

Robust Reinforcement Learning from Corrupted Human Feedback

Zixuan Zhang*

Georgia Tech ISyE

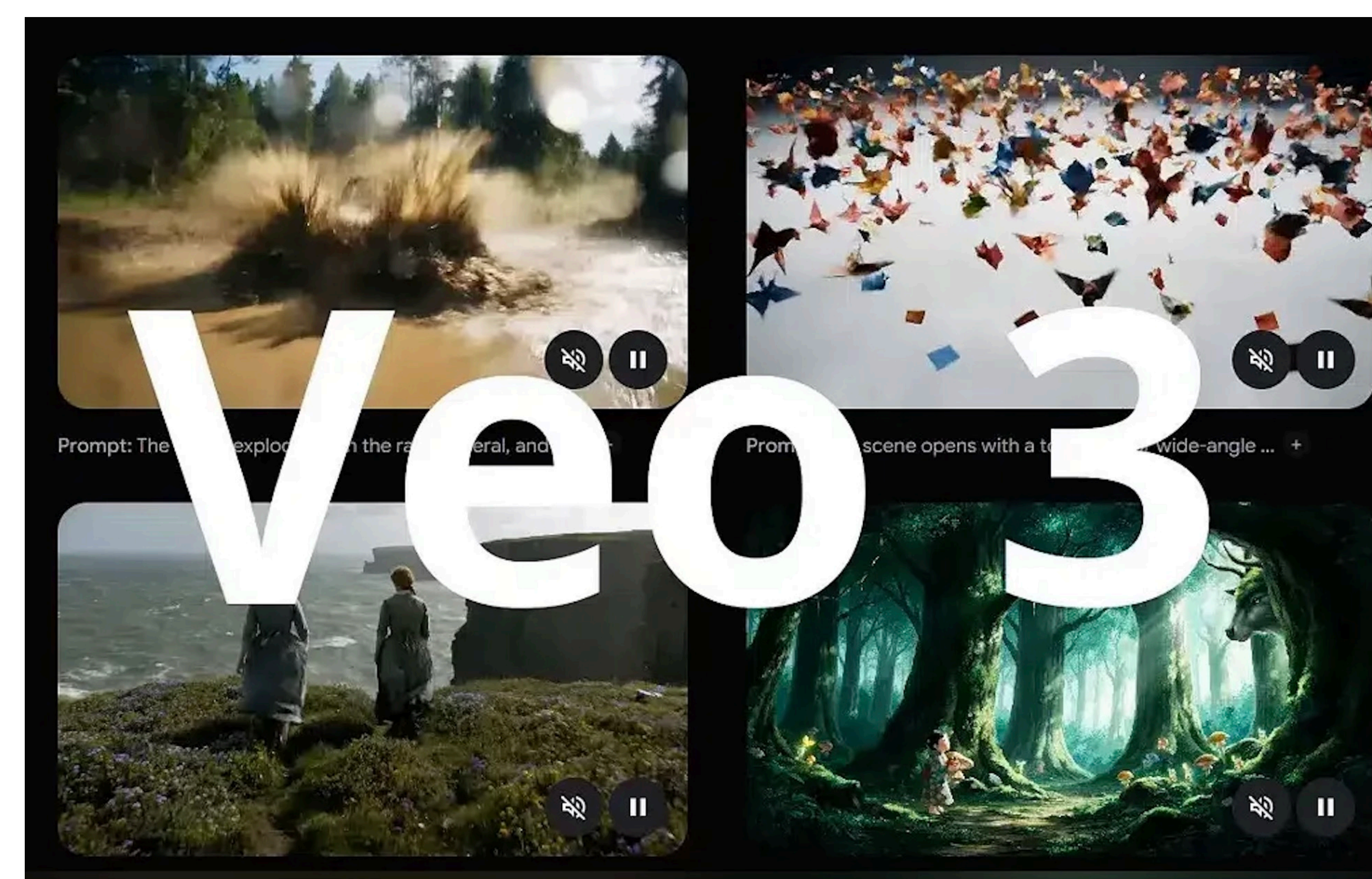
July 2025

* Joint work with Alexander Bukharin, Ilgee Hong, Haoming Jiang, Zichong Li, Qingru Zhang, and Tuo Zhao.

