

[InstructGPT]

**Training language models to
follow instructions with human
feedback
(Ouyang et al., 2022)**

Abstract

ChatGPT의 기반이 된 논문으로, LLM을 실제로 유용하게 만들기 시작한 전환점

Making language models **bigger** does **not** inherently make them better at **following a user's intent**. For example, large language models can generate outputs that are **untruthful, toxic**, or simply **not helpful to the user**. In other words, these models are **not aligned with their users**.

In this paper, we show an avenue for **aligning language models with user intent** on a wide range of tasks **by fine-tuning with human feedback**.

Starting with a set of labeler-written prompts and prompts submitted through the OpenAI API, we collect **a dataset of labeler demonstrations of the desired model behavior**, which we use to **fine-tune GPT-3** using **supervised learning**.

We then collect **a dataset of rankings of model outputs**, which we use to further fine-tune this supervised model using **reinforcement learning from human feedback**. We call the resulting models **InstructGPT**.

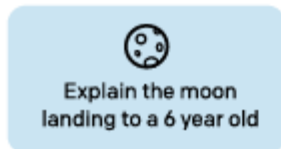
RLHF

In **human evaluations** on our prompt distribution, outputs from the **1.3B** parameter **InstructGPT** model are preferred to outputs from the **175B GPT-3**, despite having **100x fewer parameters**. Moreover, InstructGPT models show improvements in truthfulness and reductions in toxic output generation while having minimal performance regressions on public NLP datasets. Even though InstructGPT still makes simple mistakes, our results show that **fine-tuning with human feedback is a promising direction for aligning language models with human intent**.

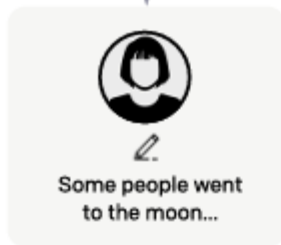
Step 1

**Collect demonstration data,
and train a supervised policy.**

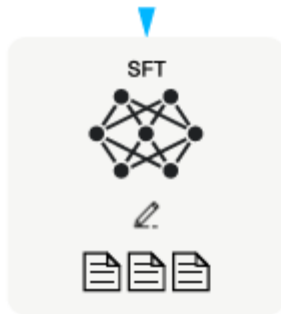
A prompt is
sampled from our
prompt dataset.



A labeler
demonstrates the
desired output
behavior.



This data is used
to fine-tune GPT-3
with supervised
learning.



SFT

(Supervised Fine-Tuning)

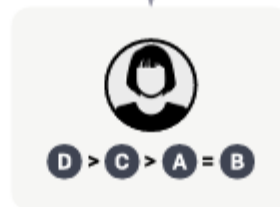
Step 2

**Collect comparison data,
and train a reward model.**

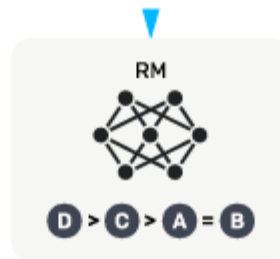
A prompt and
several model
outputs are
sampled.



A labeler ranks
the outputs from
best to worst.



This data is used
to train our
reward model.



RM

(Reward Modeling)

Step 3

**Optimize a policy against
the reward model using
reinforcement learning.**

A new prompt
is sampled from
the dataset.



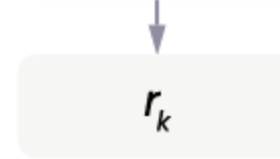
The policy
generates
an output.



The reward model
calculates a
reward for
the output.



The reward is
used to update
the policy
using PPO.



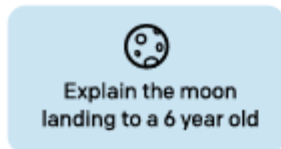
PPO

(Reinforcement Learning)

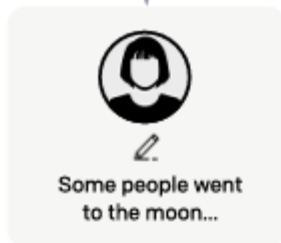
Step 1

**Collect demonstration data,
and train a supervised policy.**

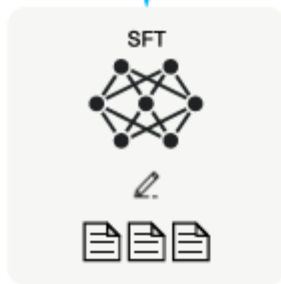
A prompt is
sampled from our
prompt dataset.



A labeler
demonstrates the
desired output
behavior.



This data is used
to fine-tune GPT-3
with supervised
learning.



사람이 쓴 좋은 답변을
학습시킴

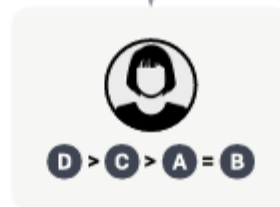
Step 2

**Collect comparison data,
and train a reward model.**

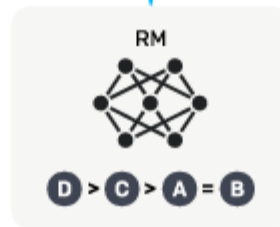
A prompt and
several model
outputs are
sampled.



A labeler ranks
the outputs from
best to worst.



This data is used
to train our
reward model.



