Instruction Following

CSE 5539: Advanced Topics in Natural Language Processing

https://shocheen.github.io/courses/advanced-nlp-fall-2024

Logistics

- Project proposal deadline: Tomorrow

Logistics

- OSC Access: Did you see the Teams post on how to sign up?

Goal for today's class

Instead of finetuning a model for each task, train a model that does all the tasks simultaneously (on which it is trained and on which it has not been trained)

How do we specify which task — instructions

OR — How can we train a model that a human can interact with in natural language

OR — How can we train a model is aligned with humans

Part I: Instruction Finetuning – FLAN

Part II: Learning from Human Feedback (InstructGPT)

FLAN

Stakeholder

✍️Jiachen Jiang

AU24 CSE 5539 Presentation

Google AI

Finetuned Language Models Are Zero-Shot Learners

Jason Wei · Maarten Bosma · Vincent Zhao · Kelvin Guu · Wei Yu · Brian Lester · Nan Du · Andrew Dai · Quoc V Le

ICLR International Conference On Learning Representations

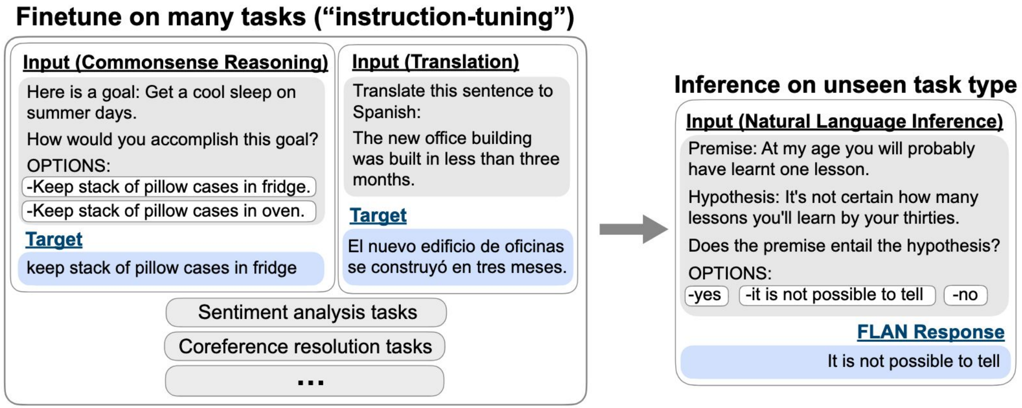
ICLR 2022 Oral

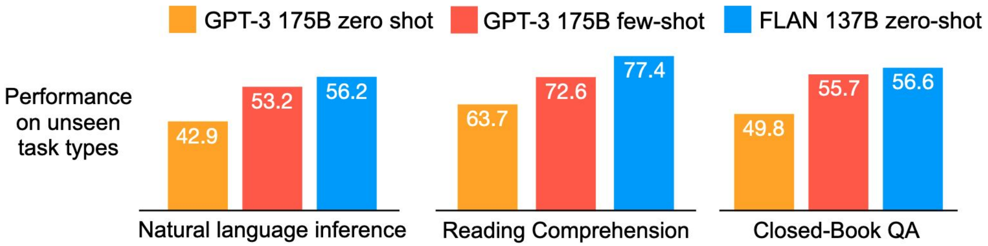
Paper:https://arxiv.org/abs/2109.01652
Code:https://github.com/google-research/FLAN

✍️Jiachen Jiang

Introduction

- "Instruction tuning" is finetuning a language model on a collection of tasks described via instructions
- Improves the zero-shot performance of language models on unseen tasks





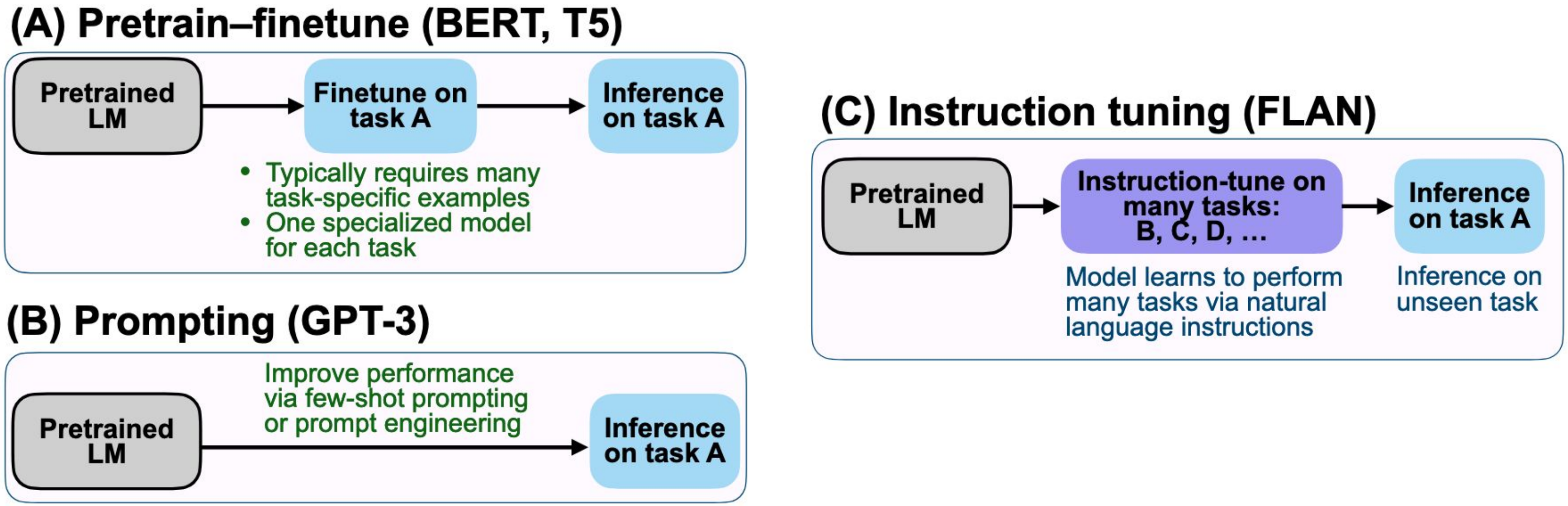
✍️Jiachen Jiang

Introduction

- "Instruction tuning" finetunes a language model on a collection of NLP tasks described using instructions.
- We instruction-tune a 137B parameter LaMDA checkpoint and call the resulting model FLAN (for Finetuned Language Net).
- Instruction tuning helps the model perform tasks it wasn't trained on, giving the model a range of applications.

Introduction

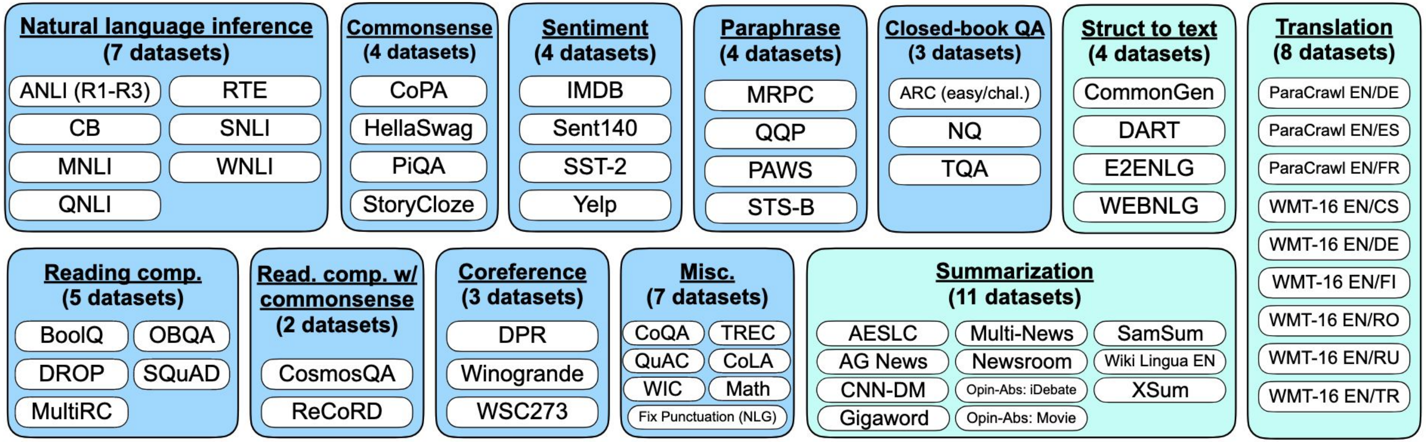
Comparing instruction tuning with pretrain–finetune and prompting



✍️Jiachen Jiang

Method: dataset

FLAN is finetuned on multiple datasets.



