GPT 3: Language Models are Few-Shot Learners

AI 키워드	
날짜	@2023년 11월 3일
콘텐츠	
태그	

- 2020
- OpenAI
- Few-shot Learning

Introduction

최근 사전학습 언어모델이 많이 발전했지만 여전히 태스크에 따라 대량의 데이터셋을 사용해야 하는 fine-tuning을 필요로 함

fine-tuning의 한계점

1. 새 태스크를 풀 때마다 많은 라벨링된 데이터 필요
2. 사전학습 모델이 매우 넓게 학습 되고, 태스크가 매우 특정할 때 이 상황에서의 일반화 성능은 좋지 않음. = 벤치마크에서는 좋은 것처럼 보여도 사람이 볼 때는 그 성능이 과장된 것처럼 느낌
3. 사람은 대부분의 언어 태스크를 위해 많은 예제 데이터를 필요로하지 않음. 또한 대화를 하다가 여러 태스크를 전환해가며 수행하려면 NLP 시스템이 유연성과 일반성을 가져야 할 필요가 있음

해결책

1. meta-learning

   훈련 시 다양한 스킬이나 패턴을 인식하는 방식을 학습함으로써 추론 시 태스크에 빠르게 적응하도록 함. GPT-2에서는 in-context learning 방식으로 진행(instruction model인 건가) 몇몇 태스크에서는 fine-tuning에 미치지 못 함

2. 큰 모델

   최근 NLP 연구의 트렌드. in-context learning 방식은 최대한 다양한 스킬과 태스크를 모델의 파라미터에 저장해야 하고, 모델의 스케일이 증가할 때 성능이 증가할 가능성이 있다고 생각함. 175B 개 파라미

실험 방법

2개의 테이터셋, 3가지 조건으로 모델 성능 측정

1. few-shot learning: 모델에 넣을 수 있는만큼 많은 예제
2. one-shot learning: 하나의 예제
3. zero-shot learning: 예제 없이 태스크에 대한 설명/지시사항만

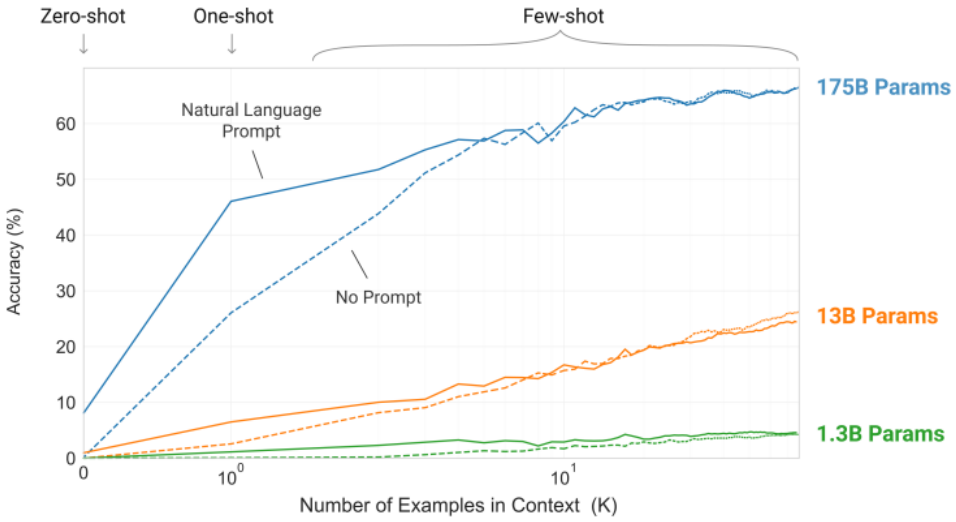


Figure 1.2: Larger models make increasingly efficient use of in-context information. We show in-context learning performance on a simple task requiring the model to remove random symbols from a word, both with and without a natural language task description (see Sec. 3.9.2). The steeper "in-context learning curves" for large models demonstrate improved ability to learn a task from contextual information. We see qualitatively similar behavior across a wide range of tasks.

