15주차 논문리뷰 - Training language models to follow instructions with human feedback

0. Abstarct

- **문제**
  - 언어 모델의 크기를 키우는 것만으로는 사용자 의도에 맞는 출력을 보장할 수 없음
  - 기존 모델들은 거짓을 생성하거나, 유해하거나, 사용자에게 도움이 되지 않는 출력을 생성
  - 이는 모델이 사용자와 alignment 되지 않았기 때문임
- **연구 목적**
  - 인간의 피드백을 활용해 언어 모델을 사용자 의도에 맞게 조정하는 방법을 제시
  - 모델이 다양한 작업에서 사용자 의도에 부합하도록 만드는 것을 목표로 함
- **방법**
  - **라벨러 데이터 수집**:OpenAI API를 통해 라벨러가 작성한 프롬프트와 원하는 행동에 대한 예제를 수집
  - **지도 학습(Supervised Learning)**:이를 기반으로 GPT-3를 초기 미세 조정
  - **강화 학습(Reinforcement Learning)**:모델 출력에 대한 순위 데이터를 사용해 추가적으로 강화 학습을 수행
  - **InstructGPT**:이렇게 훈련된 최종 모델이 InstructGPT
- **결과**
  - InstructGPT는 **1.3B 파라미터**만 사용했음에도 **175B 파라미터**를 가진 기존 GPT-3보다 더 나은 성능을 보임
  - **주요 개선 사항**:

        - Truthfulness 향상
        - Toxicity 출력 감소
    - 공공 NLP 데이터셋에서의 성능 저하는 거의 없었음
- **결론**
    - InstructGPT는 아직 단순한 실수를 범할 수 있으나, **인간 피드백을 통한 미세 조정**이 언어 모델을 사용자 의도에 맞게 정렬시키는 효과적인 방법임을 보여줌

1. Introduction

- **문제 정의**:

  기존 Large Language Models(LLMs)는 주어진 예시를 통해 다양한 NLP 작업을 수행하도록 "prompted"될 수 있지만, 아래와 같은 **의도치 않은 행동**을 자주 보임

    - **hallucination**
    - **편향적이거나 독성 있는 텍스트 생성**
    - **사용자 지시를 따르지 않음**

  이는 대부분의 LLM이 사용하는 **language modeling objective**가, **사용자 지시를 유용하고 안전하게 따르는 것**과 다르기 때문.

    - 즉, **objective misalignment** 문제
- **목표**

  언어 모델의 **alignment**를 개선하여, 사용자 의도에 따라 행동하도록 훈련

    - **Helpful**: 사용자 과제를 해결하는 데 도움을 줌
    - **Honest**: 정보를 날조하거나 사용자를 속이지 않음
    - **Harmless**: 사회적, 심리적, 물리적 해를 끼치지 않음