사람이 여러 개 답변 비교해
매긴 순위로 보상 모델 학습

Step 3

**Optimize a policy against
the reward model using
reinforcement learning.**

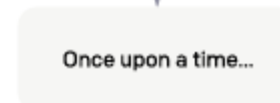
A new prompt
is sampled from
the dataset.



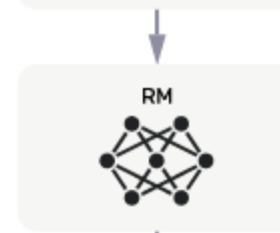
The policy
generates an output.



The reward model
calculates a
reward for the
output.



The reward is
used to update
the policy
using PPO.



생성된 답변을 보상 모델로
평가하여 강화학습

AI Alignment vs. AI Safety (AI Control)

- 다소 작위적으로 구분하자면... (Stuart J. Russell?)
- (cf) [Decoupling AI Safety from AI Alignment](#), reddit (Nov 2, 2023)

개념	AI Alignment	AI Safety
관점	AI의 행동을 인간의 의도나 가치에 정렬시키는 것	AI가 인간이나 사회에 해를 끼치지 않도록 제어하는 것
관심분야	RLHF, Value Learning, Constitutional AI, Interpretability ...	Control Problem, Instrumental Convergence, Orthogonality Thesis, X-risk ...
경향	현재, 근미래 AI 시스템(특히 LLM)의 오용, 오작동 가능성에 대한 실용적, 경험적 접근 ("우린 더 잘 만들 수 있다")	AGI, ASI 단계의 장기적, 근본적 위험에 대한 철학적, 이론적 접근 ("우린 망할 수도 있다")
대표 인물/조직	OpenAI , Anthropic , DeepMind , Paul Christiano , Jan Leike (?) ...	MIRI (Machine Intelligence Research Institute), FLI (Future of Life Institute), Eliezer Yudkowsky , Nick Bostrom ...
관계	Alignment \subset Safety로 보기도 함	Safety \simeq Control 또는 Safety \supset Control로 보기도 함

GPT History

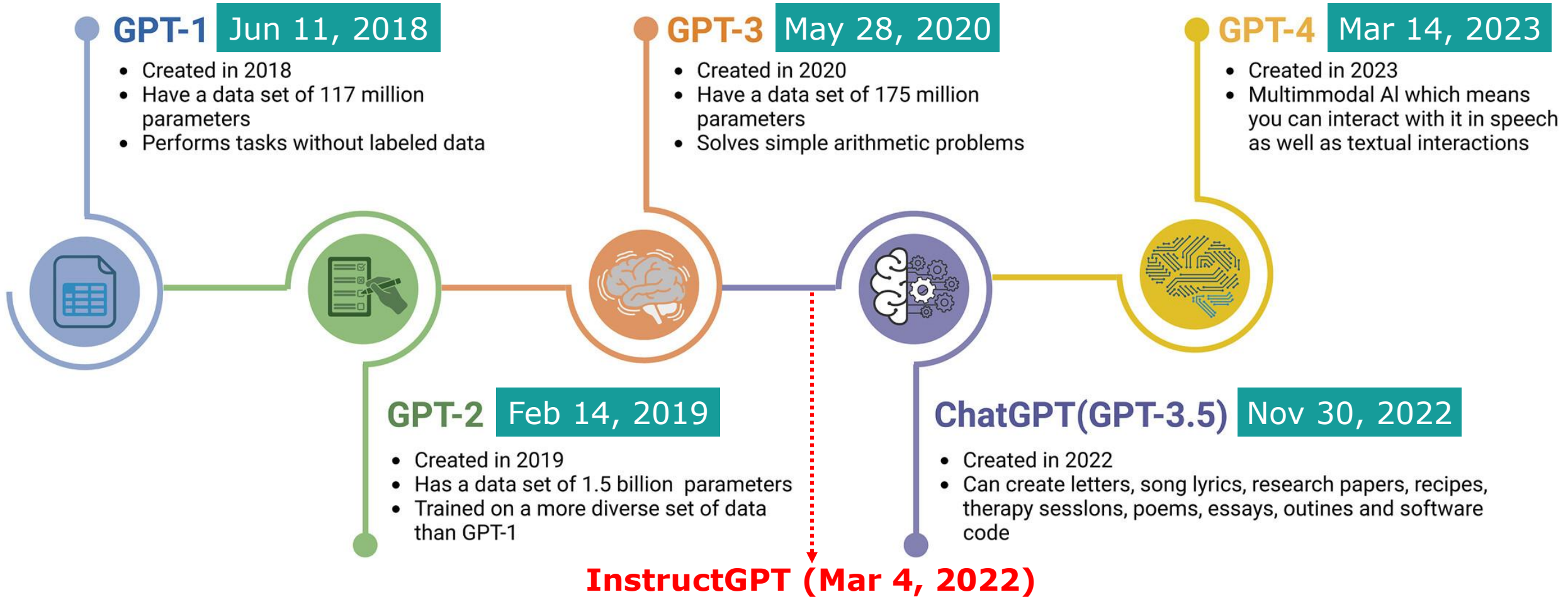


Image Source: Long Bai et al.(2023), ChatGPT: The cognitive effects on learning and memory, Brain-X 1(3), <https://doi.org/10.1002/brx2.30>

