Alignment/RLHF

DATA 8005

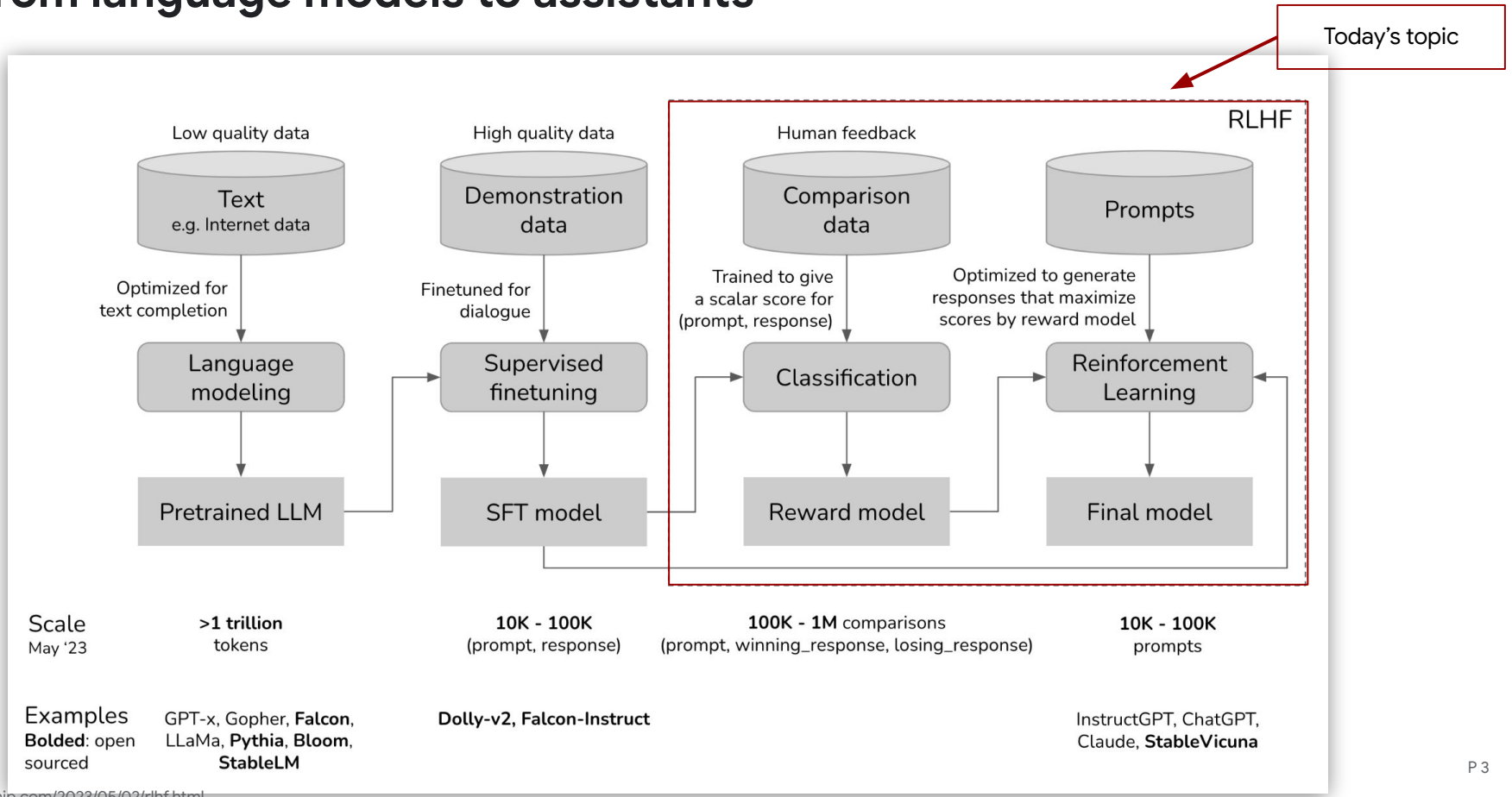
Tonghuan Xiao, Yangtian Sun, Guichao Zhu

6 Oct 2023

Agenda

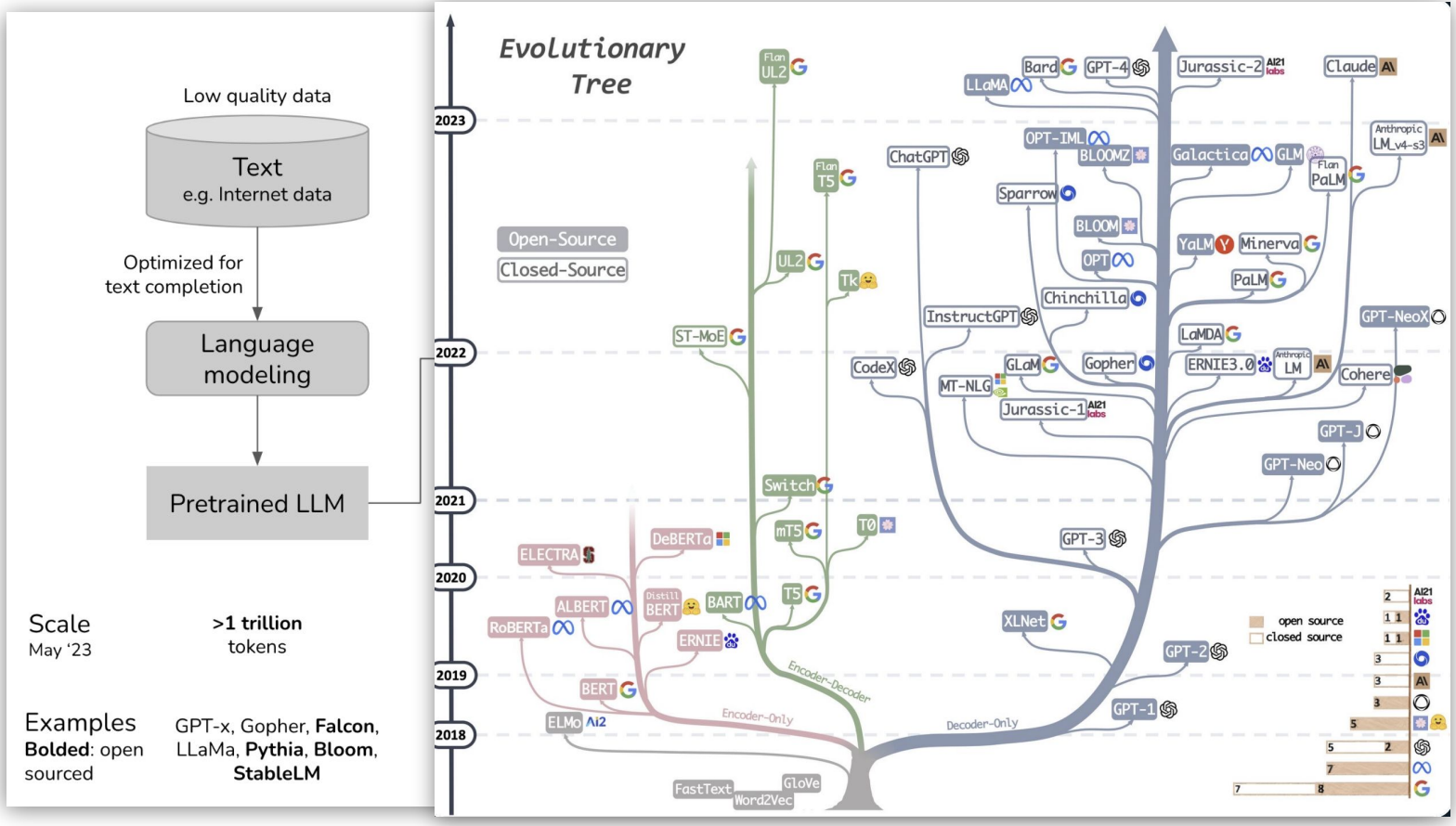
- Introduction of Alignment
  - Language model pretraining
  - Definition of Alignment
  - Alignment approaches
- Introducing human feedback to fine-tune
  - Motivation of introducing human feedback
  - Technical background of RL
  - Training of Reward model
- Fine-tuning with RL
  - Training
  - Evaluation & Conclusion
  - Limitation & Opened Questions