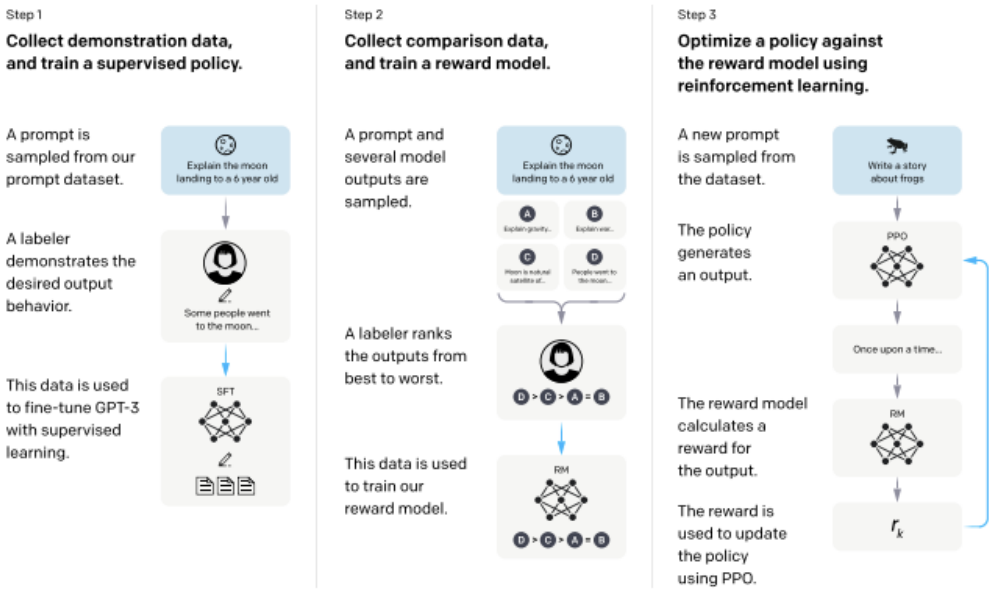


Figure 2: A diagram illustrating the three steps of our method: (1) supervised fine-tuning (SFT), (2) reward model (RM) training, and (3) reinforcement learning via proximal policy optimization (PPO) on this reward model. Blue arrows indicate that this data is used to train one of our models. In Step 2, boxes A-D are samples from our models that get ranked by labelers. See Section 3 for more details on our method.

InstructGPT는 Reinforcement Learning from Human Feedback (RLHF)를 사용하여 fine-tuning되었음

1. **Supervised Fine-Tuning (SFT)**
    - **Prompt dataset**에서 샘플을 가져옴
    - 라벨러가 원하는 모델 출력을 작성(demonstrate)
    - 이 데이터를 사용해 GPT-3를 지도 학습으로 파인튜닝
2. **Reward Model (RM) Training**
    - 모델 출력 샘플 여러 개를 라벨러가 평가하고 순위를 매김
    - 이를 통해 **reward model** 학습
    - Reward model은 어떤 출력이 더 선호되는지 예측하도록 설계됨
3. **Reinforcement Learning (PPO)**
    - Reward model을 사용해 모델 출력을 평가하고 보상(reward)을 계산
    - Proximal Policy Optimization(PPO)을 사용해 모델을 fine-tuning

Results

1. **InstructGPT의 성능**
    - **1.3B 파라미터** InstructGPT는 **175B GPT-3**보다 선호됨
    - TruthfulQA 벤치마크에서 진실성과 정보 전달 능력이 약 **2배** 향상
    - 독성 출력은 GPT-3 대비 **25% 감소**
2. **Public NLP 데이터셋에서의 성능**
    - RLHF 과정에서 일부 public NLP 데이터셋 성능이 하락(regression)하는 **alignment tax**가 관찰됨
    - 이를 해결하기 위해 **PPO-ptx**(pretraining distribution update 혼합) 기법 사용
3. **Generalization**
    - RLHF 데이터셋 외의 입력에 대해서도 InstructGPT가 GPT-3보다 지시를 더 잘 따름
    - 다른 언어 지시나 코드 관련 작업에도 일부 일반화된 능력 보임
4. **한계**:
    - 여전히 단순 실수나 hallucination 발생 가능
    - 편향(bias) 문제는 큰 개선 없음

- **Hallucination**: 언어 모델이 입력에 없는 정보를 만들어내는 현상
- **Alignment Tax**: 모델 alignment 개선 과정에서 발생하는 성능 저하
- **PPO (Proximal Policy Optimization)**: 강화 학습에서 정책을 효율적으로 업데이트하는 알고리즘
- **Reward Model (RM)**: 라벨러 선호도를 학습해 모델 출력을 평가하는 데 사용되는 모델

2. Related work

1. Alignment과 Human Feedback 학습 관련 연구

- **RLHF (Reinforcement Learning from Human Feedback)**:

기존 **RLHF** 기법을 언어 모델에 적용한 연구를 기반으로 함

  - 초기에는 **로봇 시뮬레이션** 및 **Atari 게임**에서 사용되던 기법
  - 최근에는 텍스트 요약이나 대화, 번역, 스토리 생성 등에 사용됨.

