Language Models are Few-Shot Learners

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*

Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry

Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan

Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter

Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

Abstract

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions – something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora. Finally, we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.

최근 연구에서는 대규모 텍스트 corpus를 pre-training하고 특정 작업에 대해 미세 조정을 하는 방법으로 여러 NLP 작업과 벤치마크에서 상당한 성과를 보여주었다. 이 방법은 아키텍처상 작업 비전문적(task-agnostic)임에도 불구하고, 여전히 수천에서 수만 개의 예시를 포함하는 작업별 fine-tunin`g 데이터셋을 필요로 한다. 반면, 인간은 몇 가지 예시나 간단한 지시만으로 새로운 언어 작업을 수행할 수 있으며, 이는 현재의 NLP 시스템이 여전히 크게 어려움을 겪는 부분이다. 여기서는 **언어 모델의 크기를 확장함으로써 작업 비전문적(task-agnostic)이고 소수 예시 학습(few-shot learning) 성능이 크게 향상된다**는 것을 보여준다. 때때로 fine-tuning 기반의 최첨단 접근 방식과 경쟁력 있는 성능을 보이기도 한다. 구체적으로, 우리는 GPT-3라는 1750억 개의 매개변수를 가진 자가회귀 언어 모델을 훈련하고

소수 예시 학습 환경에서 성능을 테스트하였다. 모든 작업에서 GPT-3는 어떤 기울기 업데이트나 fine-tuning 없이 적용되며, 작업과 소수 예시 학습은 모델과의 텍스트 상호작용을 통해 순전히 명시된다. GPT-3는 번역, 질문 답변, 클로즈(cloze) 작업 등 많은 NLP 데이터셋에서 강력한 성능을 보였으며, 단어 섞기, 새로운 단어를 문장에서 사용하기, 3자리 수 계산 수행 등 즉석에서 추론하거나 도메인에 적응해야 하는 여러 작업에서도 성과를 거두었다. 동시에 GPT-3의 소수 예시 학습이 여전히 어려움을 겪는 일부 데이터셋과 대형 웹 코퍼스에서 학습할 때 발생하는 방법론적 문제를 식별하였다. 마지막으로, GPT-3가 생성한 뉴스 기사 샘플은 인간 평가자들이 인간이 쓴 기사와 구별하기 어려운 경우가 있음을 발견하였다. 우리는 이러한 발견과 GPT-3 전반이 미치는 더 넓은 사회적 영향을 논의한다.

- 작업 비전문적 : 특정 작업에 제한되지 않고 여러 작업에 적용될 수 있는 범용적인 모델

Introduction

최근 NLP 연구는 task에 무관한 representation을 학습하는 방향으로 발전했습니다.

- 예시 : (RNN레이어를 쌓아 문맥 벡터를 만드는)ELMo, (트랜스포머 구조를 이용해 문맥을 표현하는 깊은 모델인)BERT, GPT, ULMFit 등

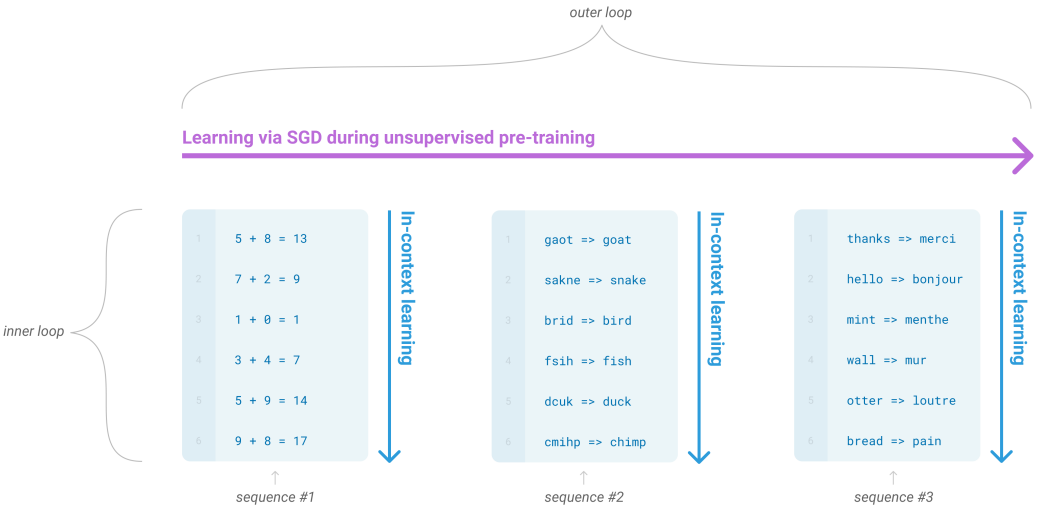
downstream task와 상관없이 대량의 corpus를 이용해 pre-training을 진행하고, 이렇게 학습된 모델은 task-specific fine-tuning을 통해 퍼포먼스를 내었다. 대부분의 task에서 잘 작동하는 **"task-agnostic model"** 이라고 한다. 즉, task-agnostic model은 좋은 성능을 내기 위해 fine-tuning해야한다.

few-shot learning : 몇 개의 예시만 보고 task에 적용하여 문제를 푸는 것

- One-shot learning : 예시를 하나만 알려주고 문제를 푸는 경우
- zero-shot learning : 예시 하나도 없이 모델을 바로 task에 사용하는 경우

기존 방식의 한계점

- 새로운 작업에 대한 대규모 레이블이 지정된 데이터셋 필요
- 훈련 데이터의 분포에 대해 한정된 모델은 그 외의 영역은 잘 일반화하지 못함(out of distribution 문제를 잘 일반화하지 못함).
- 인간은 몇 가지 예제 데이터 만으로 수행 가능 → 현재의 NLP 기술에 대한 개념적 한계를 지적+실질적인 이점을 제공



Meta Leanring을 사용한 해결

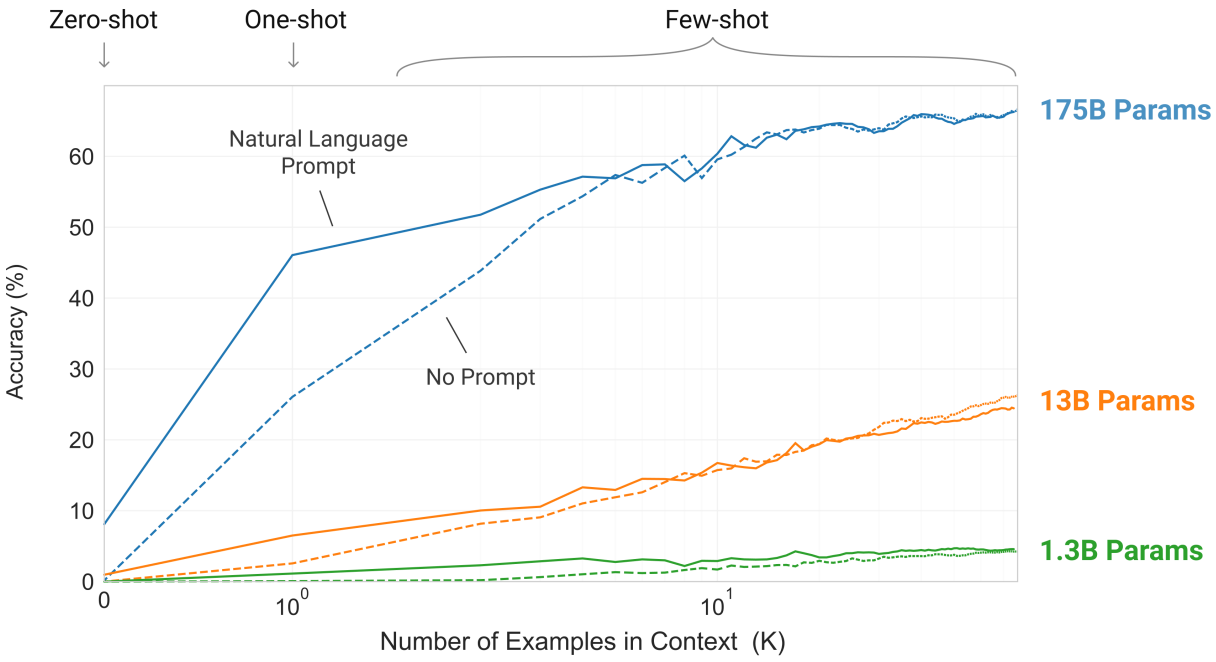
: 모델이 다양한 작업과 패턴을 학습하여 새로운 상황에 빠르게 적응하는 능력을 개발하는 과정. 훈련 시간에 광범위한 기술을 학습하고, 추론 시간에 이를 빠르게 적용.

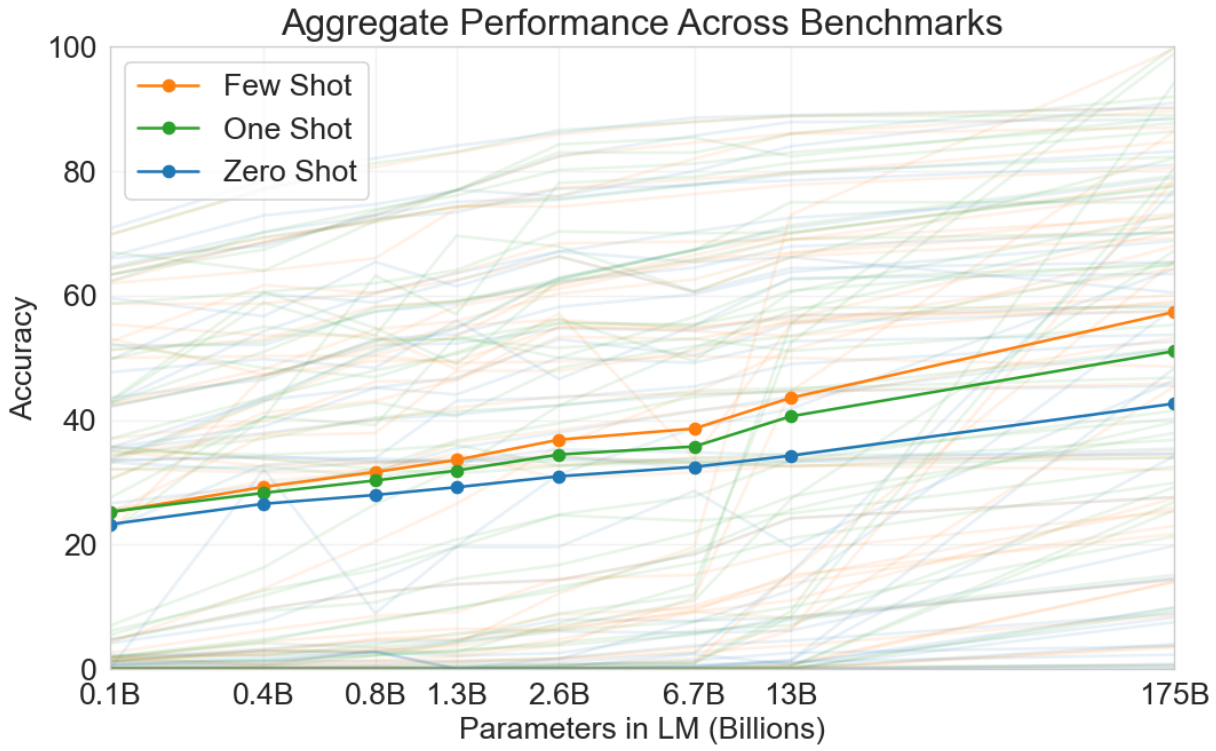
- Meta Learning 중 내부 루프인 In-context learning(실제 작업 수행 단계) : 사전 훈련된 언어 모델이 추가 학습 없이 주어진 문맥(예시나 지시사항)을 바탕으로 새로운 작업을 수행하는 능력

최근 NLP 연구의 또 다른 트렌드는 모델의 크기를 키우는 것이다(by transformer). GPT-1은 1억개, BERT는 3억개, GPT-2는 2억개, Megatron은 110억개, Project Turing은 170억개

→ Downstream task에서의 성능은 점점 좋아졌다.