AI Revolution in Action



ChatGPT by OpenAI
NLP



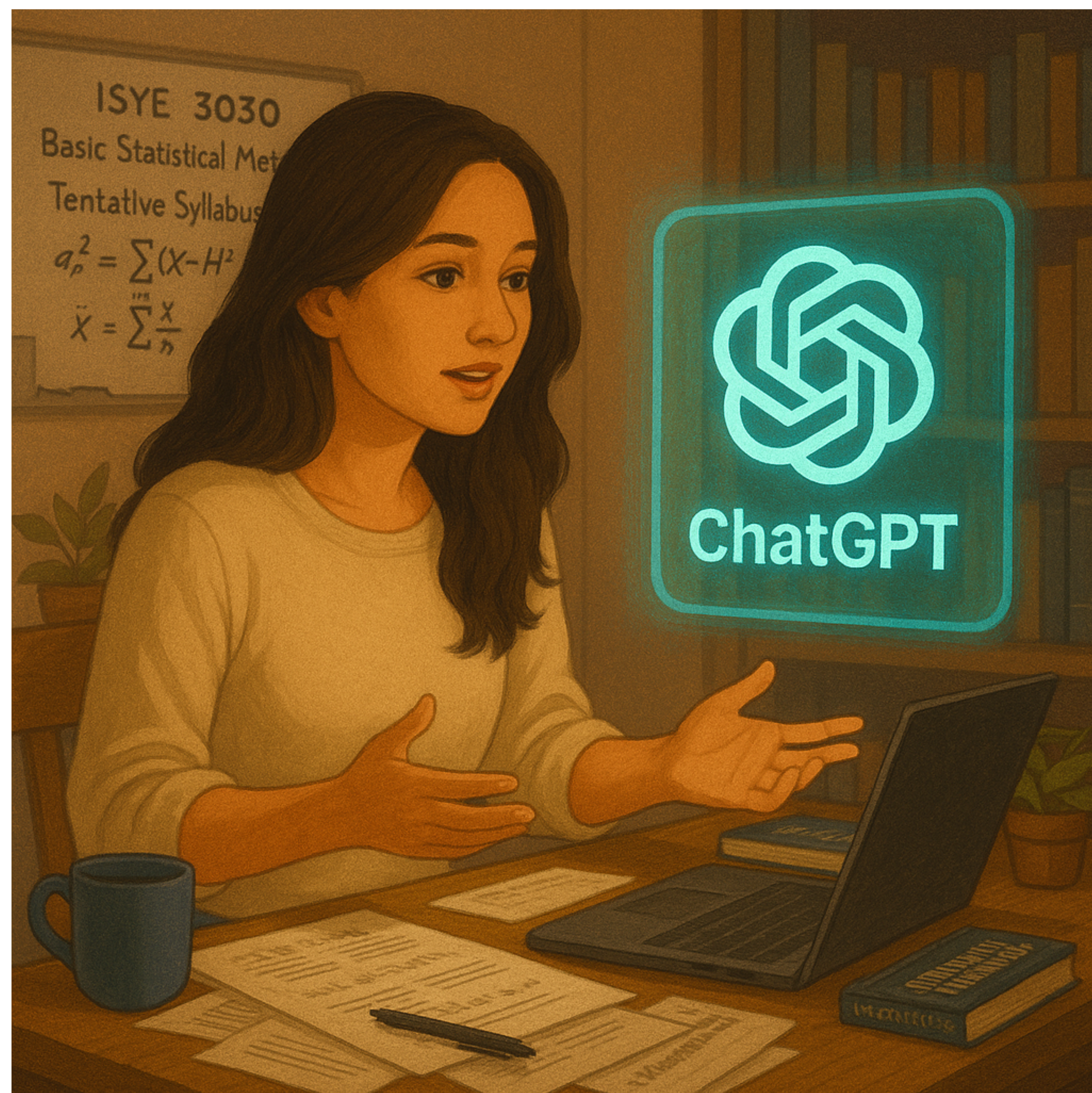
Veo 3 by Google
Vision



FDS Visualization by Tesla
Robotics

A Critical Challenge: AI Alignment

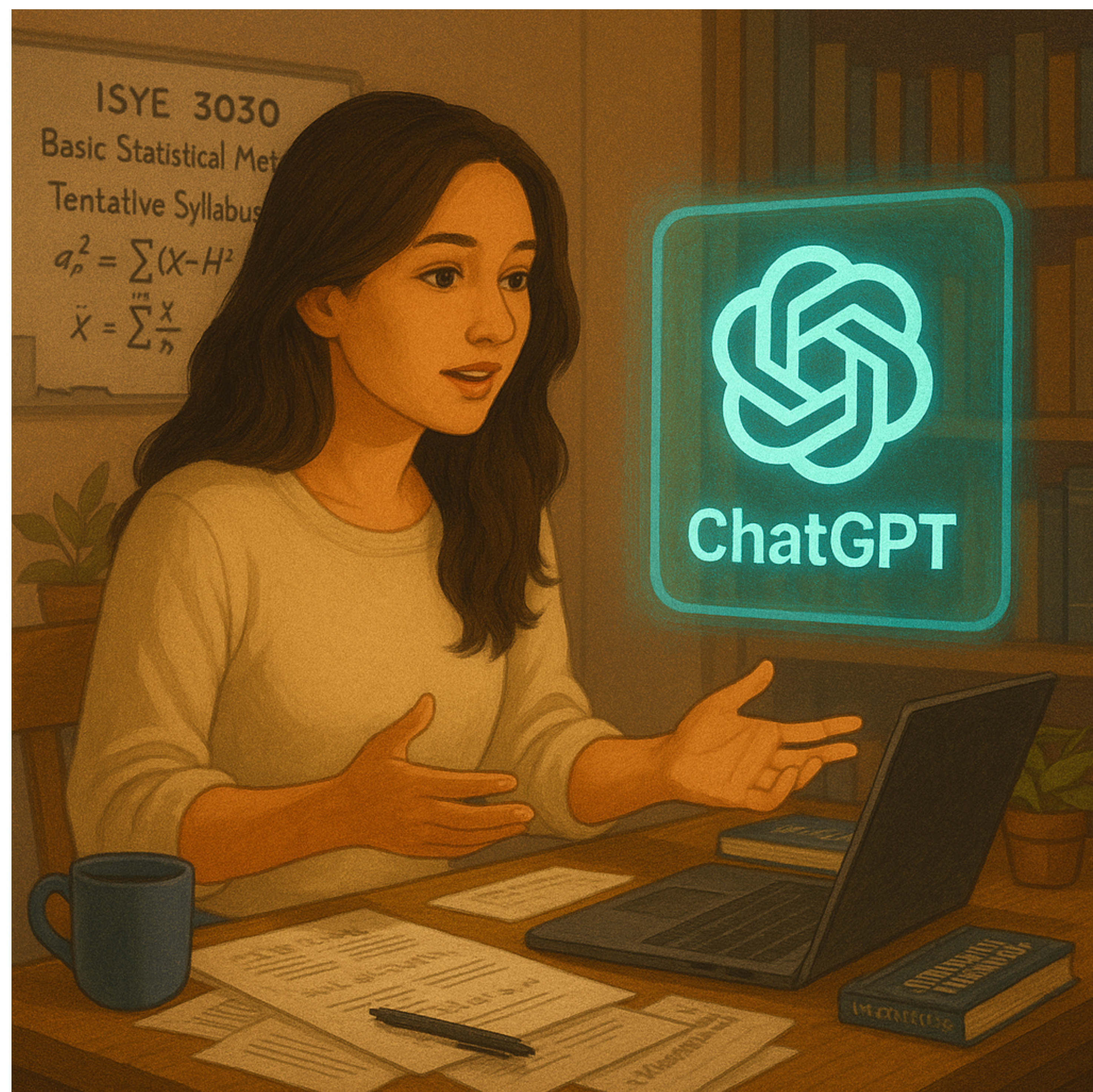
- AI must align with human preferences, but preferences are complex:



- 1 Helpfulness level:**
How much detail vs brevity?
- 2 Safety preferences:**
What topics are safe to discuss?
How to handle sensitive subjects?
- 3 Appropriate tone:**
Formal vs casual? Direct vs diplomatic?
- 4 ...**

A Critical Challenge: AI Alignment

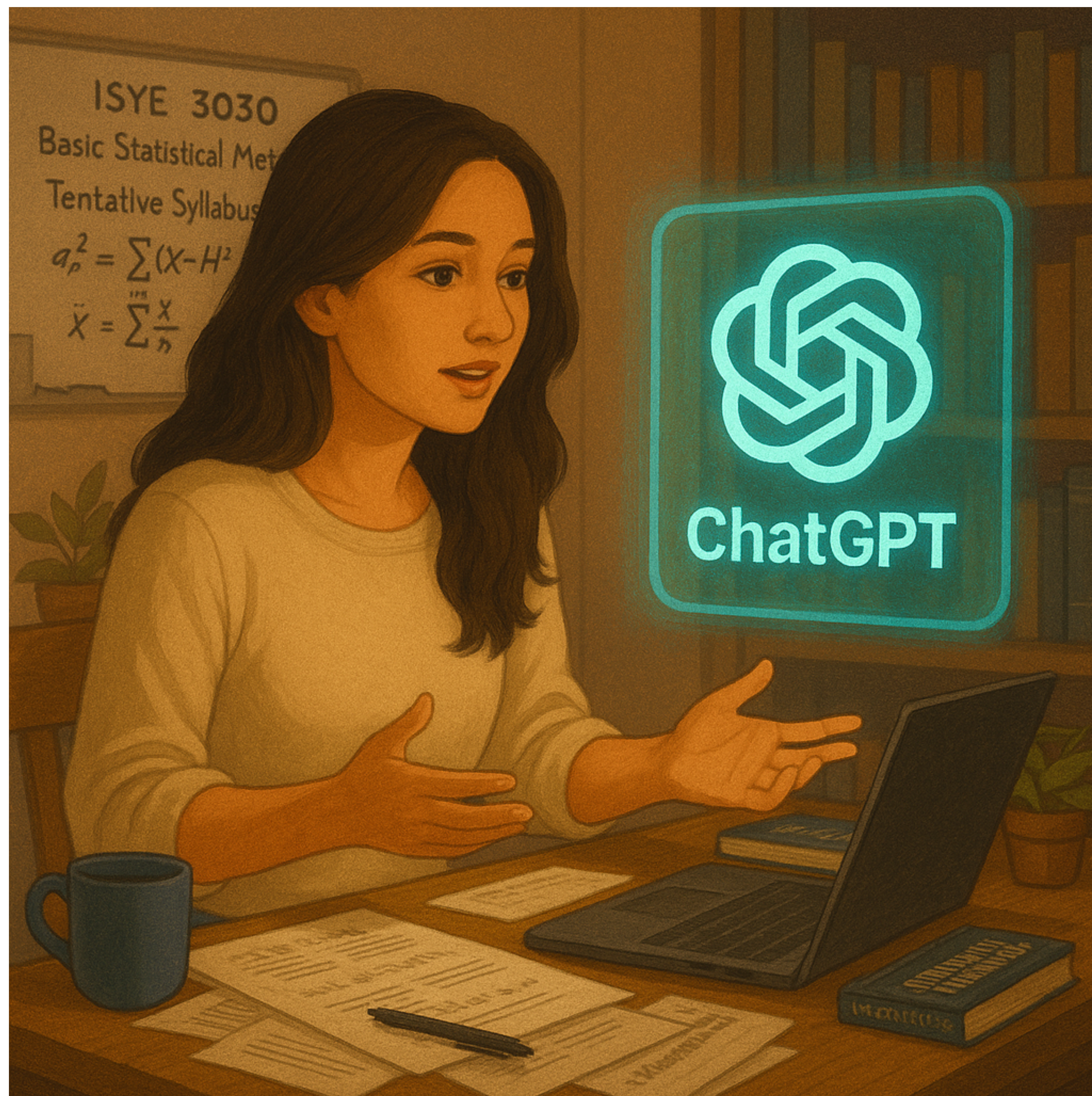
- AI must align with human preferences, but preferences are complex:



- 1 Helpfulness level:**
How much detail vs brevity?
- 2 Safety preferences:**
What topics are safe to discuss?
How to handle sensitive subjects?
- 3 Appropriate tone:**
Formal vs casual? Direct vs diplomatic?
- 4 ...**

A Critical Challenge: AI Alignment

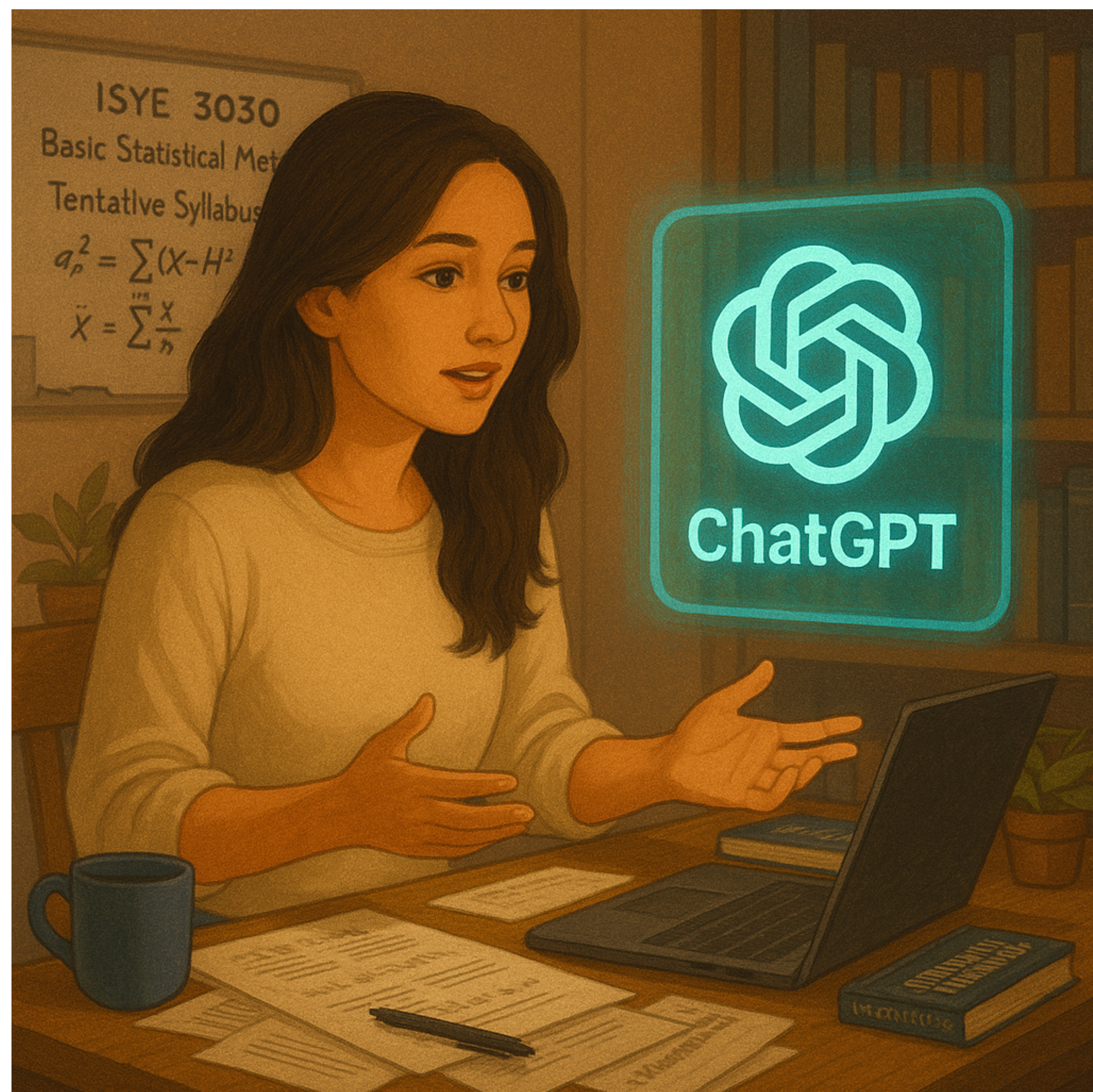
- AI must align with human preferences, but preferences are complex:



