Llama 1 모델 리뷰

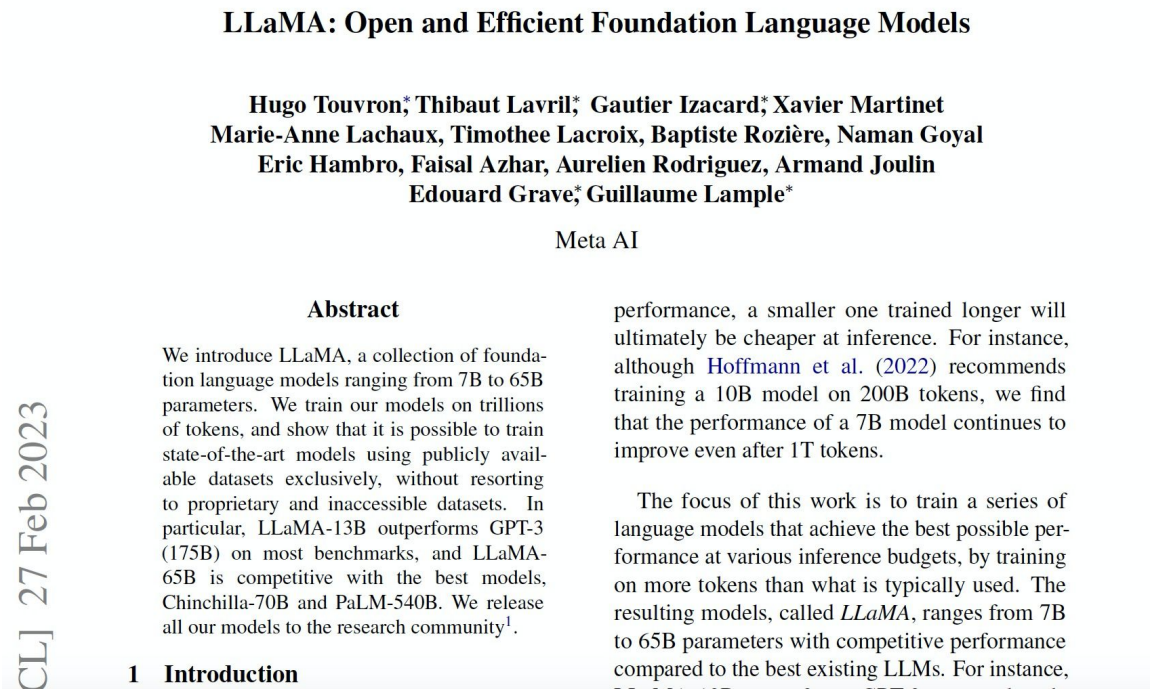
에이아이스쿨(AISchool) 대표
양진호 (솔라리스)
http://aischool.ai
http://solarisailab.com

Llama 2란?

- Llama 2는 Meta(Facebook)에서 무료로 공개한 연구와 상업적 용도로 활용할 수 있는 LLM(Large Language Model) 입니다.
- 현재 강의 촬영 시점기준(2023년 9월) 오픈소스로 공개된 LLM 중 가장 강력한 성능을 가지고 있습니다.
- Llama 2 모델을 리뷰하기에 앞서서 Llama 2 모델의 베이스가 된 Llama 1 모델에 대해 살펴봅시다.

Llama 1 Paper

- Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." arXiv preprint arXiv:2302.13971 (2023).
- https://arxiv.org/abs/2302.13971



LLaMA: Open and Efficient Foundation Language Models

Hugo Touvron*, Thibaut Lavril*, Gautier Izacard*, Xavier Martinet
Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal
Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin
Edouard Grave*, Guillaume Lample*

Meta AI

Abstract

We introduce LLaMA, a collection of foundation language models ranging from 7B to 65B parameters. We train our models on trillions of tokens, and show that it is possible to train state-of-the-art models using publicly available datasets exclusively, without resorting to proprietary and inaccessible datasets. In particular, LLaMA-13B outperforms GPT-3 (175B) on most benchmarks, and LLaMA-65B is competitive with the best models, Chinchilla-70B and PaLM-540B. We release all our models to the research community[1].

1 Introduction

performance, a smaller one trained longer will ultimately be cheaper at inference. For instance, although Hoffmann et al. (2022) recommends training a 10B model on 200B tokens, we find that the performance of a 7B model continues to improve even after 1T tokens.

The focus of this work is to train a series of language models that achieve the best possible performance at various inference budgets, by training on more tokens than what is typically used. The resulting models, called *LLaMA*, ranges from 7B to 65B parameters with competitive performance compared to the best existing LLMs. For instance,

Llama 1 Paper

- Llama 1 Paper를 같이 살펴보면서 Llama 1 모델의 디테일한 내용들을 살펴봅시다!

Abstract

- 우리는 LLaMA라는 파운데이션 모델을 소개합니다. 이 모델은 7B에서 65B까지의 파라미터를 가진 기초 언어 모델들로 구성되어 있습니다.
- 우리는 이 모델들을 수조 개의 토큰에 대해 훈련시키며, 공개적으로 이용 가능한 데이터셋만을 사용하여도 최첨단 모델을 훈련시킬 수 있다는 것을 보여줍니다. 이러한 방식은 독점적이고 접근하기 어려운 데이터셋에 의존하지 않습니다.
- 특히, LLaMA-13B는 대부분의 벤치마크에서 GPT-3 (175B)를 능가하며, LLaMA-65B는 최고의 모델인 Chinchilla-70B와 PaLM-540B와 경쟁력이 있습니다.
- 우리는 모든 모델을 연구 커뮤니티에 공개합니다.

