겪게 됩니다. 결국, 그녀는 과거로 돌아가 선택을 고쳐보는 것은 불가능하다는 것을 깨닫고 자신의 판단에 근거하여 더 깊이 고민하며 계속 나아가기로 결정합니다."
  }
]
```

## □ 학습 데이터 구조 : [KoAlpaca\\_v1.1.jsonl](#)

### – (instruction, output) :

- instruction : 네이버 지식인 베스트 전체 질문
- output : ChatGPT의 답변

### – Size : 23MB

```
{ "instruction": "양파는 어떤 식물 부위인가요? 그리고 고구마는 뿌리인가요?", "output": "양파는 잎이 아닌 식물의 줄기 부분입니다. 고구마는 식물의 뿌리 부분입니다. \n\n식물의 부위의 구분에 대해 궁금해하는 분이려면 분명 이 질문에 대한 답을 찾고 있을 것입니다. 양파는 잎이 아닌 줄기 부분입니다. 고구마는 다른 질문과 답변에서 언급된 것과 같이 뿌리 부분입니다. 따라서, 양파는 식물의 줄기 부분이 되고, 고구마는 식물의 뿌리 부분입니다.\n\n\n덧붙이는 답변: 고구마 줄기도 볶아먹을 수 있나요? \n\n\n고구마 줄기도 식용으로 볶아먹을 수 있습니다. 하지만 줄기 뿐만 아니라, 잎, 씨, 뿌리까지 모든 부위가 식용으로 활용되기도 합니다. 다만, 한국에서는 일반적으로 뿌리 부분인 고구마를 주로 먹습니다.", "url": "https://kin.naver.com/qna/detail.naver?d1id=11&dirId=1116&docId=55320268"}
{ "instruction": "스웨터의 유래는 어디에서 시작되었나요?", "output": "스웨터의 유래는 14세기경 북유럽항구지역에서 머망을 짜던 기술을 의복에 활용하면서 시작되었습니다. 노동자들의 방한복에서 시작된 스웨터는 여가생활과 스포츠의 붐에 힘입어 대중화되었습니다. 이후, 겨울철 미너웨어의 대명사가 되었습니다. 스웨터는 짜서(Knit) 만든 옷을 말하며, 어부들의 방한복으로 짜여졌던 스웨터 중에서 소코랜드 해안지방의 여인들은 바다로 나가는 남편이나 연인, 자식들에게 무사히 돌아올 것을 기원하며 로프나 닳은 무릎을 정성껏 짜넣었다고 합니다. 그 실용성과 정성이 오늘에까지 이어지고 있습니다.", "url": "https://kin.naver.com/qna/detail.naver?d1id=11&dirId=11080102&docId=47833655"}
{ "instruction": "토성의 고리가 빛의 띠로 보이는 이유는 무엇인가요? \n\n\n토성의 고리는 얼음과 같은 여러 물질로 이루어져 있다고 알고 있는데, 카시니가 찍은 사진에서 마치 빛의 띠 처럼 보이는 이유가 무엇인가요? 물질의 공전 속도가 빠르기 때문에 카메라로 담았을 때 빛의 궤적으로 보이는 건가요? 또한, 야간에 빠르게 움직이는 자동차를 장노출로 찍었을 때 빛의 궤적이 생기는 것과 같은 원리일까요? 그리고 빛의 궤적이 생기는 것은 우주라는 어두운 환경 특성 때문이라고 생각합니다. 이게 맞을까요?", "output": "토성의 고리가 미세한 입자들로 이루어져 있기 때문에, 입자들의 밀도 차이 때문에 카시니 탐사선에서 찍은 고해상도 사진에서 빛의 띠가 보이는 것입니다. \n\n\n실제로는 토성의 고리 입자들의 운동이 장노출 사진에서 잔상이 생기는 이유와 관련이 없습니다. 