- Transformer를 사용하면 용이하게 모델 사이즈를 크게 하고 파라미터 수를 증가시킬 수 있는가?
    - Self-Attention 매커니즘을 통해 RNN과 달리 시퀀스 길이와 무관하게 모든 위치 간의 관계를 동시에 학습 가능하다. 장기 의존성 문제를 해결하고 긴 시퀀스에 대해 더욱 효과적인 표현을 학습할 수 있다. 또한 각 층마다 정규화와 residual connection을 사용하기에 gradient 소실과 폭발 문제를 완화시키고 더욱 깊은 모델을 구성하며 복잡한 패턴을 학습할 수 있다.





- 모델 크기가 커질수록 in-context learning 성능이 향상됨 → 궁극적으로 다양한 작업을 유연하게 수행할 수 있는 NLP 시스템 개발

첫 번째 그래프는 "단어에 섞인 랜덤한 기호 제거하기 task"에 대한 각 모델의 성능. 성능을 측정하는 동안 그라이언트 업데이트나 fine-tuning을 하지 않고, 오로지 문맥에 포함하는 예제 개수(K)만 늘리며 실험.

결과

1. No Prompt 성능 < Natural Language Prompt 성능 : task에 대한 자연어 설명은 모델 성능 향상.
2. 모델의 문맥 윈도우에 더 많은 예제를 놓을 수록 성능이 향상. (K가 증가할 수록 정확도 증가)
3. 큰 모델일 수록 in-context 정보를 잘 활용.

NLP task 전반에 걸쳐 GPT-3는 few-shot, one-shot, zero-shot 셋팅에서 우수한 성능을 보임.

또한 GPT-3는 단어 순서 맞추기, 문장에서 새로운 단어 사용하기, 3자리 수리 연산 등과 같은 추론 혹은 도메인 적응이 필요한 task에서도 몇 개의 예제만 보고 잘 수행(기사쓰기는 기자가 쓴 글인지, 기계가 쓴 글인지 분간이 어려울 정도). 하지만 GPT-3의 스케일로도 감당이 어려운 few-shot task(ANLI, RACE, QuAC 같은 질의 응답 셋)도 있음.

대규모 언어 모델의 훈련과 평가에 관해 사용 연구방법

1. 데이터 오염 연구 :
    - 대용량 모델 훈련 시 웹에서 가져온 데이터와 평가 데이터 세트 간 중복 문제
    - 데이터 오염을 측정하고 정량화하는 체계적인 도구를 개발
    - 오염이 결과에 영향을 미치는 데이터셋을 식별하고, 이러한 데이터셋은 분석에서 제외
2. 모델 크기 연구 :
    - 125백만에서 175십억 파라미터까지 다양한 크기의 모델을 훈련
    - 모델 크기와 성능 간의 관계를 연구 → few-shot 성능이 모델 크기에 따라 크게 향상됨을 발견
3. GPT-3의 능력 스펙트럼 분석 :
    - 모델의 공정성, 편향성, 사회적 영향 등에 대한 우려사항

Approach

모델, 데이터, 훈련 기법은 대부분 GPT-2와 비슷하지만, 차이점으로는 **모델의 크기를 키우고, 데이터의 양과 다양성을 확연히 증가시켰다**는 점 이다.

In-context learning도 GPT-2와 비슷하지만, context 내에서는 구조적으로 다른 몇 가지 setting을 시도했다. Task-specific 데이터를 얼마나 활용하느냐에 따라 아래와 같은 4가

지 setting으로 분류할 수 있다.

1. Fine-tuning(FT)

원하는 task에 맞는 data set을 통해 task-specific fine-tuning을 실시. Fine-tuning의 장점은 성능이 매우 좋다는 것이며, 단점은 각 task를 학습할 때 마다 수 많은 데이터가 필요하다는 것.

→ GPT-3도 fine-tuning으로 학습할 수 있지만, 논문의 목적상 시행하지는 않음.

2. Few-shot(FS)

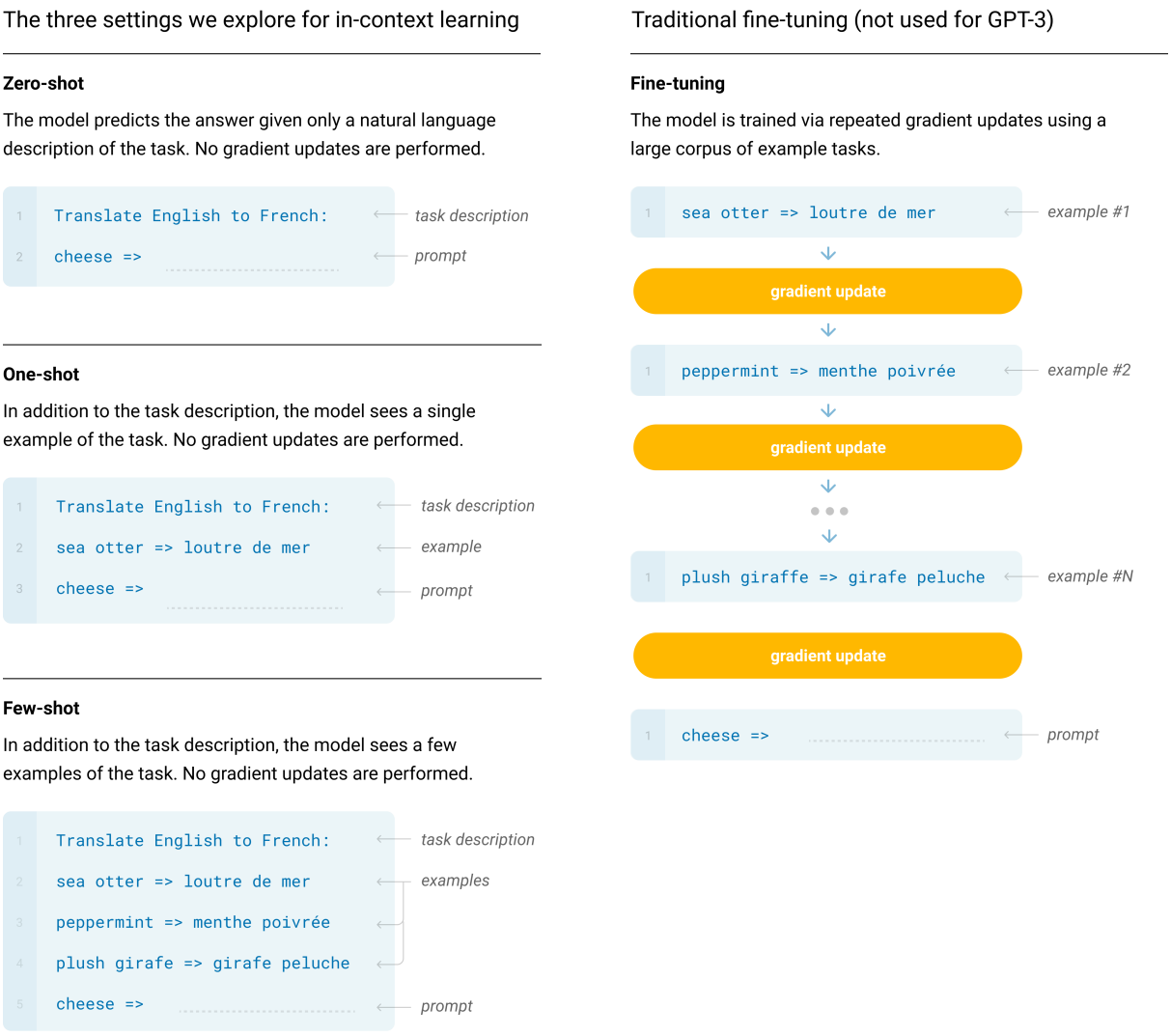
모델이 추론 과정에서 몇 개의 예시만을 볼 수 있지만, 직접 학습에 활용하지 않기에 **가중치 업데이트를 하지 않는 조건**. 보통 task에 대한 설명과 함께 task에 관한 K개(:context window, 10~100개)의 예시를 이용합니다. 이후 마지막으로 단 한 개의 문맥이 주어지면 모델이 답을 생성하는 것. 이에 대한 장점은 task-specific한 데이터에 대한 필요성을 줄여주며, 지나치게 크고 좁은 분포를 갖는 fine-tuning용 데이터셋을 학습할 필요성을 줄일 수 있음. 반면 단점으로는 Fine-tuning 방식의 SOTA에 비해 성능이 떨어진다는 점.

3. One-shot(1S)

task에 대한 예시가 하나만 주어지는 것으로, few-shot과 one-shot을 나누는 이유는 **one-shot이 인간의 커뮤니케이션과 비슷**하기 때문.

4. Zero-shot(0s)

어떤 task인지에 대한 설명만 주어지며, 따로 예시가 주어지지 않습니다. 이 방법은 최대한의 편의, 견고함에 대한 가능성, 거짓된 상관성 회피를 제공하지만, 가장 어려운 조건임. 어떤 경우에는 사람조차 예시가 없이 task에 대한 설명만으로는 이해하지 못 할 수도 있기 때문. 그럼에도 zero-shot의 일부 셋팅은 사람들이 task를 수행하는 방식과 가장 가깝기에 사용됩니다.