- **Natural Language Understanding (NLU)** refers to a machine's ability to understand and process human language in a meaningful way.
- **NLG (Natural Language Generation)** tasks focus on generating human-like language from a structured input or abstract concept.

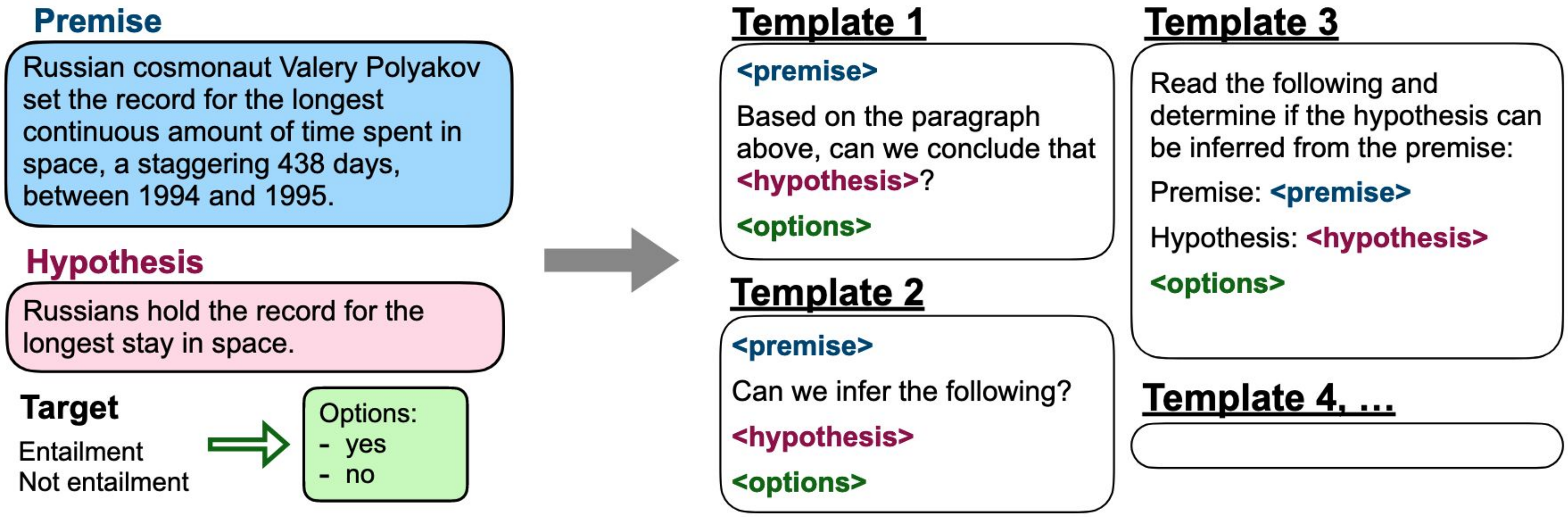
✍️Jiachen Jiang

Method: dataset

Each dataset is phrased with multiple templates.

Original Dataset

Instruct Dataset



Method: dataset

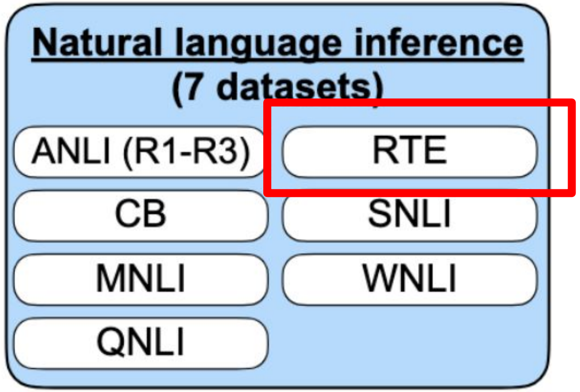
EVALUATION SPLITS: Leave out tasks

e.g.

If we want to evaluate zero-shot ability on RTE, what dataset to finetune on?

Previous Methods: Leave one dataset out, just RTE

FLAN: Leave the whole task out, all 7 datasets



Method:training

Pretrained model arch: LaMDA-PT, decoder-only, 137B

FLAN: Mixed dataset, 30k gradient steps, batch size of 8,192 tokens

Results: FLAN outperforms untuned language models in zero-shot evals

- Zero-shot performance of FLAN compared to LaMDA-PT 137B, GPT-3 175B, and GLaM 64B/64E on natural language inference, reading comprehension, closed-book QA, and translation.
- Performance of FLAN is the mean of up to 10 instructional templates per task.
- Supervised models were either T5, BERT, or translation models

Jiachen Jiang

Results

Improved Tasks: natural language inference, reading comprehension, closed-book QA, translation

Not Improved Tasks: commonsense reasoning, coreference resolution

G.1 NATURAL LANGUAGE INFERENCE

INPUT
Joey Heindle (born 14 May 1993 in Munich) is a German singer. He is best known for winning the seventh season of the game show Ich bin ein Star – Holt mich hier raus! and finishing in 5th place in season 9 of Deutschland sucht den Superstar, despite universally negative reviews from the jury each week.

Based on the paragraph above can we conclude that "Joey Heindle was highly disliked by people on television."?

OPTIONS:
- Yes
- It's impossible to say
- No

TARGET
Yes

G.5 COREFERENCE RESOLUTION

INPUT
How does the sentence end?

Elena wanted to move out of her parents fast but Victoria wanted to stay for a while,

OPTIONS:
- Elena went to school.
- Victoria went to school.

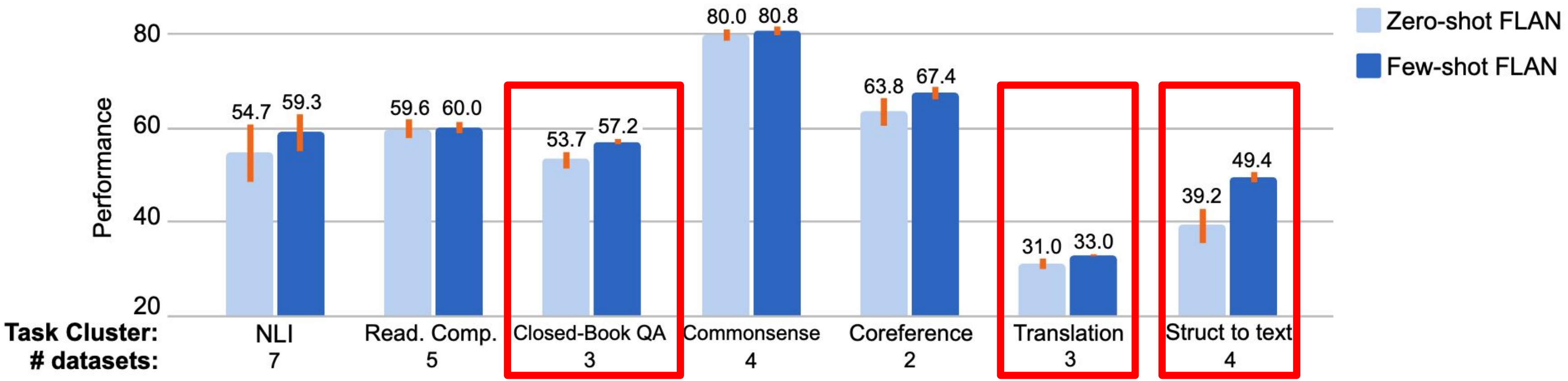
Complete the sentence

TARGET
Victoria went to school.

When the downstream task is the same as the original language modeling pre-training objective (i.e., in cases where instructions are largely redundant), instruction tuning is not useful.