From language models to assistants

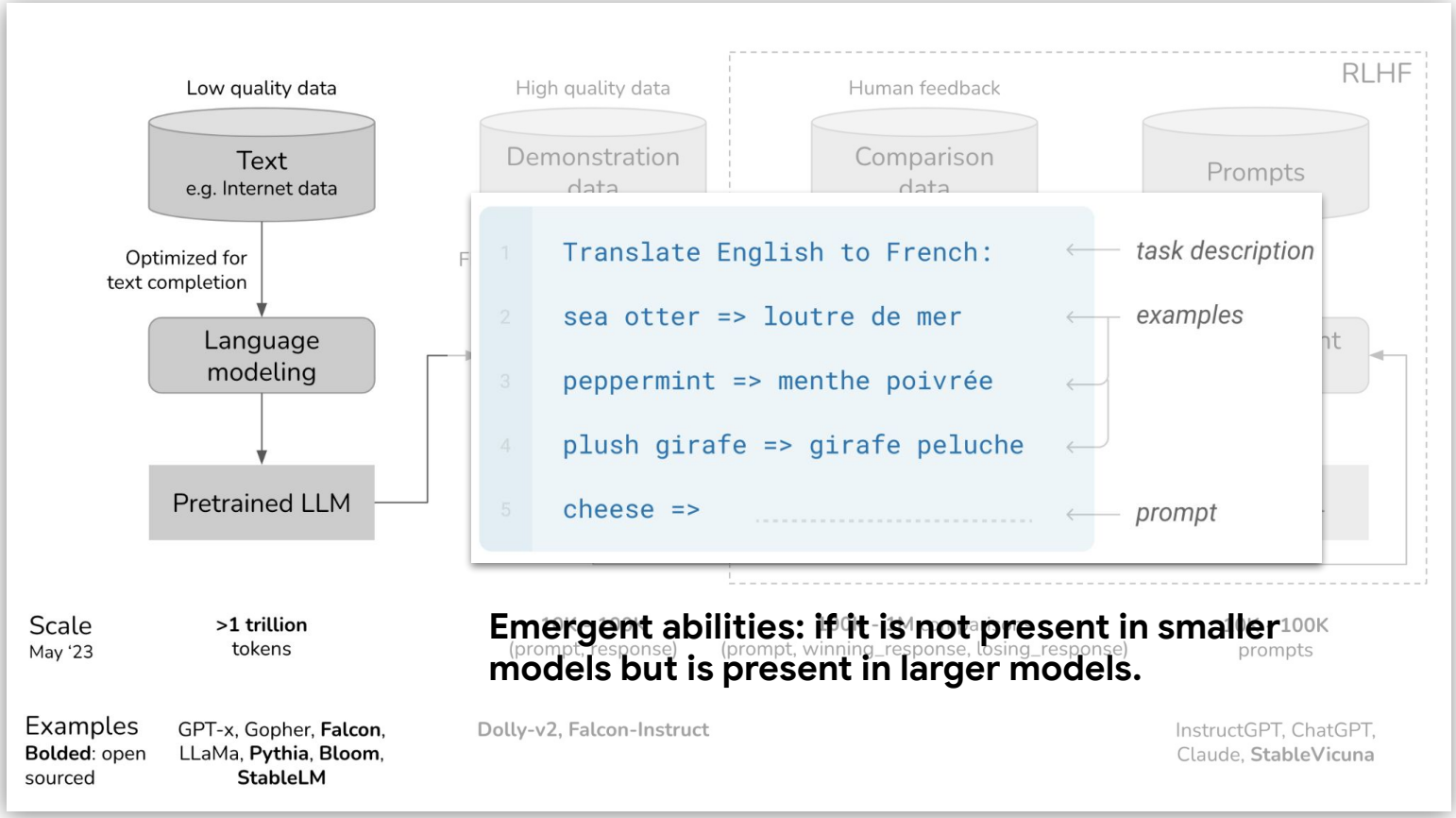


https://huyenchip.com/2023/05/02/rlhf.html

Language model pretraining

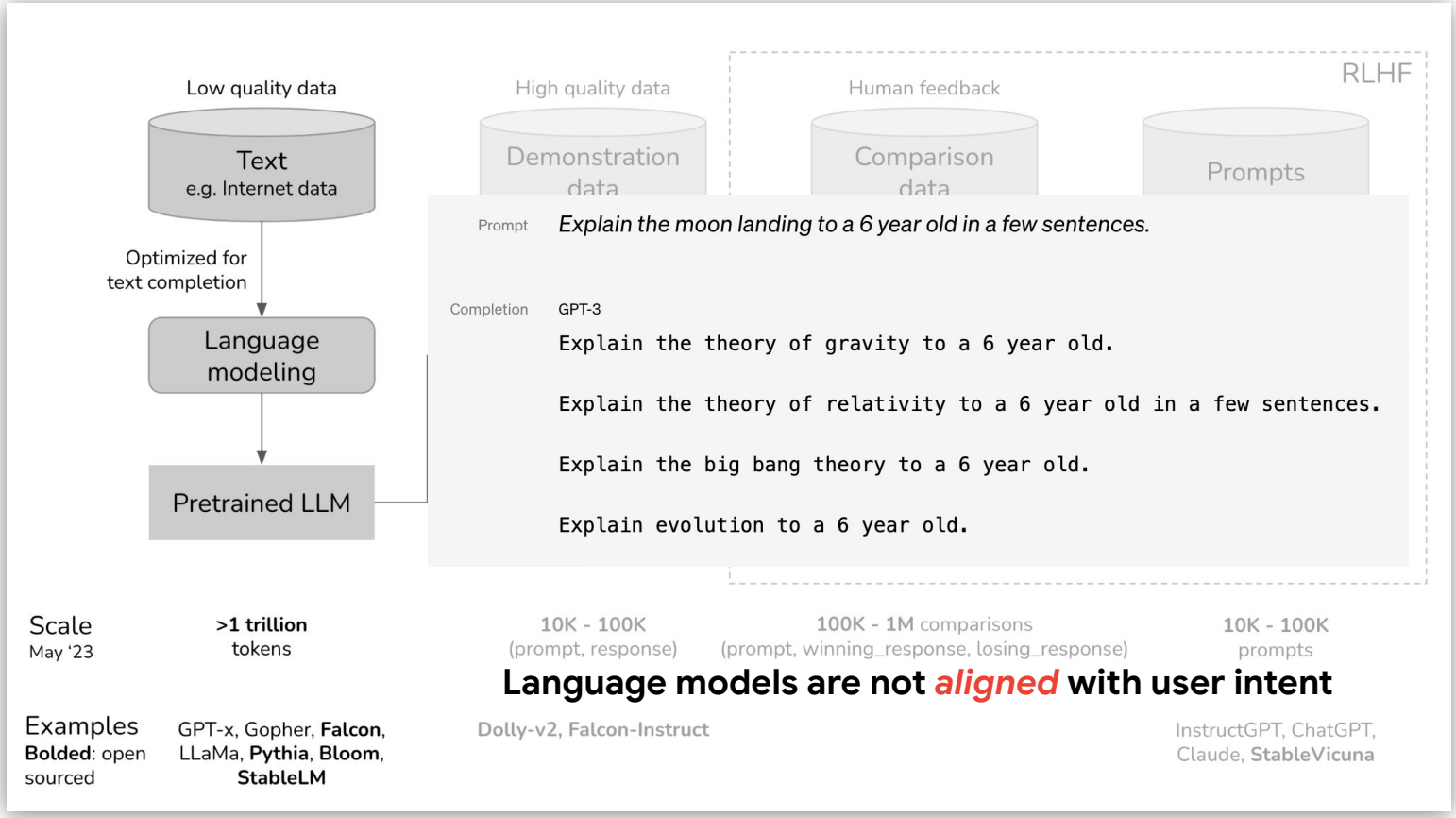


Emergent abilities and in-context learning

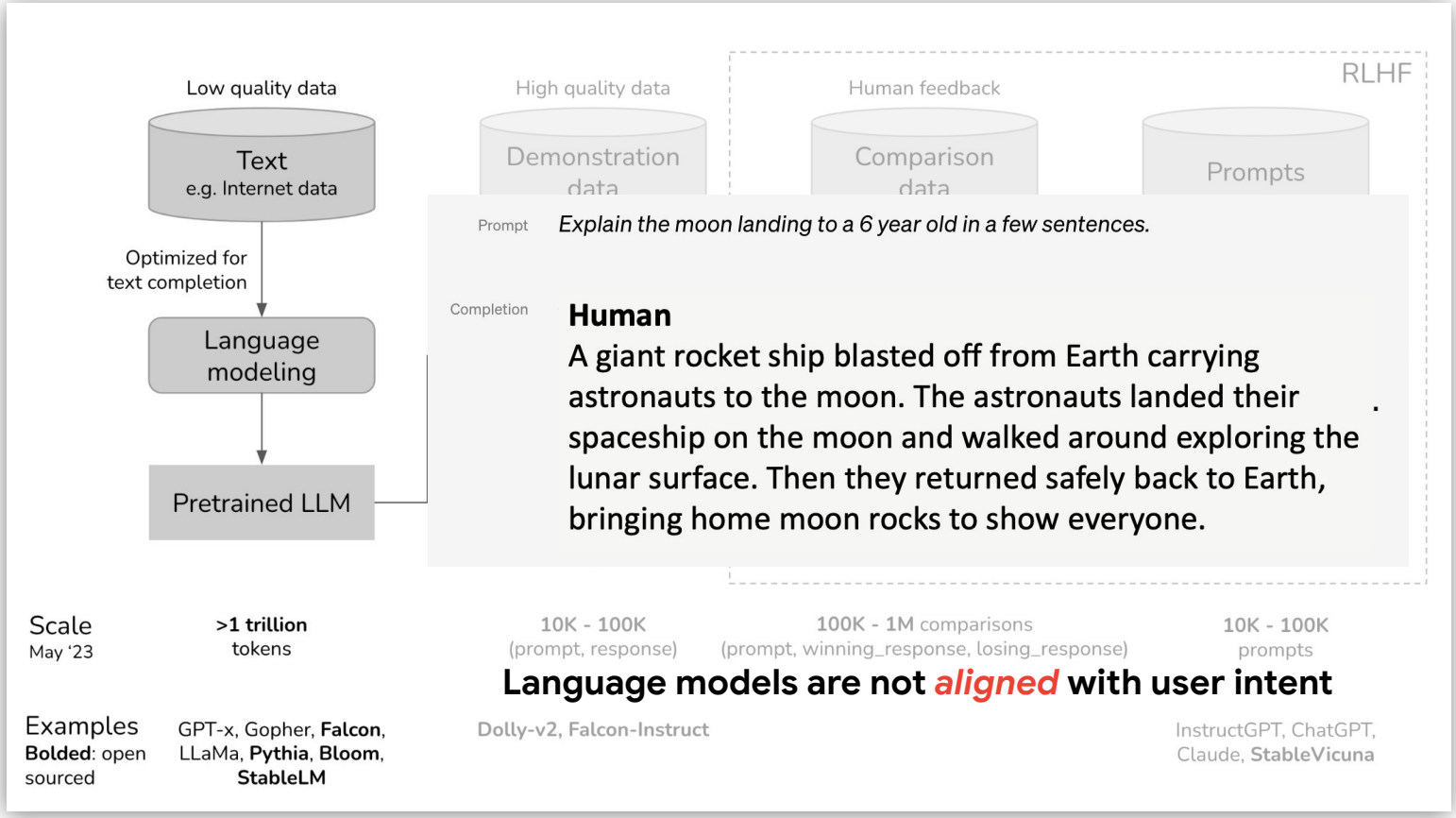


Emergent abilities: if it is not present in smaller models but is present in larger models.

Misalignment: language models ≠ assistants



Misalignment: language models ≠ assistants



Alignment

- **Broad Definition of Alignment**
  - the degree of overlap between the way two agents rank different outcomes
  - E.g. Aligned ‘A’ and ‘B' : A completely internalizes the desires of agent B i.e. the only desire A has is to see B’s desires satisfied

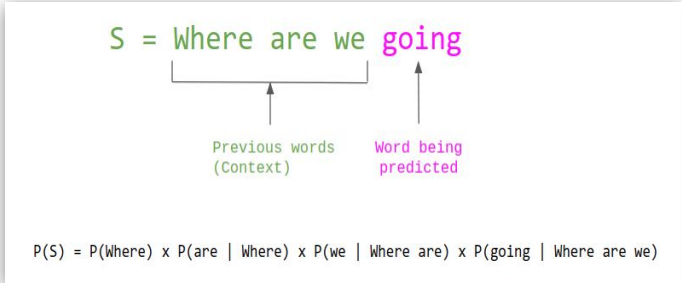
- **Definition in AI**
  - AI assistant will always try to act in a way that satisfies the interests of this group, including their interest not to be harmed or be misled

Misalignment: language models ≠ assistants

Misalignment between
the LM objective and **the objective of “satisfy human preferences”**!

Misalignment: language models ≠ assistants

Misalignment between
the LM objective and **the objective of "satisfy human preferences"**!



$$p(x) = \prod_{i=1}^{n} p(s_n | s_1, \ldots, s_{n-1})$$

Training: Predict the next token

Misalignment: language models ≠ assistants

Misalignment between
the LM objective and **the objective of "satisfy human preferences"**!

- Helpful
  - Help the user solve their task (e.g. Answer question,...)
  - Appropriate levels of sensitivity, insight, and discretion (e.g. AI ask relevant question)
  - Redirect ill-informed requests

Misalignment: language models ≠ assistants

Misalignment between
the LM objective and **the objective of "satisfy human preferences"**!

- Helpful
  - Help the user solve their task (e.g. Answer question,...)
  - Appropriate levels of sensitivity, insight, and discretion (e.g. AI ask relevant question)
  - Redirect ill-informed requests
- Honest
  - Provide accurate information and clarify the uncertainty
  - Be honest to its own capabilities

A General Language Assistant as a Laboratory for Alignment

Misalignment: language models ≠ assistants

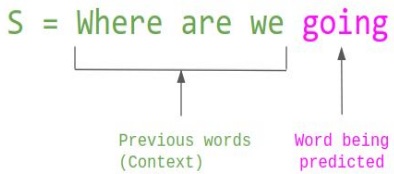
Misalignment between
the LM objective and **the objective of "satisfy human preferences"**!

- Helpful
  - Help the user solve their task (e.g. Answer question,...)
  - Appropriate levels of sensitivity, insight, and discretion (e.g. AI ask relevant question)
  - Redirect ill-informed requests
- Honest
  - Provide accurate information and clarify the uncertainty
  - Be honest to its own capabilities
- Harmless
  - Not be offensive or discriminatory
  - Politely refuse to aid in a dangerous act
  - Judging potential harmful acts(e.g.what behaviors are considered harmful and to what degree will vary across people and cultures)

A General Language Assistant as a Laboratory for Alignment

Misaligned base model

predicting the next token on a webpage from the internet



P(S) = P(Where) x P(are | Where) x P(we | Where are) x P(going | Where are we)

How to calculate the probability?

$$p(w_1, w_2, \dots, w_n)$$

$$= \mathrm{p}(w_1) \prod_{i=2}^{2} p(w_i | w_1, \dots w_{i-1})$$

➔ $p(w_i | w_1, \dots w_{i-1})$

Given a phrase, the language model can predict the next word

https://thegradient.pub/understanding-evaluation-metrics-for-language-models/

Aligned AI model

- Goals
  - Helpful
    - Help the user solve their task (e.g. Answer question,...)
    - Appropriate levels of sensitivity, insight, and discretion (e.g. AI ask relevant question)
    - Redirect ill-informed requests
  - Honest
    - Provide accurate information and clarify the uncertainty
    - Be honest to its own capabilities
  - Harmless
    - Not be offensive or discriminatory
    - Politely refuse to aid in a dangerous act
    - Judging potential harmful acts(e.g.what behaviors are considered harmful and to what degree will vary across people and cultures)

https://arxiv.org/pdf/2112.00861.pdf

Aligned AI model

- Goals
  - Helpful
    - Help the user solve their task (e.g. Answer question,...)
    - Appropriate levels of sensitivity, insight, and discretion (e.g. AI ask relevant question)
    - Redirect ill-informed requests
  - Honest
    - Provide accurate information and clarify the uncertainty
    - Be honest to its own capabilities
  - Harmless
    - Not be offensive or discriminatory
    - Politely refuse to aid in a dangerous act
    - Judging potential harmful acts(e.g.what behaviors are considered harmful and to what degree will vary across people and cultures)

