
LLaMA

Large Language Model Meta AI

인공지능 명예학회



Contents

- LLaMA 란?
- LLaMA 모델 구조
- 모델 성능과 개선점
- Instruction finetuning
- 주요 파생 모델

LLaMA 란?

LLaMA Large Language Model Meta AI

LLaMA는 이름대로 메타에서 2023년 개발한 대형 언어모델로, LLaMA-1은 2023년 2월, LLaMA-2는 2023년 7월에 공개되었다. LLaMA-1은 연구용 목적 한정으로 오픈, 2는 완전히 오픈소스로 공개되었다.

모델 구조 외적으로 다른 대형 언어모델들과의 가장 큰 차이점은 코드, 사전학습 데이터가 모두 공개된 오픈소스 모델이라는 것이다. 이때문에 다양한 파생형 모델의 탄생과 NLP 작업 성능 향상에 큰 기여를 하였다.

다양한 매개변수 버전이 존재해 가용할 수 있는 훈련 환경에 따라 선택할 수 있다.

<https://llama.meta.com/>

LLaMA 모델구조

기존에 LLM을 설계할 때 파라미터의 수가 많아야 성능이 올라간다고 여겨졌다.
하지만 모델의 크기가 무작정 큰 모델이 아니라, 모델의 크기가 상대적으로 작아도 더 많은 데이터로 훈련하면 더 좋은 성능이 나올 수 있다는 최근의 연구에 기반해 만든 모델이다.

사전학습 데이터

wikipedia, 웹 크롤링 데이터, 종합적인 도서와 논문, meta 자사 데이터,
Github, Stack Exchange 등 사용

토큰나이저

구글 SentencePiece 기반의 BPE (Byte Pair Encoding) 토큰나이저를 사용.
1.5 조개의 토큰 (LLaMA-1 기준)

LLaMA 모델구조

Transformer 아키텍처를 기반으로 제작했고, 다른 모델에서 사용된 성능 향상 기법을 적용해 발전시켰다.

정규화 기법 [GPT-3]

학습에서의 안정화 증진을 위해 각 레이어의 input에 정규화를 적용했다.

GPT-3에서도 사용한 RMS(Root Mean Squared) Normalization을 사용했다.

RMSNorm :

$$a = Wx$$

$$RMS(a) = \sqrt{\frac{1}{n} \sum_{i=1}^n a_i^2}$$

$$\overline{a_i} = \frac{a_i}{RMS(a)}$$

LLaMA 모델구조

활성화 함수

Swish 활성화 함수와 GLU 활성화 함수를 결합한 SwiGLU 활성화 함수를 사용.