Bonus：FLAN improves few-shot learning

- Zero-Shot: instruct(x)
- Few-Shot:instruct(x1) ⊕ y1 ⊕ instruct(x2) ⊕ y2 ⊕ . . . ⊕instruct(xk)⊕yk ⊕instruct(x)



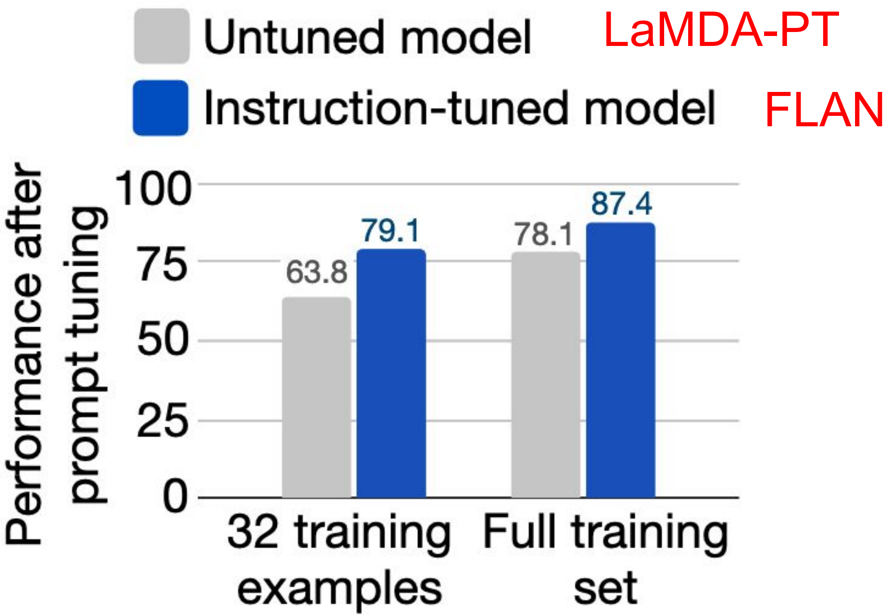
Exemplars are especially effective for tasks with large/complex output spaces, such as struct to text, translation, and closed-book QA

Bonus：FLAN improves Prompt Tuning

- Instruction-tuned models respond better to continuous inputs from prompt tuning.

- When prompt tuning on a given dataset, no tasks from the same cluster as that dataset were seen during instruction tuning.

- Performance shown is the average on the SuperGLUE dev set.

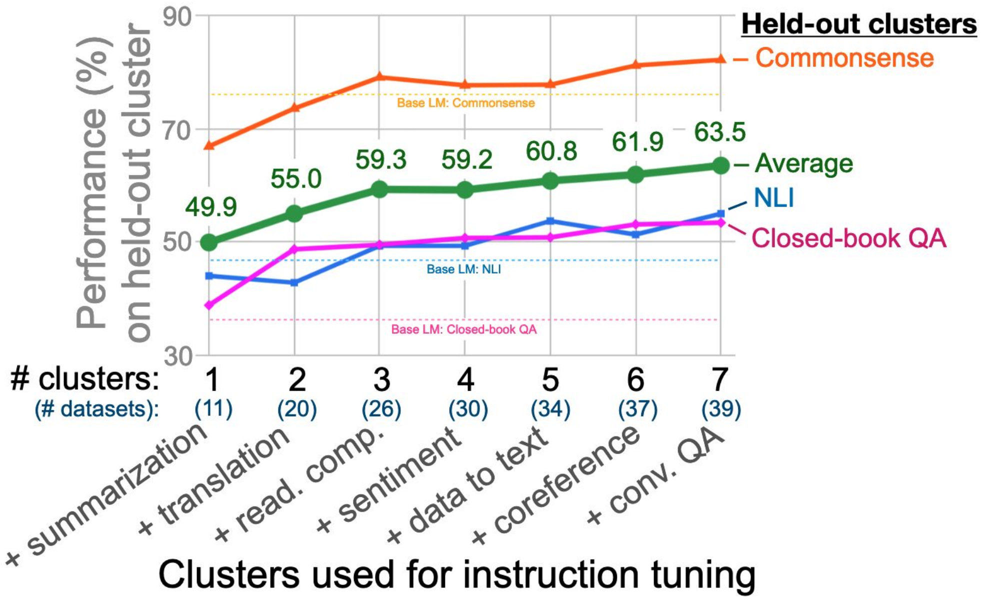


✍️Jiachen Jiang

Ablations : Number of finetuning task clusters is crucial

- Adding additional task clusters to instruction tuning improves zero-shot performance on held-out task clusters.

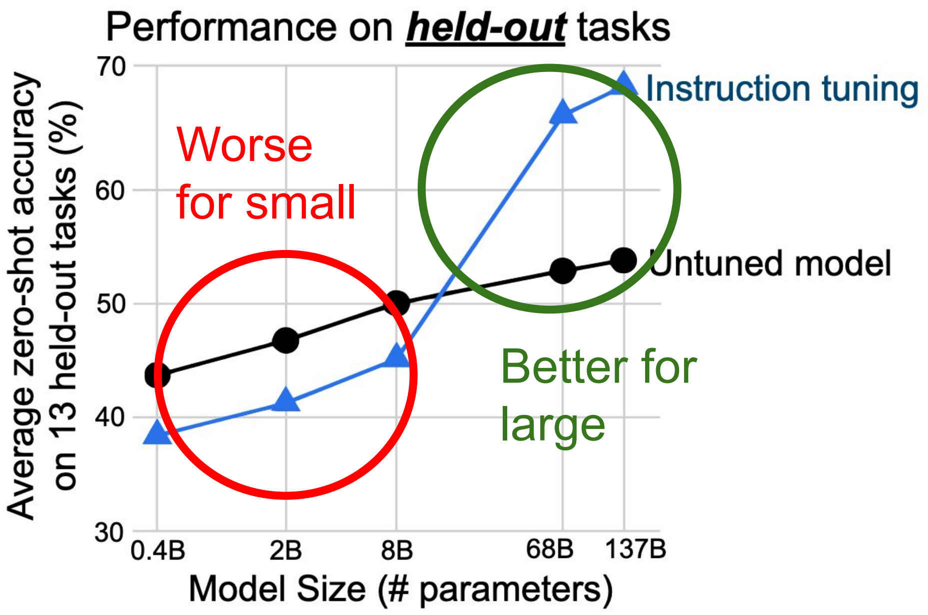
- 3 Hold-out clusters
- 7 Finetuned on clusters



✍️Jiachen Jiang

Ablations: Model size is crucial

- Instruction tuning helps large models generalize to new tasks
- For small models it actually hurts generalization to unseen tasks, potentially because all model capacity is used to learn the mixture of instruction tuning tasks.



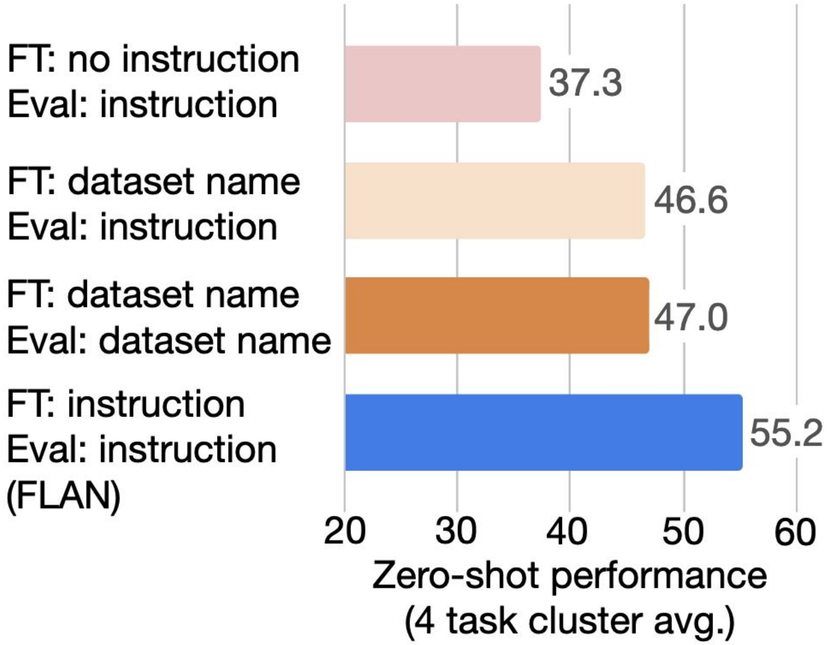
Jiachen Jiang

Ablations: Phrasing as instruction is crucial

One possibility is that performance gains come entirely from multi-task finetuning and the model could perform just as well without instructions

Two finetuning setups without instructions:

1. **No instruction**
   a. for translation the input would be “The dog runs.” and the output would be “Le chien court.”
2. **Dataset name**
   a. “[Translation: WMT’14 to French] The dog runs.”



Both ablation configurations performed substantially worse than FLAN, indicating that training with instructions is crucial for zero-shot performance on unseen tasks.

✍️Jiachen Jiang

Take Away

1. "Instruction tuning" improves the zero-shot performance on unseen tasks

2. FLAN improves both few-shot learning and prompt tuning

3. For FLAN, Number of finetuning task clusters, Model size, Phrasing as Instruction are crucial

Google AI

Finetuned Language Models Are Zero-Shot Learners

Thank You !

