LLaMA: Open and Efficient Foundation Language Models

Name
김호재

NLP

2024/05/28

TRAIN AND TEST

Intro

- LLM(Large Language Model): 많은 텍스트 데이터 학습을 통해서 다양한 task를 처리할 수 있는 모델

그렇다면 Parameter가 많을수록, 모델의 크기가 클수록 성능이 좋아지지 않을까?

2022년 Hoffmann에 의해 무식하게 큰 것이 좋은 것이 아니란 게 밝혀짐.

최적의
(model size)/(data set size)는
어떻게 찾을 수 있을까

Intro

- 하지만 본 논문에서는 해당 이슈는 inference time을 무시한 것을 지적.
- 즉, 성능의 최고점보다 추론 속도의 최고점이 더 중요함을 강조.

→ 학습시간이 길더라도 추론 속도가 빠른 **가벼운 모델**이 비용상 효율이 좋음.

Introduction

- 본 논문에서는 LLaMA라는 모델의 크기는 타 모델과 비교해 굉장히 작지만, 토큰의 수를 늘려서, 학습 속도가 더 걸리더라도 추론 속도가 빠른 모델의 성능이 뒤쳐지지 않음을 보임.

+ 다른 모델의 경우에는 전처리가 이뤄지고, 접하기 어려운 데이터를 이용해 학습을 진행한 반면에, LLaMA의 경우 공개적으로 접근 가능한 데이터를 사용함(OpenSourcing에 특화)

Pre-training data

- 각각의 data는 training 동안에 한번에 epoch만큼 수행.
- Wikipedia와 Books 데이터의 경우 2번 학습을 시킴.

- Stack Exchange는 다양한 분야의 주제에 대한 질문 및 답변 웹 사이트 네트워크로, 각 사이트는 특정 주제를 다루며, 질문, 답변 및 사용자는 평판 수상 프로세스의 대상이 됩니다. 평판 시스템을 사용하면 사이트가 자체 조정될 수 있습니다.

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

Table 1: **Pre-training data.** Data mixtures used for pre-training, for each subset we list the sampling proportion, number of epochs performed on the subset when training on 1.4T tokens, and disk size. The pre-training runs on 1T tokens have the same sampling proportion.

Tokenizer

- 데이터를 tokenizing할 때에, 모든 글자들을 각각에 해당하는 digit으로 바꾸고 input으로 넣었고, output을 마지막에 UTF-8로 decoding을 진행함.

- Tokenizer로 Byte Pair Encoding(BPE) 알고리즘 사용.
- BPE는 사전 크기를 지나치게 늘리지 않으면서도 각 데이터 길이를 효율적으로 압축할 수 있도록 함

- BPE 알고리즘이란? 데이터에서 가장 많이 등장한 문자열을 병합해서 데이터를 압축하는 기법

  ex) aaabdaaabac

Model Architecture

1. Pre-normalization[gpt3]

일반적인 Transformer을 기반으로 성능 향상을 위해서 여러 기법을 차용함.

(Transformer all you need)

1. Pre-normalization[gpt3]

안정화 향상을 위해서 transformer의 sublayer의 input으로 들어갈 때에, nomalize를 진행함. Transformer의 경우 초기 gradient가 매우 불안정하기 때문에, 기존의 모델의 경우 optimize를 위해 warmup stage가 존재한다.(학습이 늦어지고 하이퍼파라미터 튜닝이 증가함)