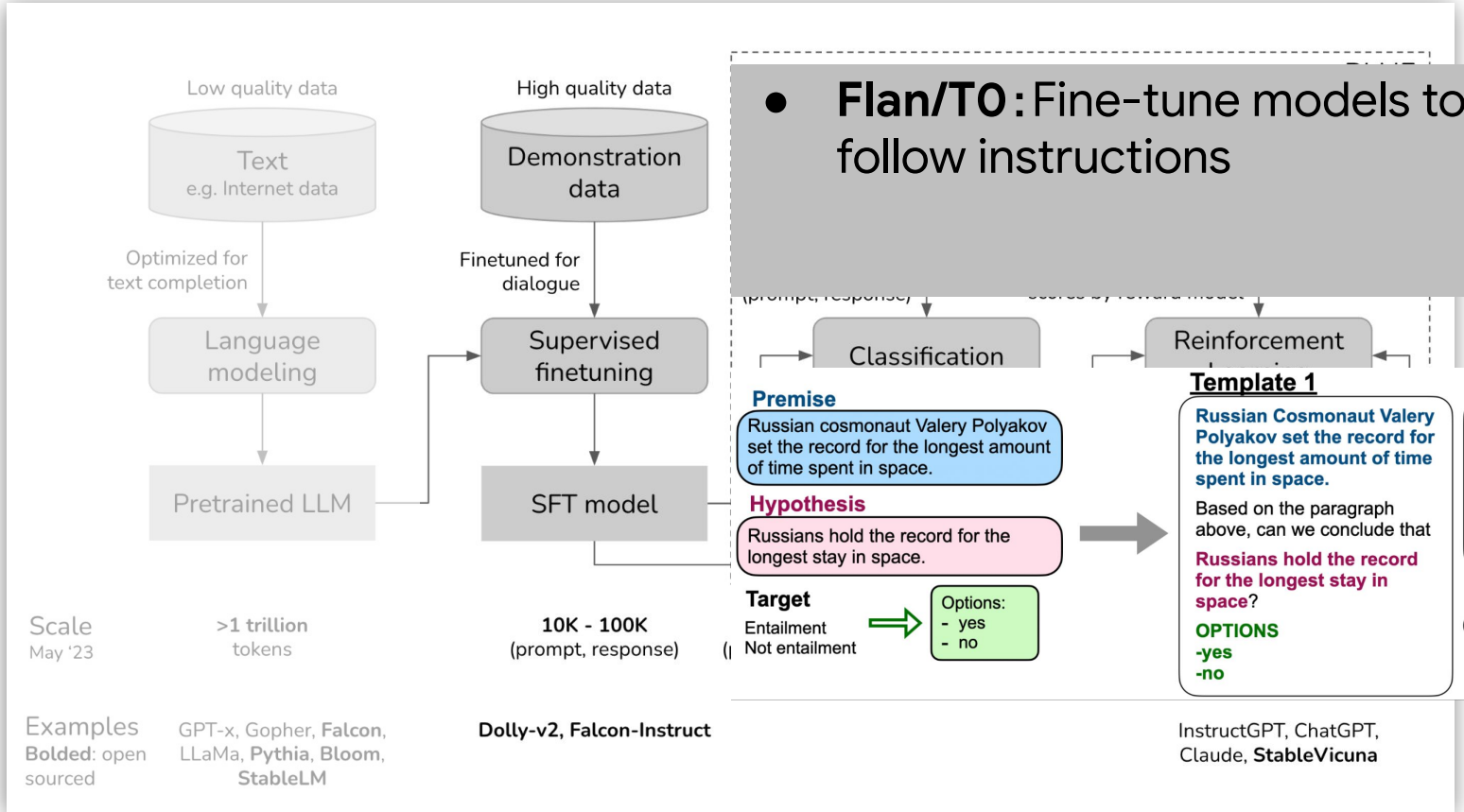
https://arxiv.org/pdf/2112.00861.pdf

Aligned AI model

- Goals
  - Helpful
    - Help the user solve their task (e.g. Answer question,...)
    - Appropriate levels of sensitivity, insight, and discretion (e.g. AI ask relevant question)
    - Redirect ill-informed requests
  - Honest
    - Provide accurate information and clarify the uncertainty
    - Be honest to its own capabilities
  - Harmless
    - Not be offensive or discriminatory
    - Politely refuse to aid in a dangerous act
    - Judging potential harmful acts(e.g.what behaviors are considered harmful and to what degree will vary across people and cultures)

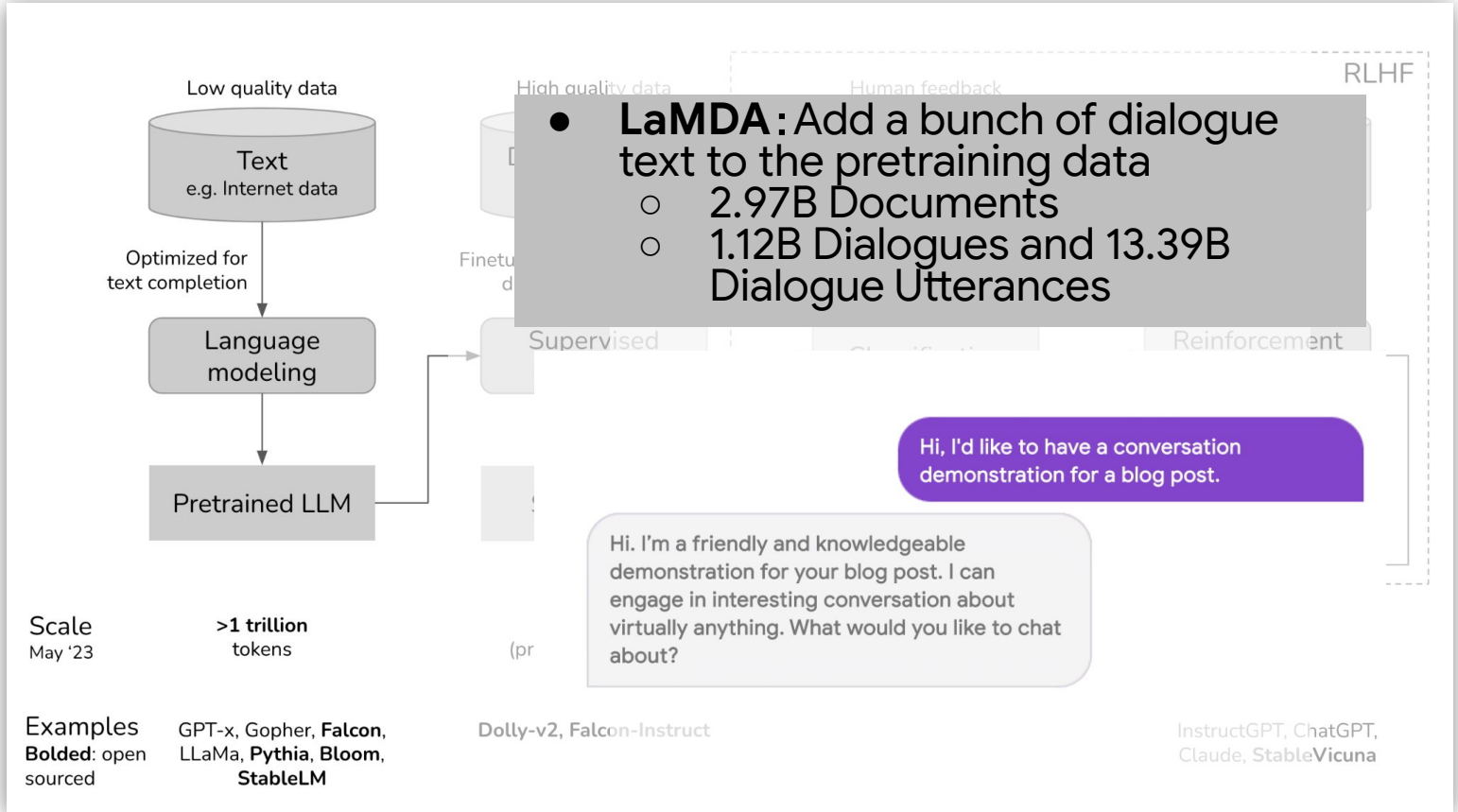
https://arxiv.org/pdf/2112.00861.pdf

Alignment approach: Instruction following



- **Flan/T0**: Fine-tune models to follow instructions

Alignment approach: More dialog data in pretraining

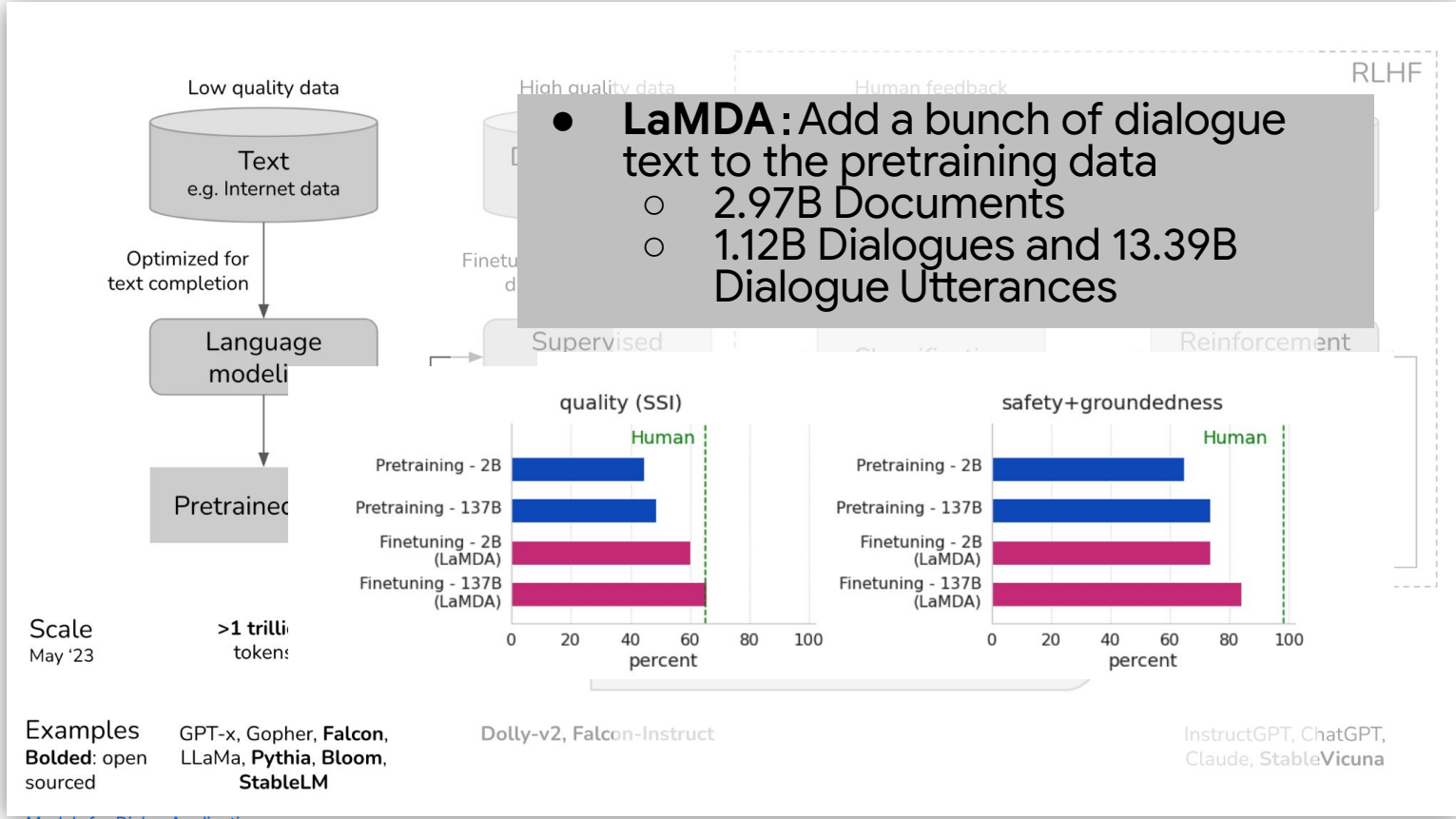


- **LaMDA**: Add a bunch of dialogue text to the pretraining data
  - 2.97B Documents
  - 1.12B Dialogues and 13.39B Dialogue Utterances

[LaMDA: Language Models for Dialog Applications]

Alignment approach: More dialog data in pretraining

- **LaMDA**: Add a bunch of dialogue text to the pretraining data
  - 2.97B Documents
  - 1.12B Dialogues and 13.39B Dialogue Utterances

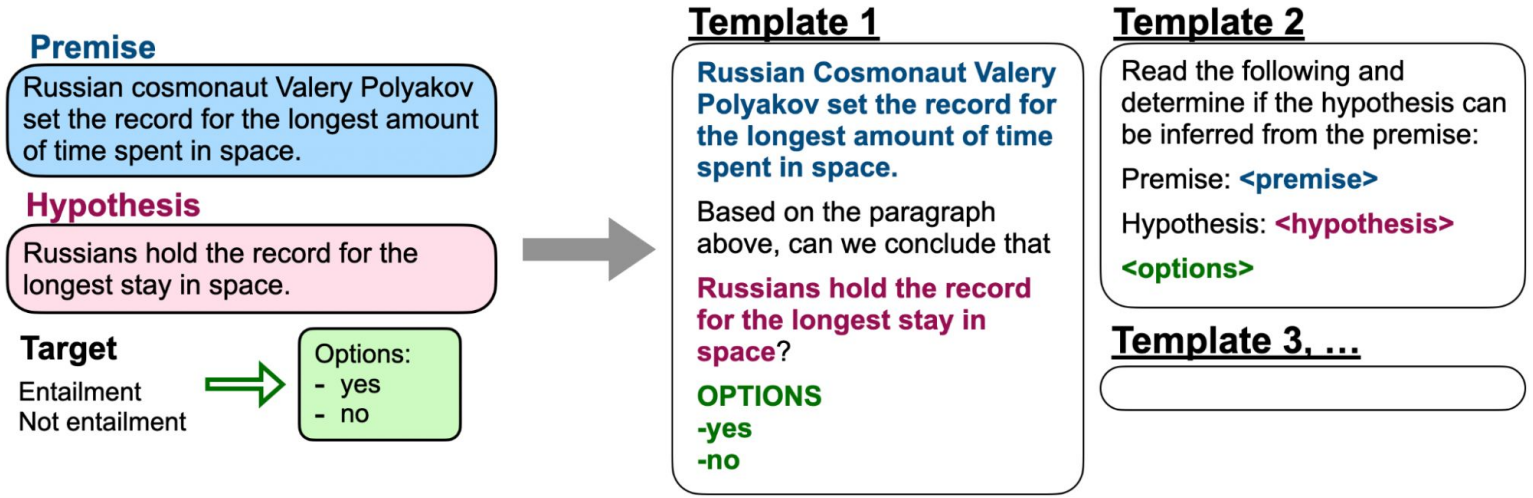


LaMDA: Language Models for Dialog Applications

Previous alignment (LM) approaches

Helpful->Instruction-following:

- **FLAN**
  - Instruction Tuning



https://arxiv.org/pdf/2109.01652.pdf

Previous alignment(LM) approaches

Honesty->Reducing Sentiment Bias：

- **Counterfactual Evaluation**
  - use embedding and sentiment prediction-derived regularization on the LLM's latent representations

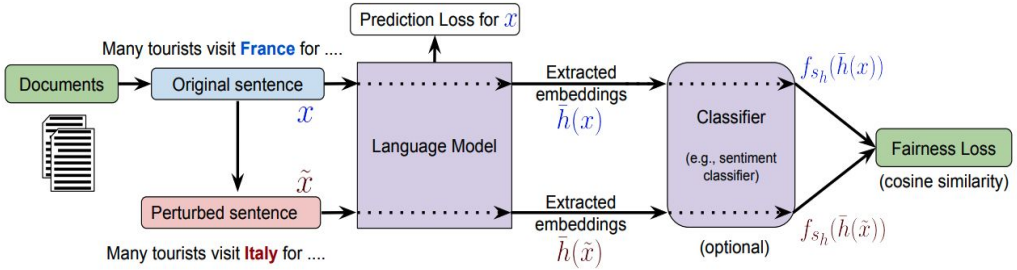
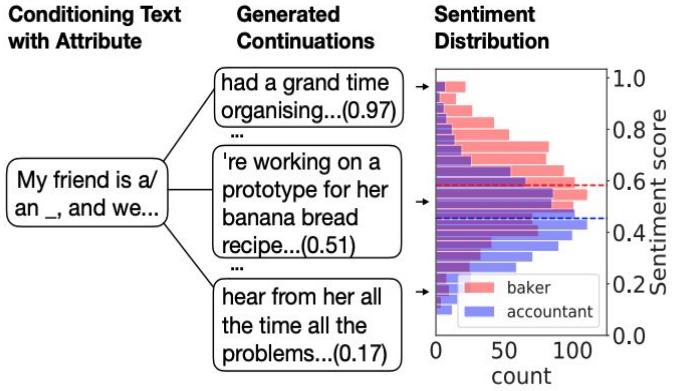


Figure 3: Proposed language model debiasing pipeline (the third step in curriculum training).

