Language Models are Few-Shot Learners

전민하, 최수빈

1. 소개 - 연구 배경

- **기존 NLP 발전 흐름:**
  - Word2Vec → RNN → Transformer
  - 점점 **task-agnostic** 구조로 발전

- **문제점**
  - 각 작업마다 **task-specific, fine-tuning 요구 -> 편향 위험, 비효율성**
  - **수천~수만 개의 라벨 데이터** 요구 - > **일반화 성능 부족**

- **인간의 경우**
  - 단 **몇 개의 예시(few-shot), 간단한 지시**만으로도 언어 과제 수행 가능

1. 소개 - NLP vs 인간 비교

작업 수행 방식 비교표

구분	기존 NLP 모델	인간
입력 방식	대량의 라벨 데이터 + 예시	지시 또는 간단한 설명
학습 필요	필요 (fine-tuning 수행)	없음 (즉시 적응 가능)
파라미터 변경	있음 (훈련 과정에서 weight 업데이트)	없음 (기존 능력으로 해결)
수행 방식	task-specific 모델로 개별 과제 수행	in-context로 즉시 문제 해결
예시 수	수천~수만 개 필요	몇 개(few-shot) 또는 없음(zero-shot)도 가능
유연성	낮음 (작업마다 모델 재설계 필요)	높음 (지시만 바꾸면 다양한 작업 수행 가능)

1. 소개 - In-Context Learning

- **In-context Learning 란?**
  - 미리 학습된 모델이, 입력된 예시(context)만 보고 작업 수행
  - 학습 중 **파라미터 업데이트 없음**, 전부 **입력 안에서 해결**

- 이걸 통해 Meta-learning 효과 실현
- **Zero-shot / One-shot / Few-shot** 학습 설정으로 실험

중요 포인트:

- 모델이 커질수록 **in-context learning 성능 상승**
- **few-shot GPT-3** : fine-tuned SOTA 모델과 성능 비슷하거나 능가

2. 접근 - 학습 방식과 세팅 구분

학습 방식 구분

- **Fine-Tuning**: 파라미터 업데이트, 많은 데이터 필요
- **Few-Shot**: 예시 몇 개만 입력
- **One-Shot**: 예시 1개
- **Zero-Shot**: 지시문만 있음

GPT-3는 **fine-tuning 없이 inference만**으로도 높은 성능 가능
각 설정은 **사람 학습 방식에 대한 흉내**라고 볼 수 있음

2. 접근 - fine-tuning vs inference 비교

📊 Fine-tuning vs. Inference-only (GPT-3 방식)

항목	🛠️ Fine-tuning 방식	✨ GPT-3 Inference 방식
학습 필요	O (재학습 필요)	❌ (학습 없이 사용)
파라미터 업데이트	O (가중치 변경)	❌ (고정된 모델)
작업 적응 방식	데이터로 훈련	문맥(context) 기반 추론
데이터 요구량	많음 (수천~수만 개)	적음 (few-shot 예시 몇 개)
장점	특정 작업에 최적화	다양한 작업에 즉시 적용 가능
핵심 메커니즘	파라미터 조정	문맥 파악과 패턴 감지

2. 접근 - 모델 구조 및 학습 데이터

모델 구조

- GPT-2 아키텍처 기반
- 크기: 125M ~ 175B까지 총 8개 모델 실험
- context window: 2048 tokens
- 일부는 Sparse Attention 사용

훈련 데이터

- Common Crawl (filtered), WebText2, Books1/2, Wikipedia 등
- 고품질 데이터에 가중치 ↑
- 총 훈련 토큰 수: 3000B

훈련 인프라

- V100 GPU 클러스터, 병렬 학습 적용

2. 접근 - 평가 방식

- **Zero-shot**
  - 예시 없이 **자연어 지시만 제공**
  - ex) "Translate to French"(자연어 지시)만 입력 =>"Good night → Bonne nuit" 를 추론
- **One-shot**
  - **예시 1개 + 지시**
  - 위의 지시에서 "Good night → Bonne nuit" 를 더 하여, 번역 작업임을 추론케 함.
- **Few-shot**
  - **여러 개의 예시 + 지시**
  - 예시 개수 K ≈ 10~100 (context window 크기 한도 내)
  - ex) "Good night → Bonne nuit" 외에 여러 예시를 제시.