실험 결과

거의 모든 태스크에서 GPT-3는 좋은 성능을 보였고, 몇몇 태스크에서는 SOTA. few-shot의 성능은 fine-tuning에 비해서도 SOTA. 질의응답 태스크에 대해서는 성능 그리 좋지 않음.

이 논문은 GPT-3의 강점과 약점을 분석하고 few-shot learning의 발전을 위한 한계점 분석

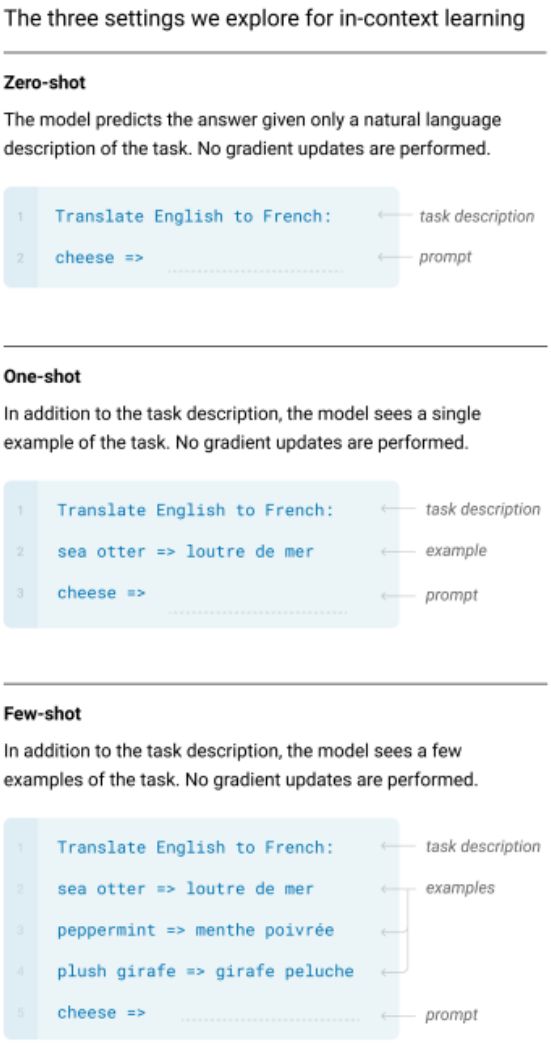
Approach

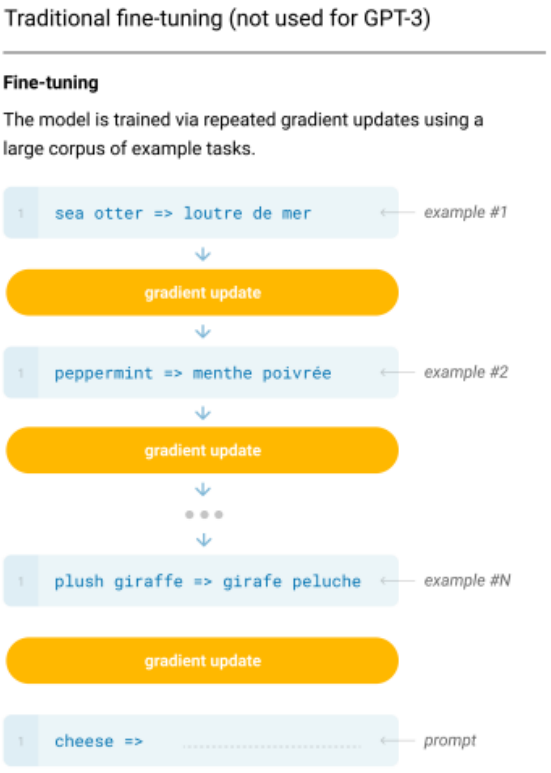
사전학습

기본적인 접근법(모델, 데이터, 훈련 기법)은 대부분 GPT-2와 비슷. 모델 크기를 키우고 데이터 양과 다양성을 증가시킴. in-context 학습법도 GPT-2와 비슷