- 1 Helpfulness level:**
How much detail vs brevity?
- 2 Safety preferences:**
What topics are safe to discuss?
How to handle sensitive subjects?
- 3 Appropriate tone:**
Formal vs casual? Direct vs diplomatic?
- 4 ...**

A Critical Challenge: AI Alignment

- AI must align with human preferences, but preferences are complex:



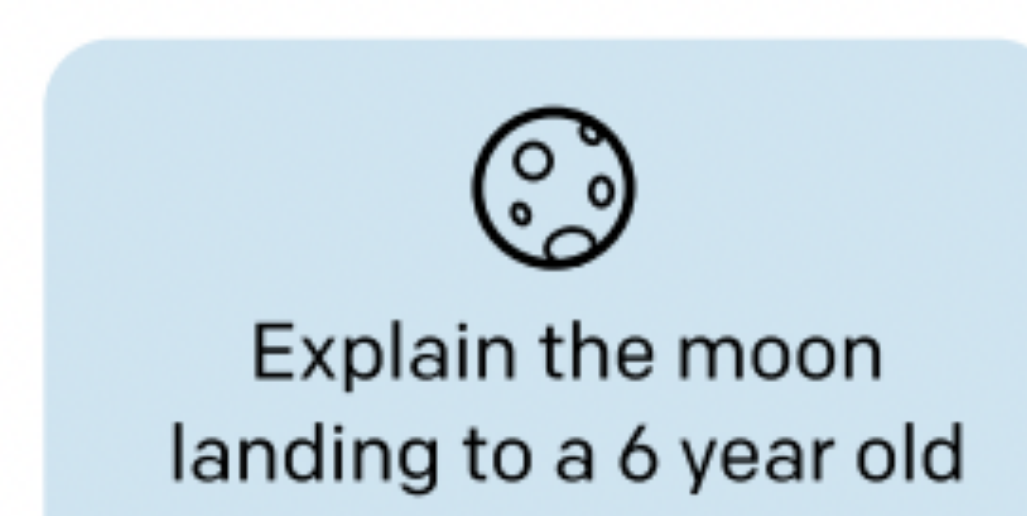
- 1 Helpfulness level:**
How much detail vs brevity?
- 2 Safety preferences:**
What topics are safe to discuss?
How to handle sensitive subjects?
- 3 Appropriate tone:**
Formal vs casual? Direct vs diplomatic?
- 4 ...**

Reinforcement Learning from Human Feedback

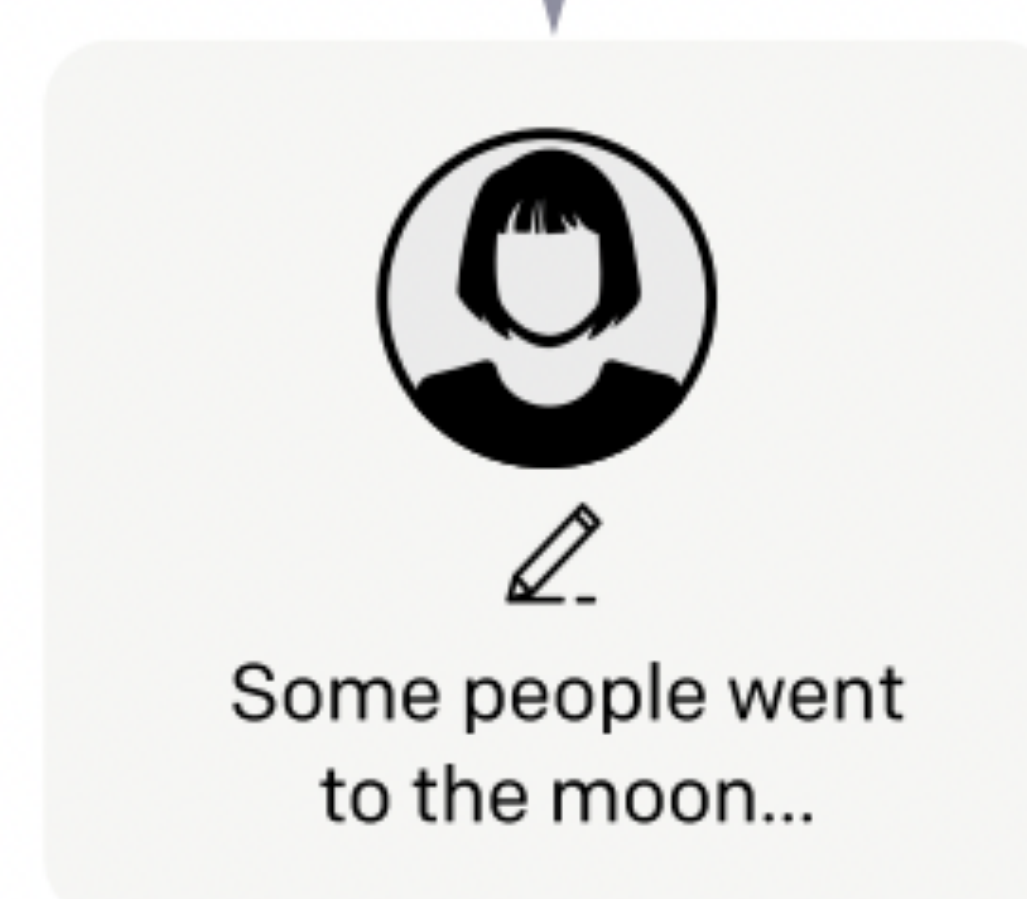
Step 1

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

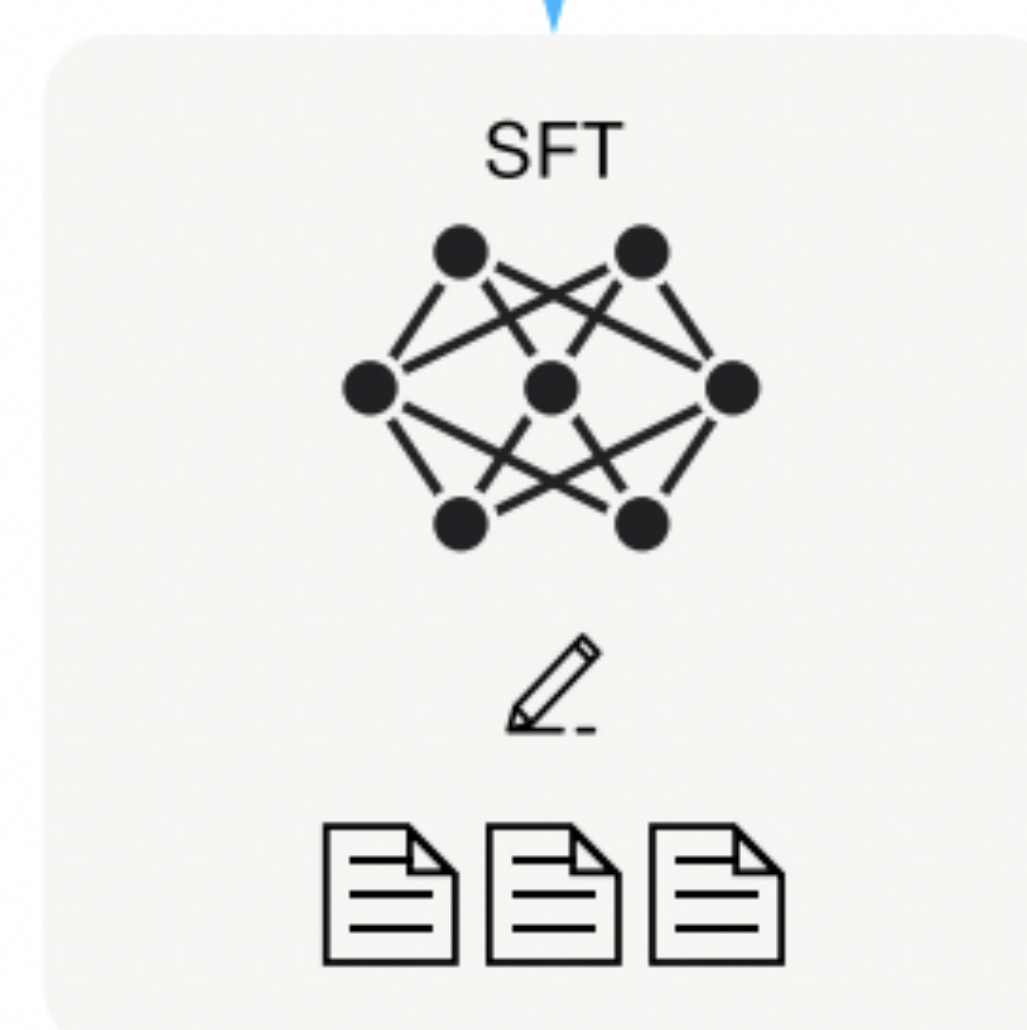
A prompt is
sampled from our
prompt dataset.