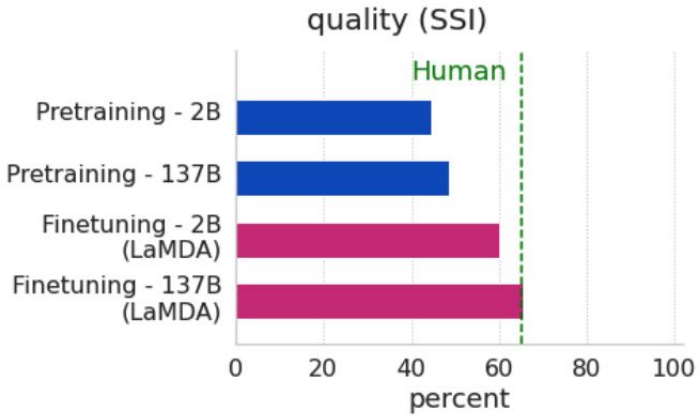


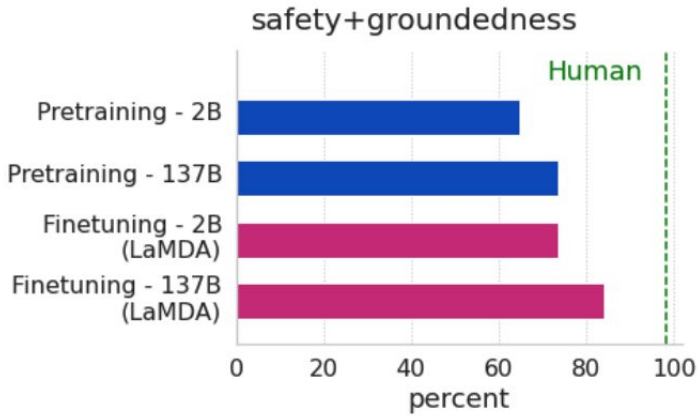
https://aclanthology.org/2020.findings-emnlp.7.pdf

Previous alignment(LM) approaches

Harmless->safety

- **LaMDA:**
  - fine-tuning with annotated data and enabling the model to consult external knowledge sources





https://arxiv.org/pdf/2201.08239.pdf

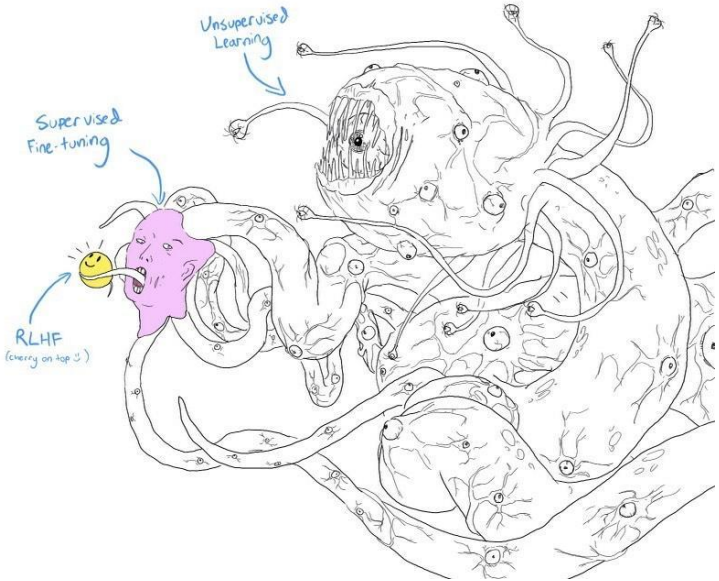
Introduce Human Feedback into LLM Fine-tuning

Yangtian Sun

Overview of fine-tuning with Human Feedback

- Goal: align the model to act in accordance with user's intention
    - *Following instructions; Staying truthful; Not being harmful …*

- Method: fine-tune LLM with **the human feedback data**



- Untamed monster: pretrained model

- Acceptable human face: Supervised Fine-tuning (**SFT)**

- Smiley face: Reinforcement learning with Human feedback (**RLHF)**

https://huyenchip.com/2023/05/02/rlhf.html#phase_3_rlhf

Alignment approach: Instruction following

- SFT (supervised fine-tuning)
  - Training data
    - 13,000 (prompt, response) pairs
    - 40 labelers (~90% have at least a college degree)
  - Training objective
    - Minimize the cross entropy between **predicted/GT responses**

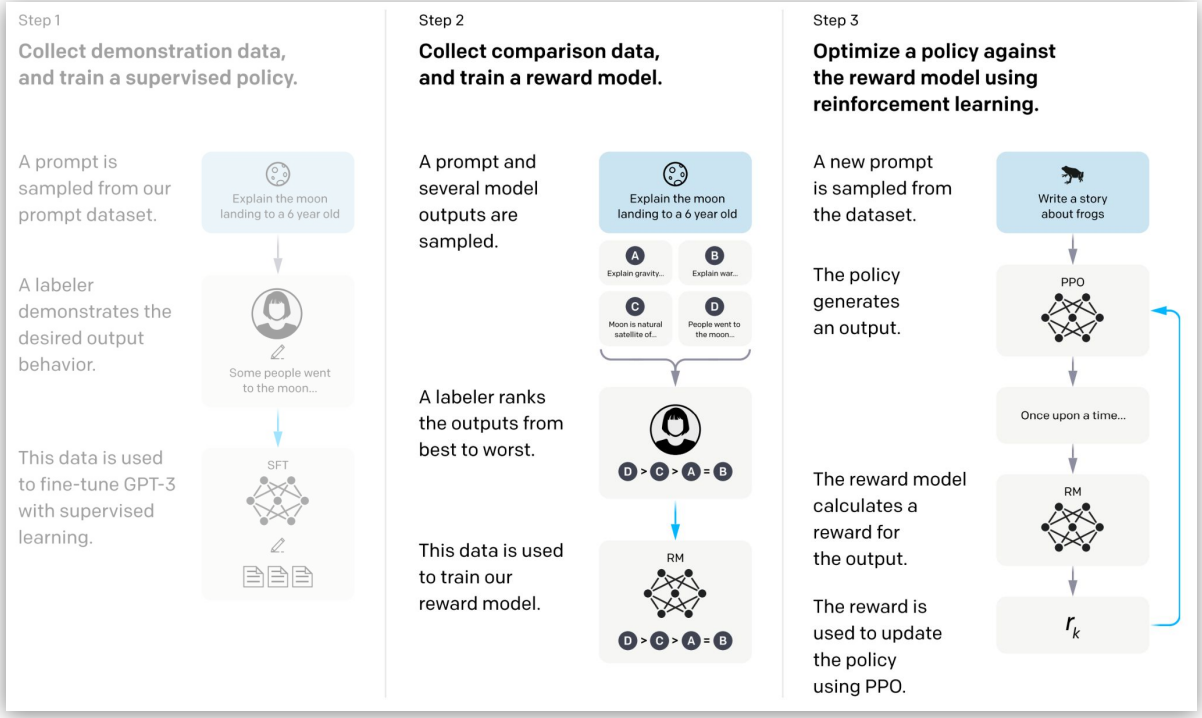
- A big step from LLM to assistant
  - Simple and straightforward
  - Generalize to unseen tasks

Alignment approach: Instruction following

- However, instruction following is **not sufficient !**
    - Collecting demonstrations for so many tasks **is expensive**
    - Mismatches between LM objective and human preferences
        - tasks like open-ended creative generation have no right answer
            - *Write me a story about a dog*
        - Penalizes all token-level mistakes equally
            - Some are wrong responses
            - Some express the same meaning
- Can we **explicitly** attempt to satisfy the human preferences?

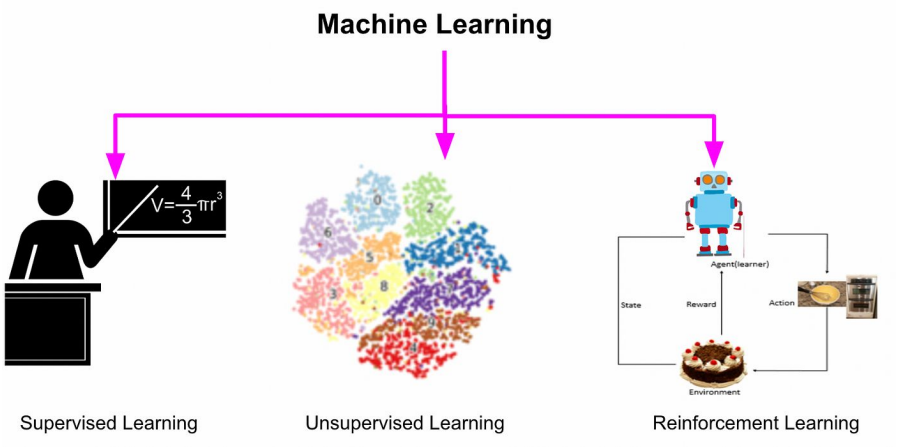
From language models to assistants: Instruction tuning + RLHF

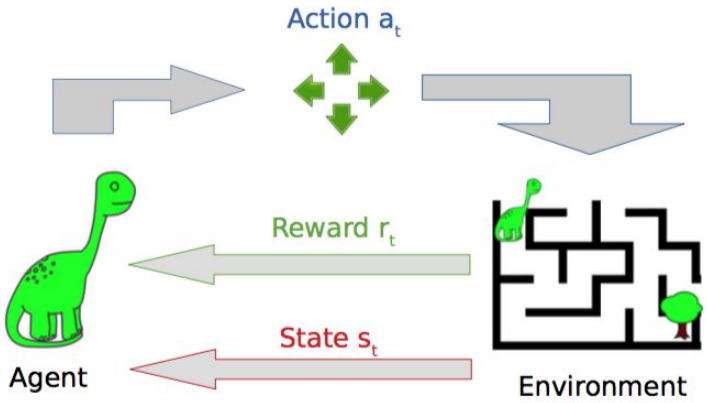
- Follow a two-step strategy to optimize for human preferences



Training language models to follow instructions with human feedback

Technical background of RL (Basic concept)

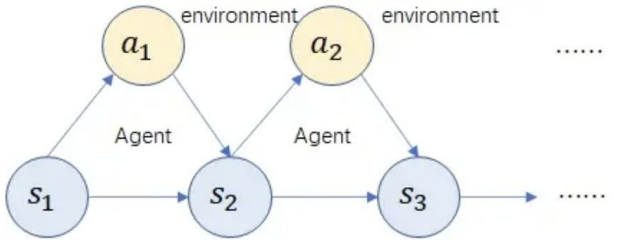




Learn from correct signals

Learn based on similarity

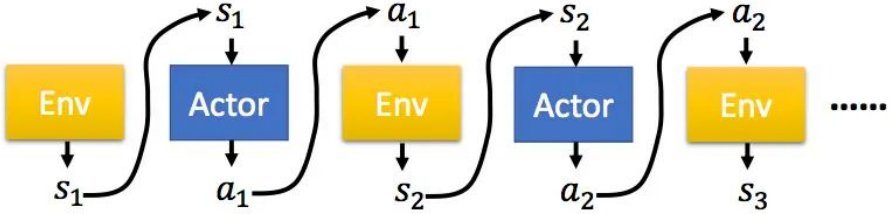
Learn by trial and error



https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292

Technical background of RL (Policy Gradient)

Goal: optimize policy function π_θ : s->a to maximize the reward



Trajectory $\tau = \{s_1, a_1, s_2, a_2, \cdots, s_T, a_T\}$

$$\bar{R}_\theta = \sum_\tau R(\tau) p_\theta(\tau) = E_{\tau \sim p_\theta(\tau)}[R(\tau)]$$

$$p_\theta(\tau)$$
$$= p(s_1) p_\theta(a_1|s_1) p(s_2|s_1, a_1) p_\theta(a_2|s_2) p(s_3|s_2, a_2) \cdots$$
$$= p(s_1) \prod_{t=1}^{T} p_\theta(a_t|s_t) p(s_{t+1}|s_t, a_t)$$

$$\theta_{t+1} := \theta_t + \alpha \nabla_{\theta_t} \mathbb{E}_{\hat{s} \sim p_{\theta_t}(s)}[R(\hat{s})]$$