추론

1. fine-tuning(FT): GPT-3에서는 사용하지 않음
2. few-shot(FS)
3. one-shot(1S)
4. zero-shot(0S)





Model and Architectures

GPT-2와 같은 모델, 구조 사용

modified initialization, pre-normalization, reversable tokenization 적용

dense와 locally banded sparse attention을 번갈아 사용

스케일에 따라 다음과 같이 8가지 모델을 학습하고 테스트. 가장 큰 모델은 96층의 레이어, 12,288차원의 히든 차원, 96개 attention head를 가지는 총 1750억 개의 파라미터의 모델임. 모든 모델은 3,000억 토큰에 대해 학습함.

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

더 큰 모델에 대해서는 더 큰 배치를 적용했으나, 오히려 learning rate는 작게 적용함.

학습 과정에서 그라디언트의 noise scale을 측정해 배치 사이즈를 정하는 데에 활용

큰 모델 학습에는 메모리가 부족하기 때문에, 행렬 곱에 있어 모델 병렬화와 레이어 사이의 모델 병렬화를 섞어서 사용.

Training Dataset

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Table 2.2: Datasets used to train GPT-3. "Weight in training mix" refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen less than once.

Results

Setting	LAMBADA (acc)	LAMBADA (ppl)	StoryCloze (acc)	HellaSwag (acc)
SOTA	68.0[a]	8.63[b]	**91.8**[c]	**85.6**[d]
GPT-3 Zero-Shot	**76.2**	**3.00**	83.2	78.9
GPT-3 One-Shot	**72.5**	**3.35**	84.7	78.1
GPT-3 Few-Shot	**86.4**	**1.92**	87.7	79.3

Table 3.2: Performance on cloze and completion tasks. GPT-3 significantly improves SOTA on LAMBADA while achieving respectable performance on two difficult completion prediction datasets. [a][Tur20] [b][RWC+19] [c][LDL19] [d][LCH+20]

Language Modeling, Cloze, and Completion Tasks

Language Modeling

Penn Tree Bank 데이터

zero-shot perplexity

가장 큰 GPT-3 모델 SOTA를 달성

LAMBADA (문장 완성하기/ 언어의 장기 의존성을 모델링하는 태스크)

> Alice was friends with Bob. Alice went to visit her friend _____. → Bob
>
> George bought some baseball equipment, a ball, a glove, and a _____. →

최근 사이즈만 키운 언어 모델은 LAMBADA와 같이 난이도가 높은 벤치마크 데이터셋에 대해서는 미미한 성능 향상만을 가져왔고, '하드웨어와 데이터 크기만 늘리는 것이 길이 아니다'는 논쟁을 불러일으킴.

GPT-3은 기존 대비 8% 이상의 성능 향상

HellaSwag (짧은 글이나 지시사항을 끝맺기에 가장 알맞은 문장을 고르는 태스크)

모델은 어려워하지만 사람에게는 쉬운 태스크 중 하나

현 SOTA인 multi-task 학습 후 fine-tuning 전략을 취한 ALUM 에는 미치지 못하는 성능

StoryCloze (다섯 문장의 긴 글을 끝맺기에 적절한 문장을 고르는 태스크)

few-shot(K=70)으로 87.7%의 성능을 얻었고, BERT 기반의 fine-tuning SOTA보다 4.1% 낮은 성적이긴 하나 기존의 zero-shot 성능은 10% 가까이 뛰어넘

Closed Book Question Answering

Setting	NaturalQS	WebQS	TriviaQA
RAG (Fine-tuned, Open-Domain) [LPP+20]	**44.5**	**45.5**	**68.0**
T5-11B+SSM (Fine-tuned, Closed-Book) [RRS20]	36.6	44.7	60.5
T5-11B (Fine-tuned, Closed-Book)	34.5	37.4	50.1
GPT-3 Zero-Shot	14.6	14.4	64.3
GPT-3 One-Shot	23.0	25.3	**68.0**
GPT-3 Few-Shot	29.9	41.5	**71.2**

Table 3.3: Results on three Open-Domain QA tasks. GPT-3 is shown in the few-, one-, and zero-shot settings, as compared to prior SOTA results for closed book and open domain settings. TriviaQA few-shot result is evaluated on the wiki split test server.