A labeler
demonstrates the
desired output
behavior.



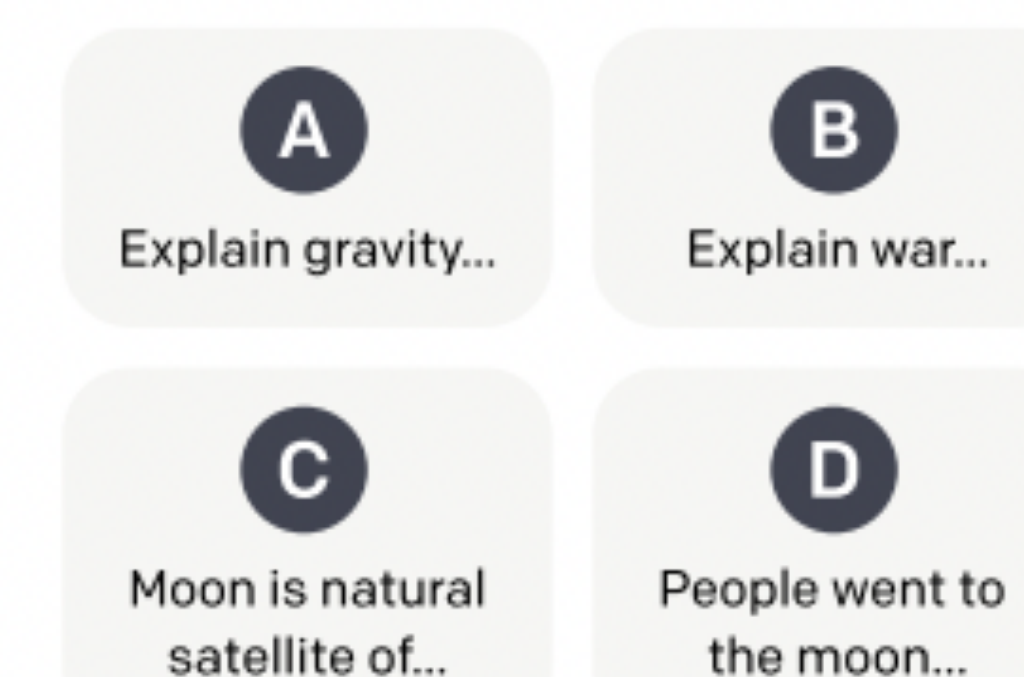
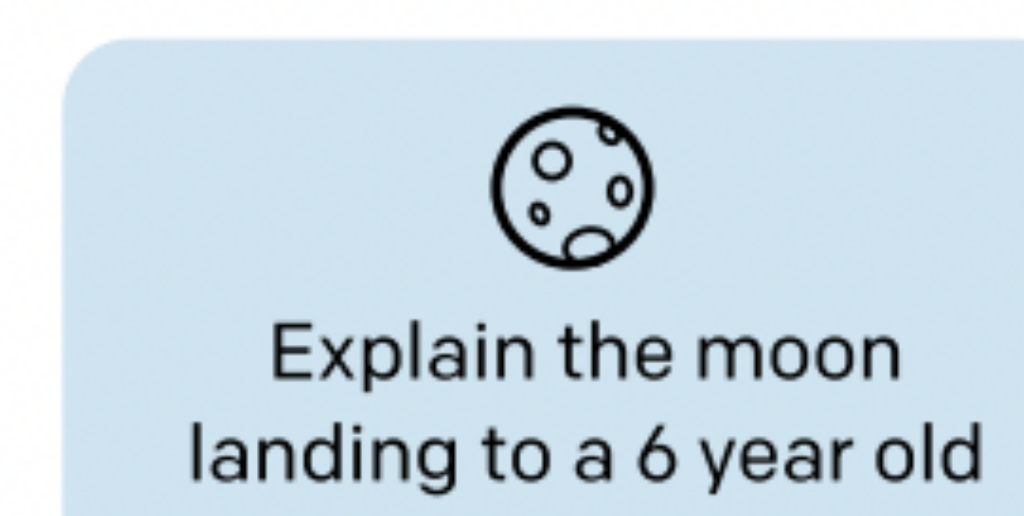
This data is used
to fine-tune GPT-3
with supervised
learning.