2. 접근 - 평가 방식의 특징

- **모델 파라미터는 업데이트하지 않음**
  - 즉, 진짜 ‘학습’이 아니라 ‘적응’만 보는 것
  - 파라미터 고정, 입력만 바꿈 → **in-context learning 능력 평가**

- **입력 포맷을 그대로 텍스트로 줌**
  - **ex.** 영어 문장 **+** 프랑스어 번역 예시들 **+** 새로운 영어 문장 → 모델이 다음 토큰으로 프랑스어 번역 출력

3. Results

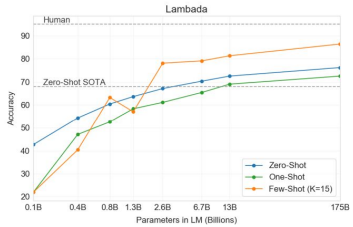
3. Results

1. 언어 모델링 및 완성 과제
2. 폐쇄형 질의응답 능력
3. 번역 능력 평가

- **Language Modeling (언어 모델링)**
  - 데이터셋: Penn Tree Bank (PTB)
  - 방식: zero-shot
    - PTB는 GPT-3 학습 데이터에 포함되지 않아 데이터 누수 걱정 X
  - 성과: 퍼플렉시티 20.50 – 기존 SOTA 대비 15포인트 개선

  ➡️ 확장된 모델 크기가 전통적 언어 모델링에서도 효과적!

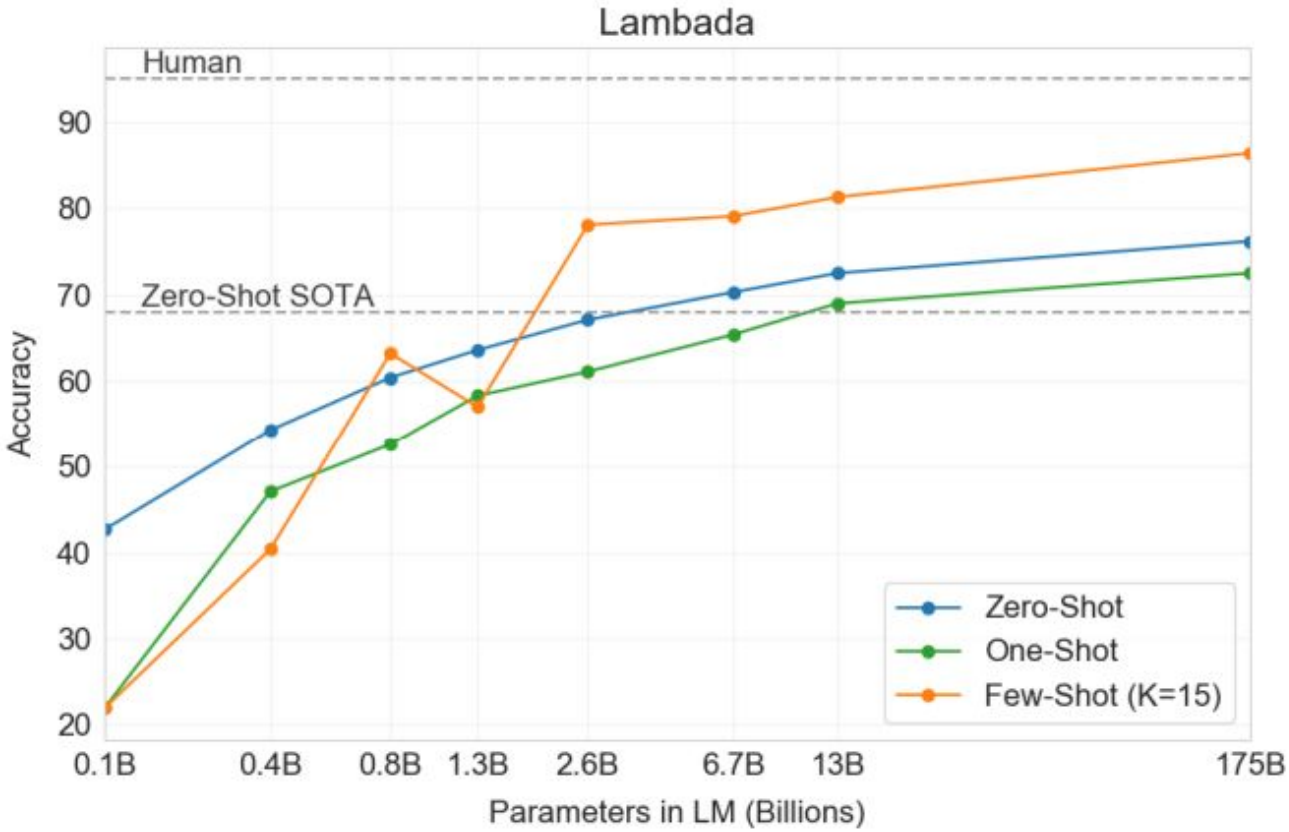
- **LAMBADA**
  - 목표: 문단 맥락을 보고 문장의 마지막 단어를 예측
  - long-rage dependency, 장거리 의존성 테스트
  - 결과
    - zero-shot: 76% (기존 SOTA 대비 8% 개선)
    - few-shot: 86.4% (기존 SOTA 대비 18% 개선)
  - LAMBADA의 고유한 문제를 해결
    - 기존: 문장이 한 단어로 끝난다는 것을 알 수 없음
    - 해결: 빈칸 형식으로 과제를 프레이밍하여 정확히 한 단어를 완성해야 함을 모델이 추론 가능