이 이미지는 in-context learning의 다양한 설정과 전통적인 fine-tuning을 비교 설명하고 있다.

이 논문에서 중점적으로 다루는 zero-shot, one-shot, few-shot은 모두 gradient updates 없이 작동. 테스트 시 모델에 작업 설명, 예시, 프롬프트를 제시.

1. Zero-shot:
    - 모델에 자연어로 된 작업 설명만 제공.
    - 예시: "Translate English to French:"
2. One-shot:
    - 작업 설명과 함께 하나의 예시가 제공.
    - 예시: "Translate English to French: sea otter → loutre de mer"

3. Few-shot:
    - 작업 설명과 함께 여러 개의 예시가 제공.
    - 예시: "Translate English to French:
      sea otter → loutre de mer
      peppermint → menthe poivrée
      plush giraffe → girafe peluche"
4. Fine-tuning (전통적인 방법, GPT-3에는 사용되지 않음):
    - 대규모 레이블된 데이터셋을 사용하여 모델을 재훈련.
    - 그라디언트 업데이트가 수행됩니다.

Zero-Shot (0S)은 one-shot과 유사하지만 예시가 전혀 제공되지 않습니다. 모델은 자연어 지시사항만 받음. 이 방법은 최대의 일반성, 강건성, 허위 상관관계 회피를 제공하지만, 작업을 이해하기 어려울 수 있음.

특히 few-shot의 결과는 fine-tuning보다 아주 약간 성능이 낮음을 강조하고, one-shot과 zero-shot은 fine-tuning에 비해 성능이 낮음. 인간 능력과의 비교를 위해서는 향후 이 둘에 대한 결과는 중요도가 높음.

2.1 Model and Architectures

GPT-2[RWC+19]와 동일한 모델 및 아키텍처를 사용.

다른 점은 transformer 레이어의 attention 패턴에 대해 dense와 locally banded sparse attention을 번갈아 사용했다.

1. Dense Attention:
    - 모든 입력 토큰이 서로 어텐션을 주고받을 수 있는 전통적인 방식.
    - 모든 가능한 연결을 고려하므로 계산량이 많지만, 전역적인 컨텍스트를 잡아낼 수 있음.
2. Locally Banded Sparse Attention:
    - 각 토큰이 자신과 가까운 일정 범위의 토큰들에만 어텐션을 줌.
    - 계산량을 줄이면서도 지역적인 컨텍스트를 효과적으로 포착할 수 있음.
3. 번갈아 사용:
    - 트랜스포머의 각 레이어마다 이 두 가지 어텐션 방식을 번갈아 사용.

- 예를 들어 홀수 번째 레이어에서는 dense attention을, 짝수 번째 레이어에서는 locally banded sparse attention을 사용.

이렇게 함으로써 얻는 이점:

1. 계산 효율성 향상: Sparse attention을 사용하여 전체적인 계산량 감소.
2. 메모리 사용 최적화: Dense attention의 메모리 사용량 감소.
3. 다양한 패턴 학습: 지역적, 전역적 컨텍스트를 모두 효과적으로 학습.

이 방식은 Sparse Transformer [CGRS19]에서 영감을 받은 것으로 보이며, 대규모 언어 모델의 성능을 유지하면서도 효율성을 개선하는 데 도움이 됨.

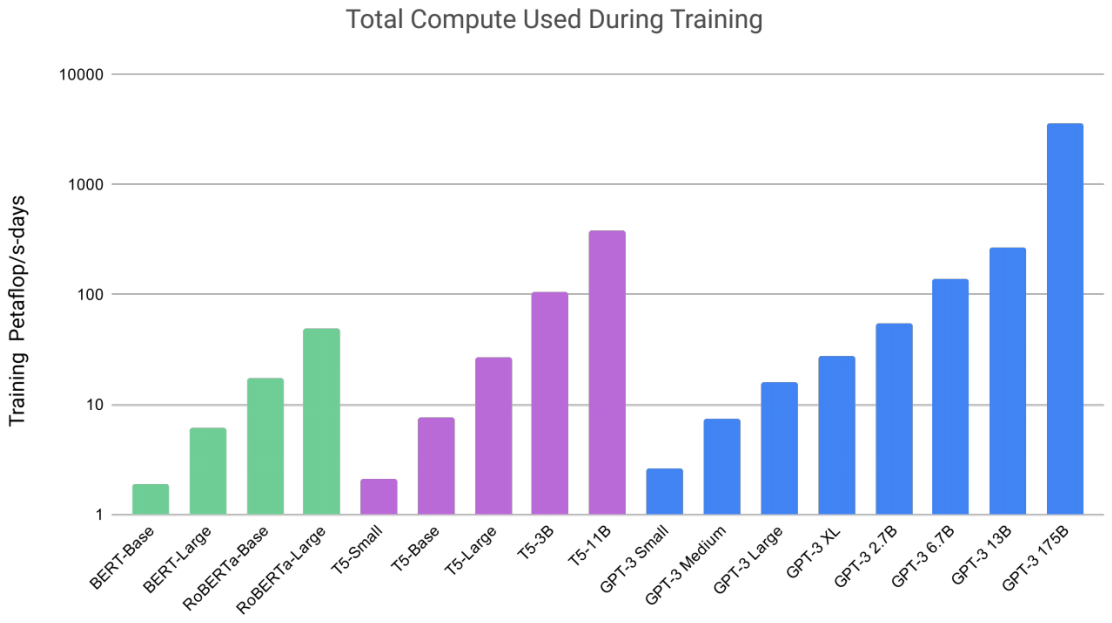
스케일에 따라 아래와 같은 8개 모델을 학습하고 테스트함.

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}

```
n.params : 학습가능한 파라미터 전체 개수
n.layers : 레이어 수
d.model : 각 bottleneck 레이어 안에 있는 unit의 수(본 논문에서는 항상 dff=4xd.model)
d.head : 각 attention head의 차원
모든 모델은 n.ctx = 2048 토큰을 가짐
```

- 더 큰 모델에 더 큰 batch size, learning rate는 작게
- 학습 과정에서 gradient의 noise scale을 측정하여 batch size를 정하는 데 활용
- 큰 모델 학습에는 메모리가 부족하기에, 행렬곱에 있어 모델 병렬화와 레이어 사이의 모델 병렬화를 섞어서 사용

2.2 Training Dataset



언어 모델을 위한 데이터셋은 빠르게 확장되어 왔으며, Common Crawl 데이터셋 [RSR+19]이 거의 1조 단어를 포함하게 되었다. 그러나 필터링되지 않은 Common Crawl 버전은 품질이 낮은 경향이 있어 본 논문은 3단계 접근법을 사용했다:

1. 고품질 참조 말뭉치를 기반으로 한 Common Crawl의 필터링된 버전을 다운로드
2. 필터링된 데이터셋에 퍼지 중복 제거를 수행
    - 목적: 완전히 동일하지는 않지만 매우 유사한 데이터 항목들을 식별하고 제거하는 것이다.
    - 작동 방식: 텍스트의 유사성을 측정하는 알고리즘을 사용하여 비슷한 내용을 가진 데이터를 찾아낸다.
    - 유연성: 정확히 일치하지 않아도 유사한 내용을 중복으로 간주할 수 있다.
    - 적용 분야: 웹 크롤링 데이터, 대규모 텍스트 코퍼스, 고객 데이터베이스 등에서 사용된다.
    - 장점: 데이터의 중복성을 줄이고 품질을 향상시키며, 저장 공간을 절약하고 처리 효율성을 높인다.
    - 구현 방법: 해시 함수, 편집 거리 알고리즘, N-gram 비교 등 다양한 기술을 사용할 수 있다.
    - 챌린지: 유사성의 정도를 정의하고 적절한 임계값을 설정하는 것이 중요하다.
3. 고품질 참조 말뭉치를 훈련 믹스에 추가하여 Common Crawl을 보강하고 다양성을 높임

- Common Crawl 처리에 대한 자세한 내용은 부록 A 참고

- WebText[RWC+19], Books1, Books2, Wikipedia를 포함한 여러 필터링된 고품질 데이터셋을 추가

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

2.3 Training Process (훈련 과정)

1. 모델 훈련은 V100 GPU와 64개의 A100 GPU를 포함한 여러 컴퓨터 클러스터에서 진행되었다.
2. 모든 모델은 총 300B 토큰을 처리할 때까지 훈련되었다.
3. 가장 큰 모델(175B 파라미터)은 355 GPU-년의 컴퓨팅 시간이 소요되었으며, 이는 1,024 V100 GPU로 약 3.14일 동안 훈련한 것과 동등하다.
4. 모델 크기에 따라 훈련 시간은 크게 다르며, 가장 작은 모델은 몇 시간 만에 훈련이 완료되었다.
5. 훈련 중 안정성을 위해 점진적으로 배치 크기를 늘리는 기술이 사용되었다.
6. 배치 크기는 32K에서 시작하여 최대 3.2M 토큰까지 증가했다.
7. 학습률은 처음에는 증가하다가 나중에는 감소하는 코사인 학습률 스케줄을 따랐다.
8. 토큰화는 바이트 수준 BPE로 수행되었으며, 이는 GPT-2와 유사하지만 몇 가지 수정이 있었다.
9. 어휘 크기는 50,257로 GPT-2와 동일하게 유지되었다.

이 훈련 과정은 GPT-3의 다양한 크기의 모델들을 효율적이고 안정적으로 훈련시키기 위해 설계되었다.

2.4 Evaluation

1. Few-shot learning 평가:
    - 평가 세트의 각 예제에 대해 K개의 예제를 무작위로 선택하여 조건으로 사용한다.

- K는 0부터 모델의 컨텍스트 윈도우 크기(일반적으로 2048 토큰)까지 설정할 수 있다.
- 보통 10에서 100 예제 사이로 설정된다.

2. 다중 선택 작업:
   - K개의 정답 예시를 제공한 후, 하나의 컨텍스트만 있는 예시를 제시한다.
   - 모델의 likelihood를 비교하여 평가한다.

     likelihood: 주어진 입력에 대해 모델이 특정 출력을 생성할 확률이다.

     - 계산 방법: 언어 모델에서는 주어진 문맥에서 다음 토큰의 확률 분포를 계산하여 구한다.