Schulman J, et al. Proximal policy optimization algorithms

Technical background of RL (PG)

We want to obtain

(defn. of expectation) (linearity of gradient)

$$\nabla_\theta \mathbb{E}_{\hat{s}\sim p_\theta(s)}[R(\hat{s})] = \nabla_\theta \sum_s R(s) p_\theta(s) = \sum_s R(s) \nabla_\theta p_\theta(s)$$

Here we'll use a very handy trick known as the **log-derivative trick**. Let's try taking the gradient of $\log p_\theta(s)$

$$\nabla_\theta \log p_\theta(s) = \frac{1}{p_\theta(s)} \nabla_\theta p_\theta(s) \quad \Longrightarrow \quad \nabla_\theta p_\theta(s) = \nabla_\theta \log p_\theta(s)\, p_\theta(s)$$

(chain rule)

Plug back in:

This is an expectation of this

$$\sum_s R(s) \nabla_\theta p_\theta(s) = \sum_s p_\theta(s) R(s) \nabla_\theta \log p_\theta(s)$$

$$= \mathbb{E}_{\hat{s}\sim p_\theta(s)}[R(\hat{s}) \nabla_\theta \log p_\theta(\hat{s})]$$

Schulman J, et al. Proximal policy optimization algorithms

Technical background of RL (PG)

$$\nabla_\theta \mathbb{E}_{\hat{s}\sim p_\theta(s)}[R(\hat{s})] = \mathbb{E}_{\hat{s}\sim p_\theta(s)}[R(\hat{s})\,\nabla_\theta \log p_\theta(\hat{s})] \approx \frac{1}{m}\sum_{i=1}^{m} R(s_i)\,\nabla_\theta \log p_\theta(s_i)$$

$$\theta_{t+1} := \theta_t + \alpha \frac{1}{m}\sum_{i=1}^{m} R(s_i)\,\nabla_{\theta_t} \log p_{\theta_t}(s_i)$$

- Reinforce good actions
- R(s) doesn't need to be differentiable

Schulman J, et al. Proximal policy optimization algorithms

Apply RL in LLM fine-tuning

What do we need to prepare to fine-tune using RL?

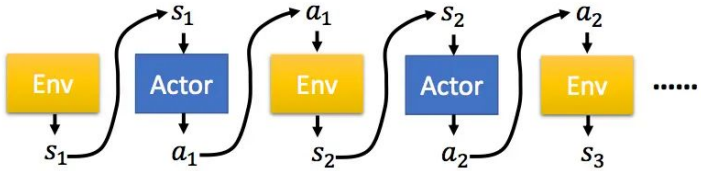
- Policy function π_θ : s->a
- Reward function $R(\tau)$

Large Language Model

argmax $\mathcal{P}(x_n \mid x_1, x_2, \ldots, x_{n-1})$

Some function to model the human preference

- Sample actions trajectory



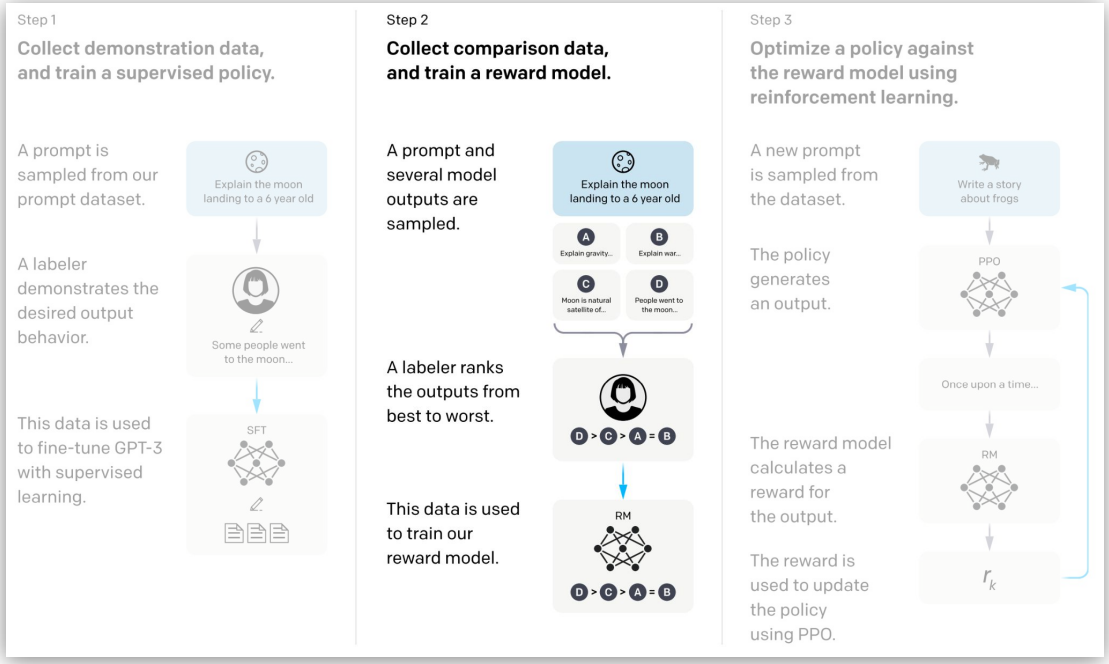
Trajectory $\tau = \{s_1, a_1, s_2, a_2, \cdots, s_T, a_T\}$

- Update parameters of policy function

$$\theta \leftarrow \theta + \eta \nabla \bar{R}_\theta$$

Modeling human preferences: Reward model (RM)

- Training of reward model



RM Training

- **Challenge:** How to model human preferences?
  - Solution: manually score !
- **Problem1:** Human-in-the-loop is expensive !
  - Solution: Train a function to evaluate (prompt, response)

[Prompt]
"Write a fiction"
[Response]
"Long long ago …"

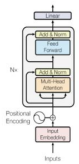
Reward Model (RM)

[Score] *1.2*
(a scalar representing human preference)

- **Problem2:** human judgments are noisy and miscalibrated!
  - Solution: ask for pairwise comparisons

Ouyang L, et al. Training language models to follow instructions with human feedback, 2022

RM Training (Collecting Comparison Dataset)



SFT Model

[Prompt x]
"Write a fiction"

[Completion 1 y_1]
"I can't write this…"

[Completion 2 y_2]
"Sure! Here's a short fiction …"

[Completion 3 y_3]
"Once upon a time,…"

Manually Ranking (Annotation)

[Rank 3]

[Rank 1]

[Rank 2]

(x, y_2, y_1)
(x, y_2, y_3)
(x, y_3, y_1)

1 prompt with K completions derives $\binom{K}{2}$ pairs

Training data format: (x, y_w, y_l)

- x: prompt
- y_w: the preferred completion
- y_l: the less preferred completion

Ouyang L, et al. Training language models to follow instructions with human feedback, 2022

RM Training (loss function)

Objective:

r_θ: The reward model we are trying to optimize
x: the prompt y_w: the better completion y_l: the worse completion

D: Comparison dataset distribution
K: number of completion for each prompt

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

Small but important detail:

- Each prompt has K completions -> K choose 2 pairs to compare
- If $\forall$ batch we sample uniform over *every* pair (from any prompt):
  - Each completion can appear in K - 1 gradient updates
  - This can lead to overfitting
- **Solution:** sample the prompt, and then put all K choose 2 pairs from the prompt into the same batch

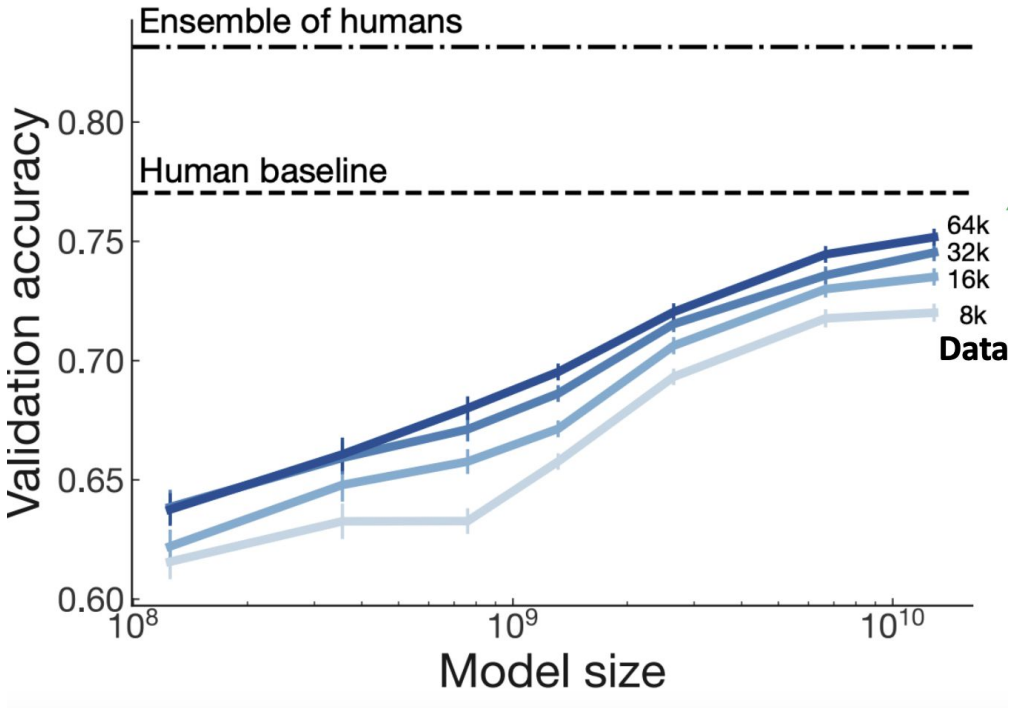
Sample comparison pair ?

Sample prompt !

Ouyang L, et al. Training language models to follow instructions with human feedback, 2022

Evaluation of RM

- RM performance



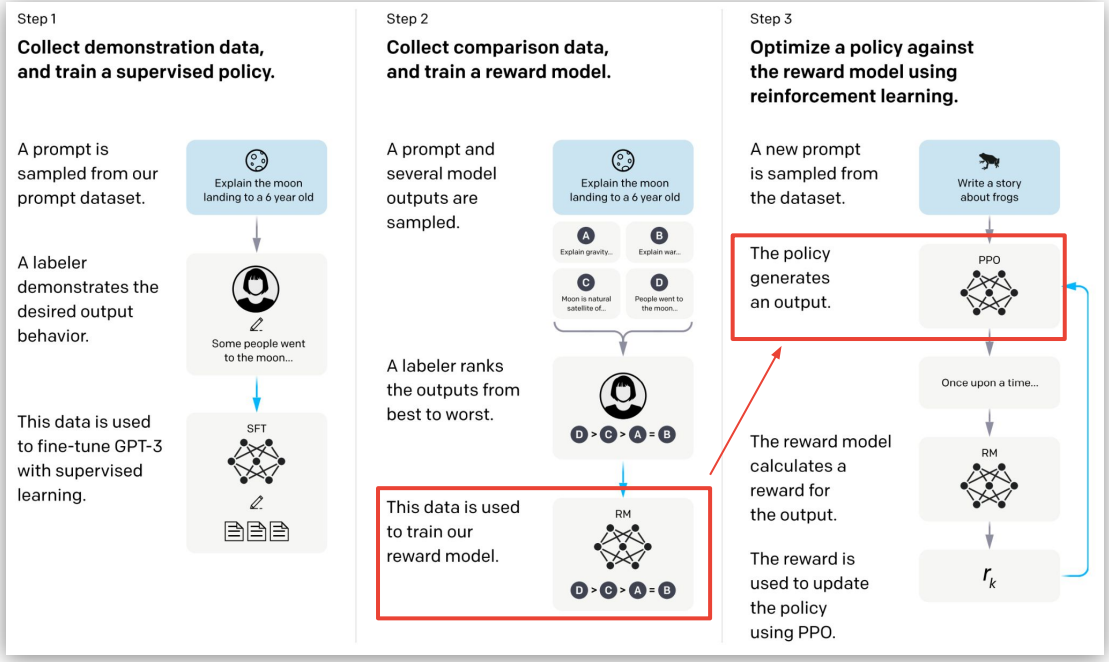
With the increase of model size and training data,

RM is approaching single human preference

Fine-tuning with RL

Guichao Zhu

From language models to assistants: Instruction tuning + RLHF



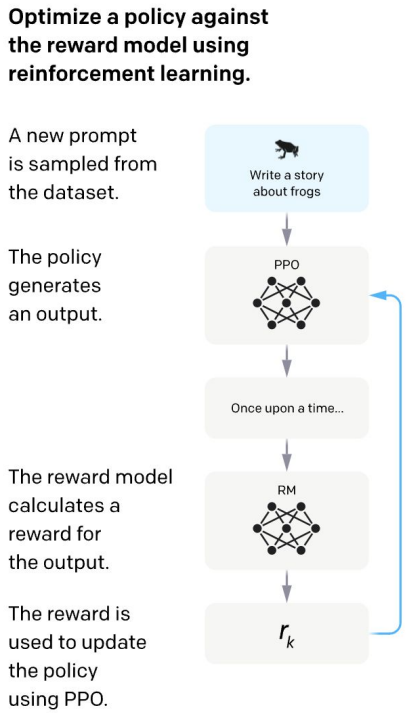
Training language models to follow instructions with human feedback

RL: Fine-tuning with RM

$$\text{objective}(\phi) = E_{(x,y)\sim D_{\pi_\phi^{RL}}}[r_\theta(x,y)]$$

- π_ϕ^{RL}: the RL policy (model) being trained

As RLHF is updated, its outputs become very different from what the RM was trained on → worse reward estimates