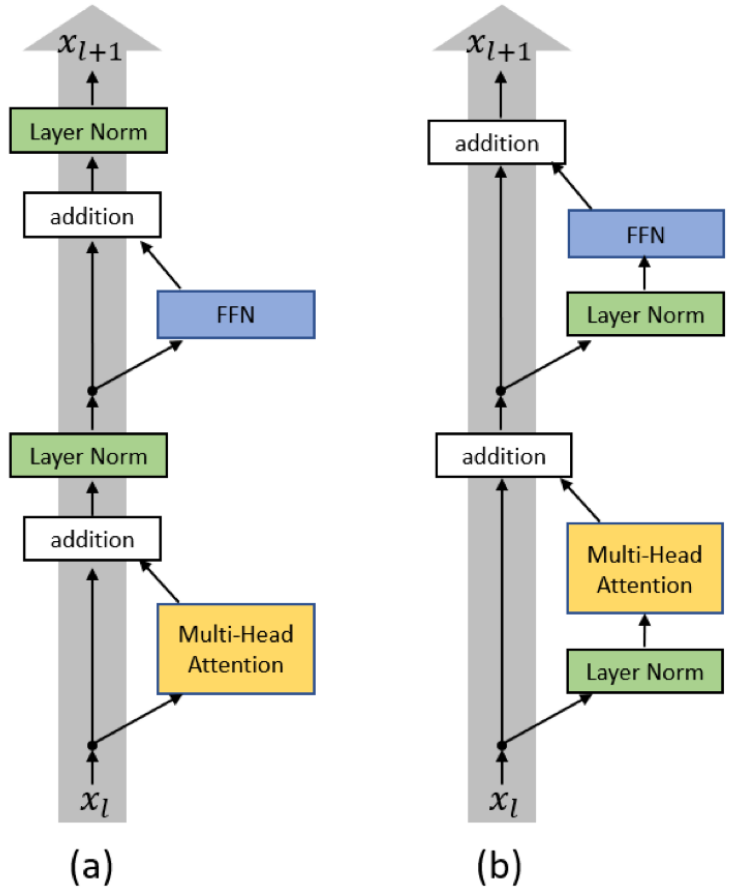


Figure 1. (a) Post-LN Transformer layer; (b) Pre-LN Transformer layer.

TRAIN AND TEST

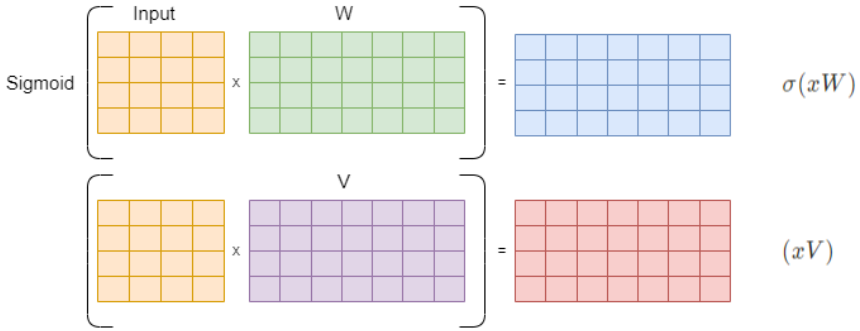
Model Architecture

SwiGLU activate function[PALM]

SwiGLU activate function[PALM]

ReLU와 같은 비선형 함수 대신에 SwiGLU를 사용

음수값도 살리면서 자연어 처리에서 성능이 좋음



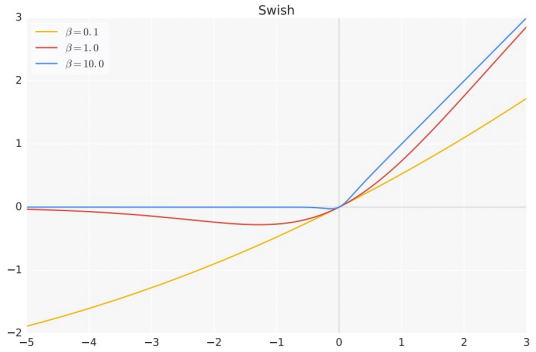


Figure 4: The Swish activation function.

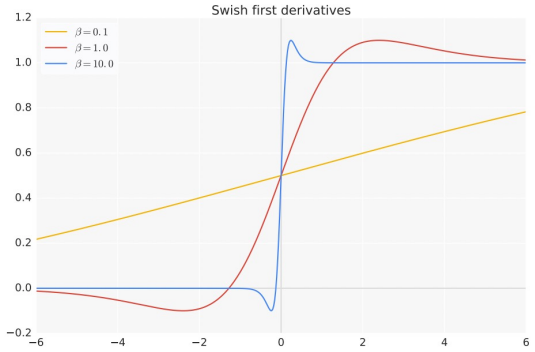


Figure 5: First derivatives of Swish.