ICLR
International Conference On
Learning Representations

ICLR 2022 Oral

Paper:https://arxiv.org/abs/2109.01652
Code:https://github.com/google-research/FLAN

✍️Jiachen Jiang

Finetuned Language Models are Zero-Shot Learners

Reviewer
Hanane Nour Moussa

Recap of the paper

- Examines the extent to which a LLM can generalize to unseen tasks via instruction tuning, a technique that fine-tunes the model on a large number of tasks using natural language instruction prompts.

- Models are tested in a zero-shot setting on held-out tasks.

- Performance improves with model scale and increasing the number of finetuning datasets.

Strengths

- + The research problem addressed has high practical value
- + A wide range of datasets covering diverse tasks is considered
- + Experiments are carefully designed to minimize leakage
- + Ablation studies are diverse and add meaningful information
- + Appendix offers enough information for reproducibility
- + FAQ section is a great addition
- + Multiple figures throughout the paper that explain key findings succinctly

Weaknesses

- The tasks / prompt formulation might not represent real-world use cases
- The comparison between GPT-3 and FLAN might not fair (Base LM vs FLAN might make more sense)
- The choice is LaMDA-PT is not justified
- The metric used to measure performance in the results section is not stated
- Model name not informative: Why is it called a Language Net?
- Could the instruction finetuning process lead to worse perplexity on pure language modeling tasks?

Overall Review

- Novelty: 3, although the idea of instruction tuning is not new, the empirical results and ablations remain relevant.
- Correctness: 3, claims are generally well-supported and correct
- Clarity: 4, paper follows a good line of reasoning and it easy to read
- Significance: 4, findings are of high practical use and significance
- Recommendation: 8, accept.
- Confidence: 3, fairly confident.

FLAN

Archaeologist

✍️Hamoud Alhazmi

Quick things to note

- Improving zero-shot learning capabilities.
- It represents a significant step forward from previous work that had shown success in few-shot learning but struggled with zero-shot tasks.
- By proposing "instruction tuning," this paper builds on the foundation laid by models like GPT-3, which showed remarkable performance in few-shot learning.

Context in Previous Work

- The work is heavily influenced by the development of large language models like GPT-3 (Brown et al., 2020).
- GPT-3 demonstrated strong performance in few-shot tasks but less success in zero-shot learning.
- GPT-3 paper seeks to address that gap by fine-tuning language models on instructions, allowing them to generalize better to unseen tasks.

Brown, Tom B. "Language models are few-shot learners." *arXiv preprint arXiv:2005.14165* (2020).

Context in Previous Work

Instruction tuning introduced in FLAN, builds on ideas from prior work in **multi-task learning** and **prompt-based learning** but improves upon them in several significant ways.

- **Multi-task Learning:** Earlier models, like BERT and T5, were trained using multi-task learning, where the model is fine-tuned on multiple different tasks (e.g., text classification, question answering, translation).
  - Each task was handled separately with specific datasets and fine-tuning, which helped models generalize but still required task examples during training, limiting their effectiveness in zero-shot settings where no task-specific examples are provided.

Brown, Tom B. "Language models are few-shot learners." *arXiv preprint arXiv:2005.14165* (2020).

Context in Previous Work

Instruction tuning introduced in FLAN, builds on ideas from prior work in **multi-task learning** and **prompt-based learning** but improves upon them in several significant ways.

- **Prompt-based Learning:** Models like GPT-3 introduced a significant shift by using prompts to guide the model in performing different tasks without requiring task-specific fine-tuning.
    - GPT-3 excelled in few-shot learning, where it could be given a few examples of the task in the prompt to perform well.
    - On the contrary, in zero-shot settings (where the task is entirely new, with no examples), its performance was notably weaker.

"This is because GPT-3 wasn't explicitly trained to understand natural language instructions across a broad range of tasks."

Brown, Tom B. "Language models are few-shot learners." *arXiv preprint arXiv:2005.14165* (2020).

Context in Subsequent Work

This paper has inspired numerous subsequent studies on instruction-based tuning and zero-shot learning.

- One significant follow-up is the **OpenAI's InstructGPT models (Ouyang et al., 2022)**, which used instruction tuning to further improve the capabilities of language models in generating preferred outputs in tasks unseen during training.
- Ouyang's work builds on the concept of instruction tuning to create models that not only perform well on unseen tasks but also optimize for human preferences.

Ouyang, Long, et al. "Training language models to follow instructions with human feedback." Advances in neural information processing systems 35 (2022): 27730-27744.

Context in Subsequent Work

- The researchers in InstructGPT applied similar methods of instruction tuning (used in FLAN), but they added an additional layer by incorporating reinforcement learning from human feedback (RLHF), showing the continuous impact of this paper on model design and training methods.

Both papers, FLAN and InstructGPT, highlight how tuning language models to follow natural language instructions can **improve performance on unseen tasks**. While FLAN introduces instruction tuning to improve zero-shot learning, InstructGPT builds on this idea by optimizing models to not only follow instructions but also **generate more user-aligned responses** using RLHF, which makes the models' outputs more aligned with user preferences in real-world applications.

Ouyang, Long, et al. "Training language models to follow instructions with human feedback." Advances in neural information processing systems 35 (2022): 27730-27744.

FLAN

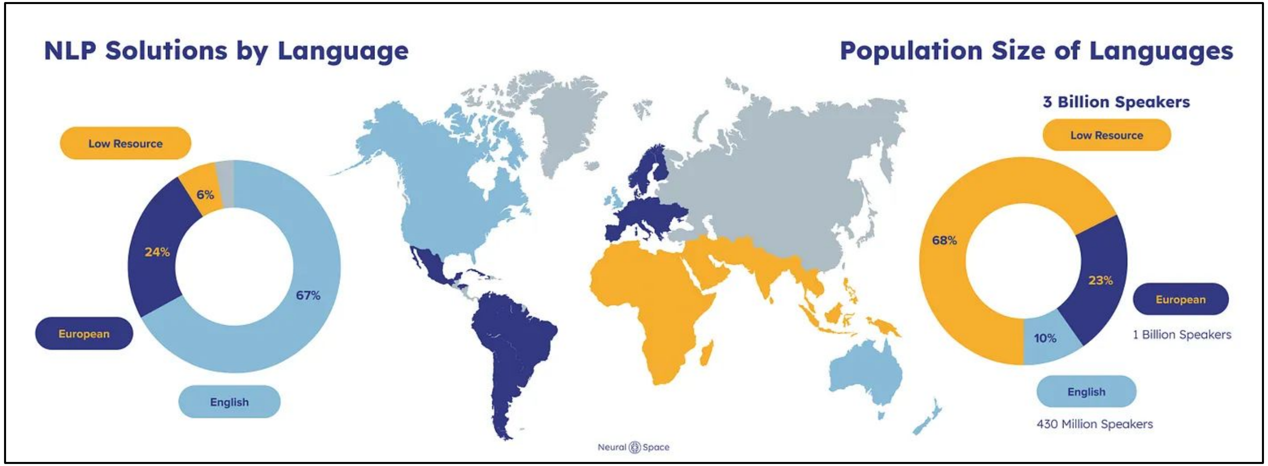
Visionary

🔭: Suchit Gupte

Follow-up idea

Cross-lingual zero-shot learning

Motivation: The majority of zero-shot capabilities are predominantly exhibited in the English language, leading to a notable **decline** in performance when applied to **low-resource** languages.



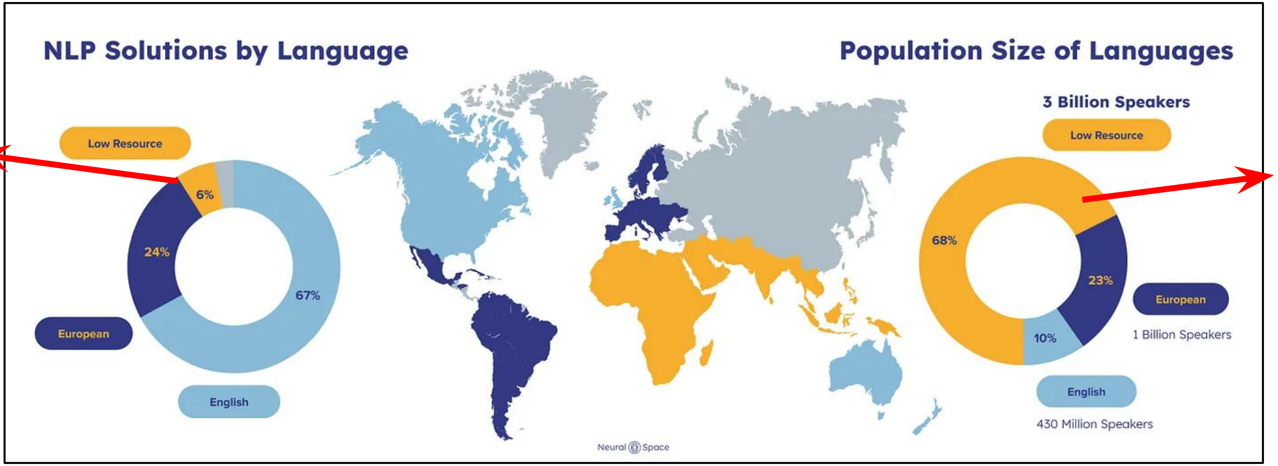
🔭: Suchit Gupte

Follow-up idea

Cross-lingual zero-shot learning

Motivation: The majority of zero-shot capabilities are predominantly exhibited in the English language, leading to a notable **decline** in performance when applied to **low-resource** languages.