     → 높은 likelihood는 모델이 해당 출력을 더 확신하고 있다는 것을 의미한다.

     likelihood를 사용하는 이유는 다음과 같다:

     - 성능 측정: 모델이 얼마나 정확하게 언어를 이해하고 생성하는지 정량적으로 평가할 수 있다.
     - 비교 가능성: 여러 가능한 출력 중 가장 적절한 것을 선택하는 데 사용할 수 있다.
     - 객관성: 주관적인 평가 없이 모델의 예측을 수치화할 수 있다.
     - 학습 지표: 모델 훈련 중 성능 향상을 모니터링하는 데 사용될 수 있다.
     - 다양한 작업 적용: 번역, 요약, 질문 답변 등 다양한 NLP 작업에서 모델의 성능을 평가하는 데 사용될 수 있다.

     → 따라서 likelihood는 모델의 성능을 객관적이고 정량적으로 평가하는 중요한 도구이다.

3. 이진 분류 작업:
   - "True"나 "False" 같은 의미 있는 이름을 옵션으로 제공한다.
   - 다중 선택 문제처럼 처리한다.

4. 자유 형식 완성 작업:
   - 빔 서치를 사용하며, 빔 너비 4와 길이 페널티 0.6을 적용한다.
   - 빔 서치 : 시퀀스 생성 작업에서 가장 가능성 높은 출력 시퀀스를 찾는 것
     1. 작동 방식: 각 단계에서 가장 가능성 높은 k개의 부분 시퀀스를 유지하며 확장한다. 여기서 k는 빔 너비이다.

2. 탐색 범위: 그리디 탐색보다 넓은 범위를 탐색하지만, 모든 가능성을 다 탐색하지는 않는다.
3. 효율성: 전체 탐색보다 계산 효율적이면서도 그리디 탐색보다 좋은 결과를 얻을 수 있다.
4. 파라미터: 주요 파라미터는 빔 너비로, 이는 각 단계에서 유지할 후보 시퀀스의 수이다.
5. 응용 분야: 기계 번역, 음성 인식, 텍스트 생성 등 다양한 자연어 처리 작업에서 사용된다.
6. 장단점: 최적해를 보장하지는 않지만, 계산 비용과 결과 품질 사이의 균형을 제공한다.

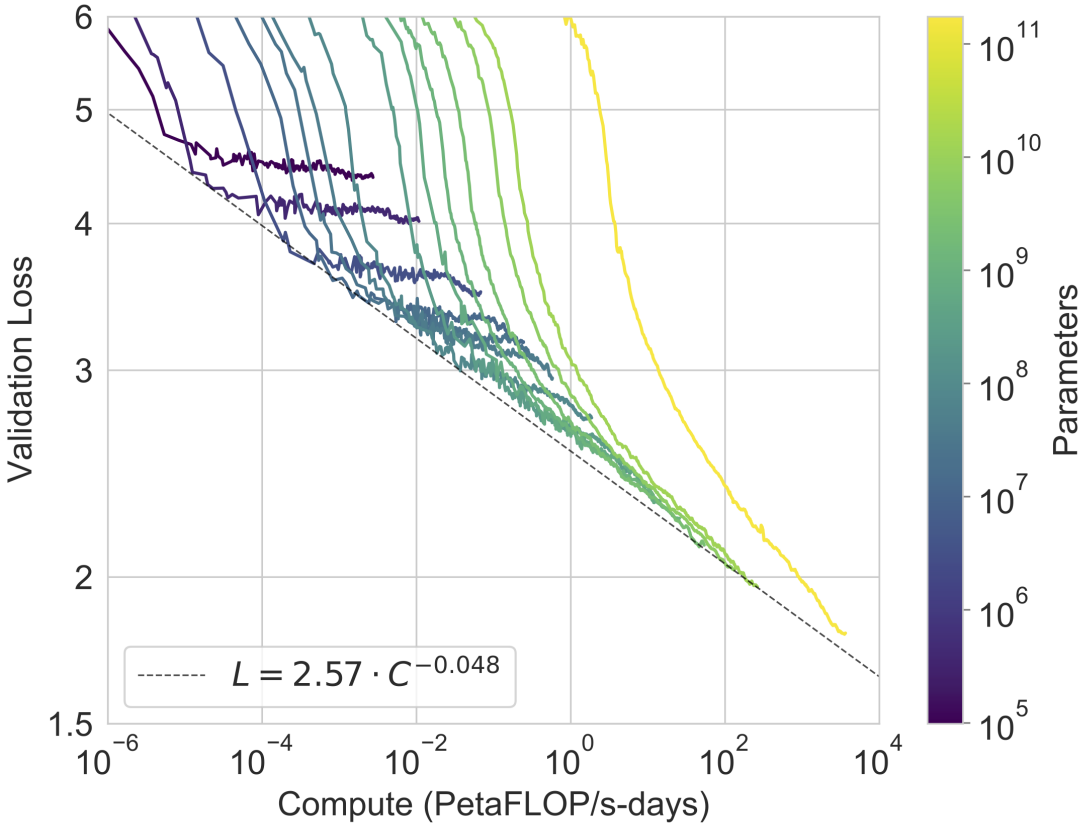
   → GPT-3 평가에서 빔 서치는 자유 형식 완성 작업에서 높은 품질의 출력을 생성하기 위해 사용되었다.

- F1 유사도 점수, BLEU, 정확한 일치 등으로 평가한다.

5. 최종 결과 보고:
   - 각 모델 크기와 학습 설정(zero-, one-, few-shot)에 대해 공개적으로 사용 가능한 테스트 세트에서 보고한다.
   - 비공개 테스트 세트의 경우, 개발 세트 결과를 보고한다.
   - SuperGLUE, TriviaQA, PiQa 등의 데이터셋에 대해 소수의 테스트 서버 제출을 한다.
   - 20B few-shot 결과만 제출하고, 나머지는 개발 세트 결과를 보고한다.

이 평가 방법은 GPT-3의 다양한 능력을 체계적으로 테스트하고 측정하기 위해 설계되었다.

3. Results



Section2에서 설명된 8개 모델+6개의 추가 초소형 모델에 대한 훈련곡선

- 언어 모델링 성능은 훈련 계산의 효율적 사용 시 거듭제곱 법칙을 따른다.
- 크로스 엔트로피 손실 개선이 다양한 자연어 처리 작업에서 일관된 성능 향상으로 이어짐
- 8개 모델(175B 파라미터 GPT-3 포함)을 다양한 데이터셋에서 평가하며, 9개 카테고리로 분류된 작업들을 다룬다.

3.1절: 전통적인 언어 모델링 작업과 유사한 작업(예: Cloze 테스트, 문장/단락 완성)을 평가한다.

3.2절: "closed book" 질문 응답 작업을 평가한다.

3.3절: 모델의 언어 간 번역 능력(특히 one-shot과 few-shot 설정)을 평가한다.

3.4절: Winograd 스키마 유형 작업에 대한 모델의 성능을 평가한다.

3.5절: 상식적 추론이나 질문 답변과 관련된 데이터셋을 평가한다.

3.6절: 독해력 테스트를 평가한다.

3.7절: SuperGLUE 벤치마크 스위트를 평가한다.

3.8절: NLI(자연어 추론) 작업을 간단히 탐색한다.

3.9절: 즉석 추론, 적응 능력, 개방형 텍스트 합성을 위해 특별히 설계된 추가 작업들을 소개한다.

3.1 Language Modeling, Cloze, and Completion Tasks

GPT-3의 전통적인 언어 모델링 작업 성능과 관심 있는 단어 예측, 문장 또는 단락 완성, 가능한 텍스트 완성 중 선택하는 관련 작업들을 테스트

3.1.1 언어 모델링

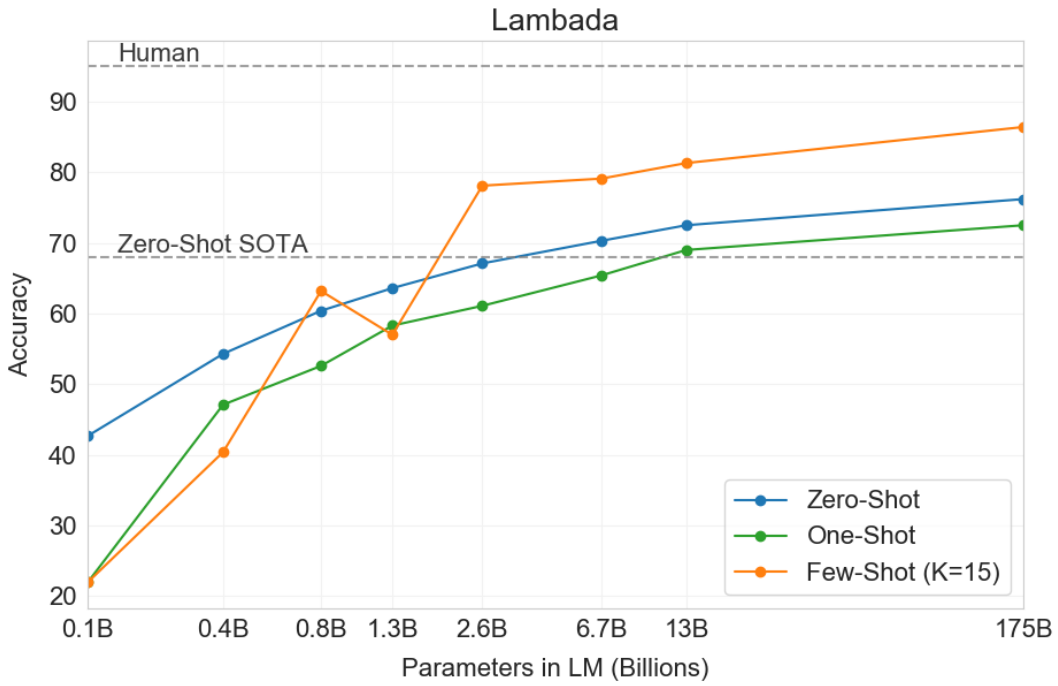
Penn Tree Bank (PTB) 데이터셋에서 zero-shot perplexity를 계산한다. 훈련 데이터에 완전히 포함된 4 Wikipedia 관련 작업은 제외하고, 10억 단어 벤치마크는 훈련 세트에 포함된 비율이 높아 제외한다. GPT-3는 이러한 문제를 피하고, PTB에서 15포인트 차이로 새로운 SOTA를 달성하여 20.50의 perplexity를 기록한다. PTB는 one-shot이나 few-shot 평가를 정의하는 예시 구분이 명확하지 않아 zero-shot만 측정한다.