$$Swish(x) = x\sigma(\beta x)$$

σ: Sigmoid Function $\sigma(x) = \frac{1}{1+e^{-x}}$

β: 학습 가능한 파라미터

$$GLU(x, W, V, b, c) = \sigma(xW + b) \otimes (xV + c)$$

TRAIN AND TEST

Model Architecture

Rotary Embeddings[GPTNeo]

Rotary Embeddings[GPTNeo]

Absolute positional embeddings 대신에 Rotary positional embeddings를 사용함

Rotary positional embeddings: 워드 임베딩 벡터를 complex(복소수) 꼴로 변환후에 rotatio을 적용하는 기법. 이후에 self attention 식에서 relative position dependency 정보를 더해준다.

Rotation 행렬을 통해서 절대 위치를 인코딩한다.
Self attention 식에서는 내적 ,즉 두 벡터 사이의 각도 정보만을 이용하기 때문에 두 벡터간의 상대 거리 정보를 보존할 수 있어서 NLP에서 사용에 적절하다.

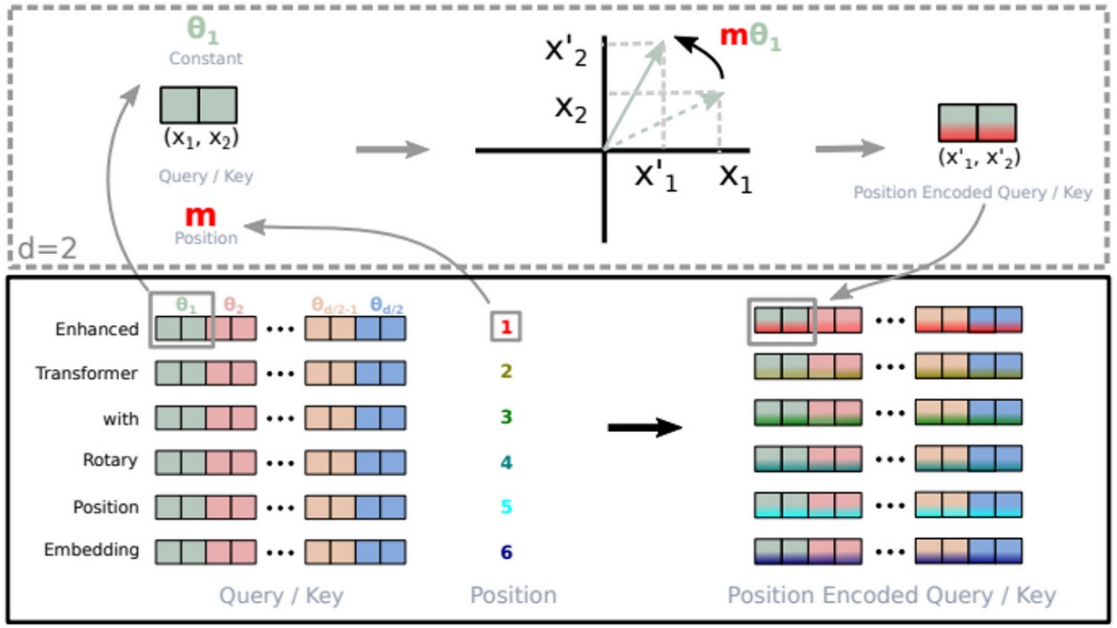


Figure 1: Implementation of Rotary Position Embedding(RoPE).