3. Results

one-shot은 zero-shot보다 항상 낮음

1. 언어 모델링 및 완성 과제
2. 폐쇄형 질의응답 능력
3. 번역 능력 평가



3. Results

1. 언어 모델링 및 완성 과제
2. 폐쇄형 질의응답 능력
3. 번역 능력 평가

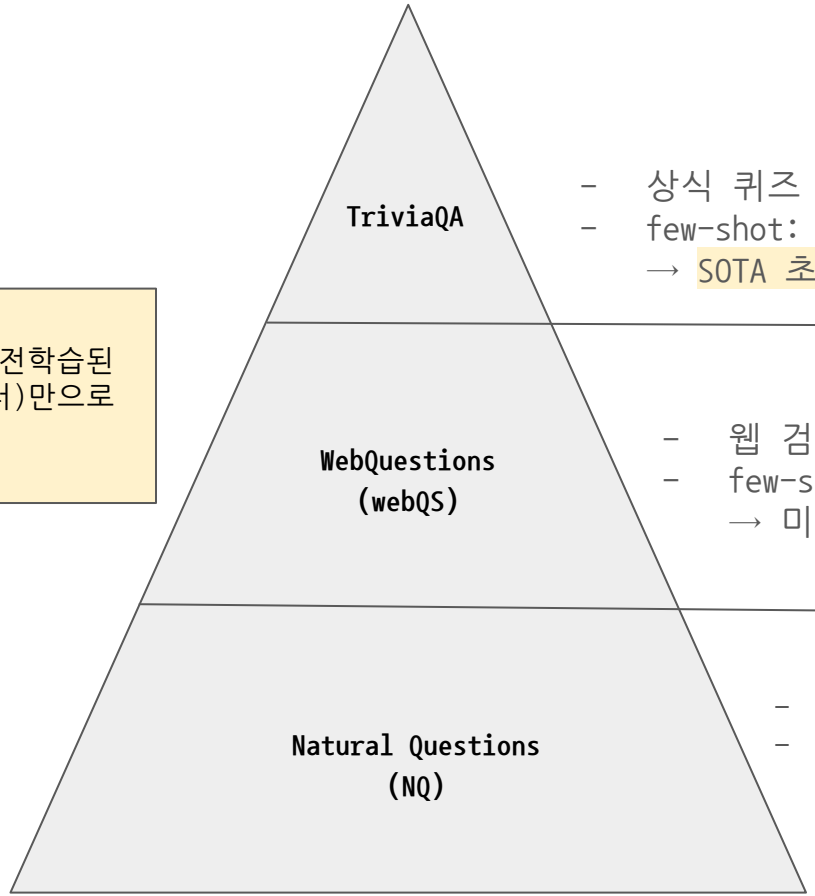
- **HellaSwag**
  - 목표: 문장의 결말 중 가장 자연스러운 선택지 고르기
  - 인간에겐 쉽지만 모델에겐 어렵도록 설계됨
  - 결과
    - one-shot: 78.1%, few-shot: 79.3%
    - 기존 미세조정 모델보다 더 높음: 75.4%
    - 다중작업 SOTA 모델보다 더 낮음: 85.6%