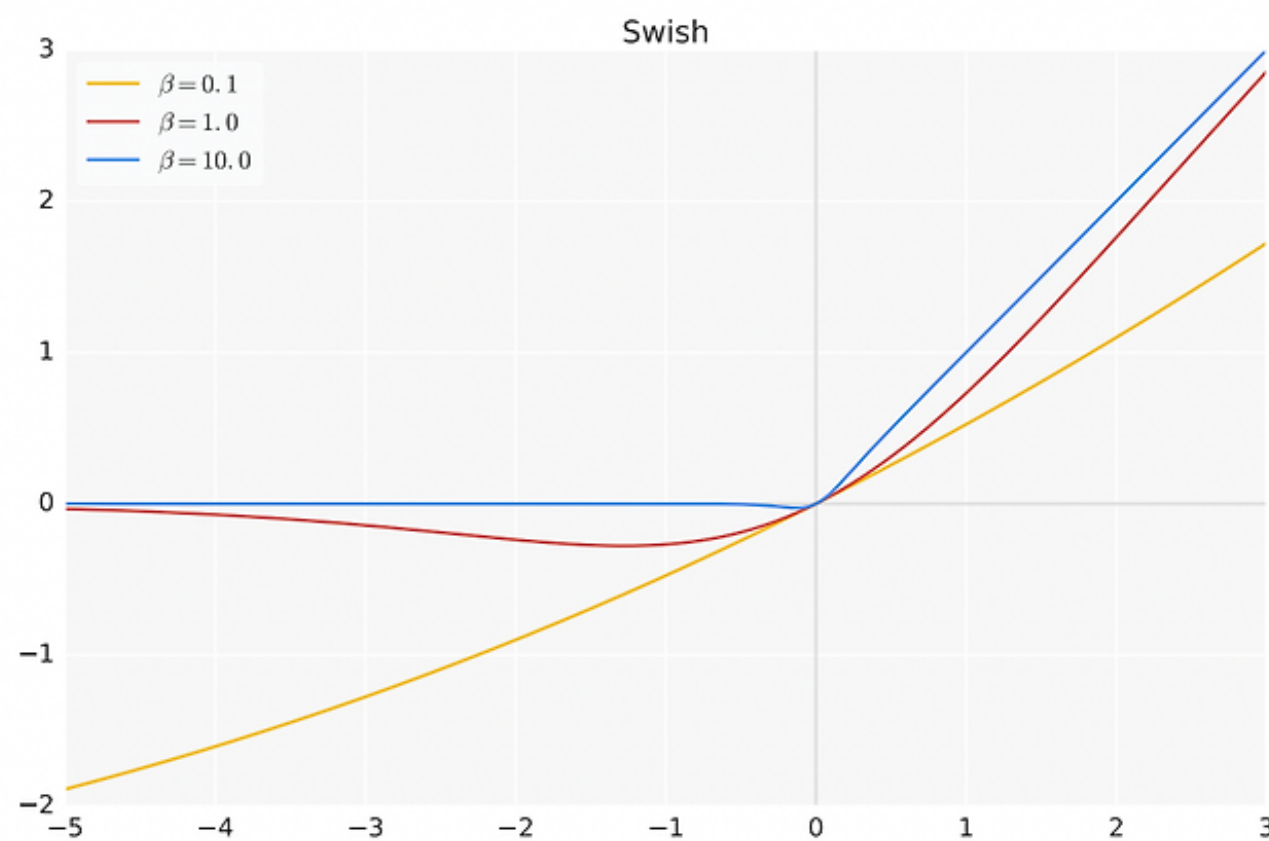


Figure 4: The Swish activation function.

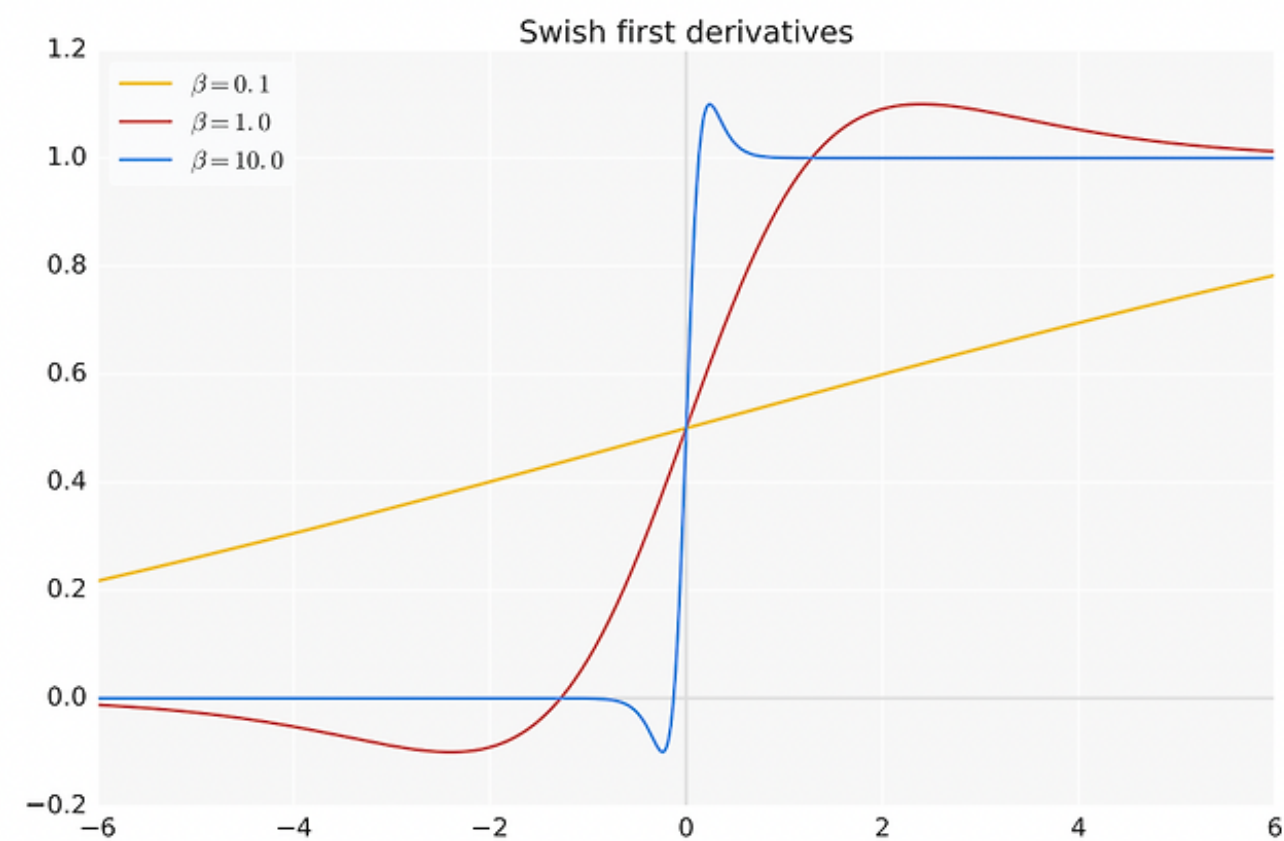


Figure 5: First derivatives of Swish.

LLaMA 모델구조

Rotary Embeddings

기존의 절대적 Positional Embedding이 아니라 Token Embedding을 회전시켜서 상대적인 위치를 표현하는 기법인 Rotary Embedding을 사용했다.

Optimizer

AdamW 사용.