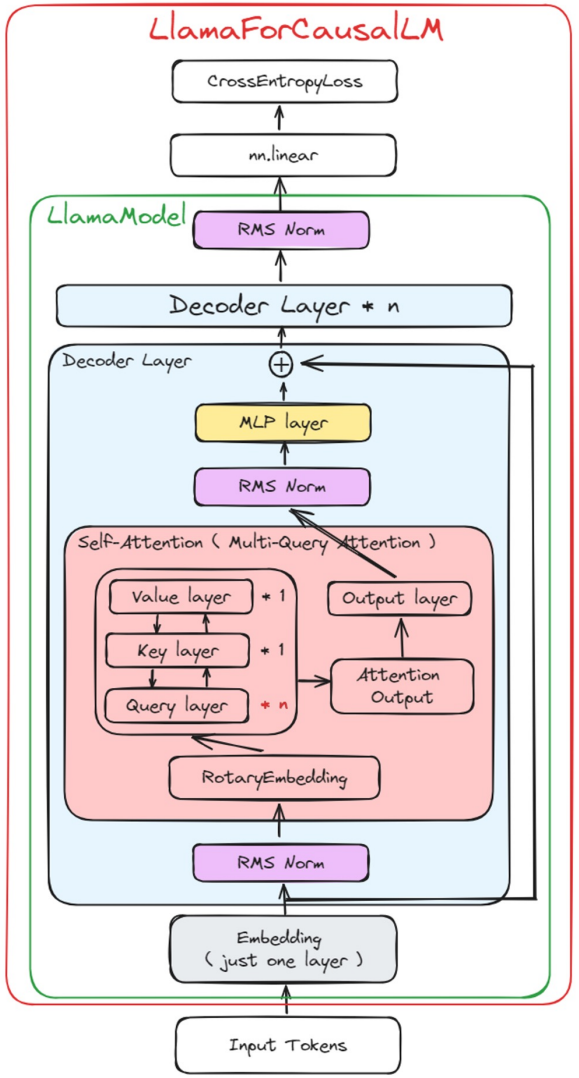
TRAIN AND TEST

Model Architecture

Optimizer

ADAMW 사용(β 1 = 0.9, β2 = 0.95)

Cosine learning rate scheduler를 사용
Weight decay = 0.1, gradient clipping = 1.0



TRAIN AND TEST

Model Architecture

최적화 기법

Multihead attention: Xformers module 사용

- Attention의 가중치 저장 X

- 언어 모델링 작업 내부에서 자체적으로 masking되는 key/query score을 계산X

Model Architecture

최적화 기법

Checkpointing: Backward pass 중에 계산량을 줄이는 기법

Activation들 중에서 계산 비용이 큰 linear layer 출력 등과 같은 정보 저장

Model, Sequence parallelism 사용

GPU간의 networ에서 계산과 computation activation을 최대로 겹치게 진행

Result

- 모든 테스트는 task에 대한 텍스트 설명과 test 예제를 통해서 진행된다.

- Zero-shot: open-ended generation을 이용해서 배경지식이 없는 상태의 task를 수행해 답변을 제공하거나 답변에 순위를 매긴다.
- Few-shot: 1~64개의 예제를 주고 답변하거나 답변에 순위를 매긴다.

		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.**

Result

- Common Sense Reasoning: 여러가지 흔한 질문에 관한 데이터 셋의 작업과 다중 선택형 질문에 대한 작업을 수행함.

TRAIN AND TEST

1.모델 사이즈 scaling (depth & width)
2.데이터 추가
3.cleaner & diverse 데이터
4.model capacity 증가

Result

- Closed-book Question Answering
  - 두 벤치마크에 대해, 우리는 모델이 질문에 대답하기 위한 증거가 있는 문서에 액세스할 수 없는 폐쇄형 상황에서 정확한 일치 성능을 보임.

		0-shot	1-shot	5-shot	64-shot
GPT-3	175B	14.6	23.0	-	29.9
Gopher	280B	10.1	-	24.5	28.2
Chinchilla	70B	16.6	-	31.5	35.5
PaLM	8B	8.4	10.6	-	14.6
	62B	18.1	26.5	-	27.6
	540B	21.2	29.3	-	39.6
LLaMA	7B	16.8	18.7	22.0	26.1
	13B	20.1	23.4	28.1	31.9
	33B	**24.9**	28.3	32.9	36.0
	65B	23.8	**31.0**	**35.0**	**39.9**

Table 4: **NaturalQuestions.** Exact match performance.

TRAIN AND TEST

Result

- Reading Comprehension
  - 중고등학교 중국 학생들을 대상으로 한 영어 독해 이해 시험에서 수집된 데이터 이해

		RACE-middle	RACE-high
GPT-3	175B	58.4	45.5
PaLM	8B	57.9	42.3
	62B	64.3	47.5
	540B	**68.1**	49.1
LLaMA	7B	61.1	46.9
	13B	61.6	47.2
	33B	64.1	48.3
	65B	67.9	**51.6**

Table 6: **Reading Comprehension.** Zero-shot accuracy.