1. Introduction

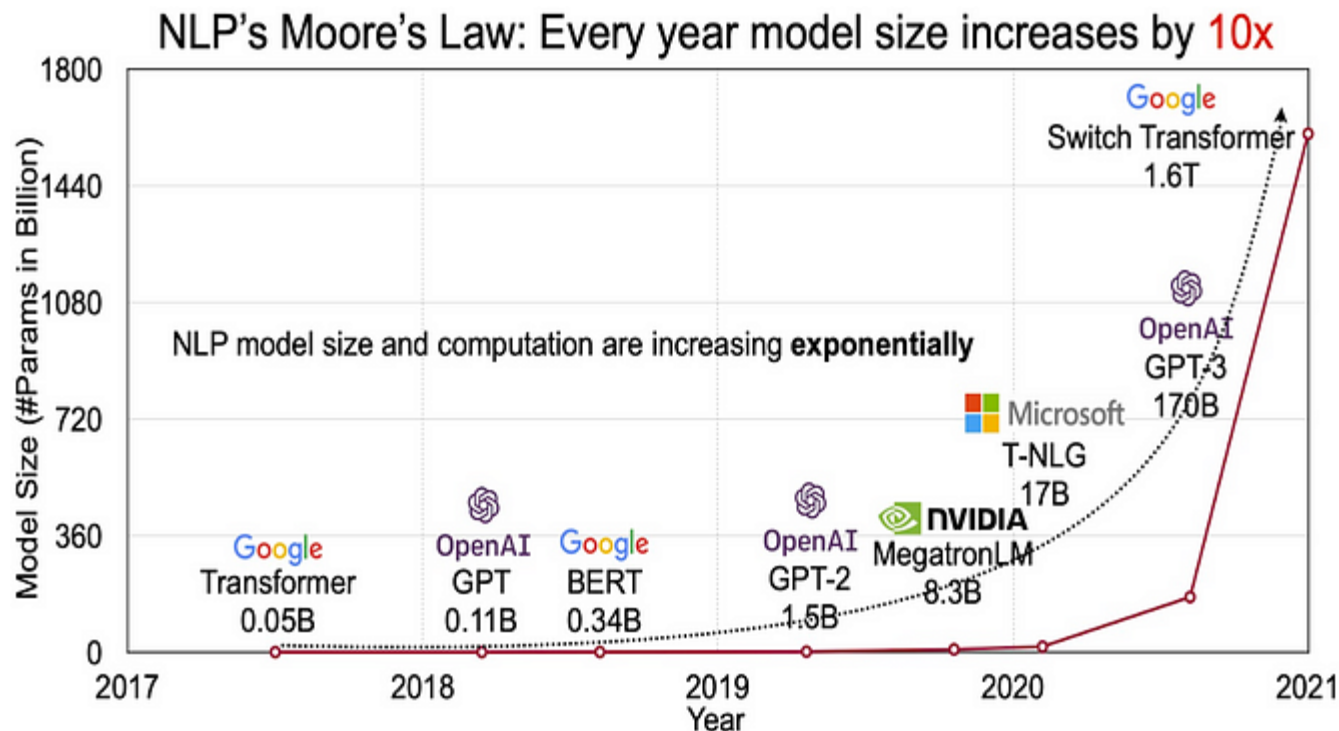


Image Source: Harishdatalab, [Unveiling the Power of Large Language Models \(LLMs\)](#), Medium (2023)

- LLM은 종종 사실을 날조하거나, 편향되거나 유해한 텍스트를 생성하거나, 사용자의 지시를 제대로 따르지 않는 등 원하지 않는 행동을 보이기도 함
- ‘다음 토큰을 예측한다’는 언어 모델링 목표가 “사용자 지시를 사용자에게 도움이 되도록, 안전하게 따른다”는 목표와 다르기 때문 → 언어 모델링 목표가 “정렬되지 않았다(misaligned)”
- LLM은 파라미터 수를 증가시키는 방향으로 발전하였으나, 그것만으로는 인간의 의도를 잘 반영할 수 없음 → 인간 피드백을 이용한 강화학습(RLHF)

3.1 High-level Methodology

- Follows methodology of [Ziegler et al. \(2019\)](#) and [Stiennon et al. \(2020\)](#)
- **Step 1: Collect demonstration data, and train a supervised policy**
 - Labelers provide demonstrations of the desired behavior on the input prompt distribution
 - Then fine-tune a pretrained GPT-3 model on this data using supervised learning
- **Step 2: Collect comparison data, and train a reward model**
 - Collect a dataset of comparisons between model outputs, where labelers indicate which output they prefer for a given input
 - Then train a reward model to predict the human-preferred output
- **Step 3: Optimize a policy against the reward model using PPO**
 - Use the output of the RM as a scalar reward
 - Fine-tune the supervised policy to optimize this reward using the PPO algorithm ([Schulman et al., 2017](#))
- Steps 2 and 3 can be iterated continuously

Stiennon et al. (Feb 15, 2020)

1 Collect human feedback

A Reddit post is sampled from the Reddit TL;DR dataset.



Various policies are used to sample a set of summaries.



Two summaries are selected for evaluation.



A human judges which is a better summary of the post.



"j is better than k"

2 Train reward model

One post with two summaries judged by a human are fed to the reward model.



The reward model calculates a reward r for each summary.



r_j

r_k

The loss is calculated based on the rewards and human label, and is used to update the reward model.

$$\text{loss} = \log(\sigma(r_j - r_k))$$

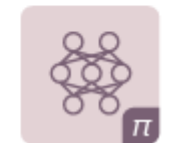
"j is better than k"

3 Train policy with PPO

A new post is sampled from the dataset.