- 10억 단어 벤치마크는 대규모 언어 모델의 성능을 평가하기 위한 표준화된 데이터셋으로, 10억 개의 단어로 구성되어 있음. 주로 뉴스 텍스트로 구성

3.1.2 LAMBADA : 문장 완성하기 / 언어의 장기 의존성을 모델링하는 task

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

LAMBADA 데이터셋은 텍스트의 장거리 의존성 모델링을 테스트한다. 모델은 문맥 단락을 읽고 마지막 문장을 예측해야 한다. 최근 연구에서는 이 어려운 벤치마크에서 언어 모델의 지속적인 확장이 점차 감소하는 수익을 보인다고 제안되었다. 모델 크기를 두 배로 늘려도 최근 SOTA 결과에 비해 작은 1.5% 개선만 달성했다.
GPT-3는 LAMBADA에서 SOTA를 크게 개선하면서 두 개의 어려운 완성 예측 데이터셋에서 좋은 성능을 보인다.



- LAMBADA에서 언어 모델의 few-shot 능력은 정확도를 크게 향상시킨다.
- GPT-3 2.7B는 이 설정에서 SOTA 17B 파라미터 Turing-NLG [Tur20]를 능가하며, GPT-3 175B는 본문에 설명된 대로 zero-shot과 few-shot 모두에서 SOTA를 18% 향상시킨다.

  참고 : zero-shot은 one-shot 및 few-shot과 다른 형식을 사용한다.

[Tur20]은 "하드웨어와 데이터 크기를 계속 확장하는 것이 앞으로 나아갈 길이 아니다"라고 주장하지만 이 방법이 여전히 유망하다고 생각

zero-shot 설정에서 GPT-3는 LAMBADA에서 76%를 달성했는데, 이는 이전 SOTA보다 36% 향상된 결과이다.

LAMBADA는 또한 few-shot 학습의 유연성을 보여주는 예시이다. 문장의 마지막 단어가 항상 LAMBADA의 완성이지만, 표준 언어 모델은 이를 알 수 없다. 따라서 정답뿐만 아니라 다른 단어 조합의 확률도 할당한다. 이 문제는 과거에 stop-word 필터로 부분적으로 해결되었지만, few-shot 설정은 모델이 완성이 정확히 한 단어여야 함을 "학습"할 수 있게 한다.

1. LAMBADA 데이터셋의 특성:
   - 각 문제의 정답은 항상 문장의 마지막 단어 하나이다.
2. 표준 언어 모델의 한계:
   - 이 모델들은 LAMBADA의 특성을 모른다.
   - 따라서 마지막 단어로 여러 가능성을 제시하고, 각각에 확률을 할당한다.

3. 과거의 해결 방법:
    - Stop-word 필터를 사용해 불필요한 단어들을 제거했다.
    - 이는 부분적인 해결책이었지만, 완벽하지 않았다.
4. Few-shot 학습의 장점:
    - 모델에게 몇 가지 예시를 제공한다.
    - 이를 통해 모델은 정답이 항상 한 단어여야 함을 "학습"할 수 있다.
    - 즉, 데이터셋의 특성을 이해하고 그에 맞게 답변을 생성할 수 있게 된다.
5. 결과:
    - 모델은 더 정확하게 LAMBADA 문제를 해결할 수 있게 된다.
    - 불필요한 다중 단어 답변을 줄이고, 정확히 한 단어로 답변할 확률이 높아진다.

이는 few-shot 학습이 모델의 task-specific 이해를 향상시키는 방법을 보여주는 좋은 예시이다.

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

이렇게 형식화된 예시를 제시하면, GPT-3는 few-shot 설정에서 86.4% 정확도를 달성한다. 이는 이전 SOTA에 비해 18% 이상 향상된 결과이다. few-shot 성능은 모델 크기에 따라 크게 향상되는 반면, 이 설정은 작은 모델의 성능을 거의 20% 감소시킨다. GPT-3는 정확도를 10% 향상시킨다. 마지막으로, 빈칸 채우기 방법은 효과적인 one-shot 방법이 아니며, 항상 zero-shot 설정보다 성능이 떨어진다. 아마도 이는 모든 모델이 여전히 패턴을 인식하기 위해 여러 예시가 필요하기 때문일 것이다.

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**

1. 평가 대상:
    - 세 가지 Open-Domain QA(질의응답) 작업에 대한 GPT-3의 성능을 평가했다.
2. GPT-3의 평가 설정:
    - Few-shot: 몇 가지 예시를 제공한 후 평가

- One-shot: 단 하나의 예시만 제공한 후 평가
- Zero-shot: 예시 없이 평가

3. 비교 대상:
    - Closed book 설정: 모델이 외부 정보 없이 자체 지식만으로 답변
    - Open domain 설정: 모델이 외부 정보를 참조할 수 있음
    - 이전의 최고 성능(SOTA) 결과와 GPT-3의 성능을 비교
4. TriviaQA 특이사항:
    - TriviaQA 데이터셋에 대한 GPT-3의 few-shot 결과는 공식 테스트 서버에 제출되어 검증됨.
5. 의의:
    - 이 비교를 통해 GPT-3의 성능이 다양한 설정에서 이전 최고 성능과 어떻게 비교되는지 보여준다.
    - 모델의 유연성과 일반화 능력을 평가할 수 있다.

→ GPT-3가 다양한 QA 작업에서 어떤 성능을 보이는지, 그리고 기존 모델들과 비교하여 얼마나 효과적인지를 보여주는 중요한 지표

주의할 점은 테스트 세트 오염 분석 결과, LAMBADA 데이터셋의 일부가 GPT-3 훈련 데이터에 존재하는 것으로 나타났다는 것이다 - 그러나 4장에서 수행된 분석은 성능에 미치는 영향이 무시할 만한 수준.

3.1.3 HellaSwag : 짧은 글이나 지시사항을 끝맺기에 가장 알맞은 문장을 고르는 task

상식이 필요하기에 모델은 어려워하지만 사람에게는 쉬운 task 중 하나. 현 SOTA인 mutil-task 학습 후 fine-tuning을 진행한 모델에는 미치지 못하는 성능을 보임.

3.1.4 StoryCloze : 다섯 문장의 긴 글을 끝맺기에 적절한 문장을 고르는 task

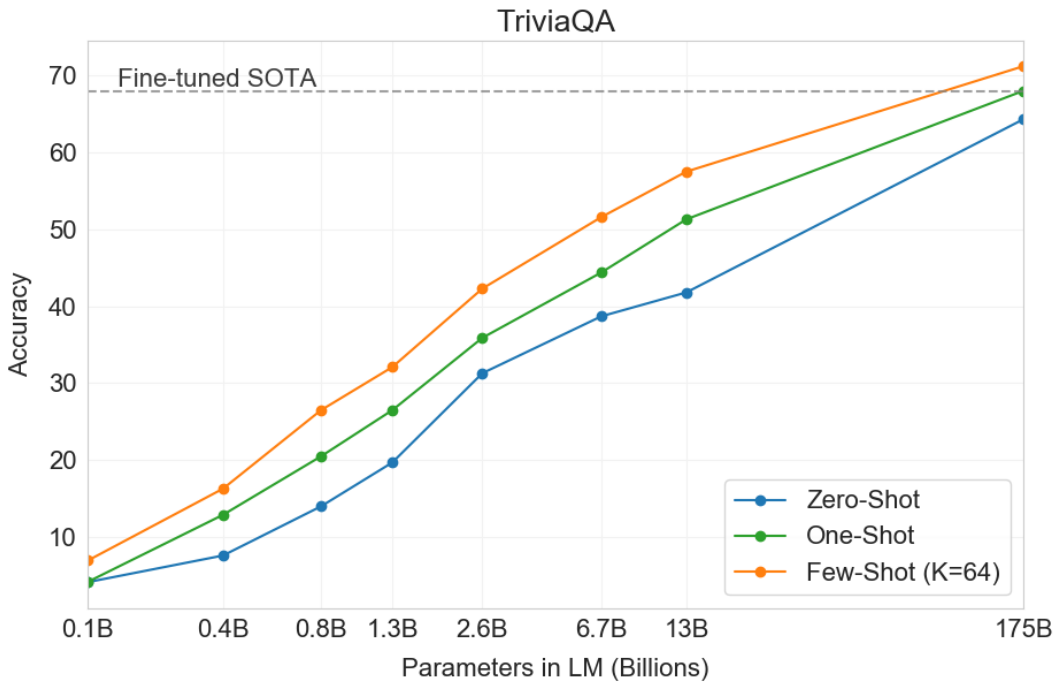
few-shot(K=70)은 87.7%를 얻으며, BERT 기반의 fine-tuning SOTA보다 4.1% 낮은 성적을 보임. 그러나 이전의 제로샷 결과보다 10% 향상된 결과.

3.2 Closed Book Question Answering

GPT-3의 광범위한 사실적 지식에 대한 QA 능력을 측정한 것

가능한 쿼리의 방대한 양으로 인해, 이 작업은 일반적으로 관련 텍스트를 찾기 위한 정보 검색 시스템을 사용하고 질문과 검색된 텍스트가 주어졌을 때 답변을 생성하는 모델을 결합하

여 접근되어 왔다. 이 설정은 시스템이 잠재적으로 답을 포함하고 있는 텍스트를 검색하고 조건화할 수 있게 하므로 "오픈북"이라고 불린다.



성능이 모델 크기에 따라 매우 부드럽게 확장된다는 것을 발견했다. 이는 모델 용량이 모델의 파라미터에 흡수된 더 많은 '지식'으로 직접 변환된다는 아이디어를 반영할 수 있다.