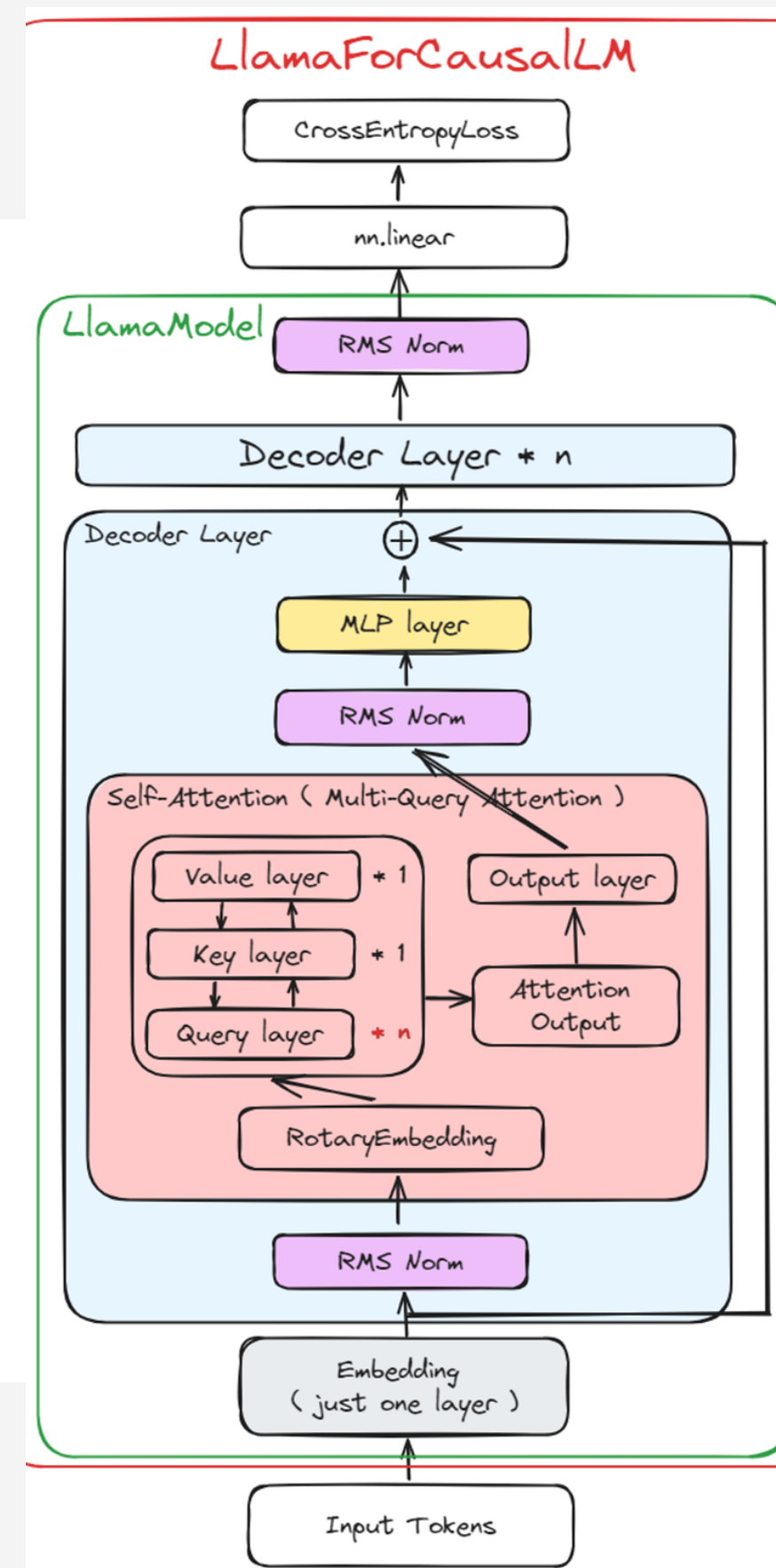
LLaMA-1과 LLaMA-2의 차이점은 토큰 개수의 변화, 파라미터 개수, 하이퍼파라미터 세부조정 등 작은 변화이므로 LLaMA-1을 기준으로 서술.

LLaMA 모델 구조

params	dimension	n heads	n layers	learning rate	batch size	n tokens
6.7B	4096	32	32	$3.0e^{-4}$	4M	1.0T
13.0B	5120	40	40	$3.0e^{-4}$	4M	1.0T
32.5B	6656	52	60	$1.5e^{-4}$	4M	1.4T
65.2B	8192	64	80	$1.5e^{-4}$	4M	1.4T

Table 2: Model sizes, architectures, and optimization hyper-parameters.

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size
GPT-3 Small	125M	12	768	12	64	0.5M
GPT-3 Medium	350M	24	1024	16	64	0.5M
GPT-3 Large	760M	24	1536	16	96	0.5M
GPT-3 XL	1.3B	24	2048	24	128	1M
GPT-3 2.7B	2.7B	32	2560	32	80	1M
GPT-3 6.7B	6.7B	32	4096	32	128	2M
GPT-3 13B	13.0B	40	5140	40	128	2M
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M



LLaMA 모델 성능과 개선점

		BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA
GPT-3	175B	60.5	81.0	-	78.9	70.2	68.8	51.4	57.6
Gopher	280B	79.3	81.8	50.6	79.2	70.1	-	-	-
Chinchilla	70B	83.7	81.8	51.3	80.8	74.9	-	-	-
PaLM	62B	84.8	80.5	-	79.7	77.0	75.2	52.5	50.4
PaLM-cont	62B	83.9	81.4	-	80.6	77.0	-	-	-
PaLM	540B	88.0	82.3	-	83.4	81.1	76.6	53.0	53.4
LLaMA	7B	76.5	79.8	48.9	76.1	70.1	72.8	47.6	57.2
	13B	78.1	80.1	50.4	79.2	73.0	74.8	52.7	56.4
	33B	83.1	82.3	50.4	82.8	76.0	80.0	57.8	58.6
	65B	85.3	82.8	52.3	84.2	77.0	78.9	56.0	60.2

Table 3: Zero-shot performance on Common Sense Reasoning tasks.