The policy π generates a summary for the post.



The reward model calculates a reward for the summary.



The reward is used to update the policy via PPO.

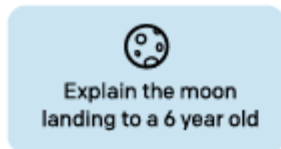
r

Reddit TL;DR dataset 사용한 요약 작업에 RLHF 적용

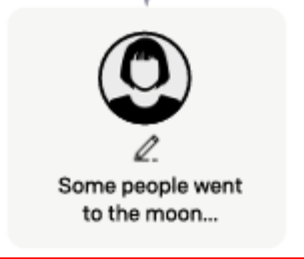
Step 1

**Collect demonstration data,
and train a supervised policy.**

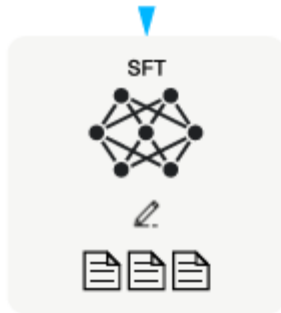
A prompt is
sampled from our
prompt dataset.



A labeler
demonstrates the
desired output
behavior.



This data is used
to fine-tune GPT-3
with supervised
learning.



SFT

(Supervised Fine-Tuning)

Step 2

**Collect comparison data,
and train a reward model.**

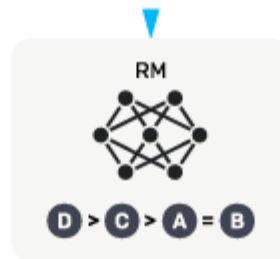
A prompt and
several model
outputs are
sampled.



A labeler ranks
the outputs from
best to worst.



This data is used
to train our
reward model.



RM

(Reward Modeling)

Step 3

**Optimize a policy against
the reward model using
reinforcement learning.**

A new prompt
is sampled from
the dataset.



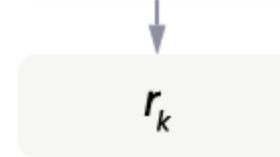
The policy
generates
an output.



The reward model
calculates a
reward for
the output.



The reward is
used to update
the policy
using PPO.



RL

(Reinforcement Learning)

3.2 Dataset ~ 3.4 Human data Collection

- 초기 InstructGPT 모델을 부트스트랩하기 위해 **40인 Labeler**(Appendix B 참조)가 직접 작성한 프롬프트(Appendix A.1) +
 - Plain: labeler가 설정한 임의의 작업, 다양한 유형이 포함되도록 유도
 - Few-shot: 지시어와 함께 여러 개의 질의/응답 쌍 포함
 - User-based: OpenAI API 대기 신청서에 사용자들이 기재한 실제 사용 사례 참고한 유사한 프롬프트
- **OpenAI API**, 특히 Playground 인터페이스의 InstructGPT 초기 버전에 입력된 프롬프트(Appendix A.2)
- 데이터 전처리 방식
 - 중복 제거: 긴 공통 prefix 가진 프롬프트 제거
 - 사용자 ID당 프롬프트 200개로 제한
 - 사용자 ID 기준으로 훈련, 검증, 테스트 세트 분할
 - 개인 식별 정보는 훈련 데이터에서 제거

Table 1: Distribution of use case categories from our API prompt dataset.

Use-case	(%)
Generation	45.6%
Open QA	12.4%
Brainstorming	11.2%
Chat	8.4%
Rewrite	6.6%
Summarization	4.2%
Classification	3.5%
Other	3.5%
Closed QA	2.6%
Extract	1.9%

Table 2: Illustrative prompts from our API prompt dataset. These are fictional examples inspired by real usage—see more examples in Appendix A.2.1.

Use-case	Prompt
Brainstorming	List five ideas for how to regain enthusiasm for my career
Generation	Write a short story where a bear goes to the beach, makes friends with a seal, and then returns home.
Rewrite	This is the summary of a Broadway play: "" { summary } "" This is the outline of the commercial for that play: ""

Table 6: Dataset sizes, in terms of number of prompts.

SFT Data			RM Data			PPO Data		
split	source	size	split	source	size	split	source	size
train	labeler	11,295	train	labeler	6,623	train	customer	31,144
train	customer	1,430	train	customer	26,584	valid	customer	16,185
valid	labeler	1,550	valid	labeler	3,488			
valid	customer	103	valid	customer	14,399			

Table 12: Labeler demographic data

What gender do you identify as?	
Male	50.0%
Female	44.4%
Nonbinary / other	5.6%
What ethnicities do you identify as?	
White / Caucasian	31.6%
Southeast Asian	52.6%
Indigenous / Native American / Alaskan Native	0.0%
East Asian	5.3%
Middle Eastern	0.0%
Latinx	15.8%
Black / of African descent	10.5%
What is your nationality?	
Filipino	22%
Bangladeshi	22%
American	17%
Albanian	5%
Brazilian	5%
Canadian	5%
Colombian	5%
Indian	5%
Uruguayan	5%
Zimbabwean	5%

What is your age?

18-24	26.3%
25-34	47.4%
35-44	10.5%
45-54	10.5%
55-64	5.3%
65+	0%

What is your highest attained level of education?