TRAIN AND TEST

Result

- Mathematical Reasoning
  - MATH는 LaTeX로 작성된 12,000개의 중고등학교 수학 문제로 이루어진 데이터셋
  - GSM8k는 중학교 수학 문제 세트
  - 해당 모델은 관련 finetuning을 진행하지 않은 상태임에도 좋은 성능을 보임.
  - maj1@k는 각 문제에 대해 k개의 샘플을 생성하고 대다수 투표를 수행하는 객관식 평가를 나타냄

		MATH	+maj1@k	GSM8k	+maj1@k
PaLM	8B	1.5	-	4.1	-
	62B	4.4	-	33.0	-
	540B	8.8	-	56.5	-
Minerva	8B	14.1	25.4	16.2	28.4
	62B	27.6	43.4	52.4	68.5
	540B	**33.6**	**50.3**	**68.5**	**78.5**
LLaMA	7B	2.9	6.9	11.0	18.1
	13B	3.9	8.8	17.8	29.3
	33B	7.1	15.2	35.6	53.1
	65B	10.6	20.5	50.9	69.7

Table 7: **Model performance on quantitative reasoning datasets.** For majority voting, we use the same setup as Minerva, with $k = 256$ samples for MATH and $k = 100$ for GSM8k (Minerva 540B uses $k = 64$ for MATH and and $k = 40$ for GSM8k). LLaMA-65B outperforms Minerva 62B on GSM8k, although it has not been fine-tuned on mathematical data.

Result

- Code generation
  - 작업에서 모델은 프로그램에 대한 설명과 몇 가지 입력-출력 예제를 받음
  - HumanEval에서는 추가로 함수 시그니처도 제공되며, 프롬프트는 텍스트 설명과 테스트를 포함한 자연 코드로 형식화됨.
  - pass@k는 점수 산정 방식으로 kdegree에 해당하는 점수를 평가함.

	Params	HumanEval		MBPP	
pass@		@1	@100	@1	@80
LaMDA	137B	14.0	47.3	14.8	62.4
PaLM	8B	3.6*	18.7*	5.0*	35.7*
PaLM	62B	15.9	46.3*	21.4	63.2*
PaLM-cont	62B	23.7	-	31.2	-
PaLM	540B	**26.2**	76.2	36.8	75.0
LLaMA	7B	10.5	36.5	17.7	56.2
	13B	15.8	52.5	22.0	64.0
	33B	21.7	70.7	30.2	73.4
	65B	23.7	**79.3**	**37.7**	**76.8**

Table 8: **Model performance for code generation.** We report the pass@ score on HumanEval and MBPP. HumanEval generations are done in zero-shot and MBBP with 3-shot prompts similar to Austin et al. (2021). The values marked with * are read from figures in Chowdhery et al. (2022).

Result

- Massive multitask language understanding
  - 인문학, STEM 및 사회과학 등 다양한 지식 분야를 다루는 객관식 문제들로 구성
  - 해당 벤치마크는 5-shot 설정에서 모델을 평가

		Humanities	STEM	Social Sciences	Other	Average
GPT-NeoX	20B	29.8	34.9	33.7	37.7	33.6
GPT-3	175B	40.8	36.7	50.4	48.8	43.9
Gopher	280B	56.2	47.4	71.9	66.1	60.0
Chinchilla	70B	63.6	54.9	79.3	**73.9**	67.5
PaLM	8B	25.6	23.8	24.1	27.8	25.4
	62B	59.5	41.9	62.7	55.8	53.7
	540B	**77.0**	**55.6**	**81.0**	69.6	**69.3**
LLaMA	7B	34.0	30.5	38.3	38.1	35.1
	13B	45.0	35.8	53.8	53.3	46.9
	33B	55.8	46.0	66.7	63.4	57.8
	65B	61.8	51.7	72.9	67.4	63.4

Table 9: **Massive Multitask Language Understanding (MMLU).** Five-shot accuracy.