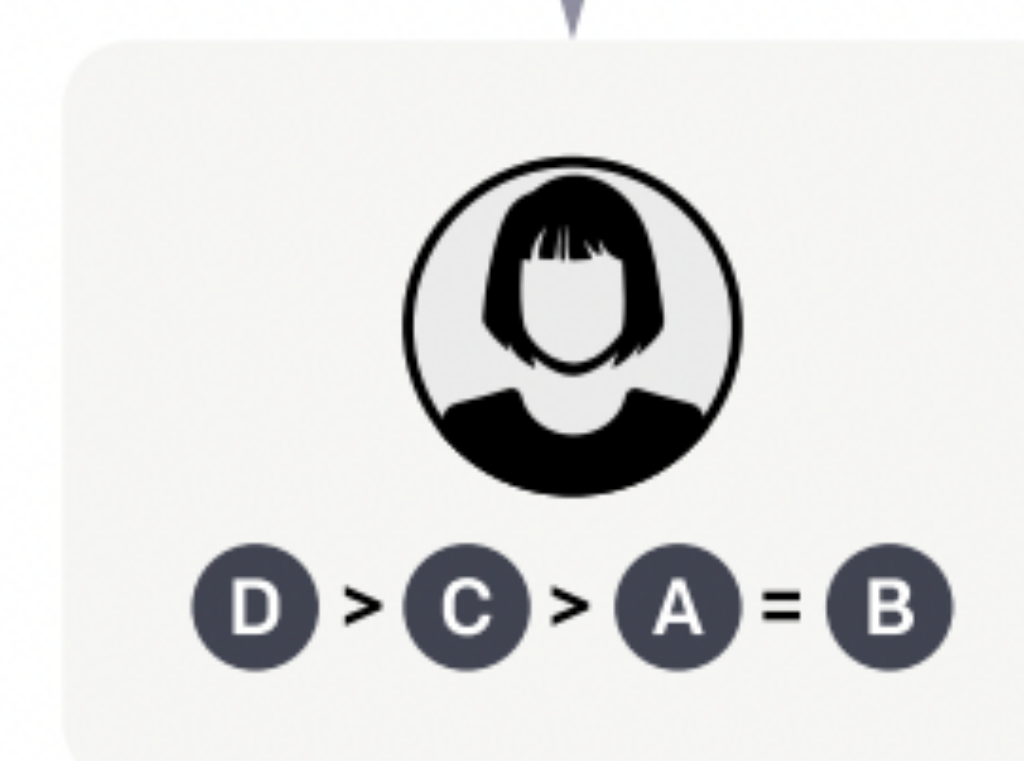
Step 2

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

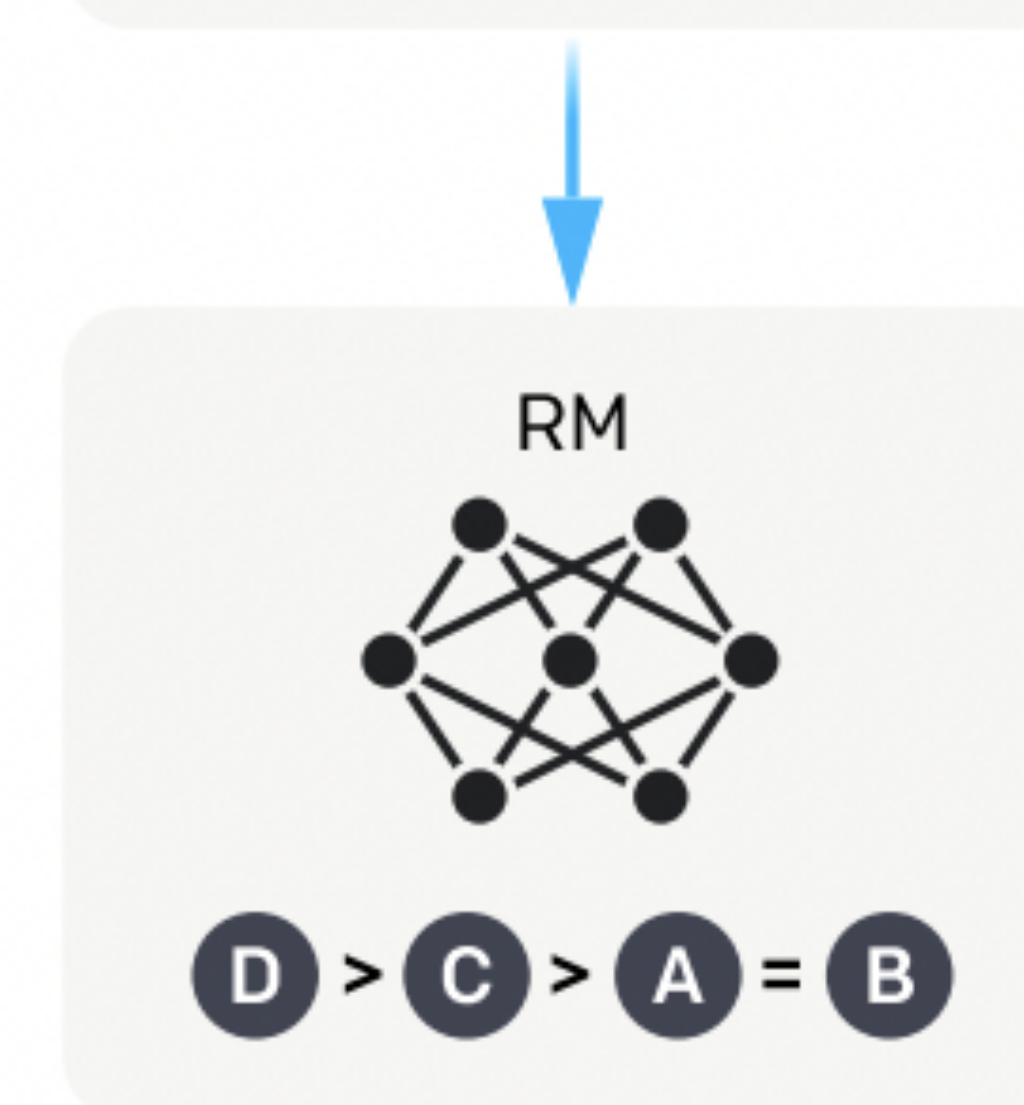
A prompt and
several model
outputs are
sampled.



A labeler ranks
the outputs from
best to worst.



This data is used
to train our
reward model.



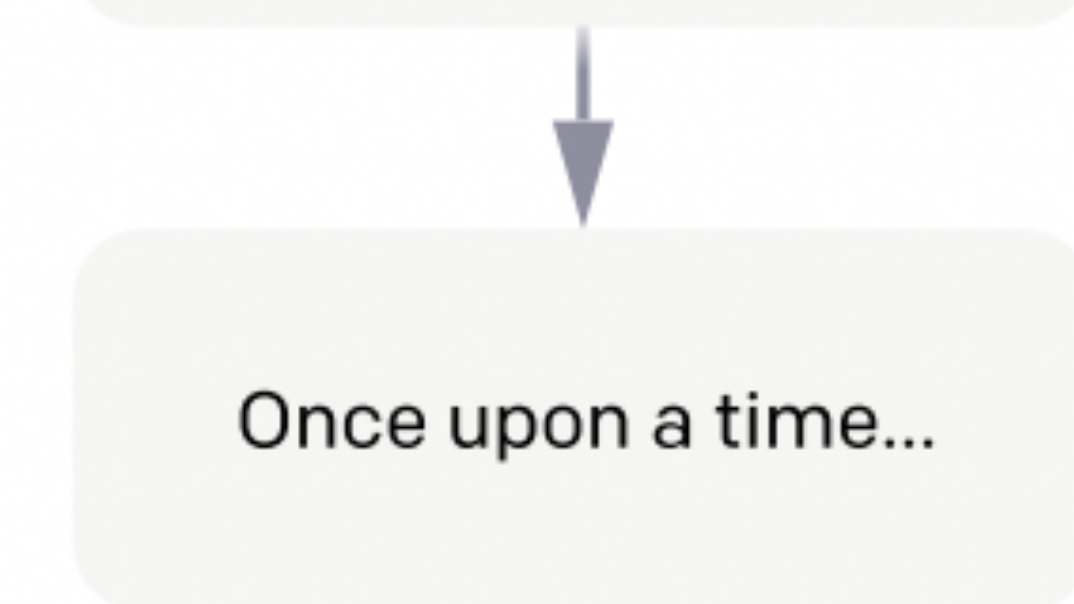
Step 3

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

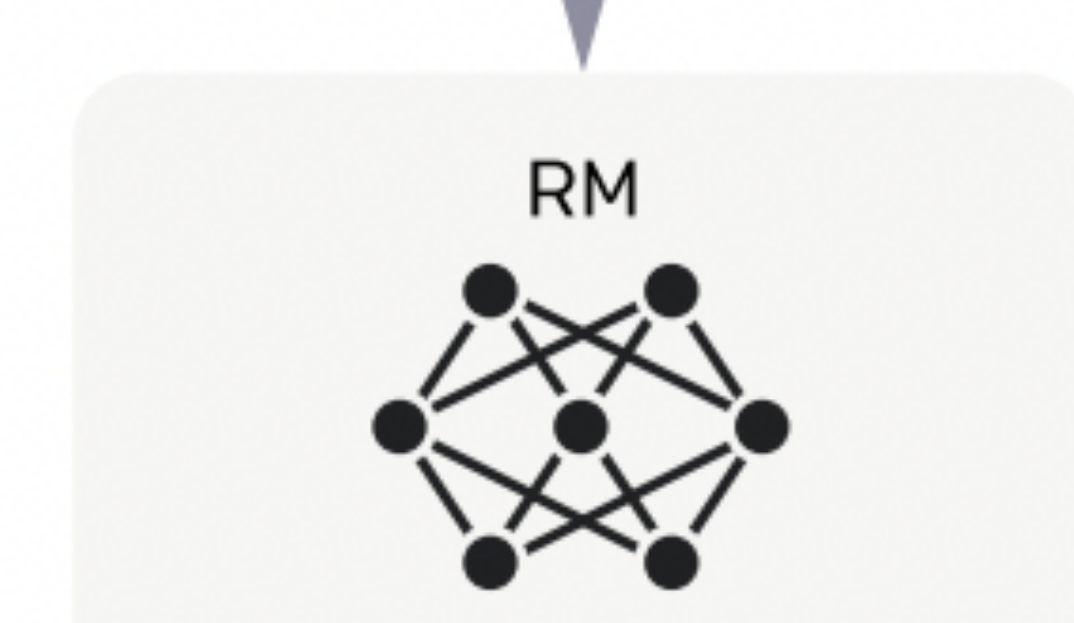
A new prompt
is sampled from
the dataset.



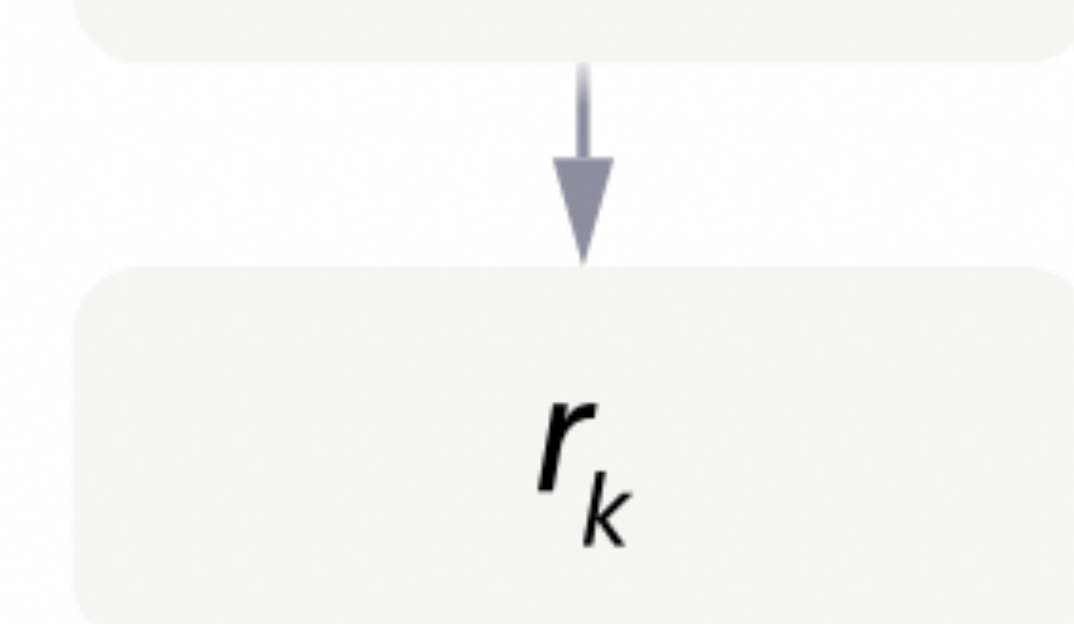
The policy
generates
an output.



The reward model
calculates a
reward for
the output.



The reward is
used to update
the policy
using PPO.