1. Introduction

- 대규모 텍스트 데이터셋에서 훈련된 대형 언어 모델(Large Language Models, LLMs)은 텍스트 지시문 또는 소수의 예시를 통해 새로운 작업을 수행할 수 있는 능력을 보여주었습니다 (Brown 등, 2020).
- 이러한 소수의 예시를 통한 학습 능력은 모델을 충분한 크기로 확장할 때 처음으로 나타났으며 (Kaplan 등, 2020), 이로 인해 이러한 모델을 더욱 확장하는 연구들이 진행되고 있습니다 (Chowdhery 등, 2022; Rae 등, 2021).
- 이러한 노력들은 더 많은 파라미터가 더 나은 성능을 이끌어낼 것이라는 가정에 기반하고 있습니다.
- 그러나 최근 Hoffmann 등(2022)의 연구에서는 주어진 컴퓨팅 예산으로 최상의 성능을 달성하는 것은 가장 큰 모델이 아니라 더 많은 데이터로 훈련된 더 작은 모델임을 보여주고 있습니다.

1. Introduction

- Hoffmann 등(2022)의 스케일링 법칙의 목적은 특정 훈련 컴퓨팅 예산에 대해 데이터셋과 모델 크기를 어떻게 가장 효과적으로 확장할지 결정하는 것입니다.
- 그러나 이 목적은 추론 예산을 무시하는데, 이는 대규모로 언어 모델을 서빙할 때 중요해집니다. 이러한 맥락에서 목표 성능 수준이 주어진 경우, 선호되는 모델은 훈련이 가장 빠른 모델이 아니라 추론이 가장 빠른 모델입니다.
- 특정 성능 수준에 도달하기 위해 큰 모델을 훈련하는 것이 더 저렴할 수 있지만, 더 오래 훈련된 작은 모델은 결국 추론에서 더 저렴할 것입니다.
- 예를 들어, Hoffmann 등(2022)은 2000억 토큰에 대해 100억 크기의 모델을 훈련할 것을 권장하지만, 우리는 70억 크기의 모델의 성능이 1조 토큰 이후에도 계속 향상된다는 것을 발견했습니다.

1. Introduction

- 이 연구의 중점은 일반적으로 사용되는 것보다 더 많은 토큰에 대해 훈련하여 다양한 추론 예산에서 최상의 성능을 달성하는 일련의 언어 모델을 훈련하는 것입니다.
- 결과적으로 나온 모델들은 LLaMA라고 부르며, 이는 70억(7B)에서 650억(65B)까지의 파라미터를 가지고 있고, 기존의 최고의 LLMs와 경쟁력 있는 성능을 보입니다.
- 예를 들어, LLaMA-13B는 크기가 GPT-3보다 10배 작음에도 불구하고 대부분의 벤치마크에서 GPT-3를 능가합니다.
- 이 모델은 단일 GPU에서 실행할 수 있으므로 LLMs의 접근성과 연구를 더욱 민주화할 것이라고 믿습니다.
- 더 높은 범위의 규모에서는, 우리의 650억 파라미터 모델도 Chinchilla나 PaLM-540B와 같은 최고의 대형 언어 모델과 경쟁력이 있습니다.

1. Introduction

- Chinchilla, PaLM, 또는 GPT-3와 달리, 우리는 오직 공개적으로 이용 가능한 데이터만을 사용하므로, 우리의 작업은 오픈 소스와 호환됩니다. 반면에 대부분의 기존 모델들은 공개되지 않았거나 문서화되지 않은 데이터에 의존하고 있습니다(예: "Books - 2TB" 또는 "소셜 미디어 대화").
- 몇 가지 예외가 존재하긴 하나, 대표적으로 OPT(Zhang 등, 2022), GPT-NeoX(Black 등, 2022), BLOOM(Scao 등, 2022), 그리고 GLM(Zeng 등, 2022)이 있지만, 이들 중 어느 것도 PaLM-62B나 Chinchilla와 경쟁력을 가지고 있지는 않습니다.

2. Approach

- 우리의 훈련 접근법은 이전 연구(Brown 등, 2020; Chowdhery 등, 2022)에서 설명된 방법과 유사하며, Chinchilla 스케일링 법칙(Hoffmann 등, 2022)에 영감을 받았습니다.
- 우리는 표준 Optimizer를 사용하여 대량의 텍스트 데이터에 대한 큰 트랜스포머를 훈련시킵니다.

2.1. Pre-training Data

- 우리의 훈련 데이터셋은 다양한 도메인을 포괄하는 여러 출처의 혼합물이며, 이는 표 1에 기록되어 있습니다.
- 대부분의 경우, 우리는 다른 대형 언어 모델(LLMs)을 훈련시키는 데 사용된 데이터 소스를 재사용합니다.
- 단, 공개적으로 사용 가능하고 오픈 소싱과 호환되는 데이터만을 사용하는 제한이 있습니다.
- 이로 인해 훈련 세트에서 그들이 차지하는 비율과 함께 다음과 같은 데이터의 혼합물이 생깁니다:

2.1. Pre-training Data

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.