LLaMA 모델 성능과 개선점

1. 오픈소스가 아닌 다른 모델에 비해 우수한 성능을 보유.
2. 다양한 튜닝 기법 적용에 최적화.
3. 오픈소스 모델이기 때문에 다양하게 모델 변형/ task 최적화 가능.
4. GPT-3, GPT-4와 같은 모델과 비교해 파라미터의 개수가 현저히 적어 다양한 환경에서 학습 가능.

결론 : 설계가 완전히 혁신적인 모델은 아니지만, 사용에 제한이 없고 성능이 준수하기 때문에 NLP 분야 발전에 상당히 기여

LLaMA 모델 성능과 개선점

1. 오픈소스가 아닌 다른 모델에 비해 우수한 성능을 보유.
2. 다양한 튜닝 기법 적용에 최적화.
3. 오픈소스 모델이기 때문에 다양하게 모델 변형/ task 최적화 가능.
4. GPT-3, GPT-4와 같은 모델과 비교해 파라미터의 개수가 현저히 적어 다양한 환경에서 학습 가능.

결론 : 설계가 완전히 혁신적인 모델은 아니지만, 사용에 제한이 없고 성능이 준수하기 때문에 NLP 분야 발전에 상당히 기여

Instruction Finetuning

Instruction tuning

사전 훈련된 언어 모델을 다양한 작업에 걸쳐 범용적으로 사용할 수 있도록 최적화하는데 중점을 둔다. 모델이 다양한 형태와 유형의 지시사항을 이해하고, 이에 대응하는 적절한 응답을 생성할 수 있도록 한다.

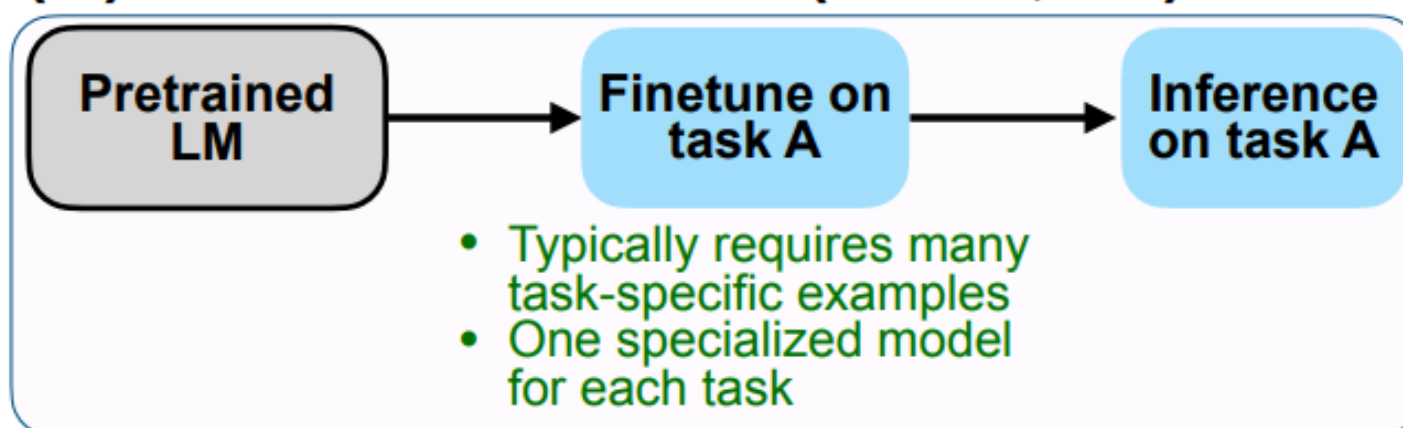
모델이 다양한 지시사항을 처리할 수 있도록 넓은 범위의 지시 기반 작업에 대해 추가학습을 수행한다. 지시사항에 대한 이해력과 응답 생성 능력을 향상시킬 수 있다.

Instruction Finetuning

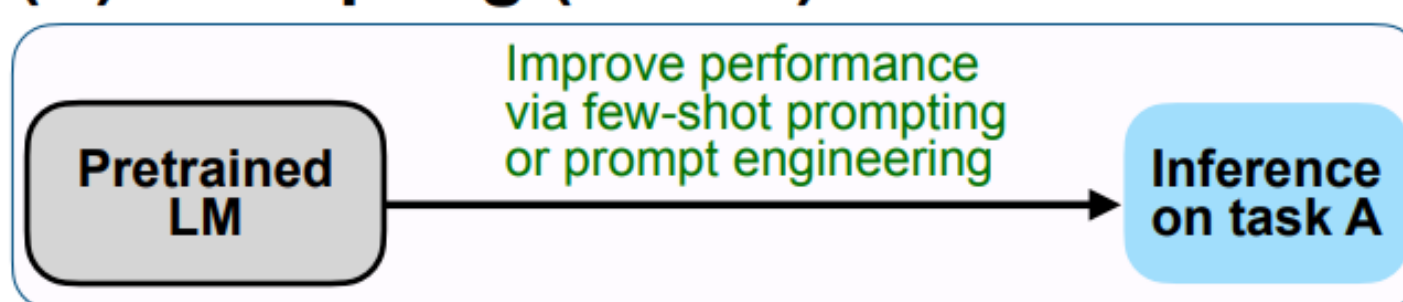
모델을 특정 작업이나 도메인에 더욱 특화되도록 조정하는 데 중점을 둔다. 모델을 특정 도메인의 데이터나 작업에 대한 지시사항으로 추가학습시킨다.