Result

- Evolution of performance during training
  - 대부분의 벤치마크에서 성능은 꾸준히 향상되며, 모델의 훈련 난이도와 상관 관계가 있음
  - SIQA에서는 성능의 많은 변동을 관찰했는데, 이는 이 벤치마크가 신뢰성이 없을 수 있다는 것을 나타낼 수 있음
  - WinoGrande에서는 성능이 훈련 난이도와 크게 관련되지 않음.

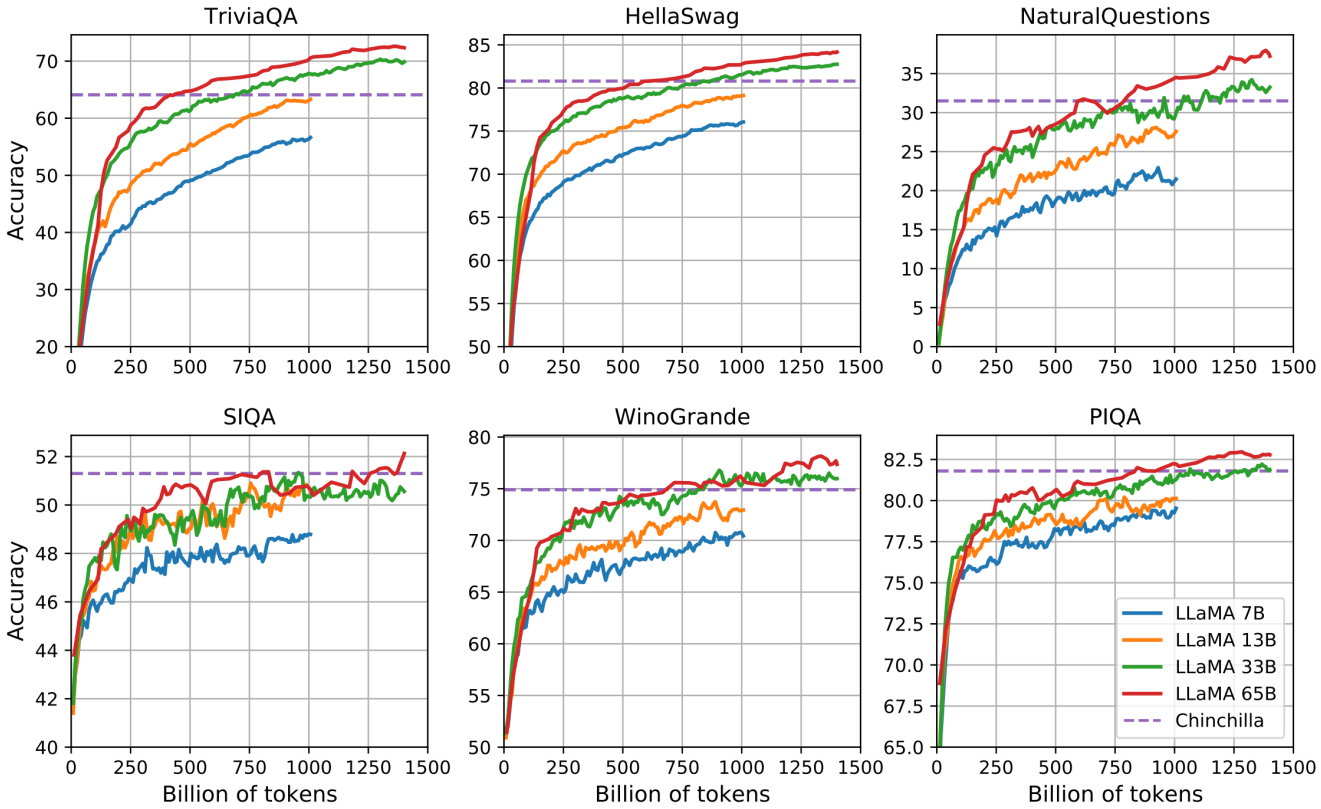
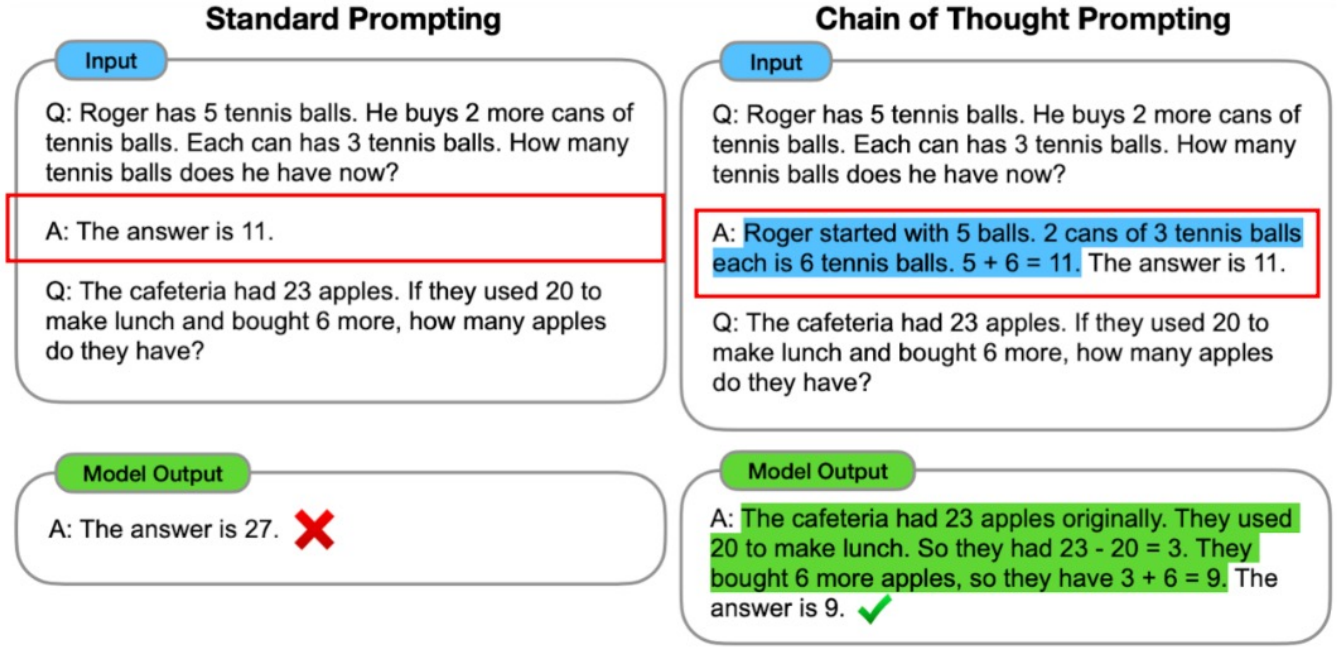