2.1. Pre-training Data

- English CommonCrawl [67%]. 우리는 2017년부터 2020년까지의 다섯 개의 CommonCrawl 덤프를 CCNet 파이프라인 (Wenzek et al., 2020)을 이용해 전처리합니다.
- 이 과정에서는 데이터를 라인 수준에서 중복 제거하고, fastText 선형 분류기를 이용해 언어를 식별하여 영어가 아닌 페이지를 제거하며, n-gram 언어 모델을 이용해 저품질의 콘텐츠를 필터링합니다.
- 더불어, 위키피디아를 참고로 사용되는 페이지와 무작위로 샘플링된 페이지를 분류하는 선형 모델을 훈련시켰고, 참고로 분류되지 않은 페이지는 제외했습니다.

2.1. Pre-training Data

- C4 [15%]. 탐색적 실험 중에 다양한 사전 처리된 CommonCrawl 데이터셋을 사용하면 성능이 향상된다는 것을 확인했습니다.
- 따라서 우리는 공개적으로 이용 가능한 C4 데이터셋 (Raffel et al., 2020)을 우리의 데이터에 포함시켰습니다. C4의 전처리 과정에도 중복 제거와 언어 식별 단계가 포함되어 있습니다:
- CCNet과의 주요 차이점은 품질 필터링에 있으며, 이는 대부분 구두점의 존재나 웹페이지에 있는 단어와 문장의 수와 같은 휴리스틱에 의존합니다.

2.1. Pre-training Data

- Github [4.5%]. 우리는 Google BigQuery에서 이용 가능한 공개 GitHub 데이터셋을 사용했습니다. 우리는 Apache, BSD 및 MIT 라이선스 하에 배포되는 프로젝트만을 유지했습니다.
- 또한, 휴리스틱을 기반으로 라인 길이나 영숫자 문자의 비율에 따라 저품질 파일을 필터링했습니다.
- 그리고 정규 표현식을 사용하여 헤더와 같은 보일러플레이트를 제거했습니다.
- 마지막으로, 파일 수준에서 정확한 일치를 기준으로 결과 데이터셋을 중복 제거했습니다.

2.1. Pre-training Data

- Wikipedia [4.5%]. 우리는 2022년 6월부터 8월까지의 기간 동안의 위키백과 덤프를 추가했습니다.
- 이 데이터는 라틴어 또는 키릴 문자를 사용하는 20개의 언어를 포함하고 있습니다: bg, ca, cs, da, de, en, es, fr, hr, hu, it, nl, pl, pt, ro, ru, sl, sr, sv, uk.
- 우리는 하이퍼링크, 댓글, 그리고 다른 형식의 보일러플레이트를 제거하기 위해 데이터를 처리했습니다.

2.1. Pre-training Data

- Gutenberg and Books3 [4.5%]. 우리는 훈련 데이터셋에 두 개의 책 말뭉치를 포함시켰습니다:
- 공공 도메인에 있는 책들을 담고 있는 Gutenberg Project와 대규모 언어 모델 훈련을 위한 공개적으로 이용 가능한 데이터셋인 ThePile의 Books3 섹션 (Gao et al., 2020).
- 우리는 책 레벨에서 중복을 제거하며, 90% 이상의 내용이 중복되는 책들을 제거했습니다.

2.1. Pre-training Data

- ArXiv [2.5%]. 우리는 ArXiv의 Latex 파일을 처리하여 우리의 데이터셋에 과학적 데이터를 추가했습니다.
- Lewkowycz et al. (2022)를 따라 첫 번째 섹션 이전의 모든 것과 참고문헌을 제거했습니다.
- 또한 .tex 파일에서 주석을 제거하고, 논문 간의 일관성을 높이기 위해 사용자가 작성한 정의와 매크로를 인라인으로 확장했습니다.

2.1. Pre-training Data

- Stack Exchange [2%]. 우리는 컴퓨터 과학부터 화학에 이르기까지 다양한 분야를 다루는 고품질의 질문과 답변이 있는 웹사이트인 Stack Exchange의 덤프를 포함했습니다.
- 우리는 가장 큰 28개의 웹사이트로부터 데이터를 보존하고, 텍스트에서 HTML 태그를 제거했습니다. 또한 답변을 점수가 높은 순서로 정렬했습니다.

2.1. Pre-training Data

- Tokenizer. 우리는 데이터를 바이트 페어 인코딩 (BPE) 알고리즘 (Sennrich 등, 2015)을 사용하여 토큰화했습니다.
- 이때 SentencePiece (Kudo and Richardson, 2018)의 구현을 사용했습니다.
- 특히, 모든 숫자를 개별 숫자로 분리하고, 알려지지 않은 UTF-8 문자를 분해하기 위해 바이트를 대체로 사용했습니다.

2.1. Pre-training Data

- 전반적으로, 토큰화 후에 우리의 전체 훈련 데이터셋은 대략 1.4T(1.4조개)의 토큰을 포함하고 있습니다.
- 대부분의 훈련 데이터에서 각 토큰은 훈련 중 한 번만 사용됩니다.
- 단, 위키피디아와 Books 도메인은 예외로, 이들에 대해서는 대략 두 번의 에포크(epoch)를 수행합니다.