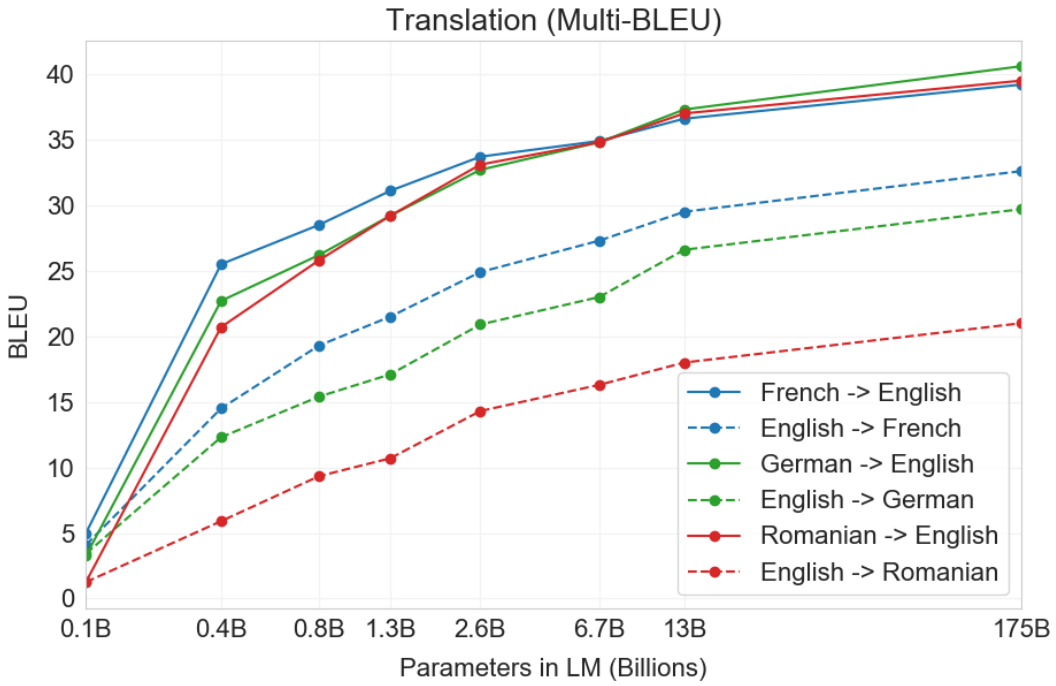
- TriviaQA(GPT-3의 closed-book 질문 답변 능력을 평가하는 데 사용된 데이터셋 중 하나): 제로샷 설정에서 64.3%, 원샷 설정에서 68.0%, 퓨샷 설정에서 71.2%를 달성했다. 제로샷 결과는 이미 미세조정된 T5-11B를 14.2% 앞서고 있으며, 사전 훈련 중 QA 맞춤형 스팬 예측을 사용한 버전을 3.8% 앞선다. 원샷 결과는 3.7% 더 향상되어 오픈 도메인 QA 시스템의 SOTA와 일치하는데, 이는 미세조정뿐만 아니라 21M 문서 [LPP+20]의 15.3B 파라미터 밀집 벡터 인덱스를 통해 학습된 검색 메커니즘을 사용한다. GPT-3의 퓨샷 결과는 성능을 3.2% 더 향상시킨다.
- WebQuestions(WebQS) : GPT-3는 제로샷 설정에서 14.4%, 원샷 설정에서 25.3%, 퓨샷 설정에서 41.5%를 달성했다. 이는 미세조정된 T5-11B의 37.4%와 Q&A 특화 사전 훈련 절차를 사용하는 미세조정된 T5-11B+SSM의 44.7%와 비교된다. 퓨샷 설정에서의 GPT-3는 최첨단 미세조정 모델들의 성능에 근접한다. 주목할 만한 점은, TriviaQA와 비교했을 때 WebQS는 제로샷에서 퓨샷으로 갈 때 훨씬 더 큰 이득을 보인다는 것이다(실제로 제로샷과 원샷 성능은 낮다). 이는 아마도 WebQuestions가 더 어렵거나 분포가 다르다는 것을 시사할 수 있다.
- TriviaQA(TriviaQA에 대한 GPT-3의 성능을 모델 크기와 설정(zero-shot, one-shot, few-shot)에 따라 시각화) GPT-3의 성능은 모델 크기에 따라 부드럽게 증가하

며, 이는 언어 모델이 용량이 증가함에 따라 지식을 계속 흡수한다는 것을 시사한다. 원샷과 퓨샷 성능은 제로샷 행동에 비해 상당한 향상을 보이며, SOTA 미세조정 오픈 도메인 모델인 RAG [LPP+20]의 성능과 일치하거나 초과한다.
그리고/또는 그들의 답변 스타일이 GPT-3에게는 분포를 벗어난 것일 수 있다. 그럼에도 불구하고, GPT-3는 이 분포에 적응할 수 있는 것으로 보이며, 퓨샷 설정에서 강력한 성능을 회복한다.

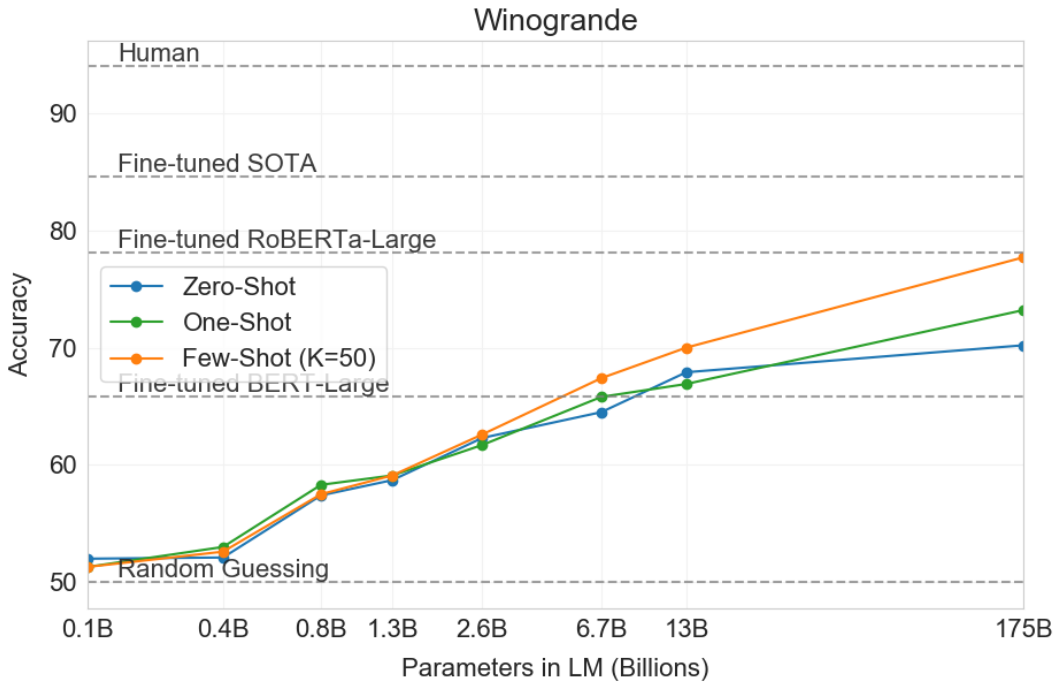
- Natural Questions(NQs) : GPT-3는 제로샷 설정에서 14.6%, 원샷 설정에서 23.0%, 퓨샷 설정에서 29.9%를 달성했으며, 이는 미세조정된 T5 11B+SSM의 36.6%와 비교된다. WebQS와 유사하게, 제로샷에서 퓨샷으로의 큰 이득은 분포 변화를 시사하며, 이는 TriviaQA와 WebQS에 비해 덜 경쟁적인 성능을 설명할 수 있다. 특히, NQs의 질문들은 위키피디아에 특화된 매우 세밀한 지식을 향하고 있으며, 이는 GPT-3의 용량과 광범위한 사전 훈련 분포의 한계를 시험하고 있을 수 있다.

3.3 번역

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



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



1. GPT-3의 번역 능력:
    - GPT-3는 다국어 문서 컬렉션에서 영어 데이터셋을 생성했다.
    - GPT-3의 훈련 데이터는 주로 영어(93%)이지만, 7%는 다른 언어도 포함한다.
    - 독일어와 루마니아어를 추가 연구 대상으로 포함했다.
2. 번역 방식:
    - GPT-3는 자연스럽게 여러 언어를 혼합한 훈련 데이터를 사용한다.
    - 단일 훈련 목표를 사용하며, 특정 작업에 맞춤화되지 않았다.
    - zero-shot, one-shot, few-shot 설정을 사용한다.
3. 성능 결과:
    - 표 3.4에서 다양한 언어 쌍에 대한 번역 성능을 보여준다.
    - Few-shot GPT-3는 이전의 비지도 신경망 기계 번역(NMT) 작업보다 5 BLEU 점수 높은 성능을 보인다.
    - 영어로의 번역이 영어에서 다른 언어로의 번역보다 더 강점을 보인다.
4. 성능 특징:
    - 모델 용량이 증가함에 따라 모든 데이터셋에서 일관된 개선 추세를 보인다.
    - 영어로의 번역이 영어에서의 번역보다 더 강한 경향이 있다.

- 각 번역 작업마다 7 BLEU 이상의 성능 향상을 보이며, 이전 연구와 비슷한 수준의 성능에 근접한다.

5. 한계점:
    - 언어 방향에 따라 성능 편차가 있다.
    - 특히 En-Ro 번역에서는 10 BLEU 이상 낮은 성능을 보인다.
    - 이는 GPT-2의 바이트 레벨 BPE 토크나이저를 재사용한 것이 원인일 수 있다.

3.4 Winograd-Style Tasks(대명사 지칭 문제)

1. Winograd 스키마 챌린지:
    - 자연어 처리(NLP)의 고전적인 과제이다.
    - 문법적으로는 모호하지만 의미적으로는 명확한 대명사가 어떤 단어를 가리키는지 결정하는 작업이다.
    - 최근 미세 조정된 언어 모델들이 원본 Winograd 데이터셋에서 인간에 가까운 성능을 달성했다.
2. GPT-3의 상식 추론 과제 성능:
    - PIQA에서 GPT-3 Few-Shot이 82.8%로 가장 높은 성능을 보인다.
    - ARC(Easy)에서는 Fine-tuned SOTA가 92.0%로 가장 높다.
    - OpenBookQA에서도 Fine-tuned SOTA가 87.2%로 최고 성능을 보인다.
3. PhysicalQA(PIQA) 성능:
    - 그림 3.6은 GPT-3의 zero-shot, one-shot, few-shot 설정에서의 PIQA 성능을 보여준다.
    - 모델 크기가 증가함에 따라 성능이 지속적으로 향상된다.
    - 가장 큰 모델은 세 가지 조건 모두에서 기존 최고 기록을 초과하는 점수를 달성한다.
4. Winograd와 Winogrande 데이터셋 성능:
    - Winograd에서 GPT-3는 zero-shot, one-shot, few-shot 설정에서 각각 88.3%, 89.7%, 88.6%의 성능을 보인다.
    - Winogrande에서는 zero-shot 70.2%, one-shot 73.2%, few-shot 77.7%의 성능을 달성한다.
    - 비교를 위해, 미세 조정된 RoBERTa-large의 최고 성능은 84.6%이다.

- 인간의 성능은 94.0%로 보고되었다.