Instruction Finetuning

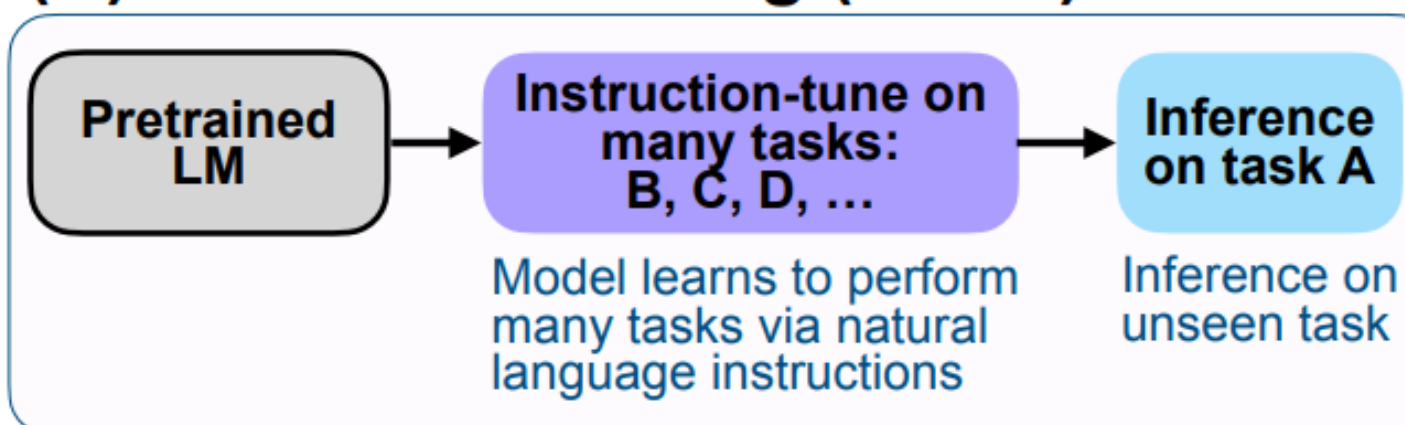
(A) Pretrain–finetune (BERT, T5)



(B) Prompting (GPT-3)



(C) Instruction tuning (FLAN)



Instruction Finetuning

Instruction tuning

사전 훈련된 언어 모델을 다양한 작업에 걸쳐 범용적으로 사용할 수 있도록 최적화하는데 중점을 둔다. 모델이 다양한 형태와 유형의 지시사항을 이해하고, 이에 대응하는 적절한 응답을 생성할 수 있도록 한다.

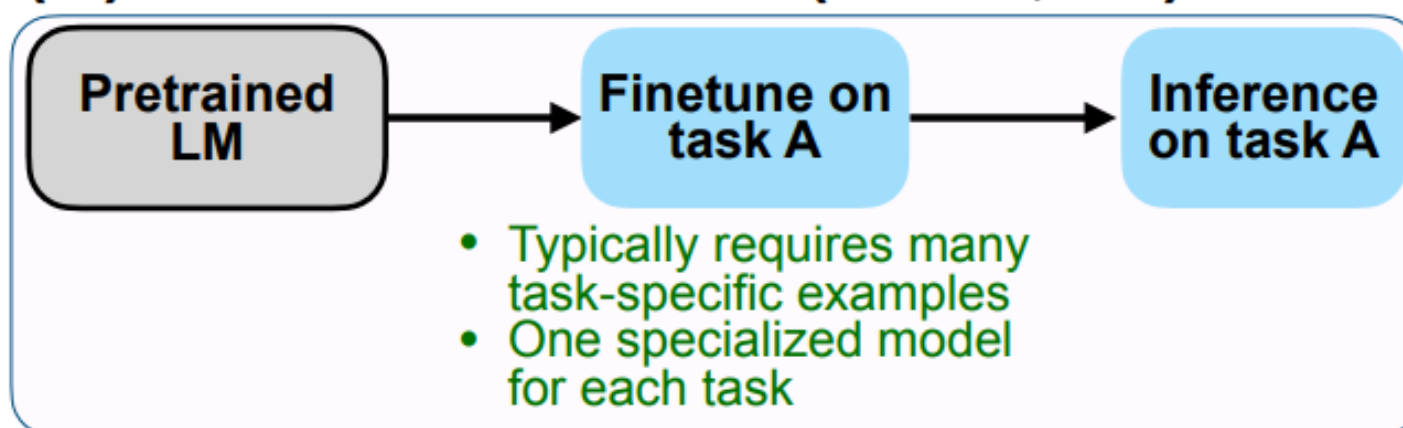
모델이 다양한 지시사항을 처리할 수 있도록 넓은 범위의 지시 기반 작업에 대해 추가학습을 수행한다. 지시사항에 대한 이해력과 응답 생성 능력을 향상시킬 수 있다.