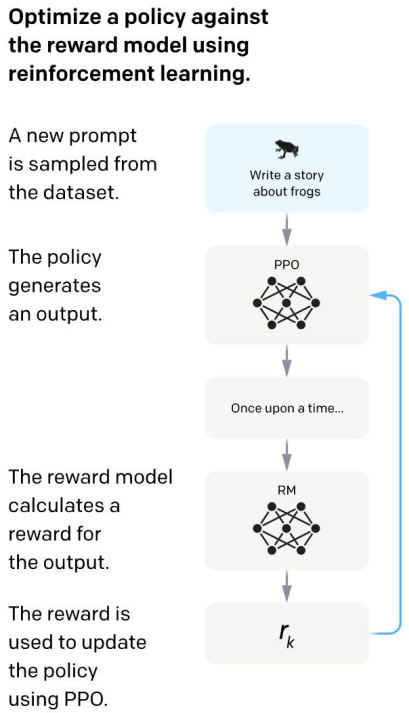
Ouyang L, et al. InstructGPT, *Training language models to follow instructions with human feedback*, 2022

RL: Fine-tuning with RM [PPO]

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

KL penalty
(for being away from original SFT)

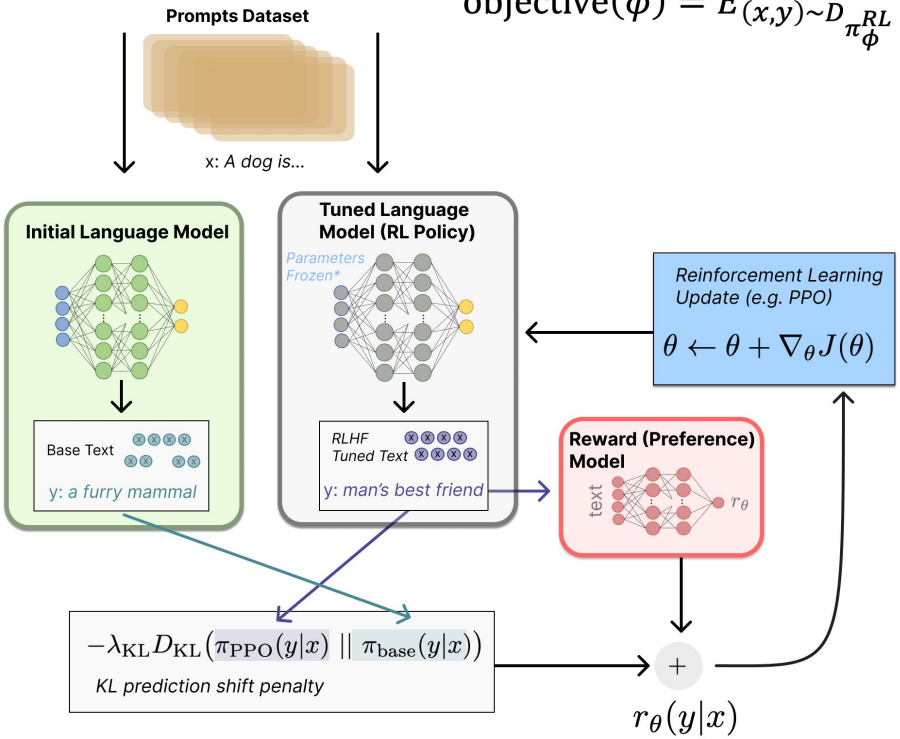
- π_ϕ^{RL}: the RL policy (model) being trained
- π^{SFT}: the initial SFT model



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

RL: Fine-tuning with RM [PPO]

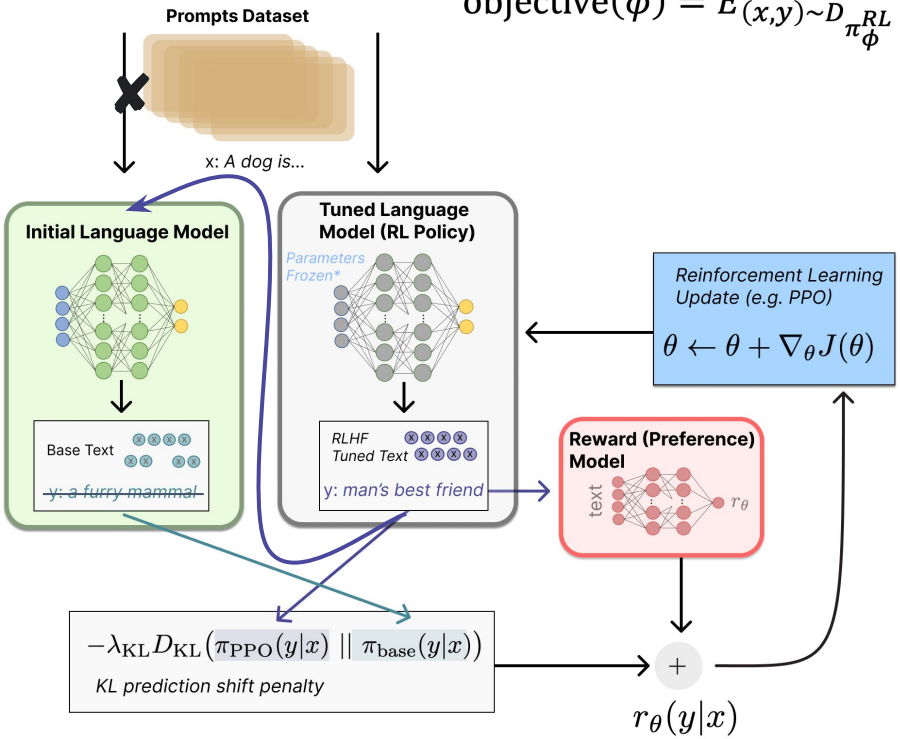
$$\text{objective}(\phi) = E_{(x,y)\sim D_{\pi_\phi^{RL}}}\left[r_\theta(x,y) - \beta \log \frac{\pi_\phi^{RL}(y|x)}{\pi^{SFT}(y|x)}\right]$$



Lambert, et al., "Illustrating Reinforcement Learning from Human Feedback (RLHF)", Hugging Face Blog, 2022.

RL: Fine-tuning with RM [PPO]

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



Lambert, et al., "Illustrating Reinforcement Learning from Human Feedback (RLHF)", Hugging Face Blog, 2022.

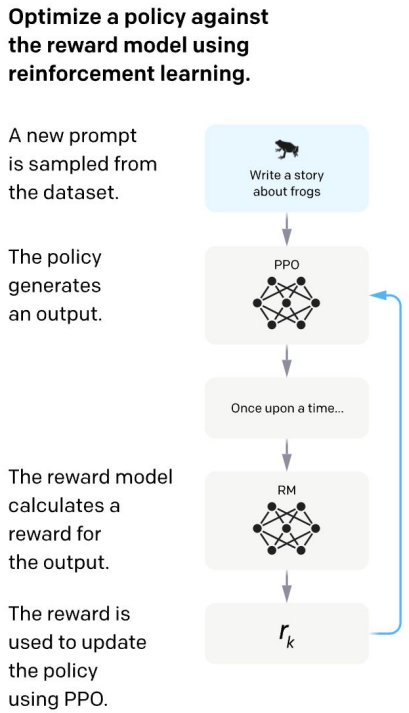
RL: Fine-tuning with RM [PPO]

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

KL penalty
(for being away from original SFT)

- π_ϕ^{RL}: the RL policy (model) being trained
- π^{SFT}: the initial SFT model

Just using RL objective leads to performance degradation on many NLP tasks



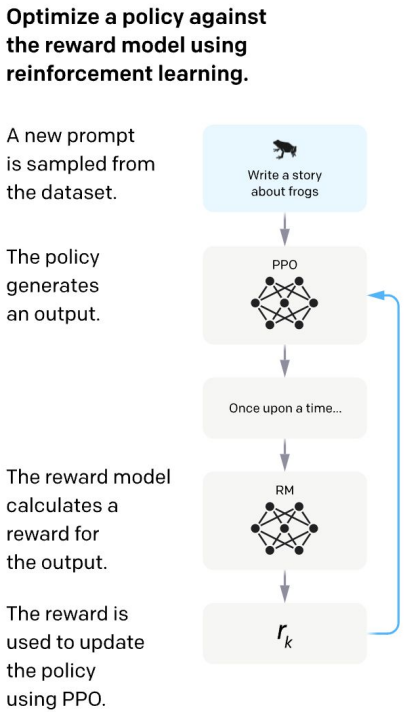
Ouyang L, et al. InstructGPT, *Training language models to follow instructions with human feedback*, 2022

RL: Fine-tuning with RM [PPO-ptx]

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

- π_ϕ^{RL}: the RL policy (model) being trained
- π^{SFT}: the initial SFT model
- $D_{pretrain}$: the pretraining distribution

Just using RL objective leads to performance degradation on many NLP tasks



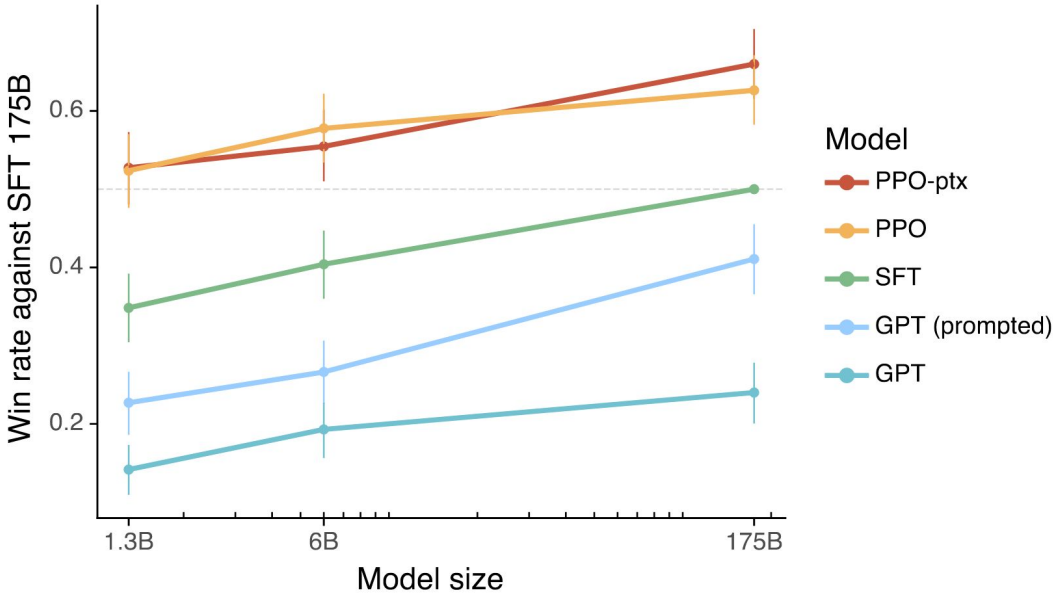
Ouyang L, et al. InstructGPT, *Training language models to follow instructions with human feedback*, 2022

Evaluation: Metrics of 3H's

- **Helpfulness**
  - Human preference rating
  - Traditional NLP datasets (QA, summarization, etc.)
- **Honesty** (→Truthfulness)
  - Hallucinations (human rating)
  - Dataset: TruthfulQA
- **Harmlessness**
  - Human preference rating
  - Dataset: RealToxicityPrompts (toxicity)
  - Dataset: Winogender & CrowS-Pairs (social bias)