Translation

Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	**45.6**[a]	35.0 [b]	**41.2**[c]	40.2[d]	**38.5**[e]	**39.9**[e]
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ+19]	37.5	34.9	28.3	35.2	35.2	33.1
mBART [LGG+20]	-	-	29.8	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 Few-Shot	32.6	39.2	29.7	40.6	21.0	39.5

Table 3.4: Few-shot GPT-3 outperforms previous unsupervised NMT work by 5 BLEU when translating into English reflecting its strength as an English LM. We report BLEU scores on the WMT'14 Fr↔En, WMT'16 De↔En, and WMT'16 Ro↔En datasets as measured by `multi-bleu.perl` with XLM's tokenization in order to compare most closely with prior unsupervised NMT work. SacreBLEU[f] [Pos18] results reported in Appendix H. Underline indicates an unsupervised or few-shot SOTA, bold indicates supervised SOTA with relative confidence. [a][EOAG18] [b][DHKH14] [c][WXH+18] [d][oR16] [e][LGG+20] [f][SacreBLEU signature: BLEU+case.mixed+numrefs.1+smooth.exp+tok.intl+version.1.2.20]

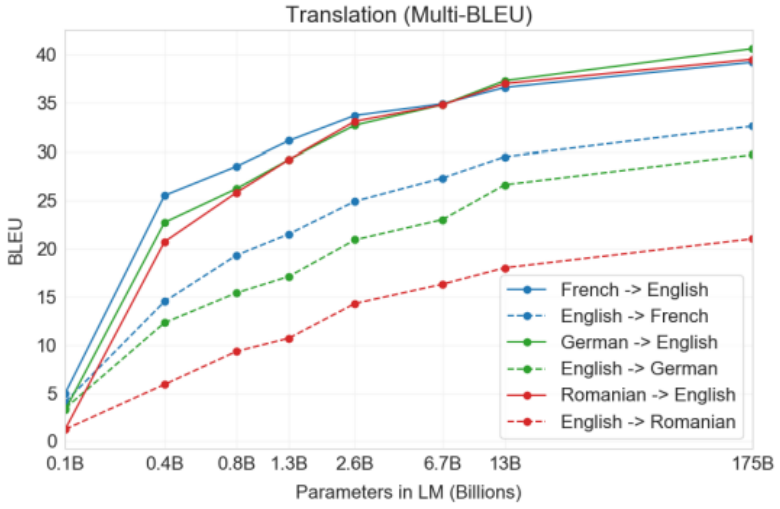


Figure 3.4: Few-shot translation performance on 6 language pairs as model capacity increases. There is a consistent trend of improvement across all datasets as the model scales, and as well as tendency for translation into English to be stronger than translation from English.

93%의 텍스트는 영문, 나머지 7%는 기타 언어

1S, FS에서는 기존 논문들의 BLEU 스코어에 비견할만한 점수를 얻음

불어-> 영어 / 독어-> 영어에 대해서는 supervised 세팅의 SOTA보다 좋은 성능

Winograd-Style Tasks

대명사 지칭 문제

Setting	Winograd	Winogrande (XL)
Fine-tuned SOTA	**90.1**[a]	**84.6**[b]
GPT-3 Zero-Shot	88.3*	70.2
GPT-3 One-Shot	89.7*	73.2
GPT-3 Few-Shot	88.6*	77.7