Instruction Finetuning

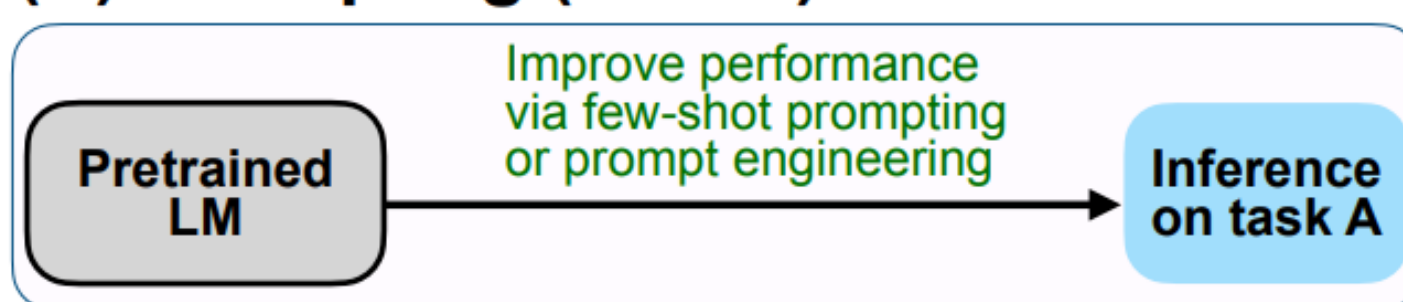
모델을 특정 작업이나 도메인에 더욱 특화되도록 조정하는 데 중점을 둔다. 모델을 특정 도메인의 데이터나 작업에 대한 지시사항으로 추가학습시킨다.

Instruction Finetuning

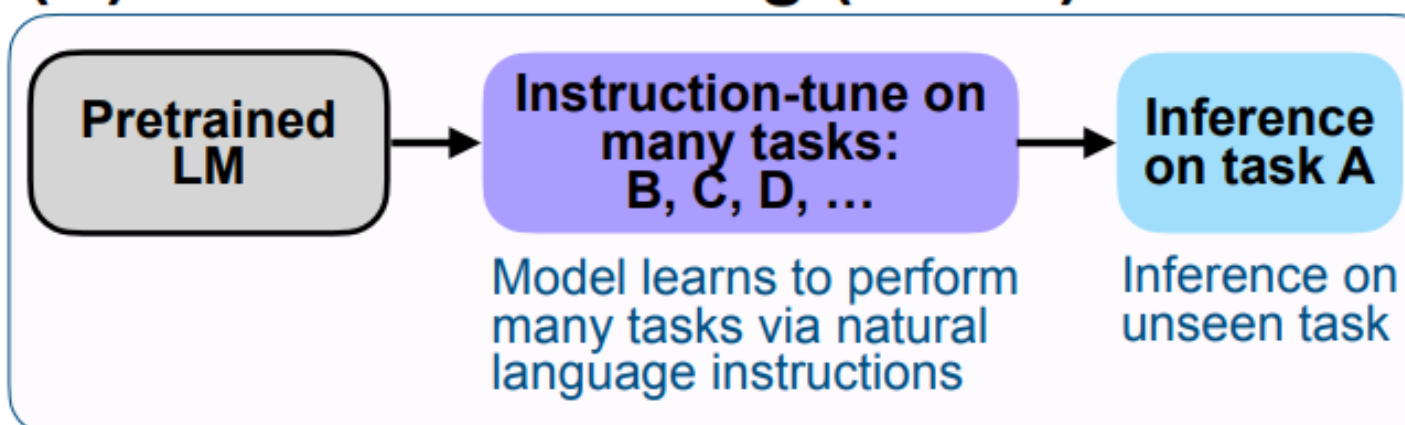
(A) Pretrain–finetune (BERT, T5)



(B) Prompting (GPT-3)



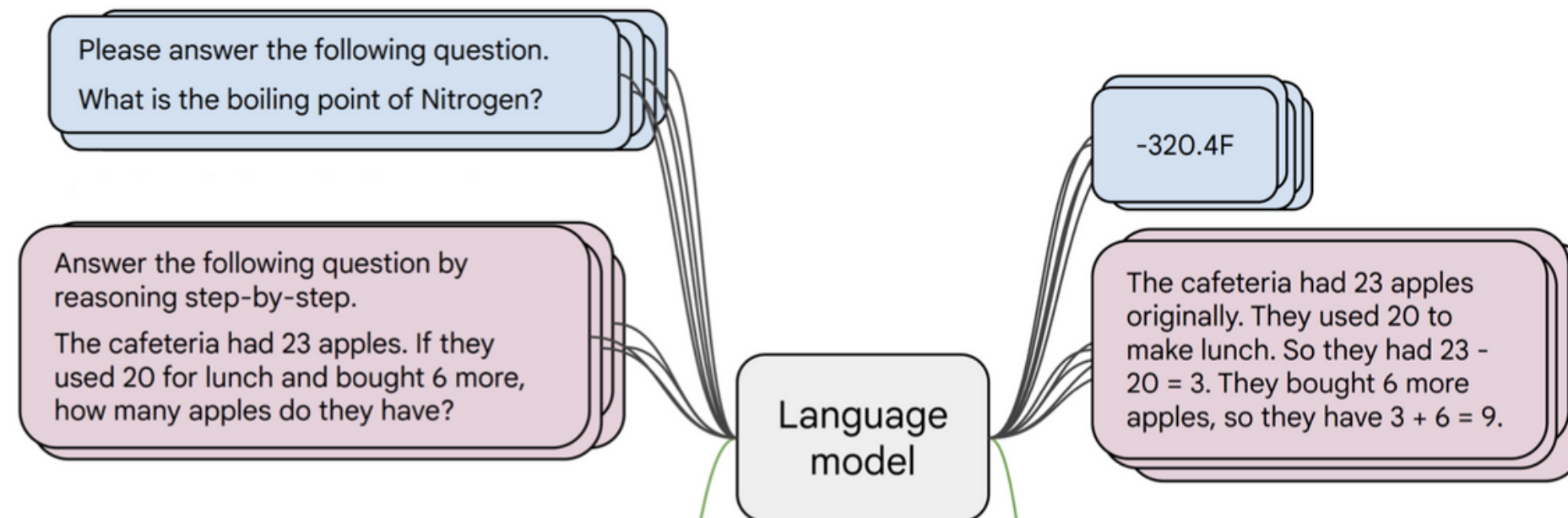
(C) Instruction tuning (FLAN)