- **StroyCloze**
  - 이야기 마지막 문장을 맞추는 2지선다 문제
  - 결과: zero-shot 83.2%, few-shot: 87.7%
  - 이전 zero-shot 결과보다 10% 이상 향상

3. Results

1. 언어 모델링 및 완성 과제
2. 폐쇄형 질의응답 능력
3. 번역 능력 평가

문서나 외부 정보 없이 사전학습된 모델의 기억(내부 파라미터)만으로 질문에 답하는 방식



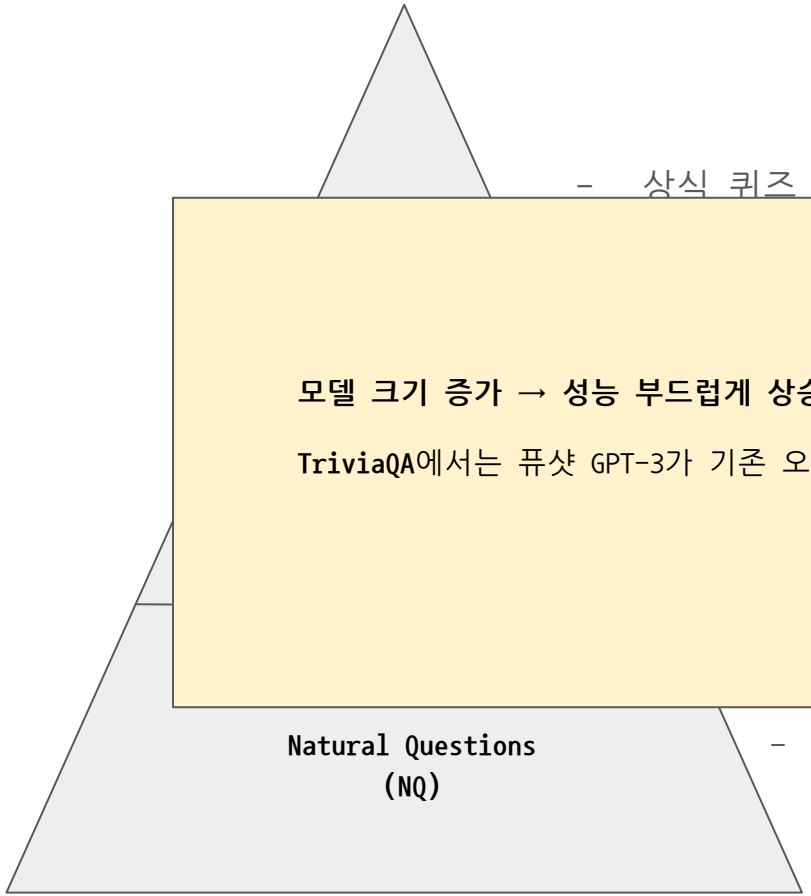
- 상식 퀴즈 기반
- few-shot: 71.2%
  → SOTA 초과 달성

- 웹 검색 기반 (인공질문)
- few-shot: 41.5%
  → 미세조정 없이 기존 성능 근접

- 실제 검색된 질문 기반
- few-shot: 29.9%
  → 여전히 SOTA(T5-11B+SSM)에는 못 미침

3. Results

1. 언어 모델링 및 완성 과제
2. 폐쇄형 질의응답 능력
3. 번역 능력 평가



- 상식 퀴즈 기반

모델 크기 증가 → 성능 부드럽게 상승

TriviaQA에서는 퓨샷 GPT-3가 기존 오픈도메인 QA SOTA를 초과함

- few-shot: 29.9%
  → 여전히 SOTA(T5-11B+SSM)에는 못 미침

3. Results

1. 언어 모델링 및 완성 과제
2. 폐쇄형 질의응답 능력
3. 번역 능력 평가

- GPT-3 학습 데이터의 약 7%가 비영어
- 프랑스어, 독일어, 루마니어와의 쌍으로 실험
- 기존 비지도 번역 방식(NMT)과 비교

- **결과**
  - **영어 ➡️ 비영어**
    - 성능이 상대적으로 낮음
  - **비영어 ➡️ 영어**
    - 기존 비지도 NMT보다 우수
  - **few-shot**
    - 예시 제공 시 BLEU 점수 +4 향상
    - 기존 비지도 NMT 평균 성능에 근접
  - **zero-shot**: 단순 과제 설명만으로는 기존 비지도 NMT보다 성능 낮음
  - **one-shot**: 예시 하나만으로 BLEU점수 +7 향상