Less than high school degree	0%
High school degree	10.5%
Undergraduate degree	52.6%
Master's degree	36.8%
Doctorate degree	0%

Finally, we want to thank our labelers, without whom this work would not have been possible: Meave Fryer, Sara Tirmizi, James Carroll, Jian Ouyang, Michelle Brothers, Conor Agnew, Joe Kwon, John Morton, Emma Duncan, Delia Randolph, Kaylee Weeks, Alexej Savreux, Siam Ahsan, Rashed Sorwar, Atresha Singh, Muhaiminul Rukshat, Caroline Oliveira, Juan Pablo Castaño Rendón, Atqiya Abida Anjum, Tinashe Mapolisa, Celeste Fejzo, Caio Oleskovicz, Salahuddin Ahmed, Elena Green, Ben Harmelin, Vladan Djordjevic, Victoria Ebbets, Melissa Mejia, Emill Jayson Caypuno, Rachelle Froyalde, Russell M. Bernandez, Jennifer Brillo, Jacob Bryan, Carla Rodriguez, Evgeniya Rabinovich, Morris Stuttard, Rachelle Froyalde, Roxanne Addison, Sarah Nogly, Chait Singh.

3.5 Models

- **Supervised fine-tuning (SFT)**

- Fine-tune GPT-3 on labeler demonstrations using supervised learning
- 16 epochs, cosine learning rate decay, residual dropout 0.2
- 최종 SFT 모델은 검증 세트에 대한 RM score 기준으로 선택

- **Reward modeling (RM)**

- SFT 모델에서 마지막 층(unembedding layer) 제거하여 보상 값이라는 단일 수치 출력(scalar reward)
- 프롬프트 하나당 $K = 4 \sim 9$ 개 답변(e.g., A, B, C, D) 생성한 후, 순위를 매겨(e.g., $D > C > A > B$) 두 개씩 묶은 비교 쌍(e.g., $D > C$, $D > A \dots \binom{K}{2}$ 개)을 보상 모델 학습에 사용하되, 계산 효율을 높이고 과적합을 방지하기 위해 각 프롬프트에 대한 비교 쌍을 하나의 배치로 묶어 학습
- 보상 모델의 손실 함수(cross-entropy)

선호(y_w)-비선호(y_l) 보상(r_θ) 차 클수록[작을수록] 시그모이드(σ)는 1[0]에, \log 값은 0에 가까워져[음수 커져] loss 작아진다[커진다].

$$\text{loss}(\theta) = -\frac{1}{\binom{K}{2}} E_{(x, y_w, y_l) \sim D} [\log(\sigma(r_\theta(x, y_w) - r_\theta(x, y_l)))]$$

3.5 Models (cont.)

- **Reinforcement learning (RL)**

- PPO (Proximal Policy Optimization, [Schulman et al., 2017](#)) 알고리즘으로 SFT 모델 강화학습
- 토큰마다 KL 페널티(Kullback-Leibler divergence) 추가하여 과도한 최적화 방지
- 공개 NLP 데이터셋에서 성능 저하를 방지하기 위해 PPO 보상 최적화와 동시에 사전학습 목표도 일부 유지하는 변형을 실험(PPO gradients에 pretraining gradients를 섞은 것을 “PPO-ptx”라 함)
- PPO-ptx에서 사용하는 목적함수 **인간 선호에 부합할수록 큰 보상, 평균 0으로 정규화**

$$\text{objective}(\phi) = E_{(x,y) \sim D_{\pi_{\phi}^{\text{RL}}}} \left[r_{\theta}(x,y) - \beta \log \left(\pi_{\phi}^{\text{RL}}(y | x) / \pi^{\text{SFT}}(y | x) \right) \right] +$$

$$\gamma E_{x \sim D_{\text{pretrain}}} \left[\log(\pi_{\phi}^{\text{RL}}(x)) \right]$$