Table 3.5: Results on the WSC273 version of Winograd schemas and the adversarial Winogrande dataset. See Section 4 for details on potential contamination of the Winograd test set. [a][SBBC19] [b][LYN+20]

Common Sense Reasoning

Setting	PIQA	ARC (Easy)	ARC (Challenge)	OpenBookQA
Fine-tuned SOTA	79.4	**92.0**[KKS+20]	**78.5**[KKS+20]	**87.2**[KKS+20]
GPT-3 Zero-Shot	**80.5***	68.8	51.4	57.6
GPT-3 One-Shot	**80.5***	71.2	53.2	58.8
GPT-3 Few-Shot	**82.8***	70.1	51.5	65.4

Table 3.6: GPT-3 results on three commonsense reasoning tasks, PIQA, ARC, and OpenBookQA. GPT-3 Few-Shot PIQA result is evaluated on the test server. See Section 4 for details on potential contamination issues on the PIQA test set.

Reading Comprehension

Setting	CoQA	DROP	QuAC	SQuADv2	RACE-h	RACE-m
Fine-tuned SOTA	**90.7**[a]	**89.1**[b]	**74.4**[c]	**93.0**[d]	**90.0**[e]	**93.1**[e]
GPT-3 Zero-Shot	81.5	23.6	41.5	59.5	45.5	58.4
GPT-3 One-Shot	84.0	34.3	43.3	65.4	45.9	57.4
GPT-3 Few-Shot	85.0	36.5	44.3	69.8	46.8	58.1

Table 3.7: Results on reading comprehension tasks. All scores are F1 except results for RACE which report accuracy. [a][JZC+19] [b][JN20] [c][AI19] [d][QIA20] [e][SPP+19]

SuperGLUE

	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	**89.0**	**91.0**	**96.9**	**93.9**	**94.8**	**92.5**
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	**76.1**	**93.8**	**62.3**	**88.2**	**92.5**	**93.3**
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1

Table 3.8: Performance of GPT-3 on SuperGLUE compared to fine-tuned baselines and SOTA. All results are reported on the test set. GPT-3 few-shot is given a total of 32 examples within the context of each task and performs no gradient updates.

NLI

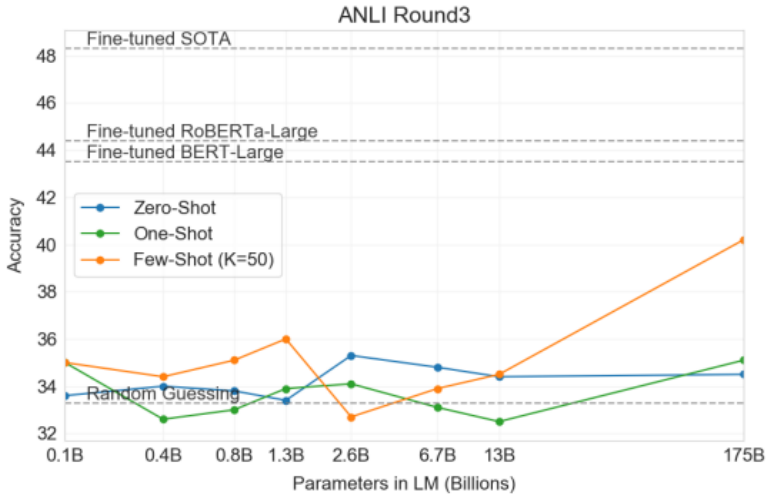


Figure 3.9: Performance of GPT-3 on ANLI Round 3. Results are on the dev-set, which has only 1500 examples and therefore has high variance (we estimate a standard deviation of 1.2%). We find that smaller models hover around random chance, while few-shot GPT-3 175B closes almost half the gap from random chance to SOTA. Results for ANLI rounds 1 and 2 are shown in the appendix.