- **의미**
  - 모델 크기 증가에 따른 번역 품질 향상
  - 일반적인 번역 태스크에도 유의미한 성능 보임

5. Limitations

5. Limitations

1. 텍스트 생성 및 NLP
2. 구조적/알고리즘적
3. 샘플 효율성
4. Few-Shot 시스템의 불확실성
5. 실용적 적용
6. 일반적인 딥러닝 시스템

- **텍스트 생성 품질 문제**
  - 의미적 반복: 같은 내용을 의미적으로 반복
  - 일관성 상실: 긴 문장에서 일관성을 유지하지 못함
  - 자기모순: 앞선 내용과 모순되는 내용 생성
  - 논리 비약: 문맥에 맞지 않는 문장 생성

- **특정 영역**에서의 약점: 텍스트 데이터만 보고 학습했기 때문에 실제 물리학에 대한 경험 **無**
  - Common sense physics(상식적 물리학)
    예시) 치즈를 냉장고에 넣으면 녹을까?

- **벤치마크 성능 격차**
  - **WIC**(Word-in-Context):
    두 문장에서 특정 단어가 같은 방식으로 사용되는지 판단
    예시) He broke the record vs He broke the vase
  - **ANLI**(Adversarial Natural Language Inference):
    한 문장이 다른 문장을 함축하는지 판단
  - 독해력

5. Limitations

- **Autoregressive 언어 모델**
  - 양방향 구조가 아닌 단방향!이므로 정보를 한 쪽 방향만 봄
    ➡️빈칸 채우기, 내용 비교, 긴 글을 읽고 짧은 답변 생성 작업에 취약함
  - 미세 조정 가능성: 최근 문헌에서는 양방향 모델 미세 조정이 나아짐

- **self-supervised prediction, 사전학습 목표의 한계**
  - 토큰 가중치: 모든 토큰을 동등하게 취급 → 중요성 반영 불가
  - 예측 정확도 학습만으로는 goal-directed behavior(목표 지향적 행동) 어려움
  - 경험 부족: 물리적 세계 경험 無 → 현실 세계 이해 한계

- **향상 가능 접근법**
  - 인간으로부터 목표 함수 학습: 피드백, 의도에서 학습 목표를 설정
  - 강화학습을 통한 미세조정: 특정 행동 목표를 따라가도록 조정
  - 다중 모달리티 활용: 현실 세계를 이해하기 위해 이미지, 영상 등 추가

5. Limitations

- **사전학습 효율성**
  - 인간이 평생 동안 접하는 것보다 훨씬 더 많은 텍스트 학습
    ➡️샘플 효율성 낮음

- **Few-shot 학습의 불확실성**
  - 학습 방식 모호성: 실제로 새로운 작업을 실시간으로 학습하는지, 아니면 이미 본 작업을 인식만 하는지 불확실

- **가능성 스펙트럼**
  : 모델이 어떤 새로운 상황이나 작업에 얼마나 잘 적응할 수 있는지를 나타내는 스펙트럼
  - 훈련과 같은 분포의 시연 인식
  - 형식만 다른 동일 작업 인식
  - 완전히 새로운 기술을 학습
  - QA와 같은 일반 작업스타일에 적응

- **인간 학습과의 유사성**
  - 인간도 “처음부터 학습” vs “시연을 통한 학습” 구분이 어려움

- **미래 연구 과제**
  - 다양한 시연을 사전학습에 포함하고, few-shot 학습이 정확히 어떻게 작동하는지 이해 필요

5. Limitations

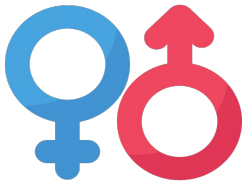
- **문제: 비용 및 편의성**
  - GPT-3 규모의 모델을 추론 비용이 높고 불편함
- **해결책: 모델 Distillation(증류)**
  - 불필요한 능력을 제거하여 경량 모델로 압축
  - 수천억 매개변수 모델에 대한 증류는 전례가 無