2.2. Architecture

- 최근 대규모 언어 모델에 대한 연구를 따라, 우리의 네트워크는 트랜스포머 아키텍처(Vaswani et al., 2017)를 기반으로 하고 있습니다.
- 우리는 이후에 제안된 다양한 개선사항을 활용하고, PaLM과 같은 다른 모델에서 사용된 것을 참고하였습니다.
- 다음은 원래 아키텍처와 주요한 차이점과, 이러한 변경에 대한 영감을 얻은 출처를 괄호 안에 표시하였습니다:

2.2 Architecture

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.**

2.3. Optimizer

- 우리의 모델은 AdamW 옵티마이저 (Loshchilov and Hutter, 2017)를 사용하여 훈련되며, 다음과 같은 하이퍼파라미터를 사용합니다: $\beta_1=0.9$, $\beta_2=0.95$.
- 우리는 코사인 학습률 일정을 사용하여, 최종 학습률이 최대 학습률의 10%가 되도록 합니다.
- 또한 가중치 감쇠로 0.1을, 그래디언트 클리핑으로 1.0을 사용합니다.
- 우리는 2,000번의 웜업 스텝을 사용하며, 모델의 크기에 따라 학습률과 배치 크기를 달리합니다 (자세한 내용은 표 2를 참조하세요).

2.4. Efficient implementation

- 65B(650억) 파라미터 모델을 훈련할 때, 우리의 코드는 80GB의 RAM을 가진 2048개의 A100 GPU에서 초당 약 380 토큰을 처리합니다.
- 이는 우리의 데이터셋에 1.4T(1.4조개) 토큰이 포함되어 있을 경우, 훈련이 대략 21일이 걸린다는 것을 의미합니다.

3. Main results

- 이전 연구(Brown 등, 2020)를 따라, 우리는 제로샷(zero-shot)과 퓨샷(few-shot) 작업을 고려하고 총 20개의 벤치마크에서 결과를 보고합니다:
- 제로샷(Zero-shot) : 작업의 텍스트 설명과 테스트 예제를 제공합니다. 모델은 열린 형태의 생성(open-ended generation)을 사용하여 답을 제공하거나, 제안된 답안을 순위로 나열합니다.
- 퓨샷(Few-shot) : 작업의 몇 가지 예제(1~64개 사이)와 테스트 예제를 제공합니다. 모델은 이 텍스트를 입력으로 받아 답을 생성하거나 다양한 옵션을 순위로 나열합니다.

3. Main results

- 우리는 LLaMA와 다른 기반 모델, 즉 공개되지 않은 언어 모델인 GPT-3 (Brown et al., 2020), Gopher (Rae et al., 2021), Chinchilla (Hoffmann et al., 2022), 그리고 PaLM (Chowdhery et al., 2022)을 비교합니다.
- 또한 오픈소스된 OPT 모델 (Zhang et al., 2022), GPT-J (Wang and Komatsuzaki, 2021), 그리고 GPTNeo (Black et al., 2022)와도 비교합니다.
- 섹션 4에서는 또한 LLaMA와 명령어 튜닝(instruction-tuned) 모델인 OPT-IML (Iyer et al., 2022) 및 Flan-PaLM (Chung et al., 2022)과도 간략하게 비교합니다.

3. Main results

- 우리는 LLaMA를 자유 형식 생성(free-form generation) 작업과 객관식(multiple choice) 작업에서 평가합니다.
- 객관식 작업에서의 목표는 주어진 문맥에 기반하여 주어진 옵션 중에서 가장 적절한 완성을 선택하는 것입니다.

3.1. Common Sense Reasoning

- 우리는 8가지 표준 상식 추론 벤치마크를 고려합니다: BoolQ (Clark et al., 2019), PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2021), ARC easy와 challenge (Clark et al., 2018), 그리고 OpenBookQA (Mihaylov et al., 2018).
- 이러한 데이터셋은 Cloze와 Winograd 스타일의 작업과 더불어 객관식 질문 응답을 포함합니다. 우리는 언어 모델링 커뮤니티에서 행한 것과 같이 제로샷 설정에서 평가합니다.
- 표 3에서는 다양한 크기의 기존 모델과 비교하고 해당 논문에서 나온 숫자를 보고합니다.
- 먼저, LLaMA-65B는 모든 보고된 벤치마크에서 Chinchilla-70B를 능가합니다만 BoolQ에서는 그렇지 않습니다. 마찬가지로, 이 모델은 BoolQ와 WinoGrande를 제외한 모든 곳에서 PaLM-540B를 능가합니다.
- 또한, LLaMA-13B 모델은 크기가 10배 작음에도 불구하고 대부분의 벤치마크에서 GPT-3를 능가합니다.

3.1. Common Sense Reasoning

		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.**

3.2. Closed-book Question Answering

- 우리는 두 가지 클로즈드-북 질문 응답 벤치마크에서 LLaMA와 기존의 대규모 언어 모델과 비교합니다: Natural Questions (Kwiatkowski et al., 2019)과 TriviaQA (Joshi et al., 2017).
- 두 벤치마크 모두에서, 우리는 모델이 질문에 답할 증거를 포함한 문서에 접근할 수 없는 클로즈드-북 설정에서 정확한 일치 성능을 보고합니다.
- 표 4에서는 Natural Questions에서의 성능을, 표 5에서는 TriviaQA에서의 성능을 보고합니다.
- 두 벤치마크 모두에서, LLaMA-65B는 제로샷과 퓨샷 설정에서 최첨단 성능을 달성합니다.
- 더 중요한 것은, LLaMA-13B도 GPT-3와 Chinchilla와 경쟁력 있는 성능을 보이고 있으며, 그 크기가 5-10배 작음에도 불구하고입니다.
- 이 모델은 추론 도중 단일 V100 GPU에서 실행됩니다.

3.2. Closed-book Question Answering

		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.