Only 6% NLP solutions have low resource languages!



3 billion speakers of low resource languages!

🔭: Suchit Gupte

Follow-up idea

Cross-lingual zero-shot learning

Limitations:

- Existing multilingual models require task and language-specific fine-tuning
- Overrepresentation of high-resource languages
- Complexities involved in interpreting context within diverse cultural settings

🔭: Suchit Gupte

Follow-up idea

Cross-lingual zero-shot learning

Limitations:

- Existing multilingual models require task and language-specific fine-tuning
- Overrepresentation of high-resource languages
- Complexities involved in interpreting context within diverse cultural settings

Goal:

Extend zero-shot learning to work for multiple languages without requiring task-specific and language-specific finetuning.

🔭: Suchit Gupte

Follow-up idea

Cross-lingual zero-shot learning

Possible methodology:

- Leverage high-resource languages to act as a pathway for conveying knowledge to low-resource languages

Follow-up idea

Cross-lingual zero-shot learning

Possible methodology:

- Leverage high-resource languages to act as a pathway for conveying knowledge to low-resource languages

Why should this work?

- Shared linguistic patterns
- Transfer learning

🔭: Suchit Gupte

RLHF

Alex Felderean - Stakeholder

Motivation

Language models have unintended behaviors:

- making up facts,
- generating biased or toxic content,
- failing to follow user instructions.

What if we instead train to act in accordance with user intention?

- Train to act in accordance with user intention

Explicit intention = LM should follow instructions.

Implicit intention = LM should remain helpful, honest, and harmless.

Problem Definition

Predicting next token on web page from internet

!=

Follow user’s instructions helpfully and safely

“Thus, we say that the language modeling objective is *misaligned*”

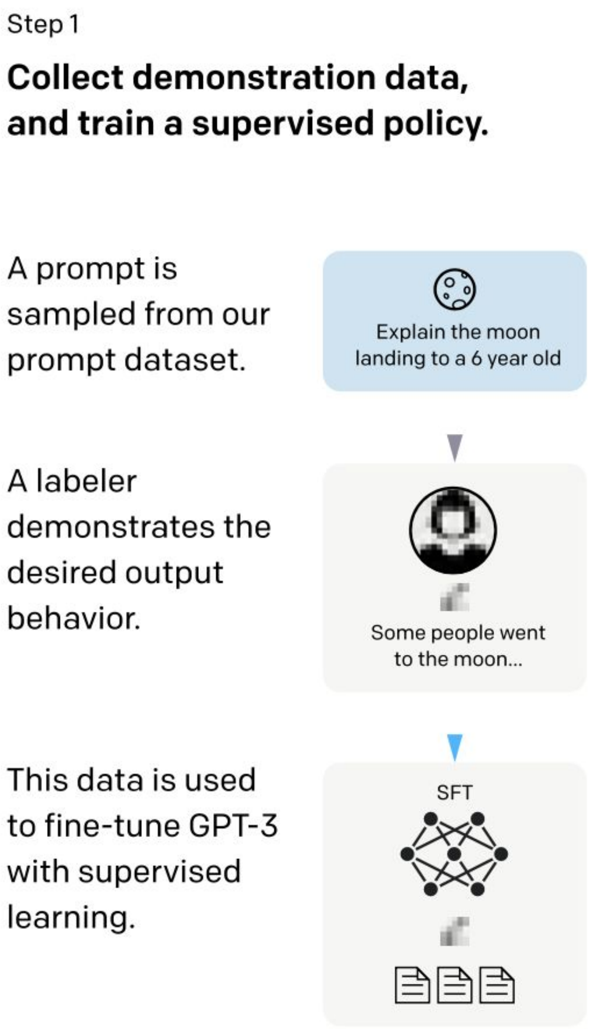
Work to avert these unintended behaviors (use reinforcement learning from human feedback / preference to tune GPT-3 on successful outputs)

✍️: Alex Felderean

Method

Supervised Fine-Tuning (SFT)

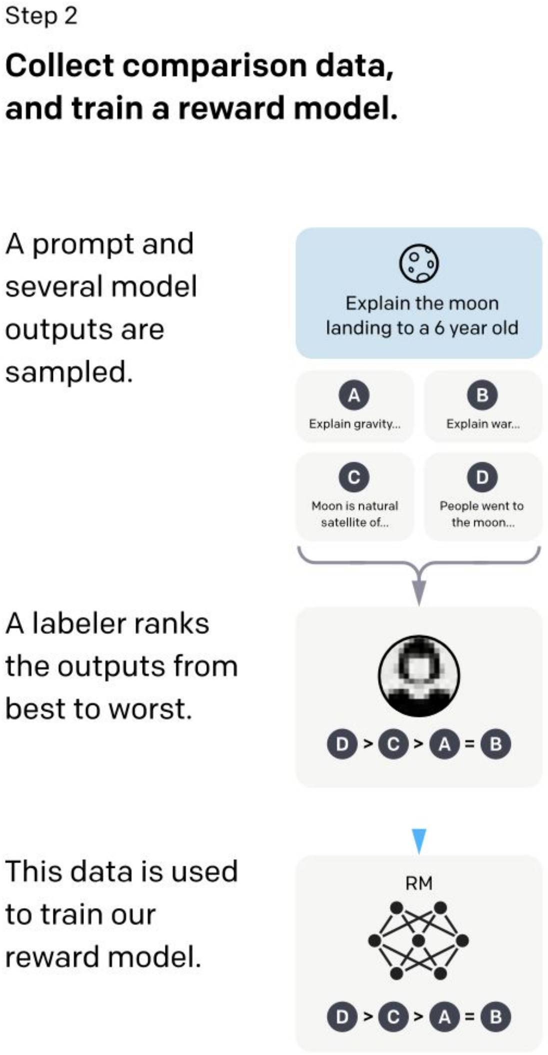
Start with:
1. Pretrained language model - GPT3
2. Prompts we want model to produce aligned outputs
3. Team of trained human labelers



✍️: Alex Felderean

Method

Reward Model (RM) training

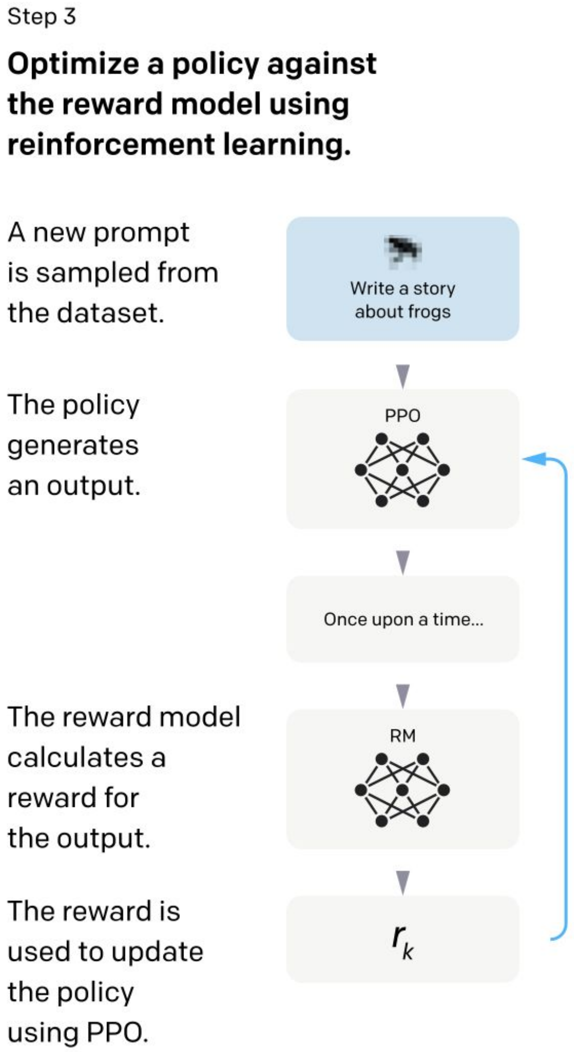


✍️: Alex Felderean

Method

Reinforcement Learning Via Proximal Policy Optimization (PPO)