Instruction Finetuning

Instruction finetuning

- **Collect examples** of (instruction, output) pairs across many tasks and finetune an LM



- **Evaluate on unseen tasks**

Q: Can Geoffrey Hinton have a conversation with George Washington?
Give the rationale before answering.

Geoffrey Hinton is a British-Canadian computer scientist born in 1947. George Washington died in 1799. Thus, they could not have had a conversation together. So the answer is "no".

Instruction Finetuning

Instruction Finetuning

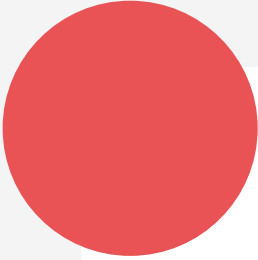
Ex) 성공회대 정보를 제공하는 생성형 챗봇을 만들고자 할 때

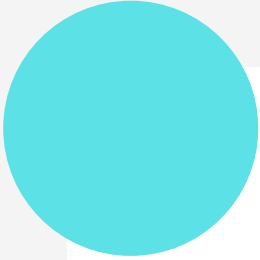
성공회대와 관련된 지시사항과 예제를 포함하는 데이터 세트를 준비해야 한다.

지시사항

- **성공회대학교에 관한 질문을 받고 그에 대해 정확한 정보로 답하는 챗봇이다.**
- **한글로 작성해라.**
- **일만관, 새천년관 ..은 성공회대의 건물이다.**

Instruction Finetuning

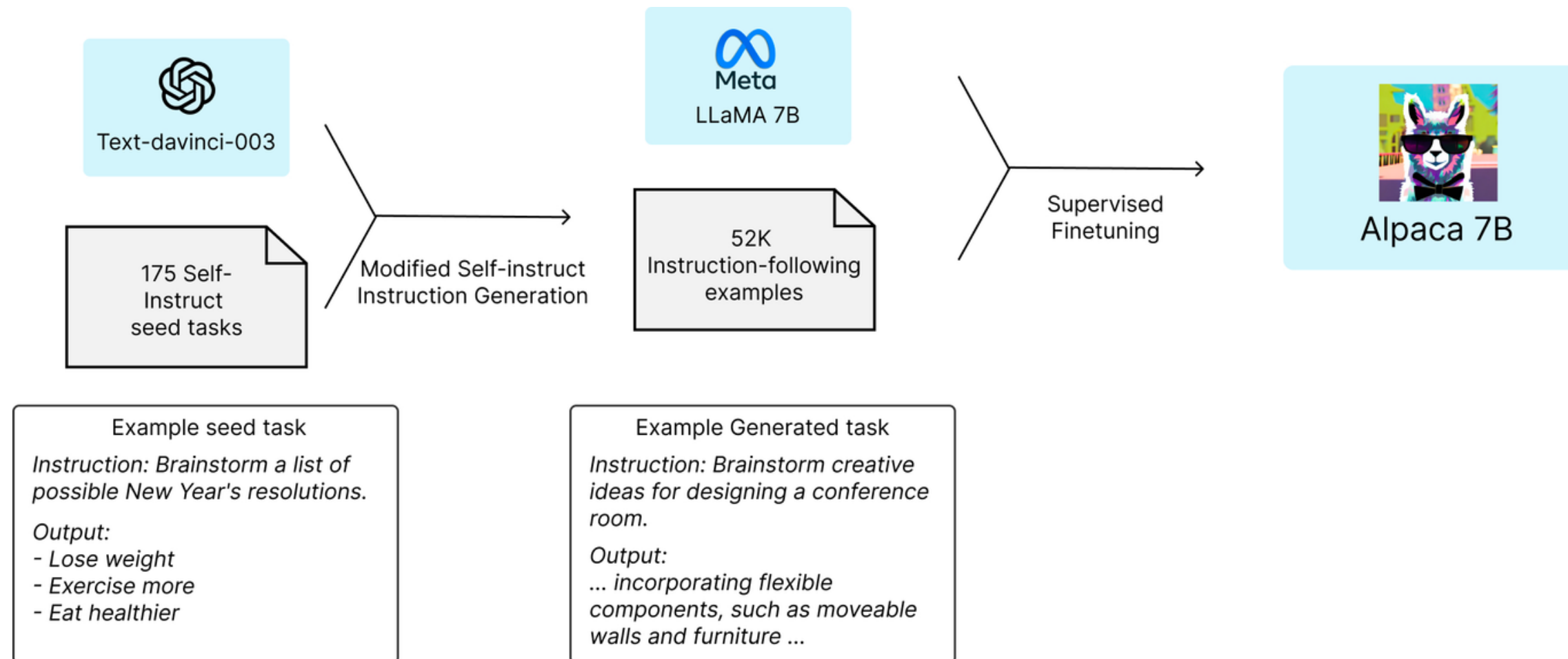
- 
- 특정 작업에 효과적으로 대응할 수 있음
 - 넓은 범위의 지식과 능력을 가지고 있는 사전 훈련된 모델을 변용 가능
 - 기존 파인튜닝보다 효율적임