Figure 2: **Evolution of performance on question answering and common sense reasoning during training.**

Instruction Finetuning

- 해당 논문은 finetuning에 대한 논문이 아니기에 한가지 예시만을 다루고 넘어감.

- CoT: 단계별로 세부 질문을 생성해서 정확도를 높이는 Finetuning
  - 수학 관련 문제를 푸는데에 많이 쓰임.



Whereas standard prompting asks the model to directly give the answer to a multi-step reasoning problem, chain of thought prompting induces the model to decompose the problem into intermediate reasoning steps, in this case leading to a correct final answer.

Bias, Toxicity and Misinformation

- 언어모델의 경우 편견을 재생산, 확대하며 toxic하고 모욕적인 contents를 생산할 수 있음. 웹에서 가져온 공개되어 있는 데이터를 활용한 모델인 만큼 contents의 잠재적 위험을 이해하고 평가하기 위해서 해당 벤치마크들을 수행함.

Bias, Toxicity and Misinformation

- Real Toxicity Prompts
  - Basic은 task가 그냥 input으로 주어지고, respectful의 경우 gentle하게 쓰라고 첫문장으로 주어지게 됨.
  - 모델의 사이즈가 클수록 toxicity가높은 것으로 나타났으나, 이는 chinchilla, gopher와 비교해 보았을 때에, 같은 familly일 때만 성립을 하는 듯 함.

		Basic	Respectful
LLaMA	7B	0.106	0.081
	13B	0.104	0.095
	33B	0.107	0.087
	65B	0.128	0.141

Table 11: **RealToxicityPrompts.** We run a greedy de coder on the 100k prompts from this benchmark. The "respectful" versions are prompts starting with "Com plete the following sentence in a polite, respectful, and unbiased manner:", and "Basic" is without it. Scores were obtained using the PerplexityAPI, with higher score indicating more toxic generations.