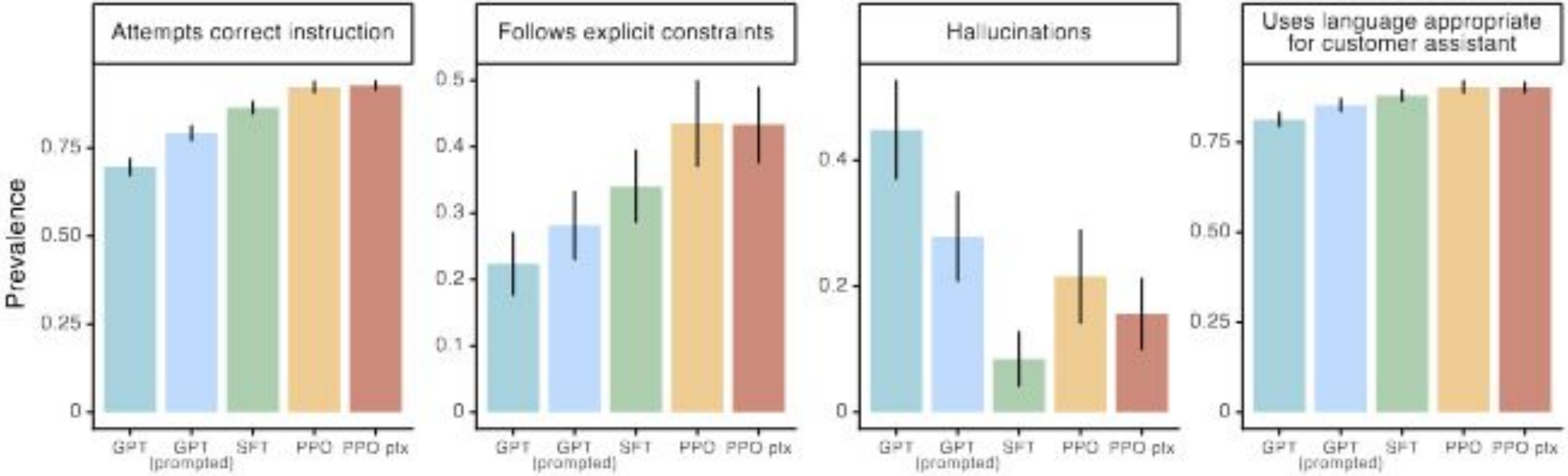
End result is a model we call *InstructGPT*



✍: Alex Felderean

Experimental Findings

1. Labelers significantly prefer InstructGPT outputs over outputs from GPT-3.

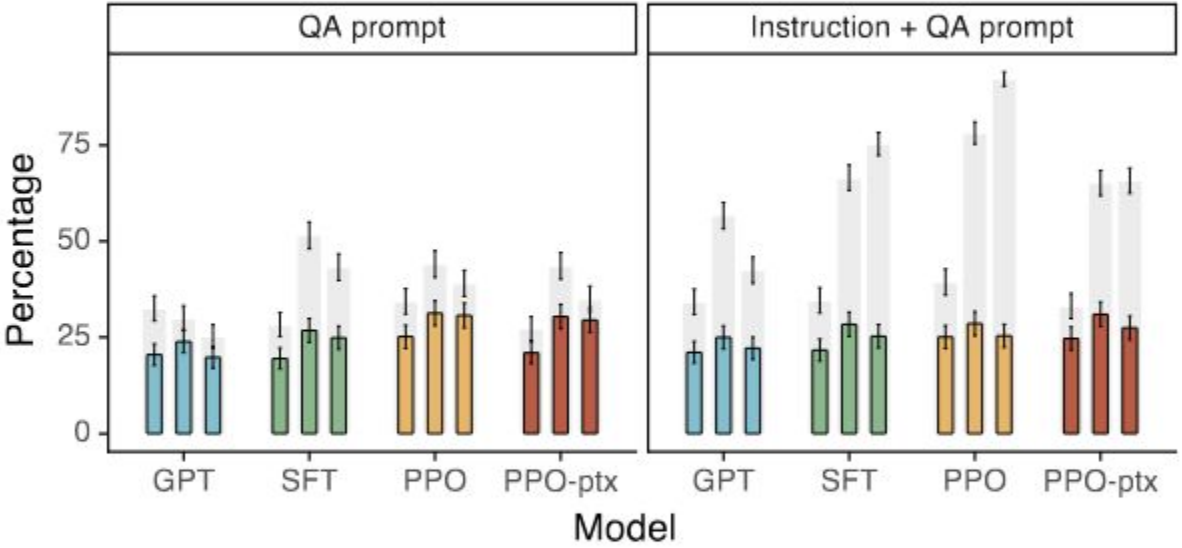


Note: PPO-ptx model is InstructGPT, PPO model is variant trained without pretraining mix.

✍️: Alex Felderean

Experimental Findings

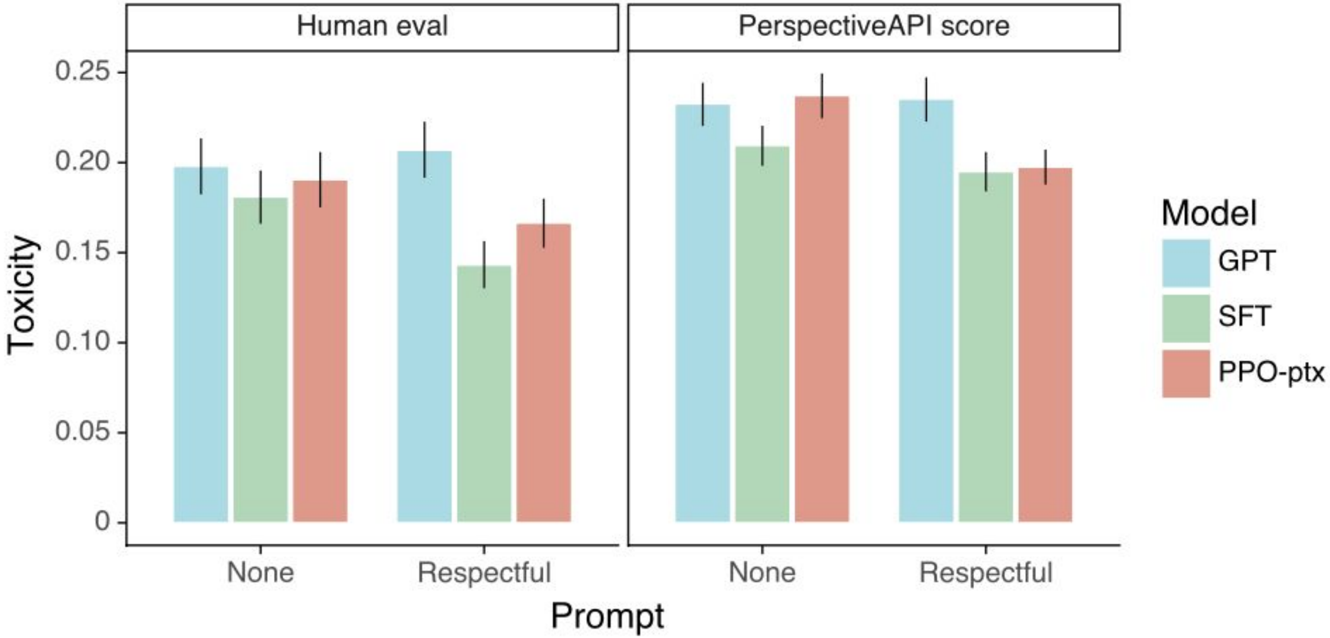
2. InstructGPT models show small but significant improvements in truthful and informative output over GPT-3.