3.2. Closed-book Question Answering

		0-shot	1-shot	5-shot	64-shot
Gopher	280B	43.5	-	57.0	57.2
Chinchilla	70B	55.4	-	64.1	64.6
LLaMA	7B	50.0	53.4	56.3	57.6
	13B	56.6	60.5	63.1	64.0
	33B	65.1	67.9	69.9	70.4
	65B	**68.2**	**71.6**	**72.6**	**73.0**

Table 5: **TriviaQA.** Zero-shot and few-shot exact match performance on the filtered dev set.

3.3 Reading Comprehension

- 우리는 우리의 모델을 RACE 읽기 이해 능력 벤치마크(Lai 등, 2017)에서 평가합니다. 이 데이터셋은 중국 중고등학생들을 대상으로 한 영어 읽기 이해 시험에서 수집되었습니다.
- 우리는 Brown 등(2020)의 평가 설정을 따르고 Table 6에서 결과를 보고합니다. 이 벤치마크에서, LLaMA-65B는 PaLM-540B와 경쟁력이 있으며, LLaMA-13B는 몇 퍼센트로 GPT-3를 능가합니다.

3.3 Reading Comprehension

		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.

3.4. Mathematical reasoning

- 우리는 우리의 모델을 두 가지 수학적 추론 벤치마크인 MATH(Hendrycks 등, 2021)와 GSM8k(Cobbe 등, 2021)에서 평가합니다.
- MATH는 LaTeX로 작성된 1.2만개의 중고등학교 수학 문제로 구성된 데이터셋입니다.
- GSM8k는 중학교 수학 문제의 세트입니다.
- 표 7에서는 PaLM과 Minerva(Lewkowycz 등, 2022)와 비교합니다. Minerva는 ArXiv와 수학 웹 페이지에서 추출한 385B 토큰으로 미세조정된 일련의 PaLM 모델이며, 반면에 PaLM이나 LLaMA는 수학 데이터로 미세조정되지 않았습니다.
- PaLM과 Minerva에 대한 숫자는 Lewkowycz 등(2022)에서 가져왔으며, maj1@k의 유무에 따라 비교합니다. maj1@k는 각 문제에 대해 k개의 샘플을 생성하고 다수결 투표를 수행하는 평가를 나타냅니다(Wang 등, 2022).
- GSM8k에서는 LLaMA-65B가 수학 데이터에 미세조정되지 않았음에도 불구하고 Minerva-62B를 능가한다는 것을 확인했습니다.

3.4. Mathematical reasoning

		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.

3.5. Code generation

- 우리는 두 개의 벤치마크인 HumanEval (Chen 등, 2021)과 MBPP (Austin 등, 2021)에서 우리 모델이 자연어 설명으로부터 코드를 작성하는 능력을 평가합니다.
- 두 작업 모두에서 모델은 프로그램의 설명을 몇 문장으로, 그리고 몇 가지 입력-출력 예시로 받습니다.
- HumanEval에서는 함수 시그니처도 받으며, 프롬프트는 텍스트 설명과 테스트 케이스가 docstring에 포함된 자연스러운 코드 형식으로 구성됩니다.
- 모델은 설명에 부합하고 테스트 케이스를 만족하는 파이썬 프로그램을 생성해야 합니다.
- 표 8에서는 코드에 미세조정되지 않은 기존의 언어 모델, 즉 PaLM과 LaMDA(Thoppilan 등, 2022)와 우리 모델의 pass@1 점수를 비교합니다. PaLM과 LLaMA은 코드 토큰의 유사한 수를 포함하는 데이터셋에서 훈련되었습니다.

3.5. Code generation

	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).

3.5. Code generation

- 표 8에서 볼 수 있듯이, 유사한 수의 매개변수를 가지고 있는 경우에도 LLaMA는 코드에 특별히 훈련되거나 미세조정되지 않은 다른 일반 모델들인 LaMDA와 PaLM보다 더 높은 성능을 보입니다.
- 13B 매개변수 이상을 가진 LLaMA는 HumanEval과 MBPP에서 모두 LaMDA 137B를 능가합니다. 또한, LLaMA 65B는 더 오래 훈련되었더라도 PaLM 62B보다 더 좋은 성능을 보입니다.
- 이 표에서 보고된 pass@1 결과는 온도 0.1로 샘플링하여 얻은 것입니다. pass@100과 pass@80 지표는 온도 0.8로 얻었습니다.
- 우리는 Chen 등(2021)이 사용한 동일한 방법을 이용하여 pass@k의 편향되지 않은 추정값을 얻었습니다.
- 코드에 특정 토큰에 미세조정을 함으로써 성능을 향상시킬 수 있습니다. 예를 들어, PaLM-Coder(Chowdhery 등, 2022)는 HumanEval에서 PaLM의 pass@1 점수를 26.2%에서 36%로 높였습니다. 코드에 특별히 훈련된 다른 모델들도 이러한 작업에서 일반 모델보다 더 나은 성능을 보입니다(Chen 등, 2021; Nijkamp 등, 2022; Fried 등, 2022). 코드 토큰에 대한 미세조정은 이 논문의 범위를 벗어납니다.