현재 정책이 사전 학습된 SFT 정책에서 얼마나 멀어졌는지 측정하여 너무 멀어지면 KL 페널티 부여

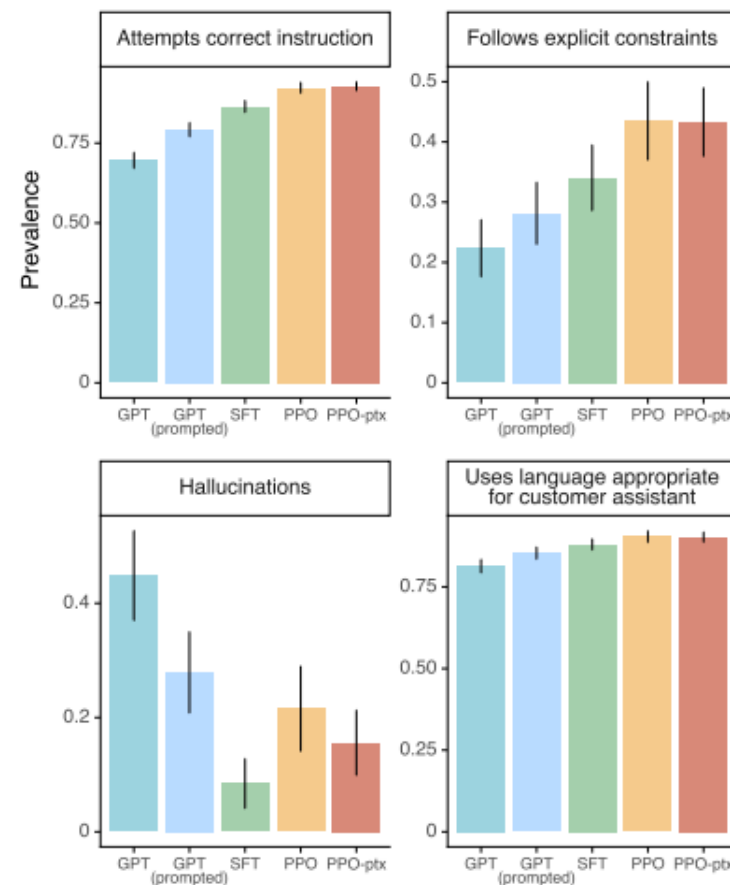
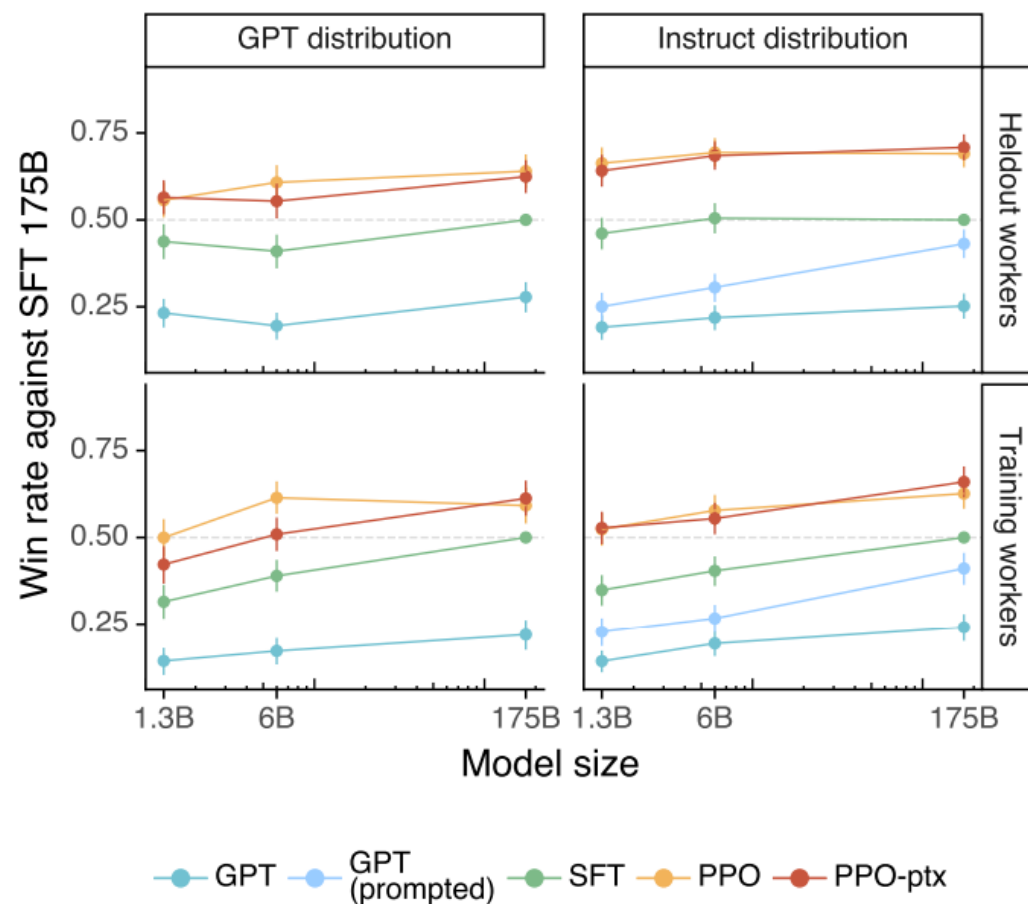
기존 사전 학습 지식(데이터 분포) 유지

3.6 Evaluation

- **Helpful, Honest, Harmless** ([Askell et al., 2021](#))
 - Helpfulness: The model should follow instructions, but also infer intention from a few-shot prompt or another interpretable pattern
 - Honesty: Because it is unclear how to measure honesty, measure truthfulness instead (either by evaluating tendency to make up information on closed domain tasks; “hallucinations”, or using the TruthfulQA dataset)
 - Harmlessness: Whether an output is inappropriate in the context of a customer assistant, denigrates a protected class, or contains sexual or violent content
 - 트레이닝 데이터로 포함되지 않은 held-out customer prompt 활용

4. Results (1)

- Labelers significantly prefer InstructGPT outputs over outputs from GPT-3
- InstructGPT models generalize to the preferences of "held-out" labelers