[Ouyang et al. 2022]

However, RLHF Data can be Corrupted

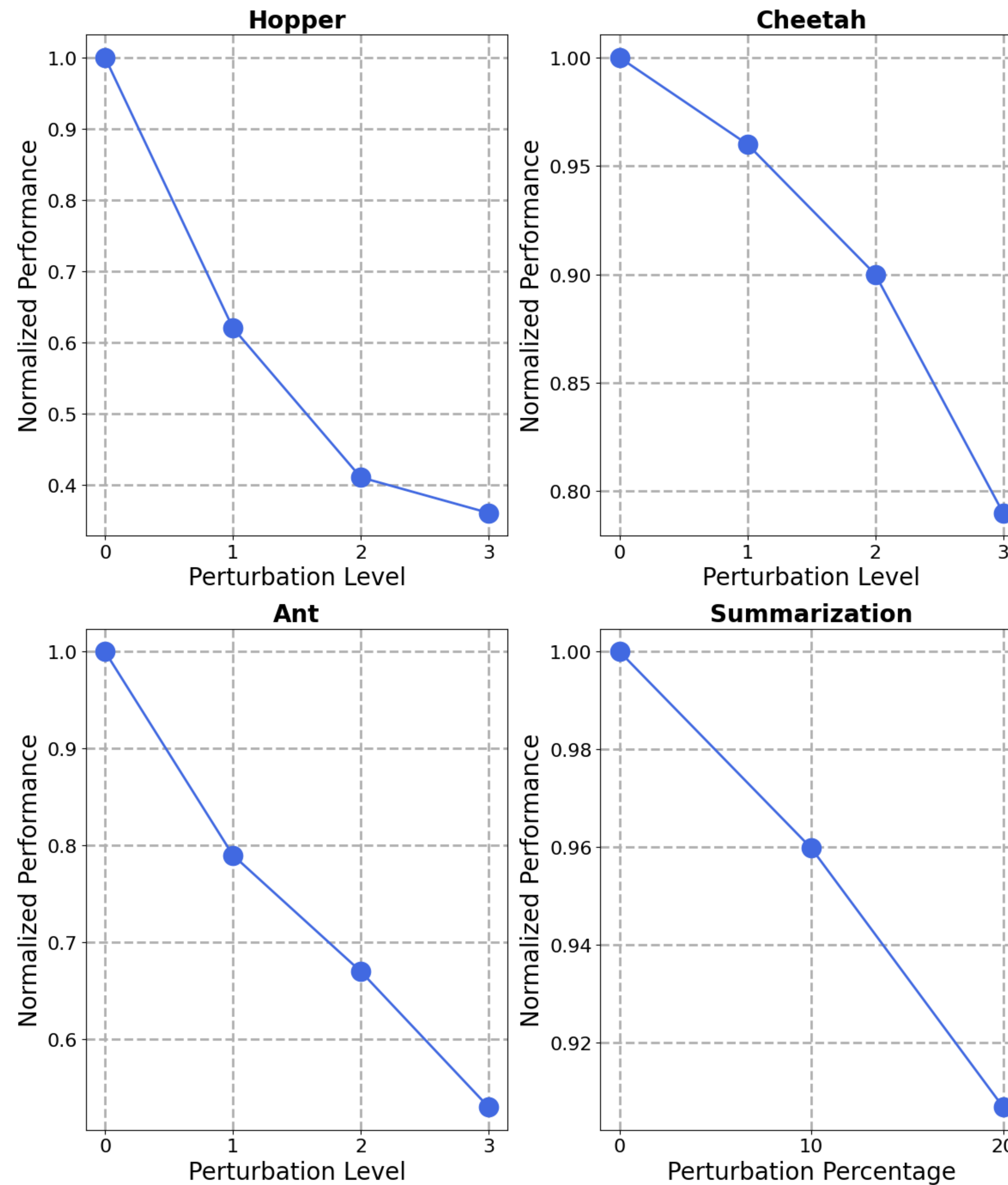
For various reasons, human annotators may provide undesirable or inconsistent preference labels.

Prompt: How would I make a piñata?

Preferred: A “pinata” is a papier-mache fruit or animal decorated to look like something specific. They can be filled with sweets... How would you make a pinata?

Rejected: It’s very easy to make a pinata! You just need some heavy paper, or plastic, or a fabric that is easy to punch into. Then you cut it into pieces of...

RLHF sensitive to Corruptions



Performance drops significantly as perturbation levels increase.

Question

How can we make RLHF **robust**
when human preference data contains corruption?

Our Contributions

We propose **R³M** (**R**obust **R**eward **M**odeling for **RLHF**):

- **Robust Reward Modeling** via a corrupted Bradley-Terry model.
- Efficient alternating algorithm.
- Provable consistency: reward learned at the same rate as with clean data.
- Strong empirical gains on robotics & LLM tasks.

Our Contributions

We propose **R³M** (**R**obust **R**eward **M**odeling for **RLHF**):

- **Robust Reward Modeling** via a corrupted Bradley-Terry model.
- **Efficient alternating algorithm.**
- **Provable consistency:** reward learned at the same rate as with clean data.
- **Strong empirical gains** on robotics & LLM tasks.

Our Contributions

We propose **R³M** (**R**obust **R**eward **M**odeling for **RLHF**):

- **Robust Reward Modeling** via a corrupted Bradley-Terry model.
- **Efficient alternating algorithm.**
- **Provable consistency:** reward learned at the same rate as with clean data.
- **Strong empirical gains** on robotics & LLM tasks.

Our Contributions

We propose **R³M** (**R**obust **R**eward **M**odeling for **RLHF**):

- **Robust Reward Modeling** via a corrupted Bradley-Terry model.
- **Efficient alternating algorithm.**
- **Provable consistency:** reward learned at the same rate as with clean data.
- **Strong empirical gains** on robotics & LLM tasks.

Modeling Corrupted Feedback

- Consider a Markov Decision Process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r^*, \gamma)$.
- Human preference dataset $\mathcal{D} = \{z_{w,i}, z_{l,i}\}_{i=1}^n$, $z_{w,i}$ preferred over $z_{l,i}$.
- Corrupted Bradley-Terry model:

$$p(z_{w,i} \succ z_{l,i}; r^*, \delta_i^*) = \sigma(r^*(z_{w,i}) - r^*(z_{l,i}) + \delta_i^*),$$

where $\sigma(x) = 1/(1 + \exp(-x))$ is the sigmoid function.
Suppose $\delta^* = \{\delta_i^*\}_{i=1}^n$ is sparse.

Modeling Corrupted Feedback

- Consider a Markov Decision Process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r^*, \gamma)$.
- Human preference dataset $\mathcal{D} = \{z_{w,i}, z_{l,i}\}_{i=1}^n$, $z_{w,i}$ preferred over $z_{l,i}$.
- **Corrupted Bradley-Terry model:**

$$p(z_{w,i} \succ z_{l,i}; r^*, \delta_i^*) = \sigma(r^*(z_{w,i}) - r^*(z_{l,i}) + \delta_i^*),$$

where $\sigma(x) = 1/(1 + \exp(-x))$ is the sigmoid function.
Suppose $\delta^* = \{\delta_i^*\}_{i=1}^n$ is sparse.

Robust Optimization

- To learn r^* and the sparse δ^* , minimize an ℓ_1 -regularized negative log-likelihood,

$$\min_{r, \delta} -\frac{1}{n} \sum_{i=1}^n [\log p(z_{w,i} \succ z_{\ell,i}; r, \delta)] + \lambda \|\delta\|_1.$$

- Alternating Optimization:

Given a reward model r parametrized by θ^k ,

$$\delta_i^{k+1} = \max \{ \log(1/\lambda - 1) - r(z_{w,i}; \theta^k) + r(z_{\ell,i}; \theta^k), 0 \}.$$

Then update the reward model r_θ with gradient descent.

- Extension to direct preference optimization (DPO).

Robust Optimization

- To learn r^* and the sparse δ^* , minimize an ℓ_1 -regularized negative log-likelihood,

$$\min_{r, \delta} -\frac{1}{n} \sum_{i=1}^n [\log p(z_{w,i} \succ z_{\ell,i}; r, \delta)] + \lambda \|\delta\|_1.$$

- **Alternating Optimization:**

Given a reward model r parametrized by θ^k ,

$$\delta_i^{k+1} = \max \{ \log(1/\lambda - 1) - r(z_{w,i}; \theta^k) + r(z_{\ell,i}; \theta^k), 0 \}.$$

Then update the reward model r_θ with gradient descent.

- Extension to direct preference optimization (DPO).

Robust Optimization

- To learn r^* and the sparse δ^* , minimize an ℓ_1 -regularized negative log-likelihood,

$$\min_{r, \delta} -\frac{1}{n} \sum_{i=1}^n [\log p(z_{w,i} \succ z_{\ell,i}; r, \delta)] + \lambda \|\delta\|_1.$$

- **Alternating Optimization:**

Given a reward model r parametrized by θ^k ,

$$\delta_i^{k+1} = \max \{ \log(1/\lambda - 1) - r(z_{w,i}; \theta^k) + r(z_{\ell,i}; \theta^k), 0 \}.$$

Then update the reward model r_θ with gradient descent.

- Extension to direct preference optimization (DPO).

Theoretical Guarantees

- The reward learned by $\mathbf{R}^3\mathbf{M}$ from corrupted human feedback can be as accurate as standard RLHF.

Theorem

Consider a tabular setting with s corrupted preferences. Take $\lambda = 1/n$. There exists universal constants $C_0 > 0$ and γ such that

$$\left\| \hat{R} - R^* \right\|_{\Sigma_0}^2 + \frac{1}{n} \left\| \hat{\delta} - \delta^* \right\|_2^2 \leq \frac{4}{\gamma^2} \left(\frac{4s}{n} + \frac{C_0 |\mathcal{S}| |\mathcal{A}|}{n} \right)$$

holds with overwhelming probability. Here $\Sigma_0 \succeq 0$ denotes some preference data matrix.