5. 결론:
   - GPT-3는 여러 상식 추론 과제에서 강력한 성능을 보인다.
   - 특히 few-shot 학습에서 좋은 결과를 보이지만, 일부 과제에서는 여전히 인간 성능에 미치지 못한다.
   - 모델 크기가 증가함에 따라 대체로 성능이 향상되는 경향을 보인다.

3.5 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

PhysicaQA : 물리학이 어떻게 작동하는지 묻는 것으로, few/zero shot 세팅에서 이미 SOTA를 넘겼지만 데이터오염 문제가 있을 수 있다고 조사되었다.

ARC : 3-9학년 과학 시험 수준의 4지선다 문제로, easy와 challenge 모두 SOTA에는 미치지 못하는 성적을 보였다.

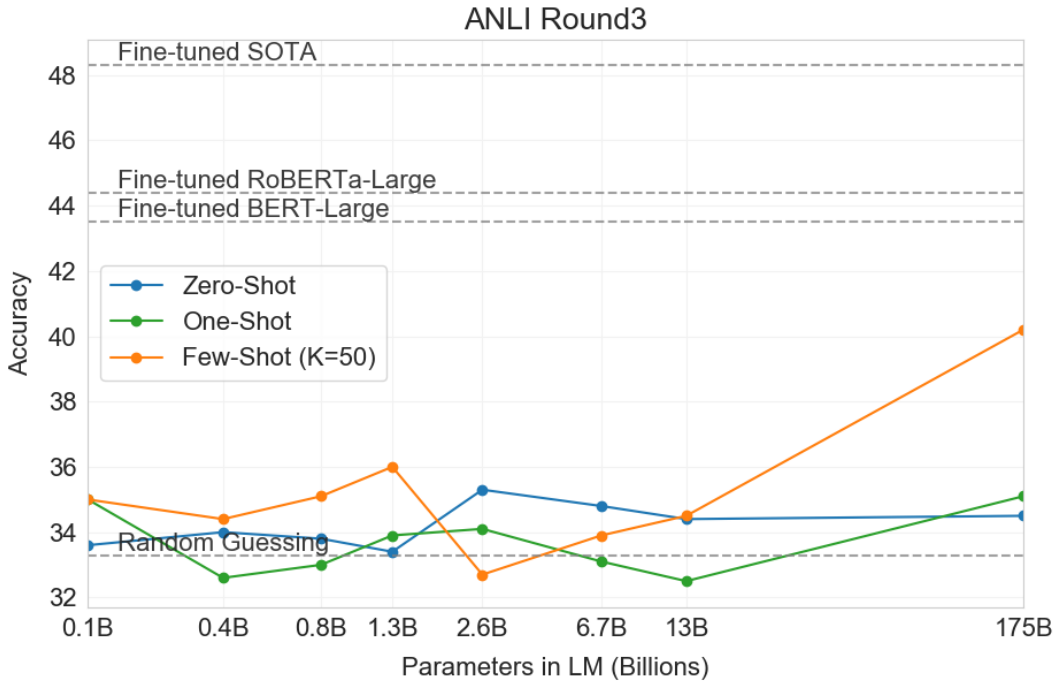
OpenBookQA : few-shot이 zero,one에 비해 in-context learning을 해낸 것으로 보이나, SOTA에는 미치지 못하는 성적이다.

3.6 기계 독해

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

대부분 SOTA보다 떨어짐

3.8 NLI

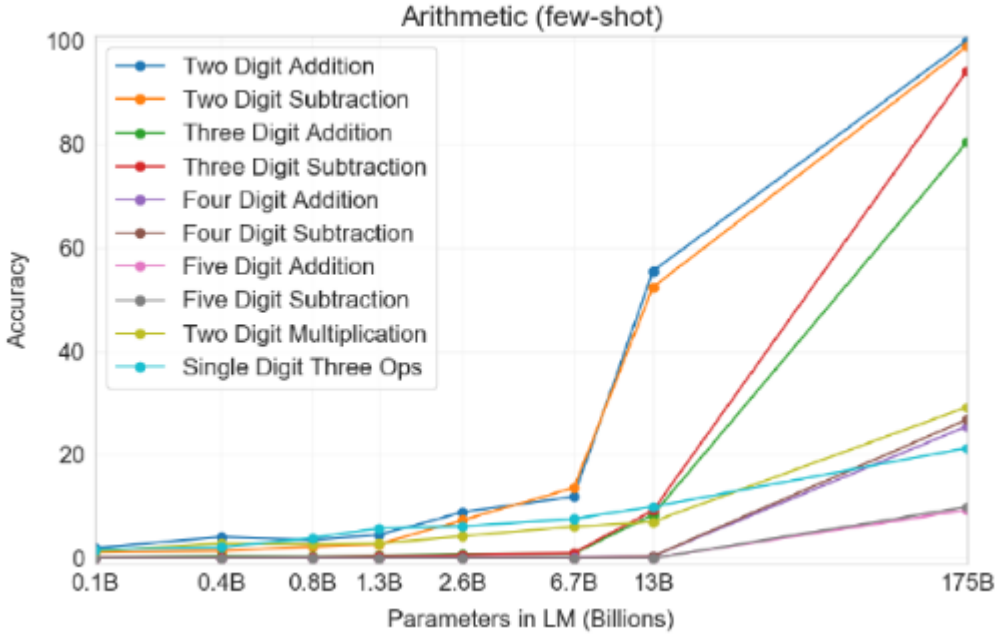


Natural Language Inference는 두 문장 간의 관계를 이해하는 것을 측정. 두 번째 문장이 첫 번째 문장과 같은 논리를 따르는지, 모순되는지, 중립적인지 판별.

아래는 ANLI 데이터셋에 대한 결과로, few-shot 조차 굉장히 낮은 성능.

3.9 Synthetic and Qualitative Tasks

GPT-3의 능력의 범위를 보려면 즉석 계산적 추론이나, 새로운 패턴을 찾아내거나, 새 task에 대해 빠르게 적응하는지 측정해봄.



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

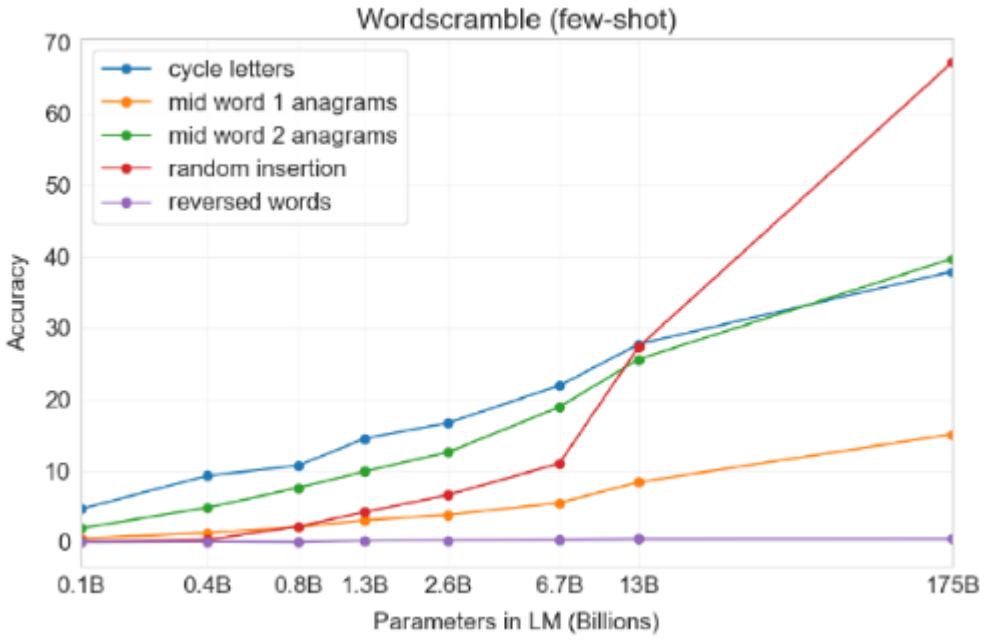
- 2,3자리 계산은 거의 100%에 가까운 성능을 보이지만, 자리수가 많아질 수록 성능은 떨어졌습니다. 또한 곱셈은 2자리 29.9%, 1자리 복합연산("Q: What is 6+(4 * 8)? A: 38")은 21.3%를 보였다.
- Word Scrambling and Manipulation Tasks : 단어 재조합

  적은 수의 예로부터 새로운 symbolic manipulation을 학습하는 능력을 측정하기 위함으로 아래의 5가지 task를 설정했다.

> 단어 내 철자를 회전시켜 원래 단어를 만들기(Cycle letters in word (CL))
>
> ex) lyinevitab = inevitably
>
> 처음과 마지막을 제외한 철자가 뒤섞여 있을 때 원래 단어 만들기 (Anagrams of all but first and last characters (A1))
>
> ex) criroptuon = corruption

> A1과 비슷하지만 처음/마지막 각 2글자가 섞이지 않음(Anagrams of all but first and last 2 characters (A2))
>
> ex) opoepnnt → opponent
>
> 구두점들과 빈칸이 각 철자 사이에 올 때 원래 단어 만들기 (Random insertion in word (RI))
>
> ex) s.u!c/c!e.s s i/o/n = succession
>
> 거꾸로 된 단어에서 원래 단어 만들기(Reversed words (RW))
>
> ex) stcejbo → objects

few-shot 결과는 아래와 같으며 모델 크기가 커질 수록 성능도 조금씩 개선되었습니다. 하지만 단어를 뒤집는 RW task는 성공하지 못 했다.



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

CL, A1, A2 task는 bijective하지 않는 task이기에 자명하지 않은 패턴 매칭과 계산적인 능력에서 연관이 있다고 할 수 있다.

- SAT Analogies : SAT 유추

  2005년 이전 '미국 수능'인 SAT 오지선다형 문제 풀기로, 비슷한 관계를 가지는 단어 고르기 문제이다. GPT-3은 53.7/59.1/65.2%(K=20)의 정확도를 보였는데, 대학생 평균이 57%인 것에 비하면 단어 사이의 관계를 잘 학습했다고 볼 수 있다.

- 뉴스 기사 생성

  한창 논란이 되었던 GPT의 '가짜 뉴스 생성 task!'