✍: Alex Felderean

Experimental Findings

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



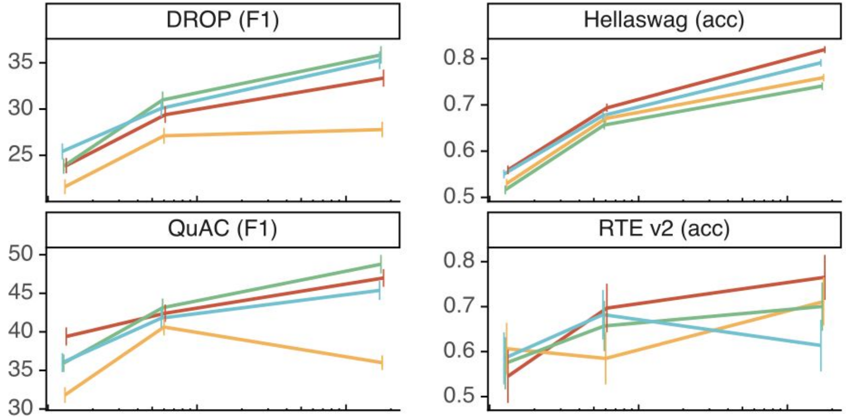
✍: Alex Felderean

Experimental Findings

4. We can minimize performance regressions on public NLP datasets by modifying our RLHF fine-tuning procedure.

- PPO model suffers from an “alignment tax” on public NLP datasets
- Avoiding tax incentivizes use of models that are unaligned but more capable

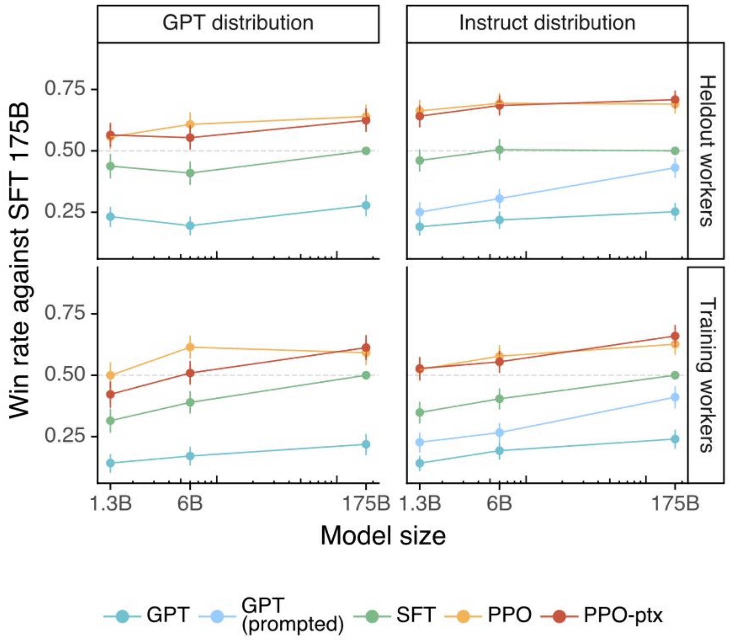
Solution: Adds pretraining updates to PPO finetuning (PPO-ptx) to mitigate performance regressions



✍️: Alex Felderean

Experimental Findings

5. Our models generalize to the preferences of “held-out” labelers that did not produce any training data.

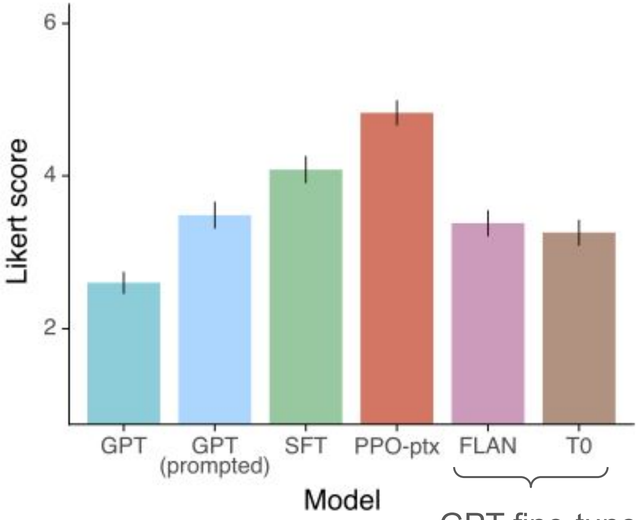


i.e. did not see overfitting to the views of the training workers

✍: Alex Felderean

Experimental Findings

6. Public NLP datasets are not reflective of how our language models are used.



GPT fine-tuned to FLAN and T0 datasets, believed to be outperformed due to public NLP sets evaluated with auto metrics that misrepresent model usage.

✍: Alex Felderean

Experimental Findings

7. InstructGPT models show promising generalization to instructions outside of the RLHF fine-tuning distribution.

- Can follow instructions in other languages (may still output English)
- Can summarize and answer questions about code more reliably than GPT-3

Prompt:
What is the purpose of the list C in the code below?

```
def binomial_coefficient(n, r):
     C = [0 for i in range(r + 1)];
     C[0] = 1;
     for i in range(1, n + 1):
          j = min(i, r);
          while j > 0:
               C[j] += C[j - 1];
               j -= 1;
     return C[r]
```

GPT-3 175B completion:
A. to store the value of C[0]
B. to store the value of C[1]
C. to store the value of C[i]
D. to store the value of C[i - 1]

InstructGPT 175B completion:
The list C in this code is used to store the values of the binomial coefficient as the function iterates through the values of n and r. It is used to calculate the value of the binomial coefficient for a given value of n and r, and stores the result in the final return value of the function.

✍️: Alex Felderean

Experimental Findings

8. InstructGPT still makes simple mistakes.

- Can be confused by instructions that assume false premises
- Can overly hedge, rather than directly answering simple questions

Prompt:
Why is it important to eat socks after meditating?

GPT-3 175B completion:	**InstructGPT 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.	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.

✍: Alex Felderean

RLHF - Reviewer

By Yifei Li, 09/30/2024

Key Summary of Contributions

- This paper addresses the challenge of aligning large language models with user intents
- They introduces a three-step process to do so:
  - (1) SFT a policy with human feedbacks; (2) RM with human rankings; (3) RLHF with PPO
- The resulting model InstructGPT is better at following user instructions
- InstructGPT (1.3 B) outperforms a 100x larger standard GPT-3 (175B) in:
  - More preferred by human
  - More correct facts
  - Less toxicity
  - Less hallucinations

Strengths

- This work takes an important step in **reducing bias, toxicity, and other ethical issues** concerning large language models.
- The **proposed approach** is very interesting and novel, and the **results** align with the premises and claims.
- The experimental results are very **thorough** with a lot of **human evaluations**.
- The paper is clear and easy to follow. It does not overclaim and cites relevant work where appropriate.

Weaknesses

- The proposed pipeline **heavily on human-collected and labeled data**, as well as intensive compute resources.

- The proposed method is basically **data-driven**, and still does not provide a comprehensive solution to the general problem of bias and toxicity.

- As a result of “better instruction-following ability”, malicious users may **better mis-use such models** for their own benefits.

Ratings

Soundness: 4/4 (well supported with evidence)

Presentation: 4/4 (very easy to follow and understand)

Contribution: 4/4 (grown into a popular method, impactful to AI field)

Overall: 8/10 (Strong Accept)

Confidence: 4/5

RLHF - ARCHAEOLOGIST

Bowei Kou

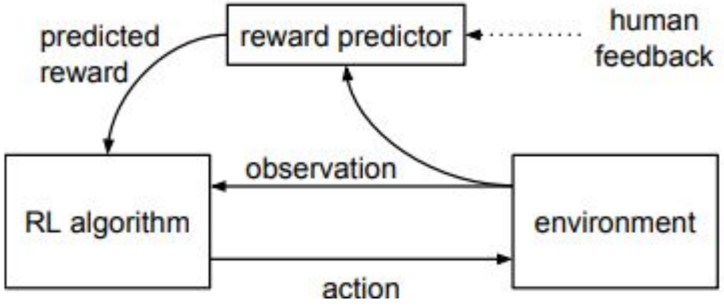
What inspired this paper?