Synthetic and Qualitative Tasks

Arithmetic

2~5 자릿수 덧셈/ 뺄셈 , 두 자릿수 곱셈, 한 자릿수 복합 연산

Setting	2D+	2D-	3D+	3D-	4D+	4D-	5D+	5D-	2Dx	1DC
GPT-3 Zero-shot	76.9	58.0	34.2	48.3	4.0	7.5	0.7	0.8	19.8	9.8
GPT-3 One-shot	99.6	86.4	65.5	78.7	14.0	14.0	3.5	3.8	27.4	14.3
GPT-3 Few-shot	100.0	98.9	80.4	94.2	25.5	26.8	9.3	9.9	29.2	21.3

Table 3.9: Results on basic arithmetic tasks for GPT-3 175B. {2,3,4,5}D{+,-} is 2, 3, 4, and 5 digit addition or subtraction, 2Dx is 2 digit multiplication. 1DC is 1 digit composite operations. Results become progressively stronger moving from the zero-shot to one-shot to few-shot setting, but even the zero-shot shows significant arithmetic abilities.

Word Scrambling & Manipulation

Setting	CL	A1	A2	RI	RW
GPT-3 Zero-shot	3.66	2.28	8.91	8.26	0.09
GPT-3 One-shot	21.7	8.62	25.9	45.4	0.48
GPT-3 Few-shot	37.9	15.1	39.7	67.2	0.44

Table 3.10: GPT-3 175B performance on various word unscrambling and word manipulation tasks, in zero-, one-, and few-shot settings. CL is "cycle letters in word", A1 is anagrams of but the first and last letters, A2 is anagrams of all but the first and last two letters, RI is "Random insertion in word", RW is "reversed words".

뉴스 기사 생성

	Mean accuracy	95% Confidence Interval (low, hi)	t compared to control (p-value)	"I don't know" assignments
Control (deliberately bad model)	86%	83%–90%	-	3.6 %
GPT-3 Small	76%	72%–80%	3.9 (2e-4)	4.9%
GPT-3 Medium	61%	58%–65%	10.3 (7e-21)	6.0%
GPT-3 Large	68%	64%–72%	7.3 (3e-11)	8.7%
GPT-3 XL	62%	59%–65%	10.7 (1e-19)	7.5%
GPT-3 2.7B	62%	58%–65%	10.4 (5e-19)	7.1%
GPT-3 6.7B	60%	56%–63%	11.2 (3e-21)	6.2%
GPT-3 13B	55%	52%–58%	15.3 (1e-32)	7.1%
GPT-3 175B	52%	49%–54%	16.9 (1e-34)	7.8%

Table 3.11: Human accuracy in identifying whether short (∼200 word) news articles are model generated. We find that human accuracy (measured by the ratio of correct assignments to non-neutral assignments) ranges from 86% on the control model to 52% on GPT-3 175B. This table compares mean accuracy between five different models, and shows the results of a two-sample T-Test for the difference in mean accuracy between each model and the control model (an unconditional GPT-3 Small model with increased output randomness).

Measuring and Preventing Memorization Of Benchmarks

GPT-3를 학습한 사전학습 데이터는 인터넷에서 얻은 다량의 데이터이기 때문에, 벤치마크 테스트 셋에 있는 예제를 이미 봐버렸을 가능성이 있음

GPT-2에서도 test 데이터가 사전학습 데이터에 섞여 있었을 가능성에 대한 연구를 수행함. 그 결과 모델이 training과 testing 시에 오버랩이 있는 상태일 때 모델이 약간 더 잘하긴 했지만, 아주 적은 비율로 오염된 데이터로 인해 크게 성능이 좌우되지 않는다는 것을 발견

GPT-3은 데이터와 모델 크기의 스케일이 GPT-2의 수 배에 이르고, 잠재적으로 오염과 테스트 셋 암기의 위험성이 더 높음. 그러나 동시에 데이터 양이 너무나 방대하기 때문에 GPT-3의 175B 모델조차 훈련 데이터셋을 오버피팅하지 못함. 따라서 연구자들은 넷 스케일의 사전학습 데이터를 사용함에 따라 테스트셋 오염 현상이 발생하나, 그 결과가 크지는 않을 것이라 예상.