3.6 Massive Multitask Language Understanding

- 대규모 다작업 언어 이해 벤치마크(MMLU), 즉 Hendrycks 등(2020)에 의해 소개된 MMLU는 인문학, STEM, 그리고 사회과학을 포함한 다양한 지식 분야의 객관식 문제로 구성되어 있습니다.
- 우리는 벤치마크에서 제공하는 예제를 사용하여 5-shot 설정에서 모델을 평가하고, 표 9에서 결과를 보고합니다.
- 이 벤치마크에서 LLaMA-65B는 평균적으로, 그리고 대부분의 도메인에서 Chinchilla-70B와 PaLM-540B 뒤에 몇 퍼센트 차이로 뒤따르고 있습니다.
- 이러한 현상의 가능한 설명은 우리가 사전 훈련 데이터로 ArXiv, Gutenberg, 그리고 Books3와 같은 제한된 양의 책과 학술 논문만 사용했기 때문일 수 있습니다.
- 이 데이터는 총합해도 177GB에 불과하며, 이러한 모델들은 최대 2TB의 책 데이터로 훈련되었습니다. Gopher, Chinchilla, 그리고 PaLM이 사용한 이런 대량의 책 데이터가 Gopher가 이 벤치마크에서 GPT-3를 능가하는 이유일 수 있으며, 다른 벤치마크에서는 비슷한 성능을 보이는 것으로 나타났습니다.

3.6 Massive Multitask Language Understanding

		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.

3.7 Evolution of performance during training

- 훈련 중에는 몇몇 질문 응답과 상식 벤치마크에서 모델의 성능을 추적하고 그 결과를 그림 2에 보고했습니다.
- 대부분의 벤치마크에서 성능은 꾸준히 개선되며, 모델의 훈련 perplexity와 상관관계가 있습니다(그림 1 참조).
- 예외적인 경우는 SIQA와 WinoGrande입니다. 특히 SIQA에서는 성능에 많은 변동성을 관찰했는데, 이는 이 벤치마크가 신뢰할 수 없을 수 있음을 나타낼 수 있습니다.
- WinoGrande에서는 성능이 훈련 perplexity와 잘 상관되지 않습니다: LLaMA-33B와 LLaMA-65B는 훈련 도중 유사한 성능을 보였습니다.

3.7 Evolution of performance during training

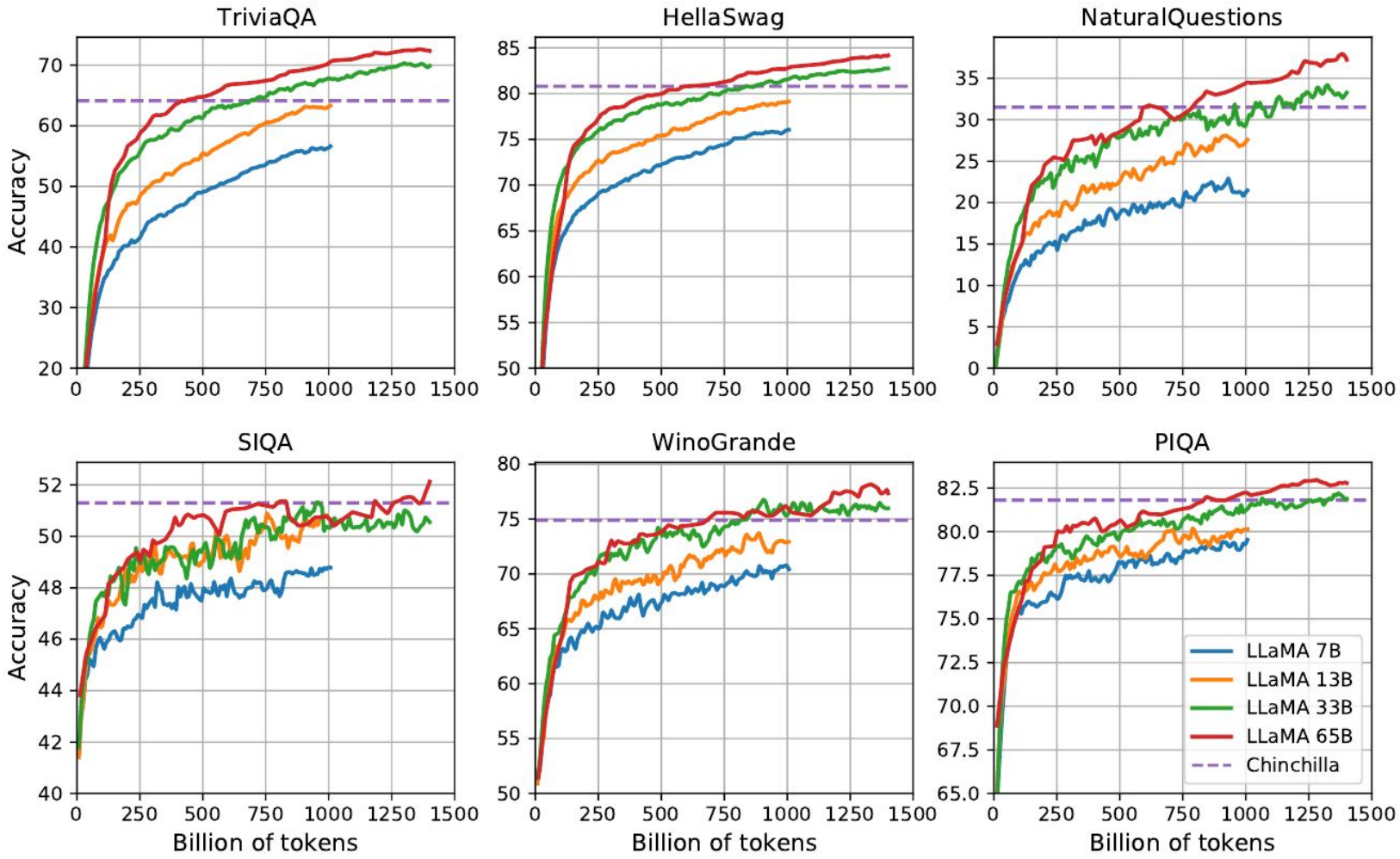


Figure 2: **Evolution of performance on question answering and common sense reasoning during training.**