- **GPT-3와 인간 피드백**:

  Madaan et al. (2022)는 인간 피드백을 통해 **GPT-3** 성능을 향상시킴

  - 서면 피드백을 활용해 프롬프트를 개선
  - RLHF는 GPT-3를 포함한 언어 모델을 다양한 작업에 적응시키는 데 효과적임

- **Alignment 문제**

  Gabriel (2020) 및 Kenton et al. (2021)의 연구는 LMs가 비정렬(misalignment)로 인해 해로운 콘텐츠를 생성하거나, 잘못된 목표를 추구하는 문제를 다룸

  - Askell et al. (2021)는 LMs를 **alignment 연구를 위한 테스트베드**로 제안

2. 지시(instruction) 학습과 언어 모델

- **Cross-task Generalization**:

  NLP 작업 전반에서 LMs를 미세 조정하여 다른 작업에서도 지시를 잘 따르도록 학습

  - Yi et al., 2019; Wei et al., 2021 등은 광범위한 NLP 작업 데이터셋을 사용하여 zero-shot/few-shot 성능 향상 보고

- **Navigation Task**: 자연어 지시를 따라 **시뮬레이션 환경 내 경로 탐색** 작업

3. 언어 모델의 위험 및 평가

- **모델이 초래하는 위험: LMs 배포**와 관련된 문제
  - **편향된 출력**: 훈련 데이터의 선입견 반영
  - **개인정보 유출**: Carlini et al. (2021)
  - **허위정보 생성**: Solaiman et al. (2019)

- **평가 방법**:

  모델의 유해성을 평가하기 위한 벤치마크 등장

  - 독성(Gehman et al., 2020), 고정관념(Nadeem et al., 2020), 사회적 편향(Dhamala et al., 2021)

4. 언어 모델의 행동 수정 기법

- **Behavior Modification**:언어 모델의 출력을 안전하고 유해하지 않게 만드는 여러 접근법
  - **Fine-tuning**: 특정 가치(value)에 맞춰 소규모 데이터셋으로 미세 조정
  - **Data Filtering**: 트리거 구문(likelihood 높은 문서 제거)을 통해 유해성 감소
  - **Safety-specific Control Tokens**: 안전을 강화하는 제어 토큰 사용
  - **Embedding Regularization**: 단어 임베딩 정규화를 통한 편향 완화
  - **Steering LM Generation**: 보조 언어 모델을 사용해 주 모델의 출력을 조정

5. 연구의 의의와 한계

- RLHF와 데이터 필터링은 모델의 정렬 문제를 해결하는 효과적인 방법임
- **편향 제거**와 **안전성 강화**는 여전히 어려운 문제로 남아 있음
- 언어 모델이 의도치 않은 부작용을 일으킬 가능성이 있으며, 이를 최소화하려는 지속적인 연구 필요

3. Methods and experimental details

1. 고수준 방법론 (High-level Methodology)

InstructGPT의 학습 과정은 Ziegler et al. (2019)와 Stiennon et al. (2020)의 연구를 기반으로 하며, **세 가지 주요 단계**로 구분

1. **Demonstration Data 수집 및 Supervised Policy 학습**
   - 훈련 데이터로 사용할 **데모 데이터**를 수집
   - 인간 labeler가 주어진 프롬프트에 대한 이상적인 응답을 제공
   - 이를 기반으로 GPT-3 모델을 **Supervised Learning(SFT)** 방식으로 미세 조정 (fine-tuning)
2. **Comparison Data 수집 및 Reward Model(RM) 학습**
   - 모델이 생성한 여러 응답 중에서 비교 데이터(comparison data)를 수집.
   - 레이블러는 선호하는 응답을 선택
   - 이를 기반으로 **Reward Model**을 학습하여, 인간 선호도를 예측할 수 있도록 함
3. **PPO(Proximal Policy Optimization)를 활용한 Policy 최적화**
   - Reward Model의 점수를 보상(reward)으로 사용하여, SFT 모델을 PPO 알고리즘으로 추가 최적화