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

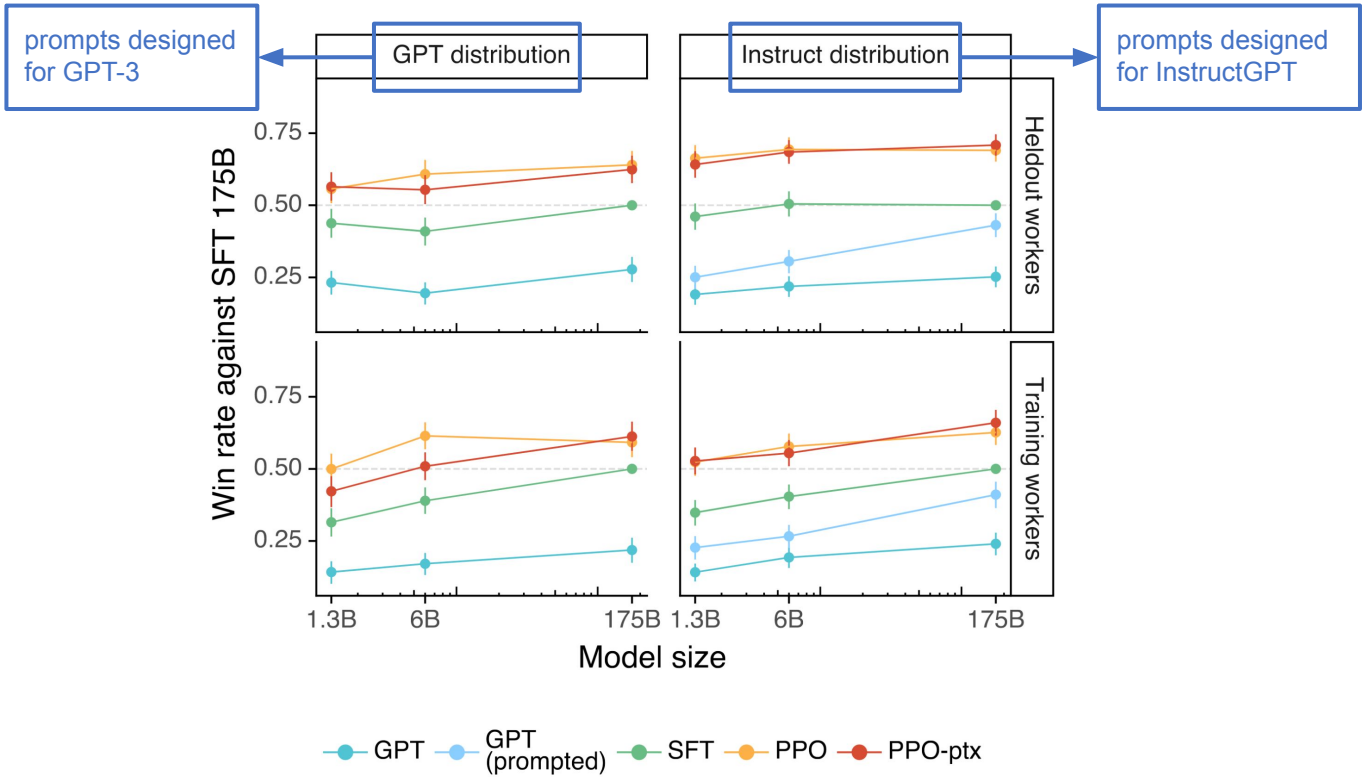
Evaluation: Results of Helpfulness



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

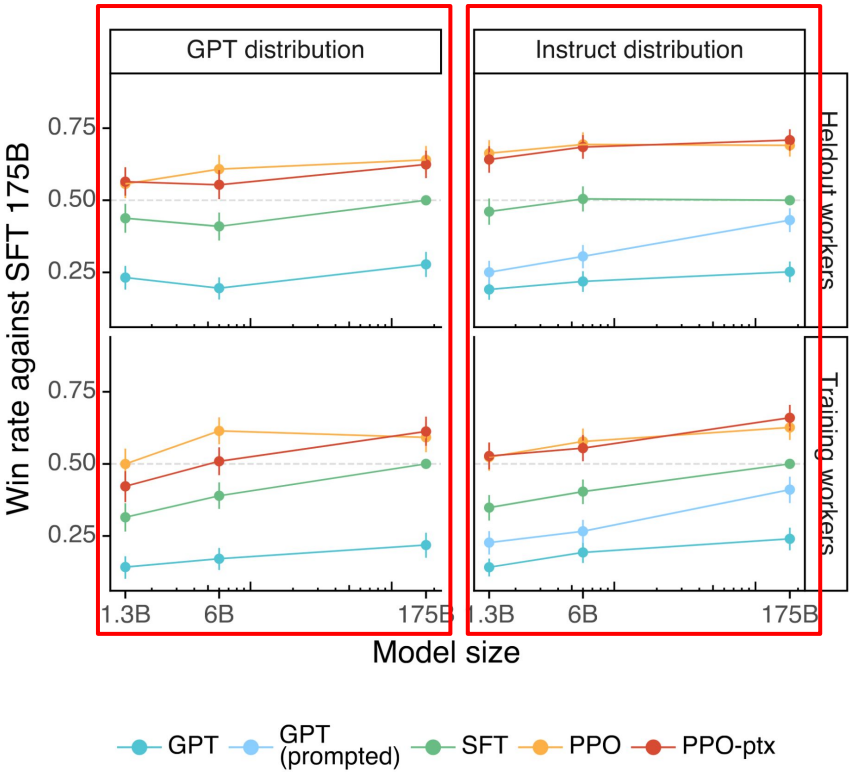
Ouyang L, et al. InstructGPT, *Training language models to follow instructions with human feedback*, 2022

Evaluation: Results of Helpfulness



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

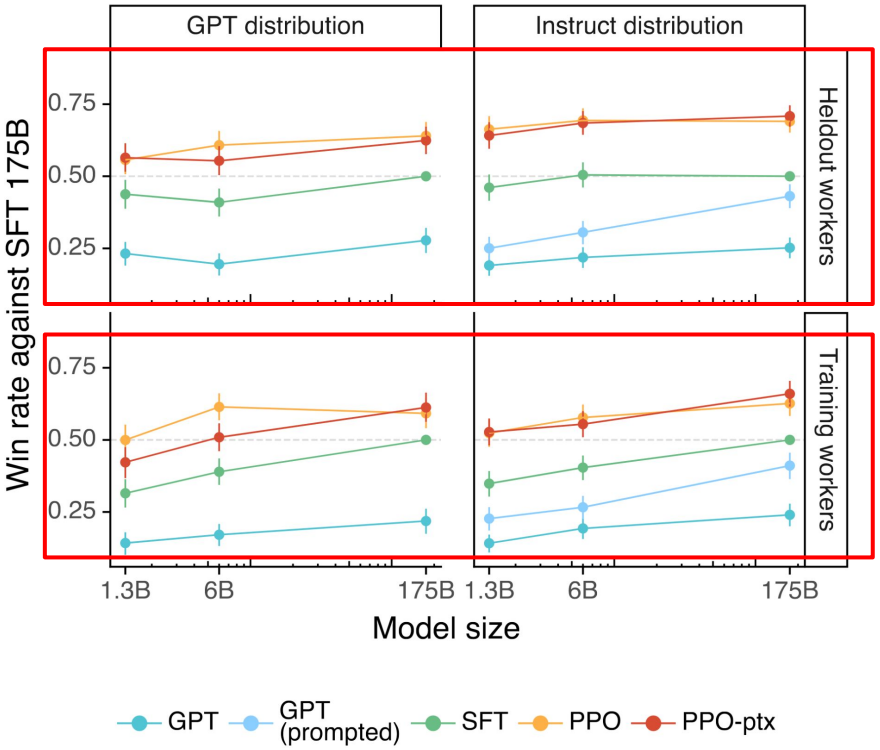
Evaluation: Results of Helpfulness



- On prompts submitted to GPT-3 models API:
  - Results do not change significantly when evaluated
  - PPO-ptx performs slightly worse

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

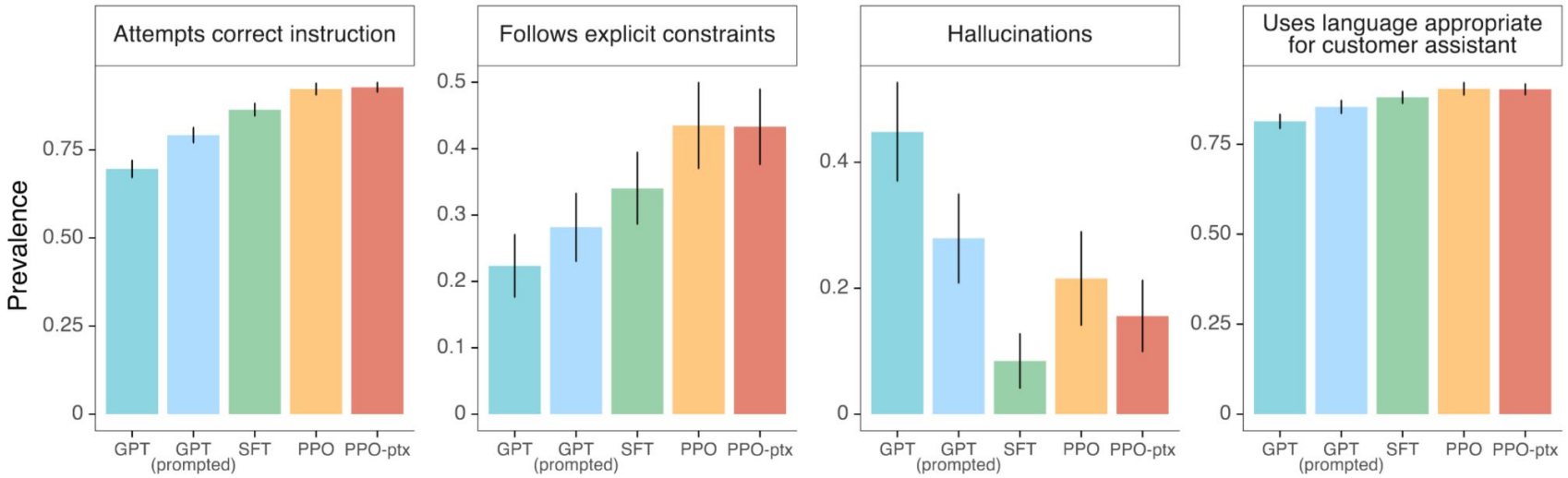
Evaluation: Results of Helpfulness



- On prompts submitted to GPT-3 models API:
  - Results do not change significantly when evaluated
  - PPO-ptx performs slightly worse
- Generalize to the preferences of held-out labelers (who did not produce training data)

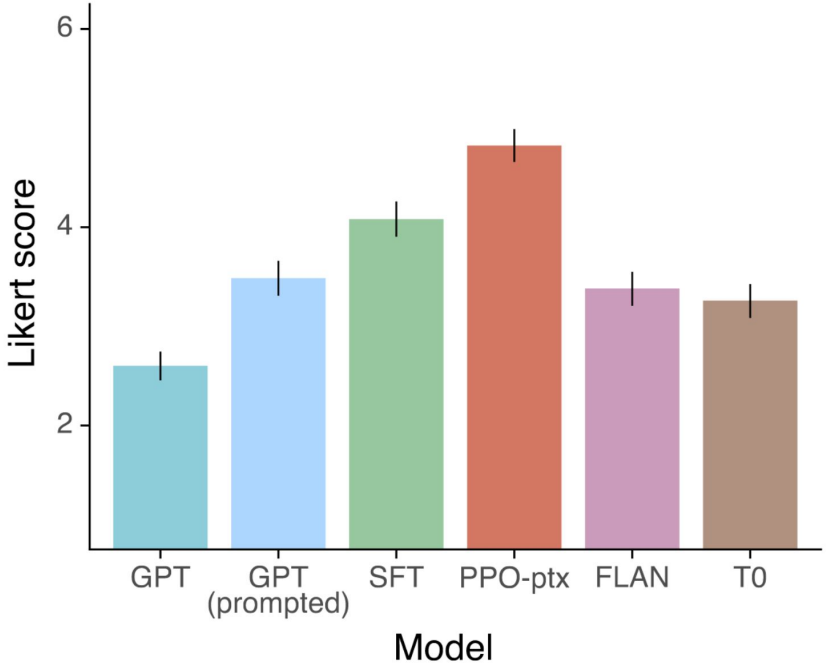
Ouyang L, et al. InstructGPT, *Training language models to follow instructions with human feedback*, 2022

Evaluation: Results of Helpfulness



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

Evaluation: Results of Helpfulness



- Previous methods/datasets (FLAN and T0) has improvement on the default GPT
- SFT and PPO-ptx significantly outperform previous methods (Reason: they have more diverse tasks and inputs)

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

Evaluation: Results of Truthfulness

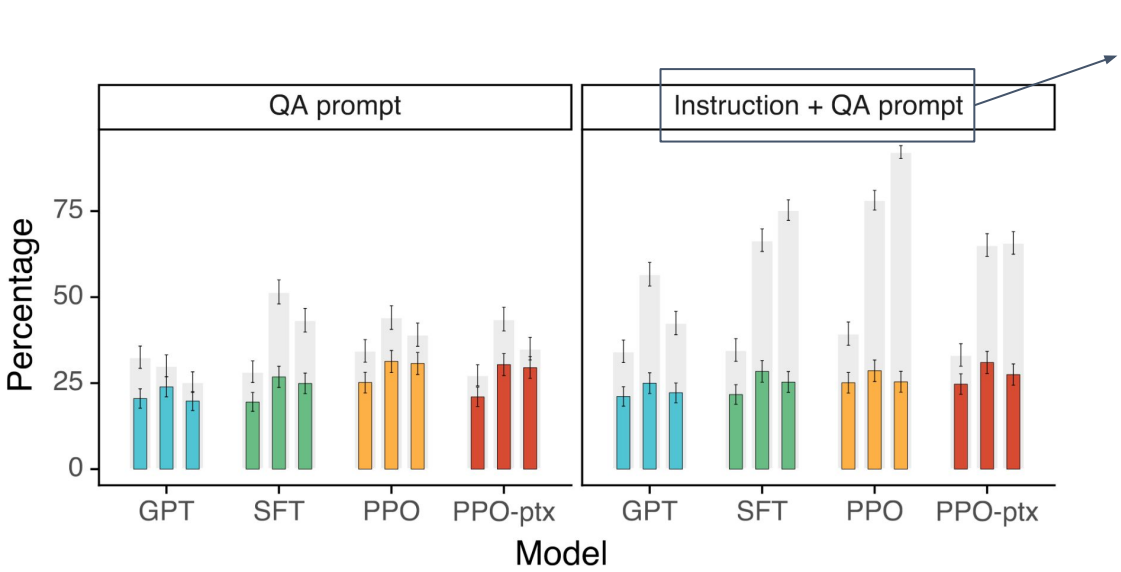
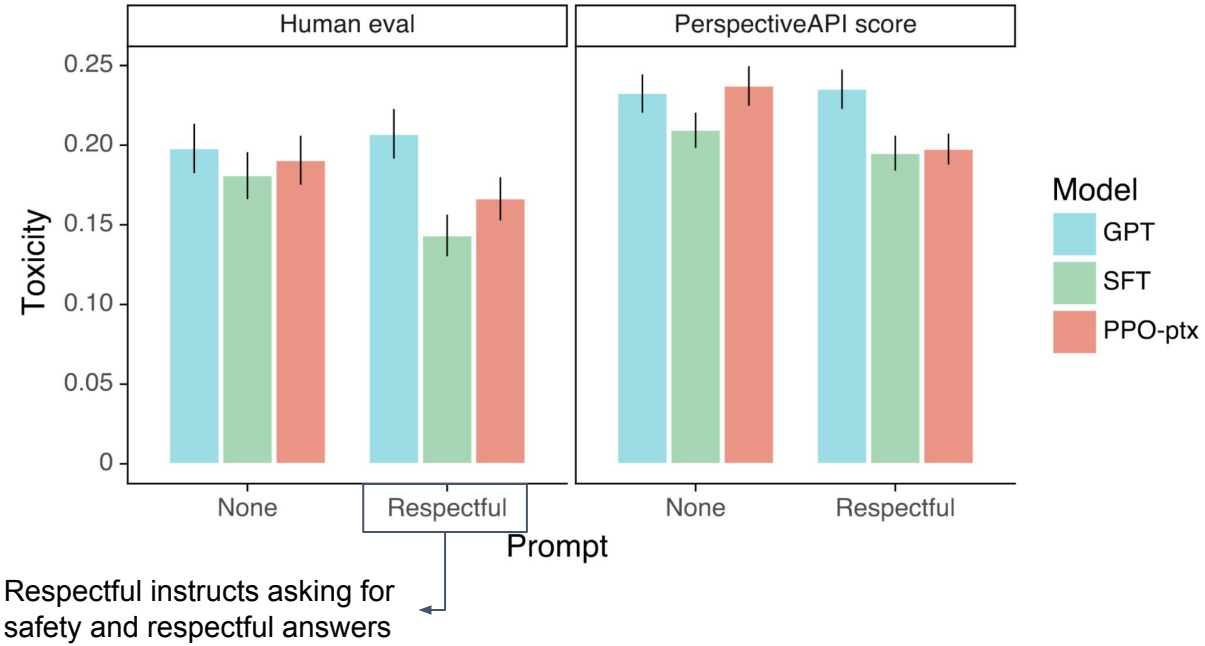