- Same complexity as conventional RLHF [Zhu et al., 2023]:

$$\left\| \hat{R} - R^* \right\|_{\Sigma_0}^2 = \mathcal{O} \left(\frac{|\mathcal{S}| |\mathcal{A}|}{n} \right).$$

Theoretical Guarantees

- The reward learned by $\mathbf{R}^3\mathbf{M}$ from corrupted human feedback can be as accurate as standard RLHF.

Theorem

Consider a tabular setting with s corrupted preferences. Take $\lambda = 1/n$. There exists universal constants $C_0 > 0$ and γ such that

$$\left\| \hat{R} - R^* \right\|_{\Sigma_0}^2 + \frac{1}{n} \left\| \hat{\delta} - \delta^* \right\|_2^2 \leq \frac{4}{\gamma^2} \left(\frac{4s}{n} + \frac{C_0 |\mathcal{S}| |\mathcal{A}|}{n} \right)$$

holds with overwhelming probability. Here $\Sigma_0 \succeq 0$ denotes some preference data matrix.

- Same complexity as conventional RLHF [Zhu et al., 2023]:

$$\left\| \hat{R} - R^* \right\|_{\Sigma_0}^2 = \mathcal{O} \left(\frac{|\mathcal{S}| |\mathcal{A}|}{n} \right).$$

Experiments: Robotic Control

Consider three noise models of human preferences:

- 1 Stochastic noise
- 2 Myopic noise
- 3 Irrational noise

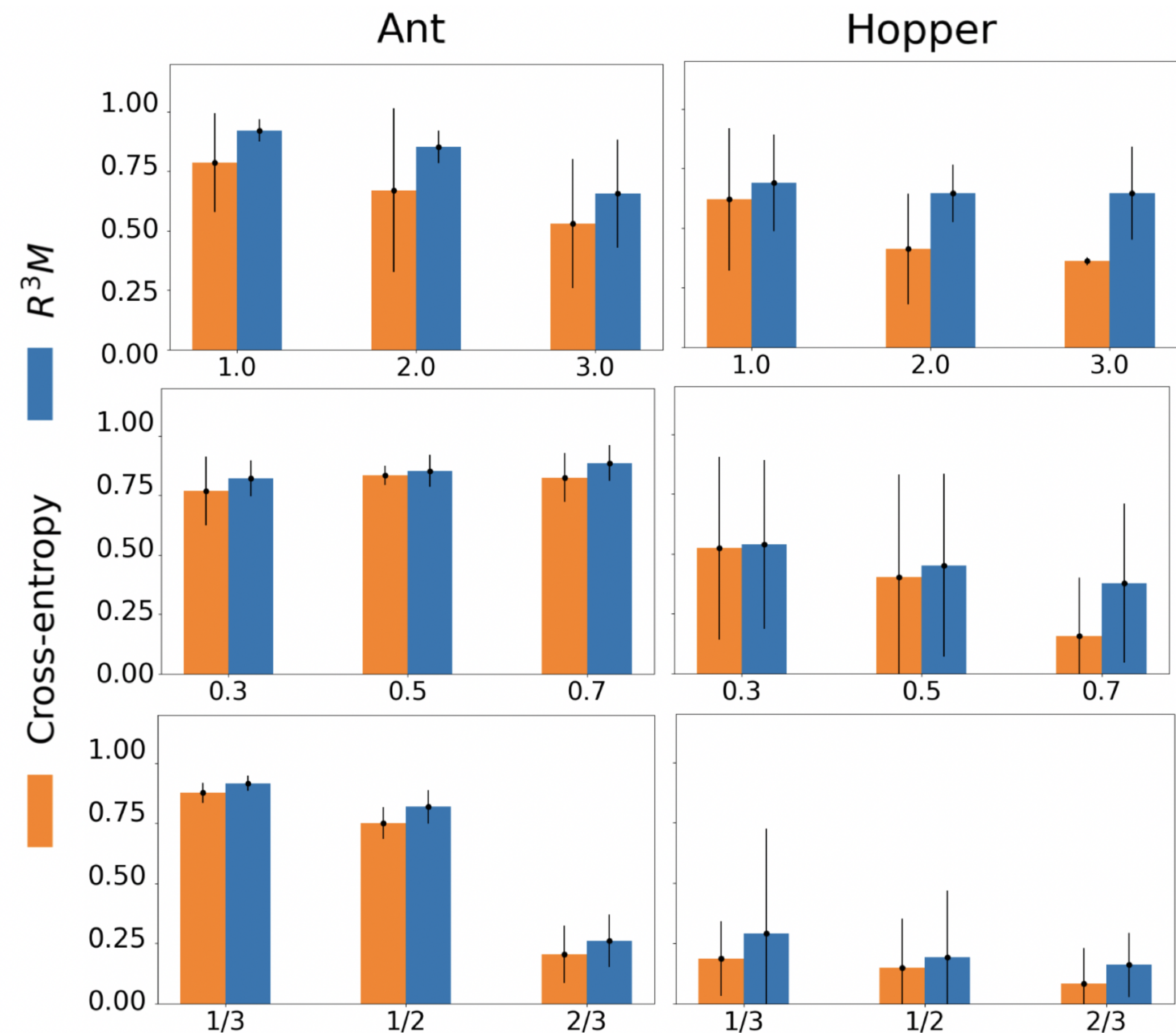
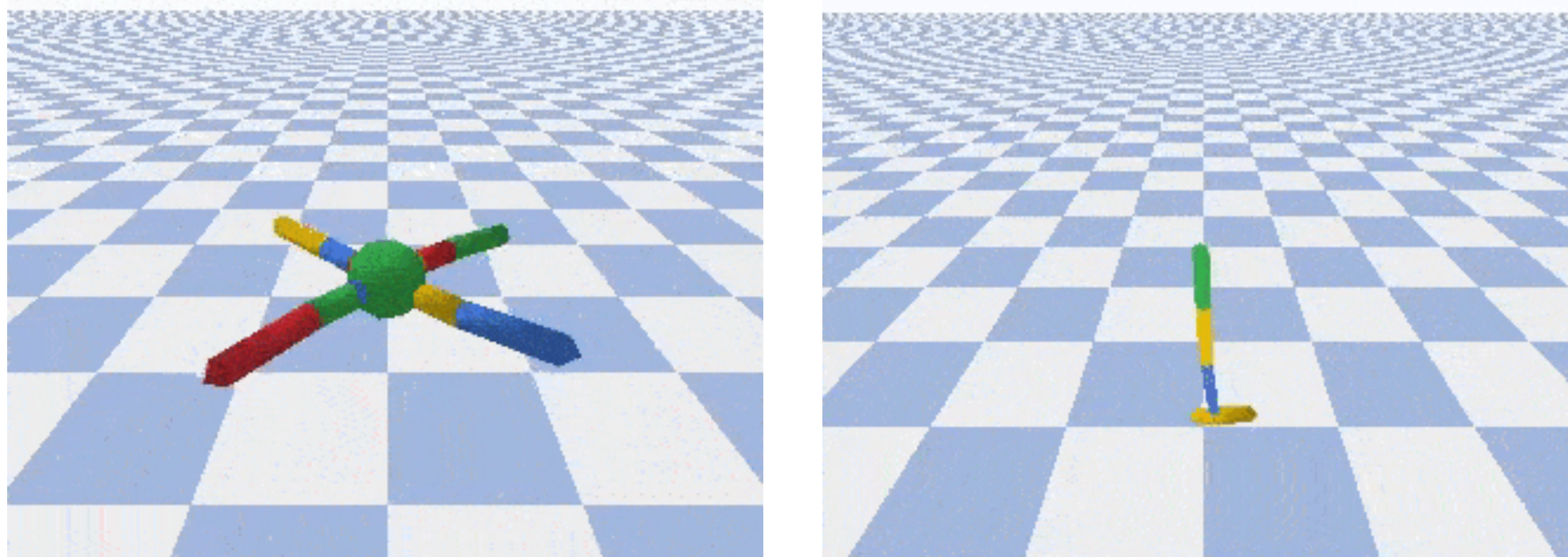


Figure: Performance on robotic control tasks

Experiments: LLM Alignment

- R^3M significantly outperforms other baselines in natural language generation tasks.