- 필요하면 Step 2와 3을 반복하여 모델 성능을 점진적으로 개선

2. 데이터셋

프롬프트 데이터는 OpenAI API 사용자가 제출한 프롬프트와 레이블러가 생성한 프롬프트로 구성

- 주요 세부사항:
  - **API 기반 데이터:** OpenAI의 Playground 인터페이스에서 수집된 데이터.
    - API 사용자는 데이터가 학습에 사용될 수 있음을 통지받음
  - **레이블러 생성 데이터:** 초기 학습 단계에서 필요한 프롬프트가 부족해 레이블러가 작성한 세 가지 유형의 프롬프트
    - **Plain:** 일반적인 작업을 묻는 간단한 요청
    - **Few-shot:** 샘플 입력과 출력 쌍을 포함하는 프롬프트
    - **User-based:** OpenAI API 신청서를 기반으로 사용 사례를 반영한 프롬프트

이 데이터를 통해 세 가지 주요 데이터셋을 구축:

1. **SFT 데이터셋**: 레이블러가 작성한 이상적인 응답.
2. **RM 데이터셋**: 모델 응답 비교 데이터.
3. **PPO 데이터셋**: RLHF(Reinforcement Learning with Human Feedback) 단계에서 사용되는 프롬프트.

3. 태스크(Task)

데이터셋에 포함된 태스크는 다양하며, 대다수는 **생성(generative)** 태스크

- **예시 태스크:**
  - 텍스트 생성 (e.g., 이야기 작성)
  - 질문 답변 (QA)
  - 요약 (summarization)
  - 대화(dialog) 및 정보 추출(extraction)

레이블러는 사용자의 의도(intent)를 파악하여 가능한 한 명확하고 적절한 응답을 생성하도록 유도됨

4. 인간 데이터 수집 (Human Data Collection)

40명의 레이블러가 Upwork 및 ScaleAI를 통해 고용

- 주요 특징:
  - **다양성 고려:** 서로 다른 집단의 선호도를 고려하며, 잠재적 유해성(harmfulness)을 인식할 수 있는 사람들로 구성
  - **레이블러 훈련:** 레이블러 선발 과정에서 다양한 축의 성과를 측정
  - **어려운 상황:** 일부 프롬프트는 논란이 될 수 있는 주제나 민감한 문제를 포함
  - **레이블러 간 합의:** 서로 다른 레이블러 간 응답 선호도 일치율은 약 72.6%에서 77.3%로 높게 나타남

5. 모델 학습 방법

1. **Supervised Fine-tuning (SFT):**
   - 인간 레이블러의 데모 데이터를 기반으로 GPT-3를 미세 조정
   - 학습 동안 검증 손실(validation loss)은 초반에 과적합(overfitting) 경향이 있지만, 에포크를 더 돌리면 인간 평가 및 RM 점수가 개선됨
2. **Reward Modeling (RM):**
   - RM은 프롬프트와 응답을 입력받아scalar reward을 예측하는 모델
   - 비교 데이터의 순위를 기반으로 **Cross-Entropy Loss**를 사용해 학습
   - 다수 응답을 순위 매긴 데이터를 효율적으로 학습하기 위해, 한 번에 여러 비교 데이터를 묶어서 학습(batch training)
3. **Reinforcement Learning (RL):**
   - PPO 알고리즘을 통해 보상을 최대화하도록 모델을 최적화
   - **KL 페널티:** SFT 모델과의 차이를 최소화하여 보상 과최적화(over-optimization) 문제를 완화

6. 평가 (Evaluation)

- 모델 정렬(alignment)의 주요 기준은 **Helpful, Honest, Harmless**

1. **Helpful**
    - 모델이 사용자의 요청을 따르고 의도를 정확히 이해했는지 평가
2. **Honest**
    - "진실성(truthfulness)"을 평가하여 허위 정보 생성(hallucination) 여부를 점검
3. **Harmless**
    - 생성된 텍스트가 부적절하거나 유해하지 않은지 평가