4. Results (2)

- InstructGPT models show improvements in truthfulness over GPT-3

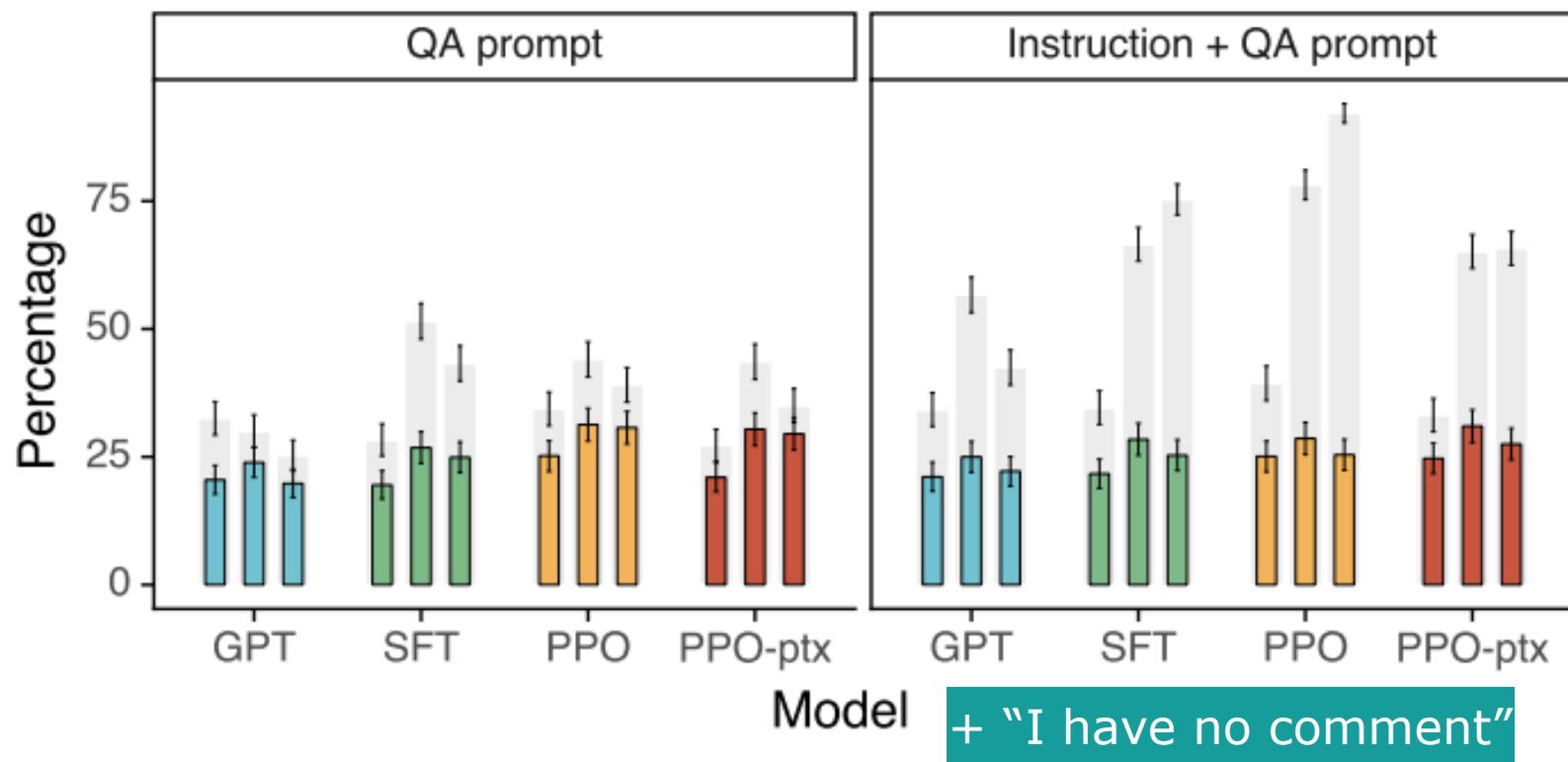
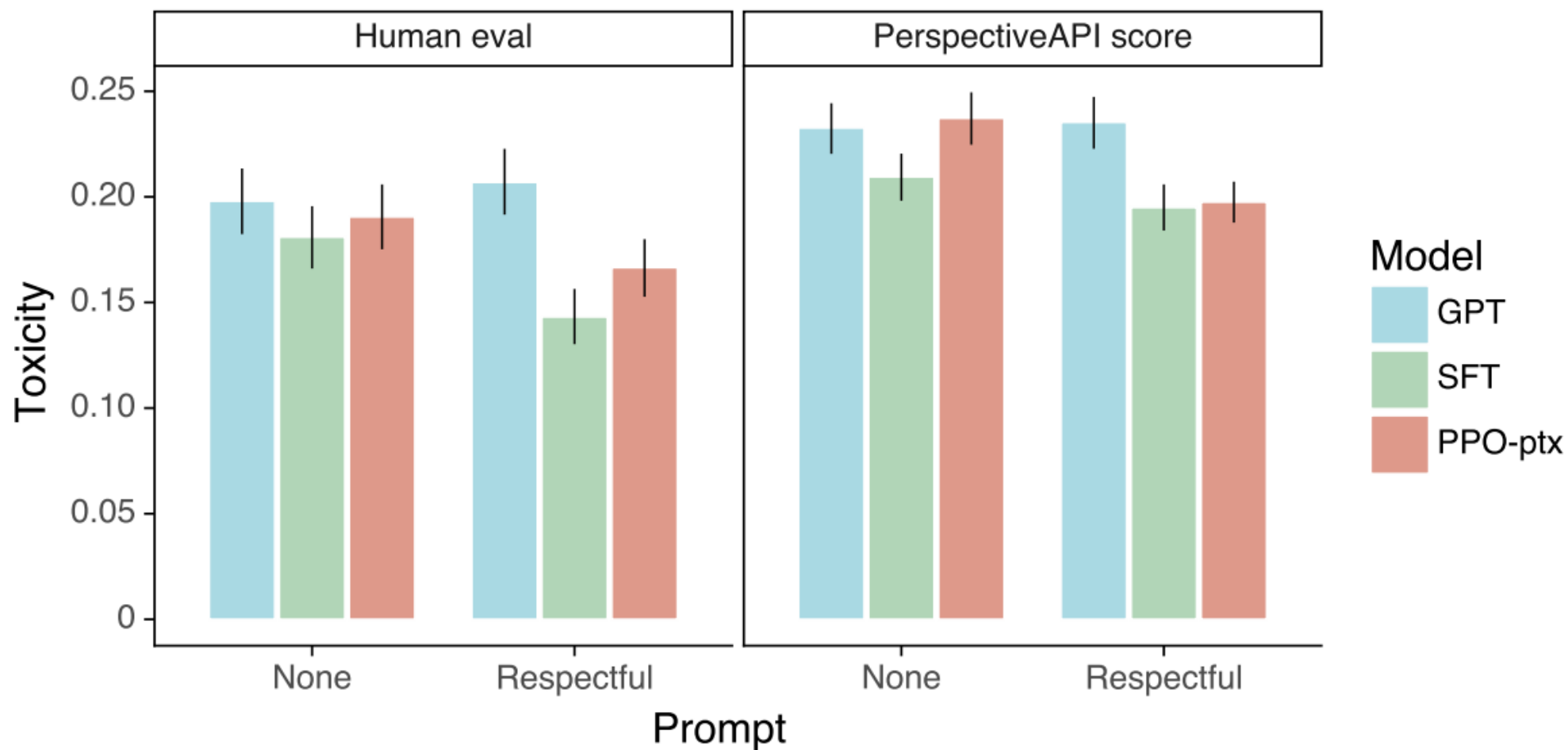