5. Bias, Toxicity and Misinformation

- 대규모 언어 모델은 훈련 데이터에 존재하는 편향을 재현하고 확대하는 경향이 있음이 여러 연구(Sheng et al., 2019; Kurita et al., 2019)에서 보여졌으며, 독성이나 불쾌한 내용을 생성할 수도 있습니다(Gehman et al., 2020).
- 우리의 훈련 데이터셋은 웹에서 얻은 데이터의 큰 부분을 차지하고 있기 때문에, 우리 모델이 이러한 내용을 생성할 가능성을 파악하는 것이 중요하다고 생각합니다.
- LLaMA-65B의 잠재적인 위험을 이해하기 위해, 독성 내용 생성과 스테레오타입 감지를 측정하는 다양한 벤치마크에서 평가를 진행합니다. 언어 모델 커뮤니티에서 일반적으로 사용되는 몇몇 표준 벤치마크를 선택했지만, 이러한 평가만으로는 이 모델들과 관련된 위험을 완전히 이해하기에는 충분하지 않습니다.

5.1. RealToxicityPrompts

- 언어 모델은 독성을 가진 언어를 생성할 수 있으며, 예를 들어 모욕, 혐오 발언 또는 위협과 같은 내용이 포함될 수 있습니다. 모델이 생성할 수 있는 독성 있는 내용의 범위는 매우 넓어, 철저한 평가가 어렵습니다.

- 최근의 여러 연구들(Zhang et al., 2022; Hoffmann et al., 2022)은 RealToxicityPrompts 벤치마크(Gehman et al., 2020)를 모델의 독성 수준을 판단하는 지표로 사용하고 있습니다. RealToxicityPrompts는 모델이 완성해야 하는 약 100,000개의 프롬프트로 구성되어 있으며, 이후에는 PerspectiveAPI 3을 통해 독성 점수가 자동으로 평가됩니다. 우리는 제3자인 PerspectiveAPI가 사용하는 파이프라인에 대한 통제권이 없어, 이전 모델과의 비교가 어렵습니다.

- 10만 개의 프롬프트 각각에 대해 우리는 우리의 모델로 탐욕적으로 생성을 하고, 그 독성 점수를 측정합니다. 프롬프트 당 점수는 0(비독성)에서 1(독성)까지 범위를 가집니다.

- 표 11에서는 RealToxicityPrompts의 기본적인 및 존중스러운 프롬프트 카테고리에 대한 평균 점수를 보고합니다. 이 점수들은 문헌에서 관찰되는 것과 "비교 가능한" 수준입니다(예를 들어, Chinchilla의 경우 0.087) 하지만 이러한 연구와 우리의 연구 방법론은 다릅니다(샘플링 전략, 프롬프트 수, API의 시간 측면에서). 모델의 크기가 커질수록 독성이 증가하는 것을 관찰하며, 특히 존중스러운 프롬프트에서 그렇습니다.

- 이는 이전 연구(Zhang et al., 2022)에서도 관찰되었으며, 크기가 다른 Chinchilla와 Gopher 사이에 차이를 보지 못한 Hoffmann et al. (2022)의 주목할만한 예외를 제외하면, 이러한 현상이 일반적입니다. 이는 큰 모델인 Gopher가 Chinchilla보다 성능이 떨어지는 것으로 볼 때, 독성과 모델 크기 사이의 관계가 모델 패밀리 내에서만 적용될 수 있다는 것을 설명할 수 있습니다.

5.1. RealToxicityPrompts

		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 decoder on the 100k prompts from this benchmark. The "respectful" versions are prompts starting with "Complete 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.

5.2 CrowS-Pairs

- 우리는 CrowS-Pairs(Nangia et al., 2020) 데이터셋을 사용하여 모델의 편향성을 평가합니다. 이 데이터셋은 다음 9가지 카테고리에서 편향성을 측정할 수 있습니다: 성별, 종교, 인종/피부색, 성적 지향, 연령, 국적, 장애, 외모, 사회경제적 지위.
- 각 예제는 스테레오타입과 반스테레오타입으로 구성되어 있으며, 제로샷 설정에서 두 문장의 복잡도를 사용하여 모델이 스테레오타입 문장을 선호하는 정도를 측정합니다. 따라서 점수가 높을수록 편향성이 높습니다.
- 표 12에서 GPT-3와 OPT-175B와 비교합니다.
- LLaMA는 평균적으로 두 모델에 비해 약간 더 유리하게 비교됩니다. 우리의 모델은 특히 종교 카테고리에서 편향이 심하며(OPT-175B에 비해 +10%), 그 다음으로는 연령과 성별이 있습니다. 이러한 편향성은 CommonCrawl로부터 비롯된 것으로 생각되며, 이는 여러 필터링 단계에도 불구하고 그렇습니다.

5.2 CrowS-Pairs

	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.

5.3 WinoGender

- 우리 모델의 성별 카테고리에 대한 편향을 더 깊게 조사하기 위해 WinoGender 벤치마크 (Rudinger et al., 2018)를 살펴봅니다. 이것은 대명사의 성별이 모델의 대용어 해석 성능에 어떤 영향을 미치는지 결정함으로써 편향을 평가하는 공동 참조 해석 데이터셋입니다.