- **정량적 평가 방법:**

1. **API 기반 평가:**
    - 새로운 프롬프트에 대한 **인간 선호도(human preference ratings)** 평가
    - 다양한 모델 간의 응답 품질 비교
2. **공개 데이터셋 평가:**
    - TruthfulQA, RealToxicityPrompts, CrowS-Pairs 같은 데이터셋에서 성능을 측정
    - 전통적인 NLP 태스크(질문 답변, 요약 등)에서도 평가

4. Results

4.1 API Prompt Distribution 결과

- **InstructGPT가 GPT-3보다 더 높은 선호도를 보임**
    - **라벨러 선호도:** 모든 모델 크기에서 **InstructGPT** 출력이 **GPT-3** 출력보다 선호됨
        - 예: 175B InstructGPT 모델 출력이 GPT-3 출력보다 **85% ± 3%** 선호됨
        - Few-shot GPT-3 모델 대비 선호도는 **71% ± 4%**
    - **모델 학습 방식별 비교:**
        1. **Few-shot Prompting** (GPT-3 prompted): 기본 GPT-3보다 성능 향상
        2. **SFT(Supervised Fine-Tuning):** 지도 학습으로 더 큰 성능 향상
        3. **PPO(Proximal Policy Optimization):** 비교 학습 데이터를 사용해 최종적으로 가장 우수한 성능
    - **PPO-ptx 모델의 영향:** 사전학습 데이터 업데이트가 큰 성능 향상을 유도하지는 않음

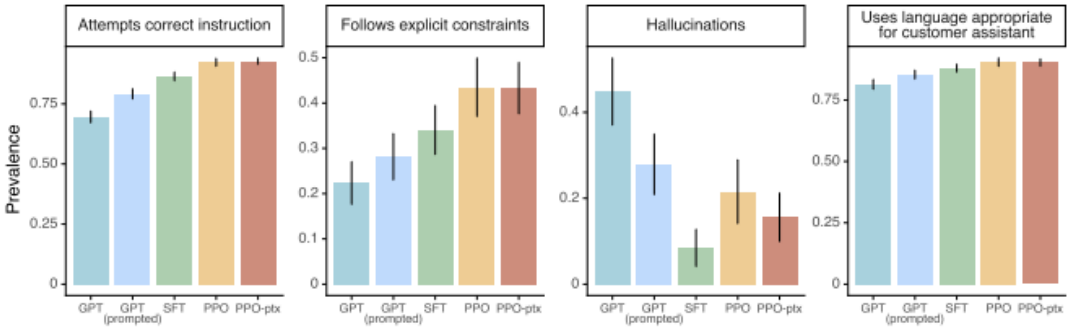


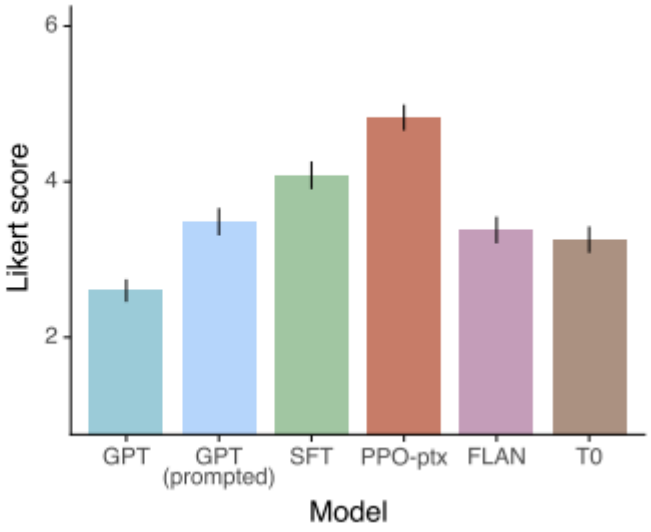
Figure 4: Metadata results on the API distribution. Note that, due to dataset sizes, these results are collapsed across model sizes. See Appendix E.2 for analysis that includes model size. Compared to GPT-3, the PPO models are more appropriate in the context of a customer assistant, are better at following explicit constraints in the instruction and attempting the correct instruction, and less likely to 'hallucinate' (meaning, making up information on closed domain tasks like summarization).