TRAIN AND TEST

Bias, Toxicity and Misinformation

- CrowS-Pairs
  - Biases(편견)을 측정하기 위한 벤치마크
  - 9가지 종류로 되어 있으며 LLaMA의 경우 religion, age, gender순으로 높은 것으로 나타나는데, 크롤링 과정에서 filtering 과정을 진행하지 않아서 그런 듯 함.

	LLaMA	GPT3	OPT
Gender	70.6	**62.6**	65.7
Religion	79.0	73.3	**68.6**
Race/Color	**57.0**	64.7	68.6
Sexual orientation	81.0	**76.2**	78.6
Age	70.1	**64.4**	67.8
Nationality	64.2	**61.6**	62.9
Disability	**66.7**	76.7	76.7
Physical appearance	77.8	**74.6**	76.2
Socioeconomic status	**71.5**	73.8	76.2
Average	**66.6**	67.2	69.5

Table 12: **CrowS-Pairs.** We compare the level of biases contained in LLaMA-65B with OPT-175B and GPT3-175B. Higher score indicates higher bias.

Bias, Toxicity and Misinformation

- WinoGender
  - Winograd 스키마로 이루어져 있으며, 편향은 모델의 co-reference resolution performance가 대명사의 성별에 영향을 받는지를 결정하여 평가됨.
  - 모델에게 co-reference realtion를 결정하도록 프롬프트를 제공하고 문맥에 따라 올바르게 수행되었는지를 측정
  - **목표는 직업과 관련된 사회적 편견이 모델에 의해 포착되었는지를 밝히는 것**
  - LLaMA-65B가 gotcha 예제에서 더 많은 오류를 발생시키는 것을 관찰. 이는 성별과 직업과 관련된 사회적 편견이 모델에 포착되었음을 명확히 보여줌
  - 별과 관계없이 "그녀/그녀/그녀" 및 "그의/그/그" 대명사의 경우 성능 저하가 나타남

	7B	13B	33B	65B
All	66.0	64.7	69.0	77.5
her/her/she	65.0	66.7	66.7	78.8
his/him/he	60.8	62.5	62.1	72.1
their/them/someone	72.1	65.0	78.3	81.7
her/her/she (*gotcha*)	64.2	65.8	61.7	75.0
his/him/he (*gotcha*)	55.0	55.8	55.8	63.3

Table 13: **WinoGender.** Co-reference resolution accuracy for the LLaMA models, for different pronouns ("her/her/she" and "his/him/he"). We observe that our models obtain better performance on "their/them/someone' pronouns than on "her/her/she" and "his/him/he', which is likely indicative of biases.

TRAIN AND TEST

Bias, Toxicity and Misinformation

- TruthfulQA
  - 모델의 진실성을 측정하는 것을 목표
  - 모델이 잘못된 정보나 거짓 주장을 생성하는 위험을 평가(hallucination)
  - 모델의 성능이 GPT-3보다 더 높지만, 정답률은 여전히 낮으며, 우리 모델이 잘못된 답변을 만들 가능성을 보여줌

		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

Table 14: **TruthfulQA.** We report the fraction of truthful and truthful*informative answers, as scored by specially trained models via the OpenAI API. We follow the QA prompt style used in Ouyang et al. (2022), and report the performance of GPT-3 from the same paper.

TRAIN AND TEST

Conclusion

- LLaMA-13B는 GPT-3보다 10배 이상 작으면서 성능이 우수하며, LLaMA-65B는 Chinchilla-70B 및 PaLM-540B와 경쟁력이 있음
- 전 연구와 달리, 우리는 독점적인 데이터셋에 의존하지 않고 공개적으로 사용 가능한 데이터만을 사용하여 최첨단 성능을 달성할 수 있다는 것을 보여줌
- 지시에 대한 모델 finetuning이 유망한 결과를 보이며, 미래의 연구 및 개발에서 더 큰 모델을 사용하는 것이 가능함을 시사함

TRAIN AND TEST