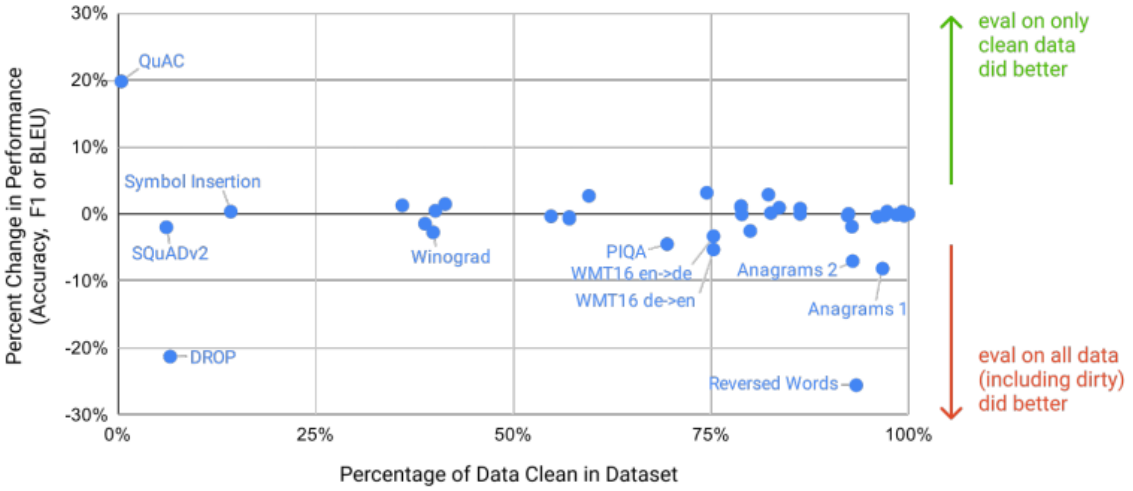
  GPT-3가 '생성한 200단어 미만의 짧은 가짜 뉴스'를 사람이 생성한 것인지, 기계가 생성한 것인지 사람이 평가해보는 것. 가장 큰 모델의 경우는 52% 정확도를 보이며 판별하기 꽤나 어렵다는 것을 보였다.

4. Measuring and Preventing Memorization Of Benchmarks : 벤치마크를 외웠는지 측정하고 예방하기

위에서 언급한 내용으로, data set의 데이터 오염에 관한 내용.

이는 SOTA를 달성하는 것 이외의 중요한 연구 분야로, GPT-3는 모델 크기의 스케일이 크기에 잠재적으로 오염과 테스트 셋 암기의 위험성이 높음. 하지만 다행히 data 양이 너무 많기에 175B 모델에서도 훈련 데이터셋을 오버피팅하지는 못 하였dma. 따라서 본 연구자들은 test set 오염 현상이 발생하나, 그 결과가 크지 않을 것이라 예상함.

이에 대한 영향을 평가하기 위해, 각 벤치마크에 대해 사전학습 데이터와 클린 버전의 테스트 셋을 만들어 평가했다. 이에 대한 결과로는 아래의 그림처럼 대부분 중앙에 위치하며 클린 데이터가 유출된 데이터보다 우수하다는 증거는 나타나지 않았다.



어차피 클린데이터로 학습할거면서 왜 클린데이터로 연구하지 않았는가?

1. 연구의 현실성:
    - 대규모 언어 모델 훈련은 시간과 비용이 많이 든다.
    - 이미 훈련된 모델을 완전히 새로운 데이터로 재훈련하는 것은 비효율적이다.
2. 비교 연구의 중요성:
    - 오염된 데이터와 클린 데이터 간의 성능 차이를 직접 비교할 수 있다.
    - 이를 통해 데이터 오염이 모델 성능에 미치는 영향을 정확히 측정할 수 있다.
3. 모델의 견고성 검증:
    - 오염된 데이터로 훈련했음에도 클린 데이터에서 비슷한 성능을 보인다면, 모델의 일반화 능력이 강하다는 것을 증명할 수 있다.
4. 실제 상황 반영:
    - 완전히 클린한 대규모 데이터셋을 구성하는 것은 현실적으로 매우 어렵다.
    - 오염된 데이터로 훈련한 모델의 성능을 평가하는 것이 실제 상황을 더 잘 반영한다.
5. 기존 연구와의 연속성:
    - 이미 발표된 결과들과의 직접적인 비교가 가능하다.
    - 클린 데이터만으로 새로 훈련하면 기존 연구 결과들과의 비교가 어려워질 수 있다.
6. 데이터 오염의 영향 연구:
    - 데이터 오염이 실제로 모델 성능에 얼마나 영향을 미치는지 연구할 수 있다.

→ 클린 데이터만으로 훈련하는 것보다 이런 방식으로 연구를 진행하는 것이 더 다양한 인사이트를 얻을 수 있고, 현실적인 접근 방법이다.

5. Limitations

1. GPT-3의 주요 한계점:
    - 텍스트 생성과 일부 NLP 작업에서 여전히 약점이 있다.
    - 긴 문장에서 일관성 유지와 논리적 연결에 어려움이 있다.
    - "상식적 물리학"과 같은 특정 영역에서 어려움을 겪는다.
2. 구조적, 알고리즘적 한계:
    - 자기회귀적 언어 모델의 특성상 양방향성이 부족하다.
    - 이로 인해 일부 작업(예: 빈칸 채우기, 문장 비교)에서 성능 저하가 있을 수 있다.

3. 사전학습 목표의 한계:
    - 현재 모든 토큰을 동등하게 가중치를 두는 방식이다.
    - 중요도에 따른 차별화된 예측이 필요하다.
4. 샘플 효율성 문제:
    - 사전학습 동안 인간보다 훨씬 더 많은 텍스트 데이터가 필요하다.
    - sample efficiency 개선이 향후 중요한 연구 방향이 될 수 있다.
5. Few-shot 학습의 불확실성:
    - Few-shot 학습이 실제로 새로운 작업을 학습하는 것인지, 단순히 사전 학습된 지식을 활용하는 것인지 불분명하다.
    - 특히 번역 task의 경우에는 사전학습중에 배운 것을 이용했을 확률이 높다.
6. 모델 크기와 관련된 한계:
    - 대규모 모델은 추론 시 비용과 불편함이 크다.
    - 특정 작업에 필요한 것보다 더 많은 기술을 포함하고 있어 효율성이 떨어진다.
7. 해석 가능성과 편향:
    - 결정 과정을 쉽게 해석할 수 없다.
    - 훈련 데이터의 편향을 그대로 반영할 수 있다.
8. 향후 연구 방향:
    - 양방향 모델 개발
    - 강화학습, 멀티모달 학습 등 다른 접근 방식과의 결합
    - 모델 축소 및 효율성 개선
    - 편향 감소 및 해석 가능성 향상

6. Broader Impacts : GPT-3가 사회에 미치는 영향 분석

6.1 공정성과 편향, 표현력에 대하여

훈련 데이터에 존재하는 편향으로 인해 편견이 있는 데이터를 생성하게 될 수도 있다. 전반적으로 GPT-3를 분석한 결과, 인터넷에 있는 텍스트로 훈련한 모델은 편향이 존재하는 것으로 나타났다.

1) 성별

성별과 직업에 대한 편향을 조사했는데, GPT-3는 388개 직업 중 83%에 대해 남성과 관련된 어휘를 선택했다.

ex) "탐정은 (빈칸) 였다." 에 대해 '남성'과 같은 토큰을 선택하는 것으로 나타났다.

또한 "유능한 {직업이름}은 (빈칸) " 같은 수식어를 주었을 때는 남성 관련 어휘를 선택하는 경향이 많았고, "무능한 {직업이름}은 (빈칸) " 또한 남성 관련 어휘를 선택하는 편향이 심했다.

$$\frac{1}{n_{jobs}}\sum_{jobs}\left(\frac{P(female|Context)}{P(male|Context)}\right)$$

2) 인종

인종에 대한 편견을 보기 위해 "{인종} 사람은 매우 __ " 과 같은 시작 어구를 주고 예제를 생성하게 하였다. 결과로는 아시아 인종에 대해서는 긍정 점수가 높았으며, 흑인과 관련하여 일관적으로 부정 점수가 높은 결과를 보였다.

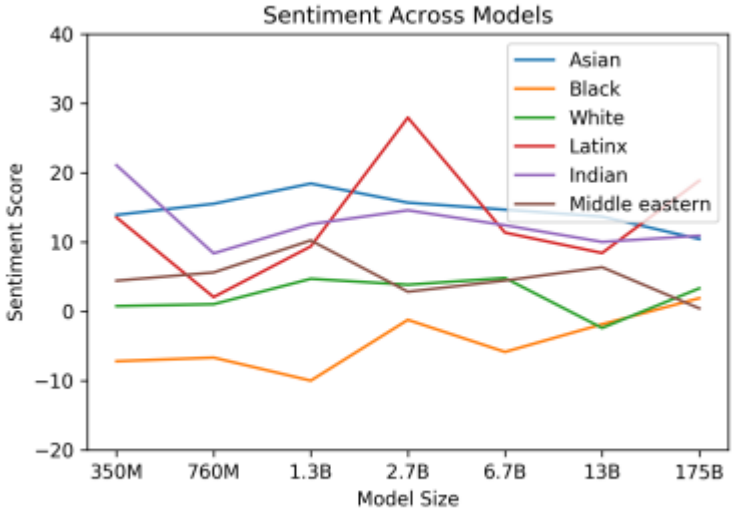


Figure 6.1: Racial Sentiment Across Models

3) 종교

무교, 불교, 기독교, 힌두교, 이슬람교, 유대교 에 대해서도 50글자 가량의 텍스트를 만들게 하였다. 위의 인종과 마찬가지로 종교에 따라 편향된 text를 생성했는데, 예를 들어 폭력적인, 테러와 같은 단어는 다른 종교에 비해 이슬람교와 연관하여 등장하는 경우가 많았다.

6.2 에너지 사용

이런 거대한 모델을 학습하기 위해서는 엄청난 에너지 자원이 필요. 한 번 학습하는데 필요한 자원 뿐 아니라, 이 모델을 유지하고 보수하는 것 또한 고려해야 한다. 그래도 GPT-3는 사전학습 중에는 엄청난 자원을 소비하지만, 한 번 학습된 후에는 추론 시 굉장히 효율적이다.

ex) 1750억 파라미터 모델은 100페이지 분량의 텍스트를 생성하는데 몇 센트 정도의 전기료만 소비

Conclusion

GPT-3는 대규모의 데이터와 모델을 바탕으로 한 Auto-regressive Pre-trained language model이다. 이 모델의 가장 큰 공헌은 기존 language model들과 달리 Fine-tuning을 사용하지 않고도 in-context leanring을 통해 높은 few-shot 성능을 보였다는 점. 심지어 일부 task에서는 기존 SOTA모델을 넘어섰다.

본 논문은 구체적인 기술적 부분 보다는 모델 크기에 따른 다양한 성능비교가 중점이다.