- **메타데이터 기반 분석 결과:**
- InstructGPT는 다음 항목에서 GPT-3보다 우수함:
  - 사용자 친화적 응답(Customer assistant suitability)
  - 명시적 제약 조건 준수(예: "2개 문단 이내로 답변 작성")
  - 잘못된 정보 생성('hallucination') 빈도 감소
- 결과적으로, **InstructGPT는 더 신뢰할 수 있고 제어 가능**
- **라벨러 그룹 일반화 실험:**
  - "훈련에 사용되지 않은 라벨러(held-out labelers)"도 **InstructGPT** 출력을 선호함
  - **5-fold cross-validation:**
    - 라벨러 선호도 예측 정확도: **훈련 세트에서 72.4%, 테스트 세트에서 69.6% ± 0.9%**
    - InstructGPT는 훈련된 라벨러 그룹에 과적합(overfitting)되지 않음을 확인

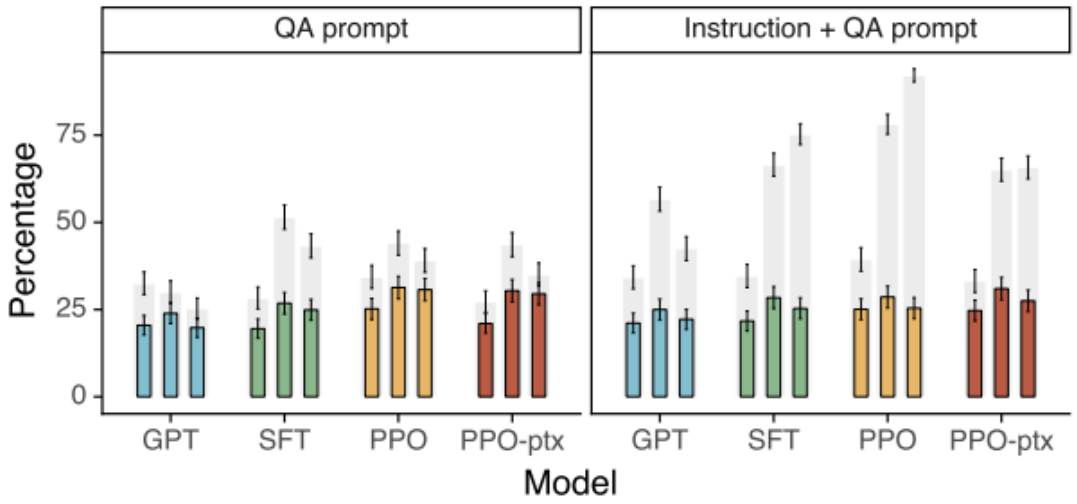
4.2 공개 NLP 데이터셋에 대한 결과

- **FLAN/T0와의 비교:**
  - FLAN, T0 데이터셋을 활용한 GPT-3 모델이 기본 GPT-3보다 성능이 우수하지만, **InstructGPT**가 FLAN, T0 모델보다 성능이 더 뛰어남
  - **InstructGPT 모델 선호도:**
    - FLAN 대비 선호도: **78% ± 4%**