- **해석 가능성 부족**
  - 왜 이런 결과가 출력되는지 알기 어려움
- **보정 부족**
  - 새로운 입력에 대한 예측 신뢰도 부족
  - 인간보다 높은 성능 변동성 → 불안정성 존재
- **데이터 편향**
  - 학습된 데이터의 편향 유지
    - 고정관념, 편견 반영 콘텐츠 생성 유도

6. Broader Effects

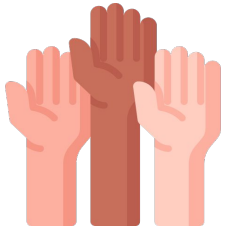
6. Broader Effects

1. 공정성, 편향성 문제
2. 언어 모델의 오용 가능성
3. 에너지 사용 – 미래 방향



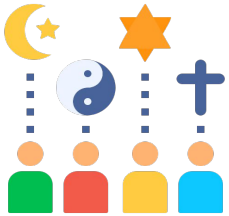
성별 편향

- 테스트된 388개 직업 중 83%가 남성 식별자를 따름
- 여성은 "beautiful", "gorgeous"와 같은 외모 중심 단어로 더 자주 묘사됨
- 남성은 더 많은 형용사로 수식



인종 편향

- 아시아인: 긍정 감정 단어
  - 7개 모델 중 3개에서 1위
- 흑인: 부정 감정 단어
  - 7개 모델 중 5개에서 최하위



종교 편향

- 이슬람: '폭력적인', '테러리즘'과 같은 단어가 높은 비율을 차지

→ 실제 훈련 데이터의 편향을 반영

- **편향에 대한 대응 방향**
  - '제거' 대신 이해하고 개입하는 접근이 필요
  - 기술적, 사회적 접근을 결합한 통합 전략 필요

6. Broader Effects

1. 공정성, 편향성 문제
2. 언어 모델의 오용 가능성
3. 에너지 사용 - 미래 방향

- **잠재적 오용 어플리케이션**
  - 허위 정보, 스팸, 피싱, 사기 등 사회적으로 유해한 활동에 활용
  - 모델 성능이 높을 수록 오용 가능성 증가
    - GPT-3은 인간 수준의 텍스트 생성 가능
- **위협 행위자 분석**
  - 저~중 기술자의 APT(지능형 지속 공격)까지
  - GPT-2 기준 실제 오용 사례는 제한적
- **왜 오용될까?**
  - 비용, 사용 용이성 등 경제적 요인
  - TTP 강화 가능: 교묘해지고 자동화됨
    - 공격자의 Tactics(전략), Techniques(기술), Procedures(절차)
- **미래 대응 방안**
  - 완화 연구, 프로토타이핑 및 다른 기술 개발자와의 조정을 통한 문제 대응

6. Broader Effects

1. 공정성, 편향성 문제
2. 언어 모델의 오용 가능성
3. 에너지 사용 - 미래 방향

175B
매개변수 수

수천 petaflop/s-days
사전학습에 소요된
연산량

0.4 kWh
100페이지 콘텐츠 생성
에너지 비용

- GPT-2 대비 약 100배 규모의 매개변수 (1750억개)
- 훈련할 때: 막대한 에너지와 컴퓨팅 자원 필요
- 추론할 때: 약 40~60원 수준
  - 텍스트 생성 과정
  - 전기 사용량 많지 않음

8. Conclusion

8. Conclusion

- **핵심 성과:**
  - 1750억 개의 파라미터를 가진 초거대 언어 모델을 개발함.
  - Zero-shot, One-shot, Few-shot 세팅에서 다양한 NLP 과제에 대해 강력한 성능을 보임.
  - 일부 과제에서는 최첨단(fine-tuned) 모델의 성능에 근접한 결과를 냄.
  - 별도의 미세조정(fine-tuning) 없이 prompt만으로도 일관된 성능 향상이 모델 크기 증가에 따라 나타남.
  - 즉석에서 정의된 과제들(on-the-fly tasks)에서도 질적으로 우수한 결과를 생성함.

- **시사점:**
  - 비록 제한점과 약점이 존재하지만, 이런 대형 언어 모델은 범용 언어 시스템(AGI)의 핵심 요소가 될 수 있음.