- 더 정확하게는, 각 문장에는 세 가지 언급이 있습니다: "직업", "참가자", 그리고 "대명사" 입니다. 이 대명사는 문장의 맥락에 따라 직업이나 참가자를 가리킵니다. 모델에게 이 공동 참조 관계를 판단하도록 지시하고 문장의 맥락에 따라 올바르게 했는지 측정합니다.

- 목표는 모델이 직업과 관련된 사회적 편견을 포착했는지 밝혀내는 것입니다. 예를 들어, WinoGender 데이터셋의 문장 중 하나는 "간호사가 환자에게 그의 근무가 한 시간 안에 끝날 것이라고 알렸다."이며, 이어서 '그의(=his)'가 누구를 가리키는지 따집니다. 그런 다음 모델로 공동 참조 해석을 수행하기 위해 '간호사'와 '환자'의 연속성에 대한 복잡성을 비교합니다. 우리는 "그녀/그녀/그녀", "그/그/그" 그리고 "그들/그들/누군가"의 3가지 대명사를 사용할 때의 성능을 평가합니다(다른 선택은 대명사의 문법적 기능에 해당합니다).

5.3 WinoGender

- 표 13에서는 데이터셋에 포함된 세 가지 다른 대명사에 대한 공동 참조 점수를 보고합니다. 우리의 모델은 "her/her/she"와 "his/him/he" 대명사보다 "their/them/someone" 대명사에 대한 공동 참조 해석을 훨씬 더 잘 수행한다는 것을 확인했습니다. 이러한 유사한 관찰은 이전 연구(Rae et al., 2021; Hoffmann et al., 2022)에서도 이루어졌으며, 이것은 성별 편향의 가능성이 높습니다. 실제로 "her/her/she"와 "his/him/he" 대명사의 경우, 모델은 문장의 증거를 사용하는 대신 직업의 대다수 성별을 사용하여 공동 참조 해석을 수행하고 있을 가능성이 높습니다.

- 이 가설을 더 조사하기 위해 WinoGender 데이터셋의 "her/her/she"와 "his/him/he" 대명사에 대한 "갓챠(gotcha)" 케이스 집합을 살펴봅니다. 이러한 경우는 대명사가 직업의 대다수 성별과 일치하지 않고, 직업이 정답인 문장들에 해당합니다.

- 표 13에서 우리의 모델, LLaMA-65B,가 갓챠 예시에서 더 많은 오류를 범하는 것을 확인할 수 있으며, 이는 모델이 성별과 직업과 관련된 사회적 편향을 잡아내고 있다는 것을 명확하게 보여줍니다. 이러한 성능 하락은 "her/her/she"와 "his/him/he" 대명사에 대해서도 나타나며, 이것은 성별과 상관없이 편향이 있다는 것을 나타냅니다.

5.3 WinoGender

	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.

5.4 TruthfulQA

- TruthfulQA (Lin et al., 2021)는 모델의 진실성, 즉 어떤 주장이 참인지를 식별할 수 있는 능력을 측정하려고 합니다. Lin et al. (2021)은 "진실"이라는 정의를 "현실 세계에 대한 글자 그대로의 진실"로 고려하며, 믿음 체계나 전통의 맥락에서만 참인 주장을 의미하지 않습니다.
- 이 벤치마크는 모델이 잘못된 정보나 허위 주장을 생성할 위험을 평가할 수 있습니다. 질문들은 다양한 스타일로 작성되어 있고, 38개의 카테고리를 다루며, 적대적으로 설계되어 있습니다.
- 표 14에서는 우리의 모델이 얼마나 진실한 모델인지 측정하기 위한 두 가지 질문에 대한 성능을 보고합니다. GPT-3에 비해 우리의 모델은 두 카테고리에서 모두 더 높은 점수를 받았지만, 정답률은 여전히 낮아, 우리의 모델이 잘못된 답을 '환영(hallucinate)'할 가능성이 높다는 것을 보여줍니다.

5.4 TruthfulQA

		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.

8. Conclusion

- 이 논문에서는 공개적으로 릴리스되며 최첨단 기반 모델과 경쟁력 있는 일련의 언어 모델을 소개했습니다. 가장 주목할 만한 것은 LLaMA-13B가 GPT-3를 능가하면서도 10배 이상 작다는 것이며, LLaMA-65B는 Chinchilla-70B와 PaLM-540B와 경쟁력을 가집니다.
- 이전 연구와 달리, 독점 데이터셋을 사용하지 않고도 공개적으로 이용 가능한 데이터만을 사용하여 최첨단 성능을 달성할 수 있다는 것을 보여줍니다.
- 이러한 모델을 연구 커뮤니티에 공개함으로써 대규모 언어 모델의 발전을 가속화하고, 그들의 견고성을 향상시키고 악성 및 편향과 같은 알려진 문제를 완화하는 노력에 도움이 될 것이라고 기대합니다.
- 추가로, Chung 등(2022)과 같이 이러한 모델을 지침에 대한 미세 조정(finetuning)이 유망한 결과를 가져온다는 것을 확인했으며, 이를 미래의 연구에서 더 깊게 조사할 계획입니다.
- 마지막으로, 우리는 규모를 키우면서 지속적인 성능 향상을 보았기 때문에, 미래에는 더 큰 모델을 더 큰 사전 훈련 데이터셋에서 훈련시켜 릴리스할 계획입니다.