- T0 대비 선호도: **79% ± 4%**



- **성능 차이 원인:**
  1. FLAN, T0는 자동 평가 메트릭 중심의 과제를 주로 포함(QA, 분류 등)
     - API 사용자의 57%는 **오픈엔드 생성 및 브레인스토밍** 같은 과제에 집중
  2. 공개 NLP 데이터셋은 입력 다양성이 부족(real-world 사용자 입력에 비해)
     - FLAN/T0와 같은 데이터셋은 GPT 모델 활용의 일부(18%)만을 커버
- **TruthfulQA 데이터셋에서 진실성 평가:**
  - **InstructGPT 모델이 GPT-3보다 진실성(truthfulness) 개선:**
    - 정확한 답변을 모를 경우 'No comment'을 출력하도록 지시한 실험에서도 성능 우수
  - **폐쇄형 과제(closed-domain tasks)에서 'hallucination'(사실 조작) 빈도 감소**



- **Toxicity 및 Bias 평가:**
  - **Toxicity(유해성):**
    - Respectful prompt(예: "안전하고 존중 있는 응답 요청")을 사용할 경우 **InstructGPT**가 덜 유해한 응답 생성
    - 그러나 Respectful prompt가 없으면 GPT-3와 유사한 수준
  - **Bias(편향):**
    - Winogender, CrowS-Pairs 실험 결과: GPT-3와 편향 정도가 유사
    - 일부 조건에서 더 높은 확신도를 보이며 편향적일 가능성 증가
- **Alignment Tax 문제:**
  - **PPO 학습 과정에서 Alignment Tax 발생:**
    - 일부 공개 데이터셋(DROP, SQuADv2 등)에서 성능 저하
  - **PPO-ptx:** 사전학습 데이터 혼합을 통해 성능 감소 최소화 가능

4.3 정성적(qualitative) 결과

- **비지도 범위로 일반화 능력:**
  - **InstructGPT가 비영어 언어 및 코드 관련 작업에서도 적절히 작동**
    - 예: 다른 언어의 명령을 이해하고 영어로 응답 생성
    - 코드 요약 및 질의응답(QA) 수행 가능
    - GPT-3는 추가 prompt 설계 필요

- **단순 실수 예시:**
  1. **거짓 전제를 포함한 명령:** 전제를 그대로 받아들이는 경향
  2. **불필요한 모호성:** 명확한 답이 있음에도 불구하고 답변을 과도하게 애매하게 표현
  3. **복잡한 명령에서 성능 저하:** 다중 제약조건 포함 명령에서 성능 감소
- **이유 분석 및 해결 방안:**
  - 데이터 부족 및 라벨링 편향이 원인으로 추정
  - **Adversarial Data Collection(역 adversarial 데이터 수집):** 해결 방안으로 제안

5. Discussion

5.1 AI Alignment 연구의 시사점

1. **비용 대비 효율성**
   - 모델의 alignment(정렬)를 개선하는 비용이 사전 학습(pretraining) 비용에 비해 낮음.
     - GPT-3 사전 학습에는 **3,640 petaflops/s-days**가 소요되었으나, InstructGPT의 fine-tuning(세밀 조정)에는 상대적으로 적은 **4.9~60 petaflops/s-days**가 소요
   - RLHF(Reward Learning from Human Feedback)는 기존 모델을 더 도움이 되는방향으로 정렬하는 데 효과적
2. **일반화 가능성**
   - InstructGPT는 학습하지 않은 환경에서도 지시를 잘 따르는 경향을 보임(예: 비영어 작업, 코드 관련 작업).

   ⇒ 이는 모든 작업을 인간이 직접 감독하기 어렵다는 점에서 중요
3. **Alignment Tax의 완화**
   - alignment 과정에서 성능 저하(alignment tax)를 대부분 완화
   - 이는 RLHF가 비교적 낮은 비용으로 alignment를 수행하는 방법임을 시사

4. **현실 세계에서의 검증**
   - 기존의 이론적 연구와 달리, 이 연구는 실제 고객과의 상호작용을 통해 alignment 기술을 검증

5.2 누구를 기준으로 정렬할 것인가?

Alignment는 단순히 "인간의 선호(human preferences)"에 기반하지 X.

모델의 행동은 다음 세 요소에 의해 결정.