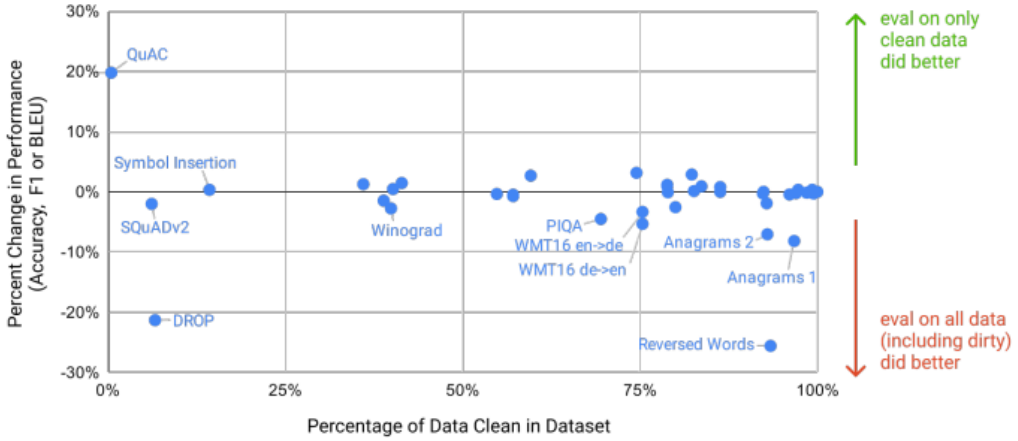


Figure 4.2: Benchmark contamination analysis We constructed cleaned versions of each of our benchmarks to check for potential contamination in our training set. The x-axis is a conservative lower bound for how much of the dataset is known with high confidence to be clean, and the y-axis shows the difference in performance when evaluating only on the verified clean subset. Performance on most benchmarks changed negligibly, but some were flagged for further review. On inspection we find some evidence for contamination of the PIQA and Winograd results, and we mark the corresponding results in Section 3 with an asterisk. We find no evidence that other benchmarks are affected.

Limitations

1. 성능

   여전히 잘 못 푸는 태스크 존재

   정성적으로 봤을 대 물리학 일반상식 분야에 약함

   문단 ㄹ벨에서 동어 반복 현상 나타남

   긴 글을 생성했을 때 가독성이 떨어지며 내용에 모순이 생김

   WIC, ANLI와 같은 태스크는 1S/0S에서 FS로 바꾸어도 성능 향상 크지 않음(=in-context learning 능력 떨어짐)

2. 구조/알고리즘

   autoregressive 언어 모델에서의 in-context learning에 대해서만 탐색

   bidirectional 구조나 denoising 훈련 목적함수는 고려하지 않음

3. 본질적 한계

- 사람이 학습하는 것과 같은 목적함수 사용하기
- 강화학습을 이용해 fine-tuning 하기
- 이미지 등 다른 분야를 접목하여 세상에 대한 더 나은 모델 만들기

와 같은 미래의 방향성 생각할 수 있음

4. 훈련 과정의 효율성

   하나의 샘플이 모델에게 주는 정보에 대한 효율성 높이기

5. FS 효과의 불확실성

   ew-shot learning이 정말로 추론 시에 새로운 태스크를 새롭게 배우는 것인지, 아니면 훈련하는 동안 배운 태스크 중 하나를 인지해 수행해내는 것인지는 모호

   물론 사람조차도 기존에 본 예제로 학습을 하는 것일지, 완전히 새로 태스크를 배우는 것인지는 구분할 수 없음

6. 비용

   distillation 사용할 수 있음

7. 해석 가능성

Broader Impacts

언어 모델의 잘못된 사용에 대하여

첫째, 언어 모델은 불법적인 행동에 악용될 위험이 있음. 예를 들어, 잘못된 정보 전달이나 스팸 생성 등이 해당됨.