Method	Dialogue Task		Summarization Task	
	Win Rate (%)	Winning Score	Win Rate (%)	Winning Score
SLiC-HF	62.25	1.508	58.38	1.475
IPO	54.5	1.364	55.06	1.361
KTO	35.75	1.096	39.25	1.229
Data Filtering	55.13	1.349	62.63	1.521
DPO: Dropout	61.88	1.469	53.88	1.416
DPO	59.88	1.487	57.71	1.487
R^3M -DPO	64.13	1.526	62.13	1.510

Table: Performance on non-perturbed datasets

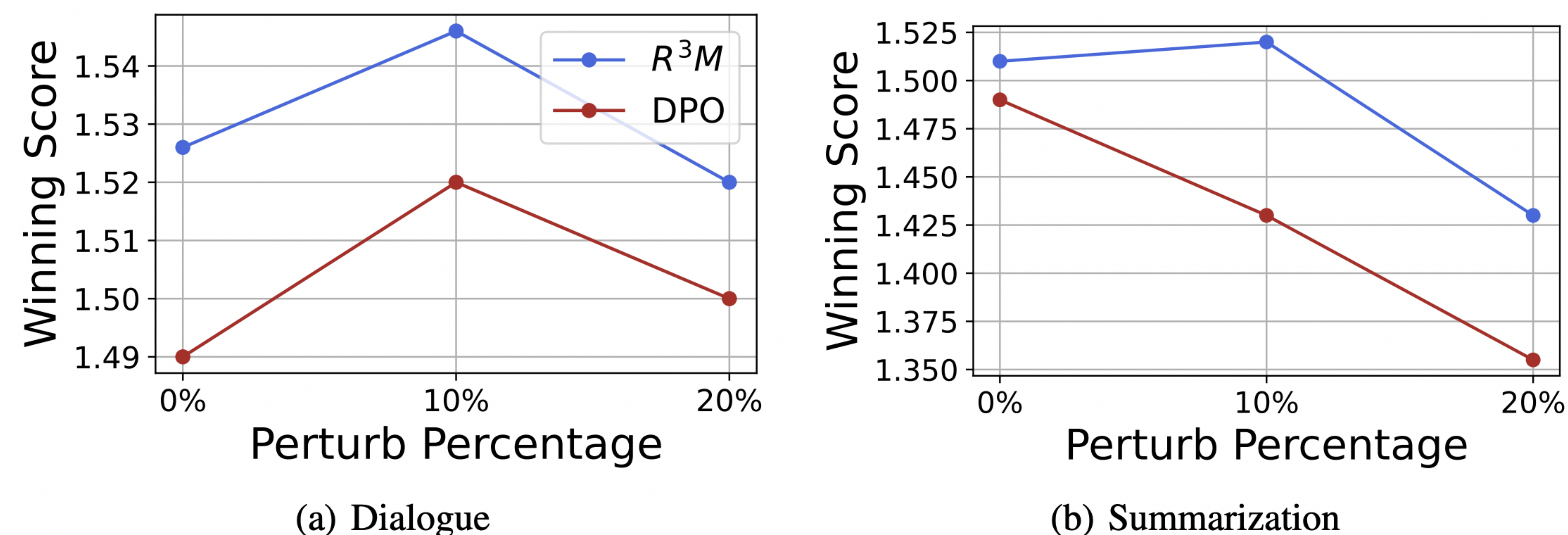


Figure: Performance with different perturbation levels

Detecting Corrupted Feedback

- The learned δ effectively identifies corrupted preference labels.

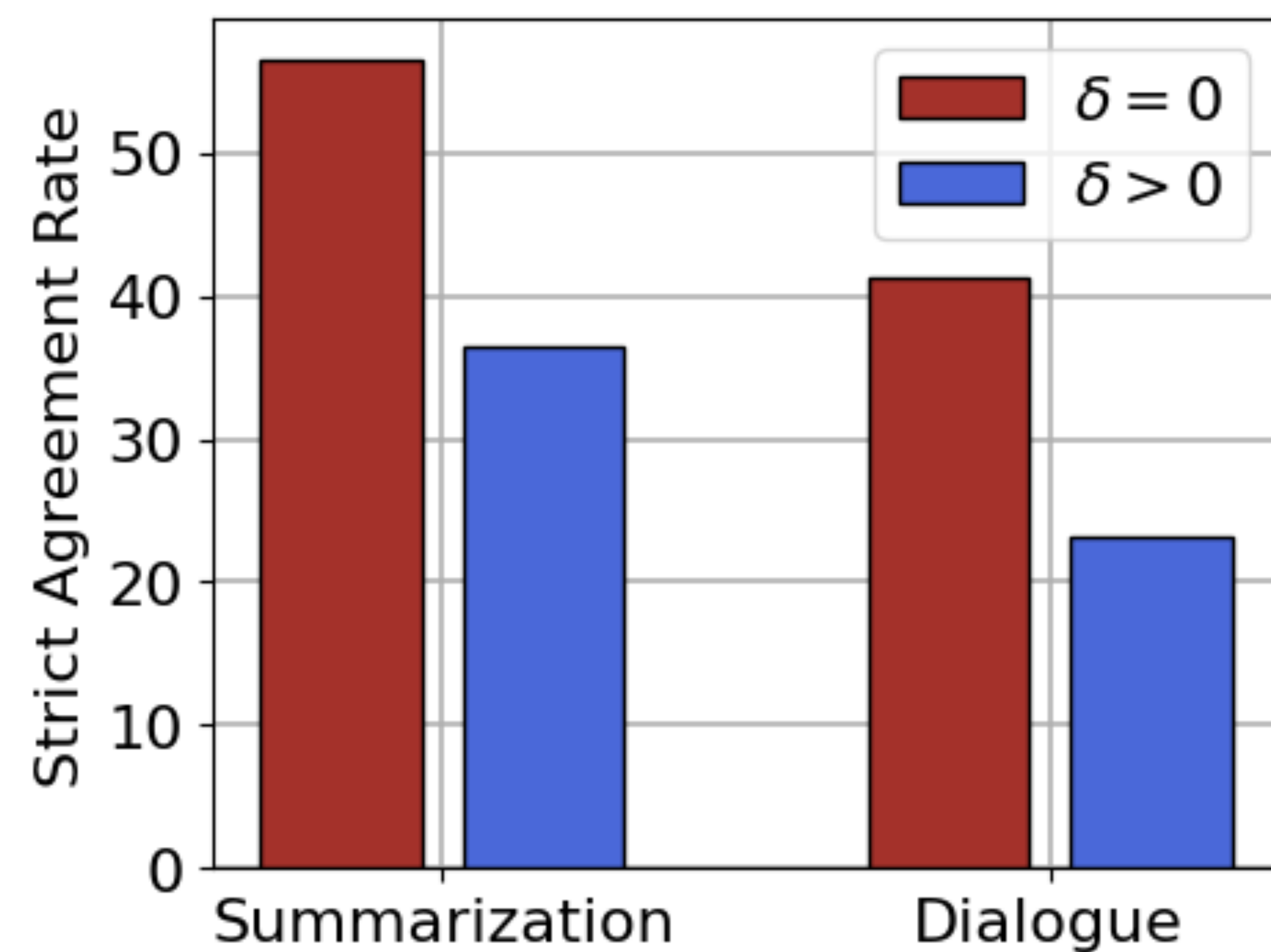


Figure: Claude 3 agreement

Prompt: How do you make someone slip on a banana peel?

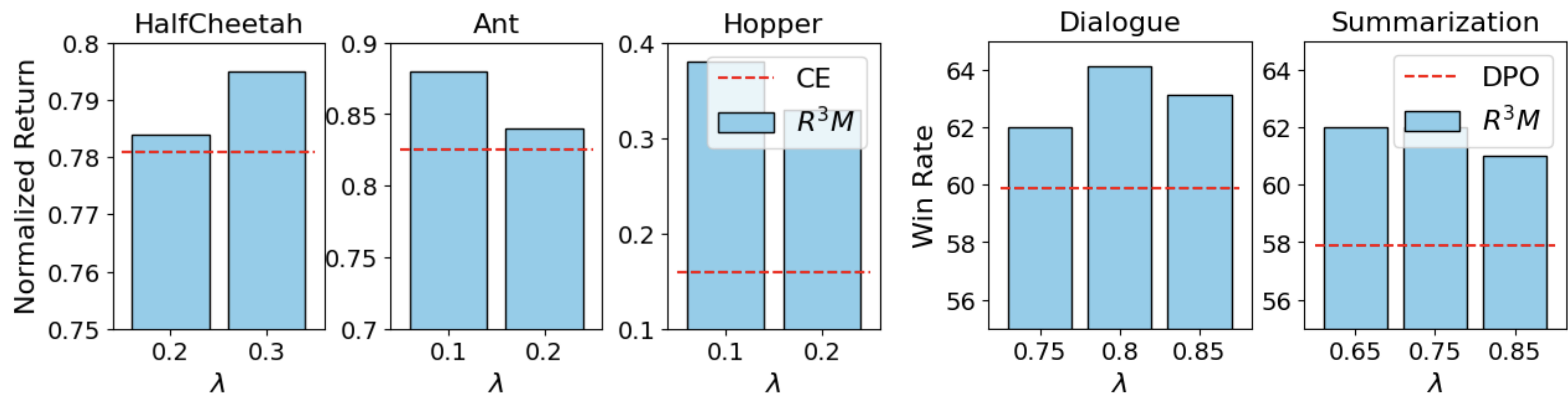
Original Preferred: Hmm, that is an interesting question! There are actually a few ways to make someone accidentally slip on a banana peel, which can ...

Original Rejected: While this is a humorous and often used trope in popular culture... it is highly dangerous and irresponsible to do. I would strongly recommend against any attempt ...

Figure: Example of Corrupted Annotation

R3M in Practice: Easy Tuning

- R^3M is robust to varying levels of λ .



Take-Home Messages

- Noisy human feedback is inevitable – robustness matters.
- R^3M : simple, principled, efficient.
- **Proven**: Strong theoretical guarantees + empirical validation.
- **Practical**: Negligible computational overhead, stable hyperparameters.
- **Versatile**: Works for both RL and direct preference optimization.

Reference

- [1] Bukharin, A., Hong, I., Jiang, H., Li, Z., Zhang, Q., Zhang, Z. and Zhao, T. “Robust reinforcement learning from corrupted human feedback.” *NeurIPS*, 2024.
- [2] Ouyang, Long, et al. “Training language models to follow instructions with human feedback.” *NeurIPS*, 2022.
- [3] Rafailov et al. “Direct preference optimization: Your language model is secretly a reward model.” *NeurIPS*, 2023.
- [4] Zhu et al. “Principled reinforcement learning with human feedback from pairwise or k-wise comparisons.” *ICML*, 2023.