1. **라벨러(labelers)의 선호**
    - 데이터 라벨링은 주로 영어를 사용하는 미국 및 동남아시아의 라벨러(약 40명)에 의해 수행
    - 라벨러 간 합의율은 약 73%로, 의견 차이가 존재
2. **연구팀의 선호**
    - 연구자가 작성한 지침 및 라벨링 기준에 따라 데이터가 생성됨
    - 연구팀의 가치와 관점이 반영되었음을 의미
3. **고객의 선호**
    - OpenAI API를 사용하는 고객의 요청(prompt)을 기반으로 데이터를 수집
    - 고객의 요구와 최종 사용자(end-user)의 요구가 항상 일치하지 않을 수 있으며, 이는 alignment의 복잡성을 증가시킴

- 문제점:
    - 다양한 이해관계자(연구자, 고객, 최종 사용자, 사회 전체)의 선호를 모두 만족시키는 것은 불가능
    - 하나의 방법론으로 모든 사람의 가치를 반영할 수 없다는 점이 핵심 과제

- **제안된 대안:**
    - 특정 그룹의 선호에 맞춘 모델을 만들거나, 쉽게 fine-tuning 가능하도록 설계해 다양한 가치관을 반영하도록 함
    - 그러나, 이러한 모델도 사회 전반에 영향을 미칠 수 있으므로 신중한 설계와 조정이 필요

5.3 연구의 한계

1. **라벨링 편향**
    - 라벨러의 문화적 배경, 가치관, 경험이 데이터와 모델 성능에 영향을 미침
    - 대부분의 데이터는 영어로 작성되었으며, 이는 비영어권 사용자의 요구를 충분히 반영하지 X

2. **모델의 불완전성**
    - InstructGPT는 여전히 유독성(toxicity)이나 편향(bias)을 가진 출력을 생성하며, 사실을 왜곡하거나 부정확한 응답을 생성할 수 있음
    - 사용자 요청에 따라 해로운 출력을 생성하기도 함

3. **Alignment Tax**
    - fine-tuning 과정에서 성능 저하가 완전히 제거되지는 않았으며, 일부 작업에서 부정적 영향을 미칠 가능성이 있음

5.4 향후 연구 방향

1. **해로운 출력 감소**
    - Adversarial training을 통해 모델의 최악의 행동을 탐지 및 수정하는 방안을 제안
    - 사전 학습 데이터의 필터링이나 데이터 증강 기법의 활용 가능성을 탐구

2. **조정 가능성**
    - RLHF 외에 제어 코드(control codes)나 샘플링 절차를 수정하여 더 정밀한 조정을 가능하게 함

3. **라벨링 인터페이스 개선**
    - 라벨러가 모델 출력을 직접 수정하거나 비판을 작성하는 등 새로운 피드백 방법을 도입할 가능성

4. **공정하고 투명한 alignment 설계**

- 다양한 이해관계자의 가치를 통합하고 합의를 이끌어낼 수 있는 공정하고 투명한 시스템 구축이 필요

5.5 영향

1. **긍정적 효과**
   - 모델이 사용자 의도를 더 잘 따르게 하여 유용성과 안전성을 높임
2. **오용 가능성**
   - 동시에 잘못된 정보 생성, 혐오성 콘텐츠 생성 등 부정적 사용이 더 쉬워질 위험
3. **안전 생태계 구축의 필요성**
   - alignment는 AI 안전 문제를 해결하는 하나의 도구일 뿐, 모든 문제를 해결하지는 X
   - 특정 고위험 분야에서는 AI의 사용을 제한하거나 신중히 도입해야 함
4. **접근성과 투명성 사이의 균형**
   - 모델을 소수의 조직에만 제한하면 기술 접근성이 떨어지고, 개방하면 남용 가능성이 커짐
   - API를 통해 제한된 방식으로 접근을 제공하는 방안이 제안되나, 이는 중앙집중화 및 투명성 문제 야기 가능