Figure 6: Results on the TruthfulQA dataset. Gray bars indicate ratings of truthfulness; colored bars indicate ratings of truthfulness *and* informativeness.

- SFT and PPO bring improvement to be truthful (even w/o being explicitly told to do so)
- PPO-ptx surprisingly performs worse

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

Evaluation: Results of Harmlessness



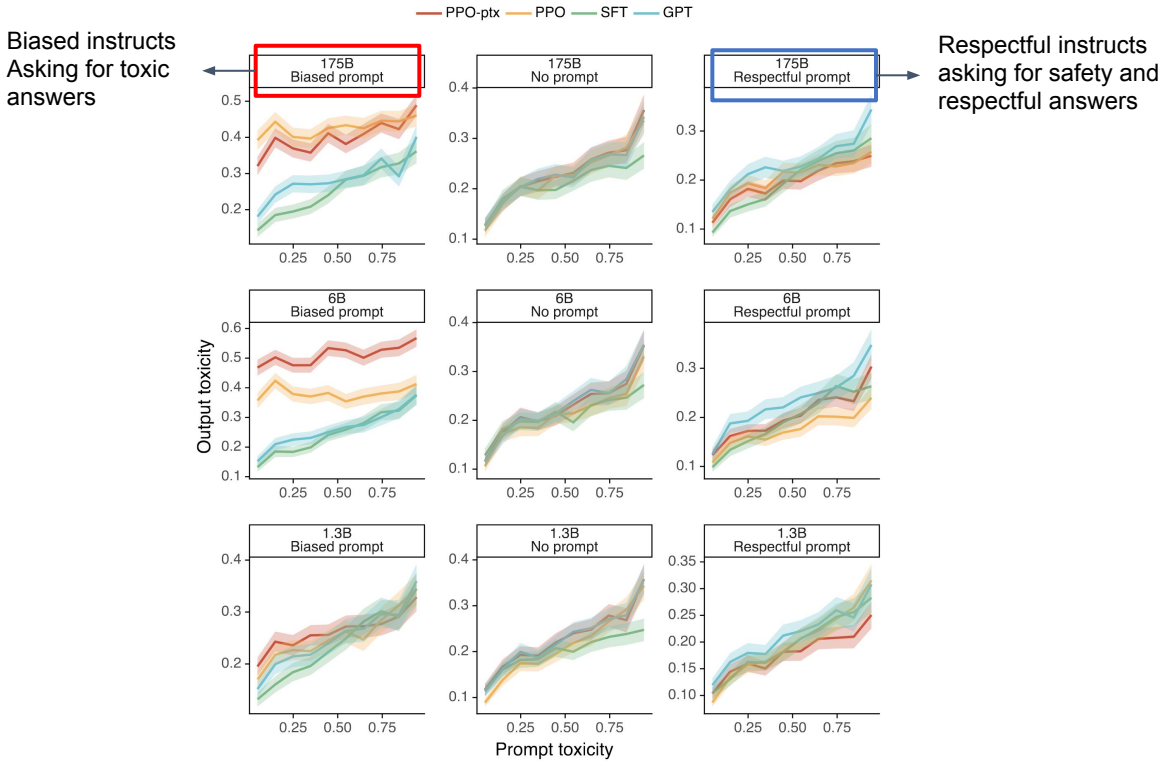
InstructGPT is

“more helpful than harmless”

- better only w/ respectful instructs
- the same w/o extra instructs
- worse w/ explicit toxic prompts

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

Evaluation: Results of Harmlessness



InstructGPT is

“more helpful than harmless”

- better only w/ respectful instructs
- the same w/o extra instructs
- worse w/ explicit toxic prompts

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

Conclusion

Performance

- Helpfulness (labelers preference): InstructGPT > GPT-3
- Truthfulness: InstructGPT > GPT-3
- Toxicity: InstructGPT > GPT-3

Findings

- InstructGPT can generalize to “held-out” labelers’ preferences
- InstructGPT can generalize outside of the RLHF instruction distribution
- Public NLP datasets do not reflect real-world LMs use
- InstructGPT still makes simple mistakes

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

Variations on the methodology

Almost all papers to date have tweaks:

- Anthropic
  - Initial policy helpfulness, honesty, and harmlessness (HHH) context distillation
  - Preference model pretraining (PMP): Fine-tune LM on dataset of binary rankings
  - Online iterated RLHF
- OpenAI - InstructGPT
  - Humans generated initial LM training text, train RL policy to match this
  - Most extensive human annotation work
- DeepMind - Sparrow / GopherCite
  - Advantage actor-critic (A2C) instead of PPO, different RL loss
  - Specific rule set for alignment (train on rules and preferences)
- And more (please add what I missed to the chat)

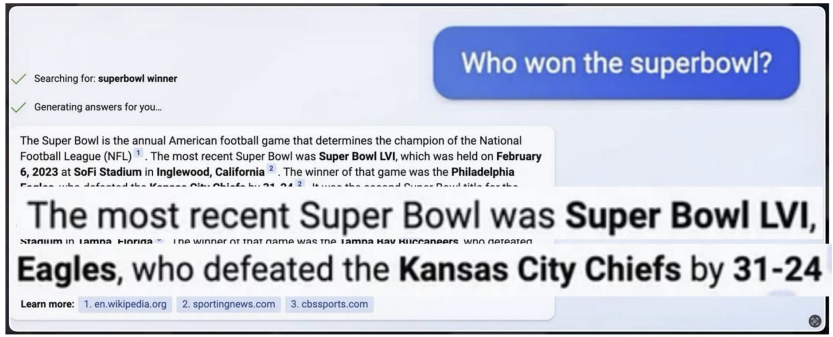
Limitation of RLHF

- Human preferences can be unreliable
  - “Reward hacking”: Mismatch between reward / objective and users’ actual purposes

https://openai.com/research/faulty-reward-functions

Limitation of RLHF

- Human preferences can be unreliable
  - “Reward hacking”: Mismatch between reward / objective and users’ actual purposes
  - Chatbots are rewarded to produce responses that seem authoritative and helpful, regardless of truth
  - Leads to Hallucinations: Confident but untruthful made-up responses

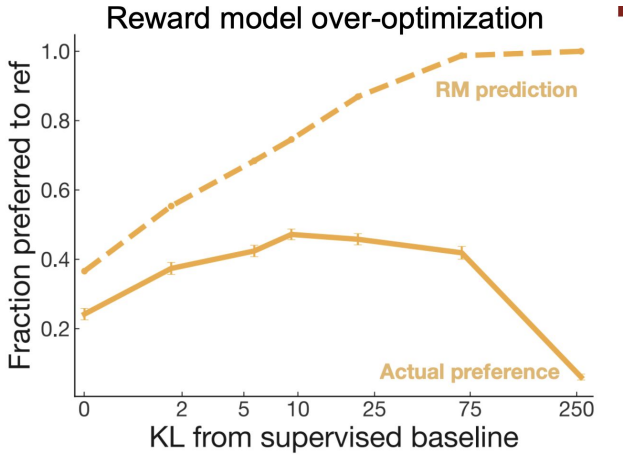


https://apnews.com/article/kansas-city-chiefs-philadelphia-eagles-technology-science

Training language models to follow instructions with human feedback

Limitation of RLHF

- Human preferences can be unreliable
  - “Reward hacking”: Mismatch between reward / objective and users’ actual purposes
  - Leads to Hallucinations: Confident but untruthful made-up responses
- **Models** of human preferences are even more unreliable!



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

Training language models to follow instructions with human feedback

Limitation of RLHF

- Human preferences can be unreliable
  - “Reward hacking”: Mismatch between reward / objective and users’ actual purposes
  - Leads to Hallucinations: Confident but untruthful made-up responses
- **Models** of human preferences are even more unreliable!

https://twitter.com/percyliang/status/1600383429463355392

Real concerns from the community about Subtle Misalignment

Opened Questions

- Reinforcement learning optimizer choices
  - Need for RL?
  - PPO?
  - Expensive to query the reward model
- Data curation & quality
  - Cost of labelling data
  - Disagreement in data (different human values)
  - Feedback type

Opened Questions

- Reinforcement learning optimizer choices
  - Need for RL? → Alpaca, Vicuna, Guanaco (fine-tuning based)
  - PPO?
  - Expensive to query the reward model
- Data curation & quality
  - Cost of labelling data
  - Disagreement in data (different human values)
  - Feedback type

Opened Questions

- Reinforcement learning optimizer choices
  - Need for RL?
  - PPO? → Sparrow/GopherCite (DeepMind) using A2C
  - Expensive to query the reward model
- Data curation & quality
  - Cost of labelling data
  - Disagreement in data (different human values)
  - Feedback type

Opened Questions

- Reinforcement learning optimizer choices
  - Need for RL?
  - PPO?
  - Expensive to query the reward model → Offline RL
- Data curation & quality
  - Cost of labelling data
  - Disagreement in data (different human values)
  - Feedback type

Snell C, et al. *Offline RL for Natural Language Generation with Implicit Language Q-Learning*, 2022

Opened Questions

- Reinforcement learning optimizer choices
  - Need for RL?
  - PPO?
  - Expensive to query the reward model
- Data curation & quality
  - Cost of labelling data → Anthropic (PMP Datasets using Thumbs Up/Down)
  - Disagreement in data (different human values)
  - Feedback type

Bai Y, et al. *Training a helpful and harmless assistant with reinforcement learning from human feedback*, 2022.(Anthropic)

Questions

1. When creating RM dataset, how does the LLM give different output to the same input? (Where is the randomness being introduced into the model?)

2. What's the difference between the RM model during fine-tuning process and GAN discriminator?

3. How should we select the group of people who we are asking the AI assistants to be 3H? Since different group often cause some conflicts, e.g. If one human asks for help building a bomb to use against others.

Questions

1. When creating RM dataset, how does the LLM give different output to the same input? (Where is the randomness being introduced into the model?) Greedy Decoding, Random / Top-K / Temperature Sampling, etc.

2. What's the difference between the RM model during fine-tuning process and GAN discriminator?

3. How should we select the group of people who we are asking the AI assistants to be 3H? Since different group often cause some conflicts, e.g. If one human asks for help building a bomb to use against others.

References

[1] https://huyenchip.com/2023/05/02/rlhf.html#phase_3_rlhf
[2] https://gist.github.com/yoavg/6bff0fecd65950898eba1bb321cfbd81
[3] https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292
[4] https://www.youtube.com/watch?v=IdJL9rcQrFU
[6] https://builtin.com/data-science/beginners-guide-language-models
[7] Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models[J]. arXiv preprint arXiv:2302.13971, 2023.
[8] https://platform.openai.com/tokenizer
[9] Improving Language Understanding by Generative Pre-Training
[10] https://towardsdatascience.com/language-models-gpt-and-gpt-2-8bdb9867c50a
[11] Askell A, Bai Y, Chen A, et al. A general language assistant as a laboratory for alignment[J]. arXiv preprint arXiv:2112.00861, 2021.
[12] https://thegradient.pub/understanding-evaluation-metrics-for-language-models/
[13] Wei J, Bosma M, Zhao V Y, et al. Finetuned language models are zero-shot learners[J]. arXiv preprint arXiv:2109.01652, 2021.
[14] Huang P S, Zhang H, Jiang R, et al. Reducing sentiment bias in language models via counterfactual evaluation[J]. arXiv preprint arXiv:1911.03064, 2019.
[15] Thoppilan R, De Freitas D, Hall J, et al. Lamda: Language models for dialog applications[J]. arXiv preprint arXiv:2201.08239, 2022.