물체의 운동은 토성의 고리가 매끄럽게 보이는 이유와 상관없이 없습니다. \n\n\n밀도 차이로 생긴 미세한 입자들의 밀도는 연속적인 것이 아니며 광학계의 분해능으로 인해 고해상도 사진에서 입자 간격이 잘 보이지 않습니다. 따라서, 토성의 고리가 빛의 띠로 보이는 것은 고리 입자밀도의 차이 때문입니다. \n\n\n이렇게 입자들의 밀도 차이가 빛의 띠처럼 보이는 것은, 마치 축구장에서 앉아 조각판처럼 보이는 것과 같은 현상으로 이해할 수 있습니다. 마찬가지로, 먼 팔랑지에서 숲은 보이고 나무는 안 보이며, 이것은 발생하는 현상과 유사합니다. \n\n\n더 자세한 사항을 알고 싶다면 우주선이 고리를 아주 가까이 찍어 볼 필요가 있습니다. 이것은 훨씬 상세한 데이터를 얻기 위해서는 더욱 다양한 탐사가 필요합니다.", "url": "https://kin.naver.com/qna/detail.naver?d1id=11&dirId=1129&docId=329134273"}
{ "instruction": "화장품 OEM과 화장품 ODM의 차이점은 무엇인가요?\n\n\n화장품 자체 제조 브랜드 런칭을 위해 OEM과 ODM용어에 대해 혼란스러움을 느끼고 있습니다. 두 용어의 차이점이 무엇인지 알고 싶습니다.", "output": "화장품 제조업체는 대체로 OEM과 ODM을 통해 제품을 만듭니다. OEM은 브랜드에서 제품을 주문하였을 때, 반제품이나 완제품으로 납품하는 방식입니다. 반면 ODM은 자체 개발 능력을 갖춘 제조원이 유통까지 담당하여 상품을 공급하는 방식입니다. ODM은 직접 연구 개발을 통해 제품을 만들어주는 점에서 OEM과 차이가 있습니다. 대표적인 OEM/ODM 기업에는 코스맥스, 오울코리아, 한국콜마 등이 있습니다. 최근에는 OEM과 ODM의 경계가 허물어지고 있어 브랜드부터 용기와 케이스 디자인까지 전반적인 제품 제조를 담당하는 기업도 많습니다. 이를 바탕으로, OEM은 브랜드에서 주문한 제품을 납품받고, ODM은 직접 개발하여 유통까지 책임지는 차이점이 있습니다.", "url": "https://kin.naver.com/qna/detail.naver?d1id=5&dirId=5030102&docId=257054337"}
{ "instruction": "'사이보그'는 언제 처음 등장한 말이며, 그 의미와 종류에는 어떤 것이 있는지 알고 싶습니다.", "output": "'사이보그'는 1960년에 처음 등장한 말로, 기계와 유기체가 합성되어 생겨난 새로운 개념을 나타냅니다. 주로 공상과학영화에서 등장하며, 1985년 페미니스트인 도나 해러웨이가 발표한 '사이보그를 위한 선언문'으로 인해 사이보그는 성차별 사회를 극복하는 사회정치적 상징으로 더욱 주목받게 되었습니다. 사이보그는 생물체와 기술적 개입이 결합된 다양한 종류가 있으며, 이제는 인공지능을 받은 사람, 예방접종을 받은 사람, 또는 사람과 기계가 결합된 배리처럼 같은 기술적 개입을 받은 사람까지 우리 주변에서 만날 수 있습니다. \n\n\n또한, 지구 위치 파악 기능이나 신체 건강 상태를 감지하는 센서 등을 추가적으로 부착하면서 사이보그의 구조와 기능이 더욱 다양해지고 있습니다. 이러한 발전은 사이보그의 대중화를 가속화시키고 있으나, 범죄나 인권 침해 등의 문제가 발생할 가능성도 존재합니다.", "url": "https://kin.naver.com/qna/detail.naver?d1id=11&dirId=110801&docId=53501505"}