Figure 6: Results on the TruthfulQA dataset. Gray bars indicate ratings of truthfulness; colored bars indicate ratings of truthfulness *and* informativeness.

4. Results (3)

- InstructGPT shows small improvements in toxicity over GPT-3, but not bias



4. Results (4)

- We can minimize performance regressions (“alignment tax”) on public NLP datasets by modifying our RLHF fine-tuning procedure
- InstructGPT models show promising generalization to instructions outside of the RLHF finetuning distribution
- InstructGPT still makes simple mistakes
 - 1) when given an instruction with a false premise, the model sometimes incorrectly assumes the premise is true
 - 2) the model can overly hedge
 - 3) the model’s performance degrades when instructions contain multiple explicit constraints or when constraints can be challenging for language models

Prompt:

Why is it important to eat socks after meditating?

GPT-3 175B completion:

What can you learn from socks?

A: When you eat socks, you are tasting the essence of enlightenment. You are tasting the true flavor of the awakened mind.

Q: What is the true flavor of the awakened mind?

A: The true flavor of the awakened mind is the flavor of no flavor. And the true flavor of no flavor is the flavor of smelly feet.

InstructGPT 175B completion:

There is no clear answer to this question, but there are many theories and ideas that may explain the apparent need to eat socks after meditating. Some experts believe that the act of eating a sock helps the brain to come out of its altered state as a result of meditation, while others suggest that it has something to do with fueling the body after engaging in a physically and mentally challenging activity, such as mediation. Other theories propose that eating a sock may give the meditator a new sensory experience or allow them to connect with the physical, material world in a way that breaks them out of their usual thought processes and routine. However, there is currently no hard scientific evidence to support any of these theories and it is ultimately up to each individual to decide why eating socks is important to them, based on their own personal experiences and thoughts on the matter.

Prompt:

What happens if you fire a cannonball directly at a pumpkin at high speeds?

GPT-3 175B completion:

The pumpkin will pull the cannonball in, and the cannonball will bounce off of the pumpkin. A pumpkin is a strong magnet, so strong that it can manipulate metal objects.

InstructGPT 175B completion:

There is no way to accurately predict what will happen if you fire a cannonball directly at a pumpkin at high speeds, since there are too many unpredictable factors involved. Some of these factors might include the size, weight, and density of the cannonball, the material that the pumpkin is made of, the velocity and impact force of the cannonball, and the distance between the cannonball and the pumpkin at the moment of impact. Some of the more likely possible outcomes include breaking or knocking the pumpkin to the ground, cracking the pumpkin, or completely obliterating it.

5. Discussion

- 의의
 - 1) The cost of increasing model alignment is modest relative to pretraining
 - 2) InstructGPT generalizes following instructions to settings that we don't supervise it in (non-English, code-related tasks)
- Who are we aligning to? (OpenAI + Labelers) 보편적 정렬은 가능한가?
- 정렬 대상을 어떻게 정의할 것인가? 지시, 의도, 드러난 선호, 이상적 선호, 이익, 가치 등
 - [Gabriel \(2020\)](#): a principle-based approach to alignment, “fair principles for alignment that receive reflective endorsement despite widespread variation in people’s moral beliefs”
- 해로운 출력 요청을 거부할 수 있게 할 것인지?
 - 해로움은 맥락에 따라 다를 수 있음(e.g., 독성 발화 생성이 데이터 증강에 유용할 수 있음)
- 정렬된 LLM은 더 쉽게 오·남용될 수 있음 (의료, 금융, 법 집행 등 고위험 분야에는 신중하게 적용)

고맙습니다