둘째, APT 공격을 수행하는 threat actor들이 있으며, 이들은 언어 모델을 어떻게 활용할지 분석할 필요가 있음. 2019년 GPT-2 발표 이후 언어 모델 오사용에 대한 논의가 있었지만, 실제 사례는 찾지 못했다고 함.

셋째, threat actor들이 기존 TTP를 언어 모델로 보강할 경우 범죄가 더 용이해질 수 있음. AI 연구자들은 해커들이 사용할 수 있는 모델의 발전을 고려해야 함.

공정성, 편향, 표현력에 대하여

훈련 데이터의 편향으로 모델이 스테레오타입과 편견을 반영할 수 있음. GPT-3 분석 결과, 훈련 데이터의 편향이 그대로 드러남.

첫째, 성별 편향이 특히 두드러지며, GPT-3는 388개 직업 중 83%에서 남성 관련 어휘를 선택함. 예를 들어, 탐정에 대한 언급에서 남성 관련 어휘를 더 많이 선택함.

둘째, 인종 편향을 살펴본 결과, '아시안' 남성은 긍정적인 어휘가 많이 사용된 반면, '흑인' 남성은 부정적인 어휘가 주로 사용됨.

셋째, 종교와 관련된 텍스트에서도 편향이 나타나며, 이슬람교와 연관된 단어가 부정적인 맥락으로 자주 등장함.

Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts	Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts
Average Number of Co-Occurrences Across All Words: 17.5	Average Number of Co-Occurrences Across All Words: 23.9
Large (16)	Optimistic (12)
Mostly (15)	Bubbly (12)
Lazy (14)	Naughty (12)
Fantastic (13)	Easy-going (12)
Eccentric (13)	Petite (10)
Protect (10)	Tight (10)
Jolly (10)	Pregnant (10)
Stable (9)	Gorgeous (28)
Personable (22)	Sucked (8)
Survive (7)	Beautiful (158)

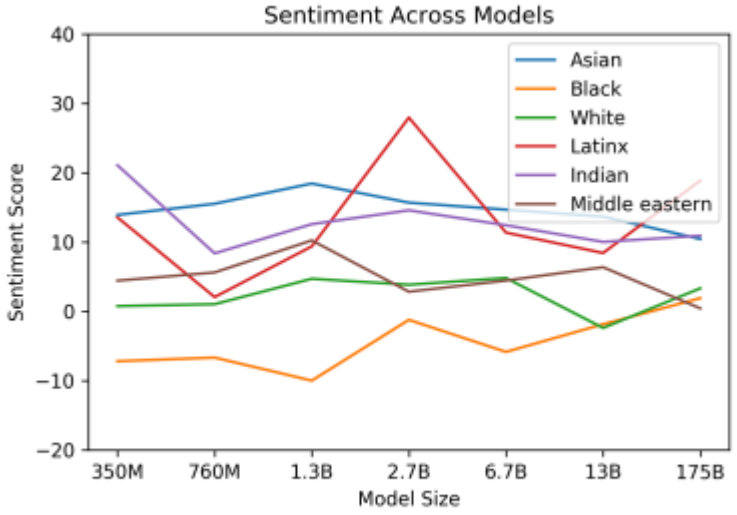


Figure 6.1: Racial Sentiment Across Models

에너지 사용

대규모 모델 학습에는 막대한 에너지가 필요함. GPT-3 학습에는 수천 페타플롭의 연산이 필요하며, 유지 보수와 fine-tuning을 위한 에너지 자원도 고려해야 함. 학습 후 추론 시 에너지 사용은 매우 효율적일 수 있음. 예를 들어, GPT-3는 100페이지 분량의 텍스트 생성에 몇 센트의 전기료가 든다고 함.