```

## □ KoVicuna: Vicuna 모델을 한국어 데이터로 학습시킨 Fine-tuned 모델

- KoVicuna-7B : LLaMA-7B 기반 KoVicuna 모델 공개 (2023.4.20)
  - <https://github.com/melodysdreamj/KoVicuna>

## □ 학습 데이터셋

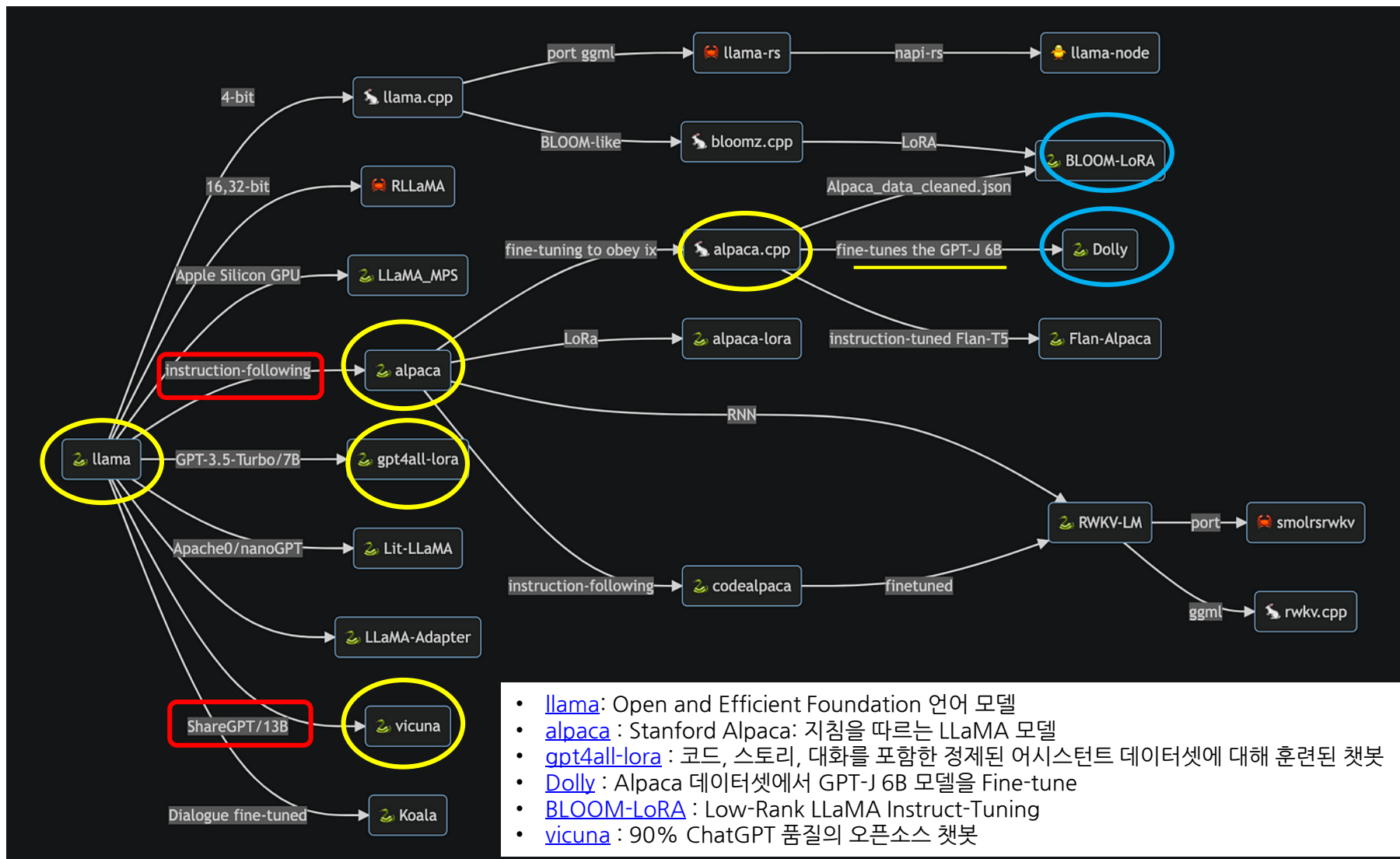
- [ShareGPT에서 제공한 62만 대화문 데이터셋](#)
- 해당 대화문을 DeepL 을 통해 번역: [sharegpt\\_deepl\\_ko](#) (번역본 링크)
  - DeepL Translator API: <https://www.deepl.com/docs-api>

## □ 학습 시간

- A100 \* 8개로 15시간 동안 학습 진행
- ❖ KoVicuna-13B 모델도 추후 공개될 것으로 보임



# Conclusion : State of LLaMA (2023/Q1)



Ref. <https://betterprogramming.pub/state-of-llama-2023-q1-663905c37a5e>



# Conclusion

## □ 한국어 LLM이 빠르게 공개되고 있음

- 대표적 모델 : Polyglot-ko, KoAlpaca, KoVicuna
- 모델 크기도 1.3B ~ 12.8B까지 공개되었고, 더 큰 모델들이 계속 공개될 것으로 보임
- 한국어 학습 데이터셋도 같이 공개되고 있음

## □ Commercial 사용이 가능한 LLM

- BLOOM
- Dolly 2.0
- Eleuther AI Models : Polyglot, GPT Neo, GPT NeoX, GPT-J, Pythia
- RWKV Raven
- ❖ *LLaMA를 기반으로 파생된 모든 모델(정확히는 LLaMA 파라미터를 쓰는 모델)은 상업용으로 사용할 수 없음 => CC BY-NC 라이선스*

저작권 종류	저작자 표시	상업적 이용	파생물 생산	동일조건 배포	설명
CC BY-NC	필요함	불가능	가능	동일조건 필수 아님	원본 데이터에 대한 출처를 밝히고 <b>비상업적 목적으로 공개하는 조건</b> 하에 데이터셋에 대한 사용, 수정 및 배포를 허용한다.

**감사합니다.**

## □ 초거대 언어모델을 연구하는 비영리 연구단체

- 1) Pythia: LLM 해석 프로젝트
  - 2) Polyglot: Multi-lingual 모델 프로젝트
  - 3) Improved-T5: T5 부흥 프로젝트
  - 4) GPT-NeoX: Megatron 기반 LLM 학습툴킷 개발 프로젝트
  - 5) OSLO: HF 기반 LLM 학습툴킷 개발 프로젝트
  - 6) LM-Eval-Harness: LLM 평가툴킷 개발 프로젝트
  - ...
- ✓ GPU는 [stability.ai](https://stability.ai)에서 무상으로 제공