- 
- 사전 훈련된 모델의 성능이 뛰어나야 제대로 훈련됨
 - 데이터 제작의 어려움
 - open-ended 문제의 경우 채점의 기준이 애매함

LLaMA 파생모델

Alpaca

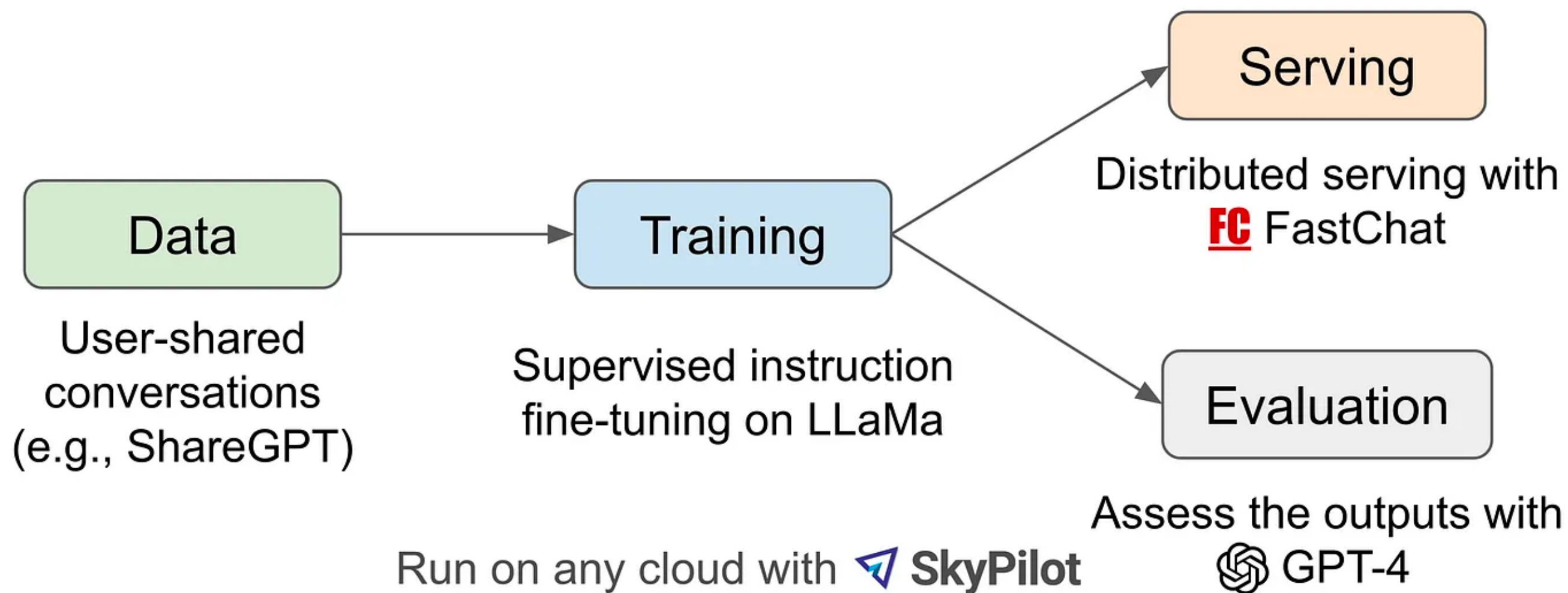
스탠퍼드 대학에서 공개한 모델로, LLaMA 7B를 사전학습 모델로 사용해 파인튜닝했다. 언어모델로 생성한 Instructions을 데이터로 사용해 LLaMA의 지시 이해 능력과 논리적인 답변 생성 능력을 향상시켰다. Solar와 같은 모델도 Alpaca 기반으로 추가 학습되었다.



LLaMA 파생모델

Vicuna

버클리 대학교 등 총 4개 대학에서 Alpaca에 영감을 받아 공동으로 제작한 오픈소스 챗봇이다. SharedGPT(사용자 프롬프트와 ChatGPT의 답변을 공유할 수 있는 사이트)의 대화를 기반으로 LLaMA를 챗봇으로 파인튜닝했다.



참고 자료

LLaMA : <https://arxiv.org/pdf/2302.13971.pdf>

self-instruct: <https://github.com/yizhongw/self-instruct>

**Open source LLMs : LLaMa, Alpaca, Vicuna (With PEFT, LoRA):
<https://velog.io/@srk/Open-source-LLMs-LLaMa-Alpaca-Vicuna-With-PEFT-LoRA>**