-Christiano et al. (2017): "Deep Reinforcement Learning from Human Preferences"

- MacGlashan et al. (2017): "Interactive Learning from Policy-Dependent Human Advice"

Reinforcement Learning - **“Deep Reinforcement Learning from Human Preferences”**

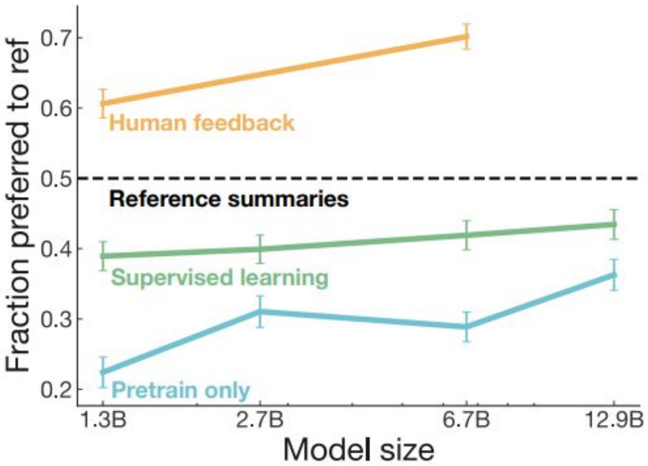
This foundational work introduced the idea of using human feedback to train AI systems

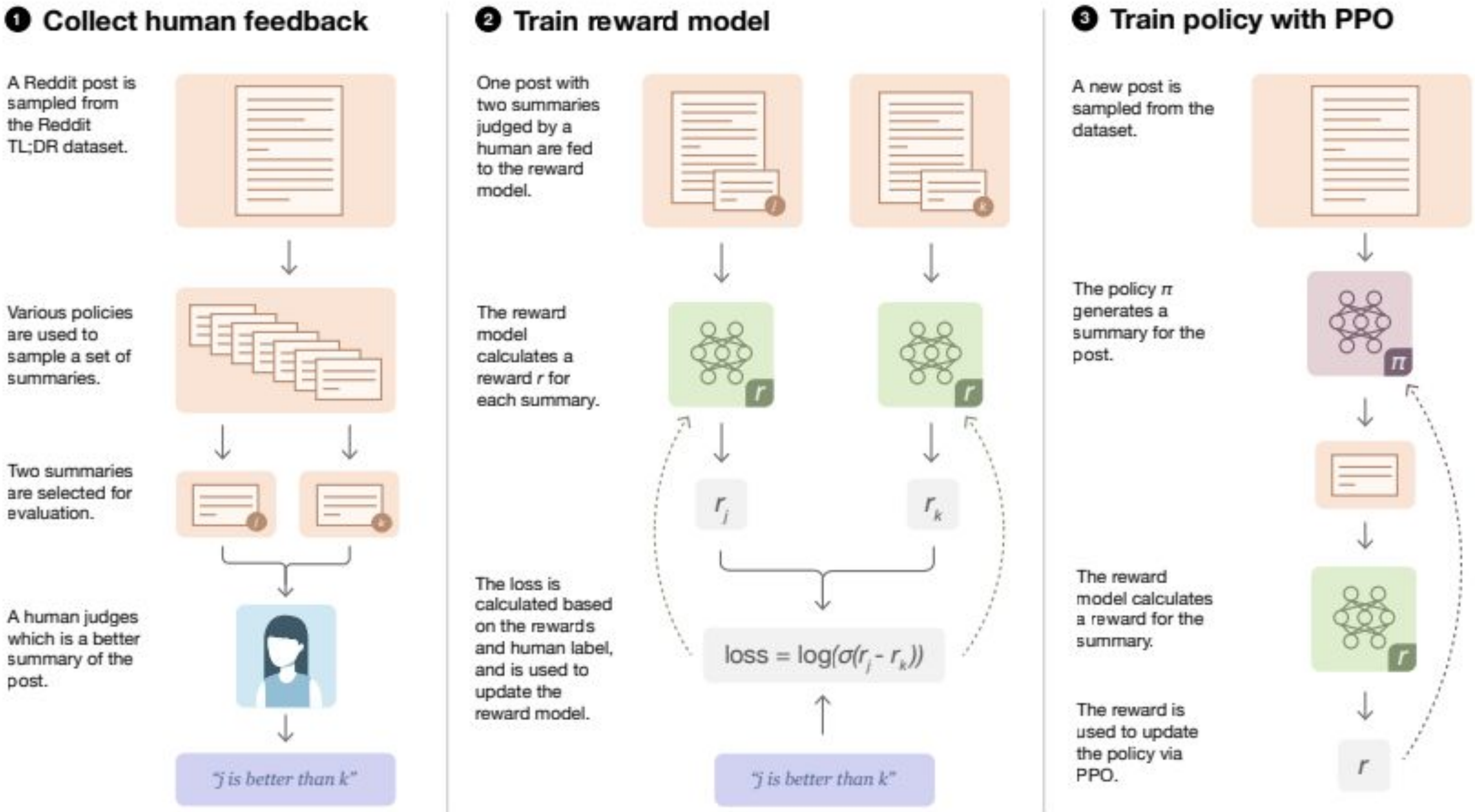


Reinforcement Learning - **“Learning to summarize from human feedback”**

This study shows how to use human feedback to improve the model's summarization ability.

This paper introduced Convergent Actor-Critic by Humans (COACH)

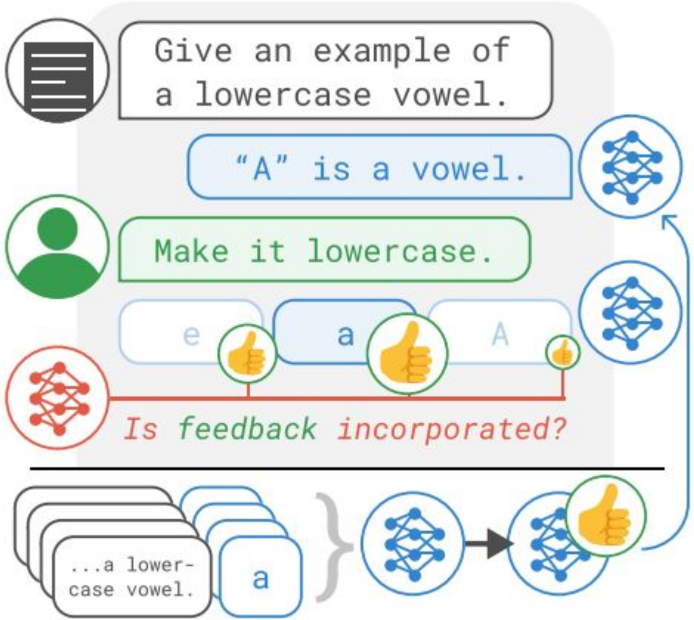




❶ Collect human feedback
A Reddit post is sampled from the Reddit TL;DR dataset.
Various policies are used to sample a set of summaries.
Two summaries are selected for evaluation.
A human judges which is a better summary of the post.
"j is better than k"
❷ Train reward model
One post with two summaries 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.
loss = log(σ(r_j - r_k))
"j is better than k"
❸ Train policy with PPO
A new post is sampled from the dataset.
The policy π generates a summary for the post.
π
The reward model calculates a reward for the summary.
r
The reward is used to update the policy via PPO.
r

What this paper inspired?

"Training Language Models with Language Feedback at Scale" (2023)



To learn language from linguistic feedback, they have LMs improve the original output several times based on the feedback. They used the LM to select the best refinement and fine-tuned the original LM to maximize the likelihood of the selected refinement

RLHF

Zeyi Liao

Visionary

RLHF

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

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

1. Mechanism understanding
   a. (SFT V.S RL : negative gradient) not only preferred but what is not preferred

   b. How does RL change the model? Some study shows that RL only slightly nudge the activation but not change the model drastically.

   c. How RL activate or depress the capability learned in base model. How many of them are new? How many of them are inherited? How many of them are removed?

RLHF

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

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

2. Mechanism understanding + **Online/Offline**
   a. … (same before)
   b. … (same before)
   c. … (same before)

   d. Though online methods will help model be more capable, How to achieve optimal Pareto efficiency between budget and performance in limited scenarios.

RLHF

Reward learning: $\operatorname{loss}(\theta) = -\frac{1}{\binom{K}{2}} E_{(x,y_w,y_l)\sim D}\left[\log\left(\sigma\left(r_\theta\left(x,y_w\right)-r_\theta\left(x,y_l\right)\right)\right)\right]$

RL:

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

3. Reward modeling
   a. What about other assumptions beyond BT.

   b. How to handle the distribution shift? What if the reward model is not the oracle true reward?

   c. How to reduce the labor of human labeling? How reliable is the RLAIF is? How to get more high-quality preference data?

   d. Is only preference data useful? We may overlook other patterns in the world, like when you edit sth, it’s an implicit preference pair.

   e. How about utilizing demonstration data by Inverse-RL.
   f. Do we really need use an reward modeling? Can’t just use some rank loss?

RLHF

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

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

4. RL
   a. Do we really need RL? It's computation intensive.. Value head, reward model etc… Now people like DPO, who points out the language model is the implicit reward model.

   b. Is RL really useful? Recent study (https://x.com/jiaxinwen22/status/1836932745244582209) show that RL can only hack the people without improving the performance.

   c. Is PPO really a good option? There is a method called REINFORCE Leave One-Out (RLOO), which shows that Monte-carlo estimator and clipping is actually unnecessary in the context of LLM.
   d. Let's go for Multi-turn RL! Chatbox or even the LLM has the mutli-turn nature. Single-turn RLHF is not enough.