

AlpacaFarm

A Simulation Framework for Methods that Learn from Human Feedback

Yann Dubois*, Xuechen Li*, Rohan Taori*, Tianyi Zhang*, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto



Xuechen Li, 10/05/23

Paper: <https://arxiv.org/abs/2305.14387>

Code: https://github.com/tatsu-lab/alpaca_farm

Blog: <https://crfm.stanford.edu/2023/05/22/alpaca-farm.html>

From GPT3 to InstructGPT and ChatGPT

```
Prompt Explain the moon landing to a 6 year old in a few sentences.

Completion GPT-3
Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

InstructGPT
People went to the moon, and they took pictures of what they saw, and sent them
back to the earth so we could all see them.
```

Instruction-following models are more user-friendly

This work studies building these models and creates a simulator for RLHF

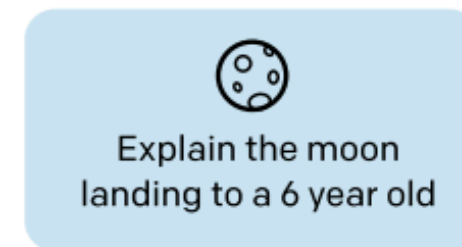
Building instruction-following models

Step 1: Supervised Fine-Tuning (SFT)

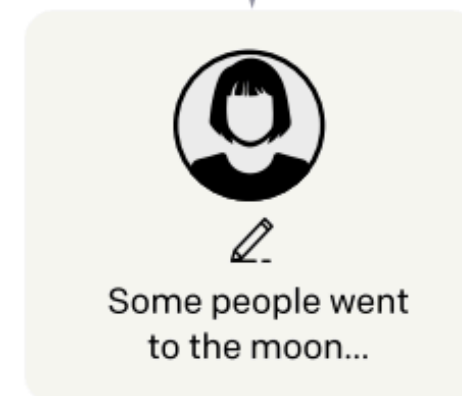
Step 1

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

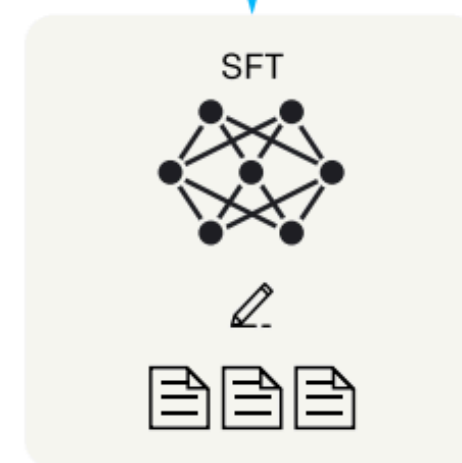
A prompt is
sampled from our
prompt dataset.



A labeler
demonstrates the
desired output
behavior.



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



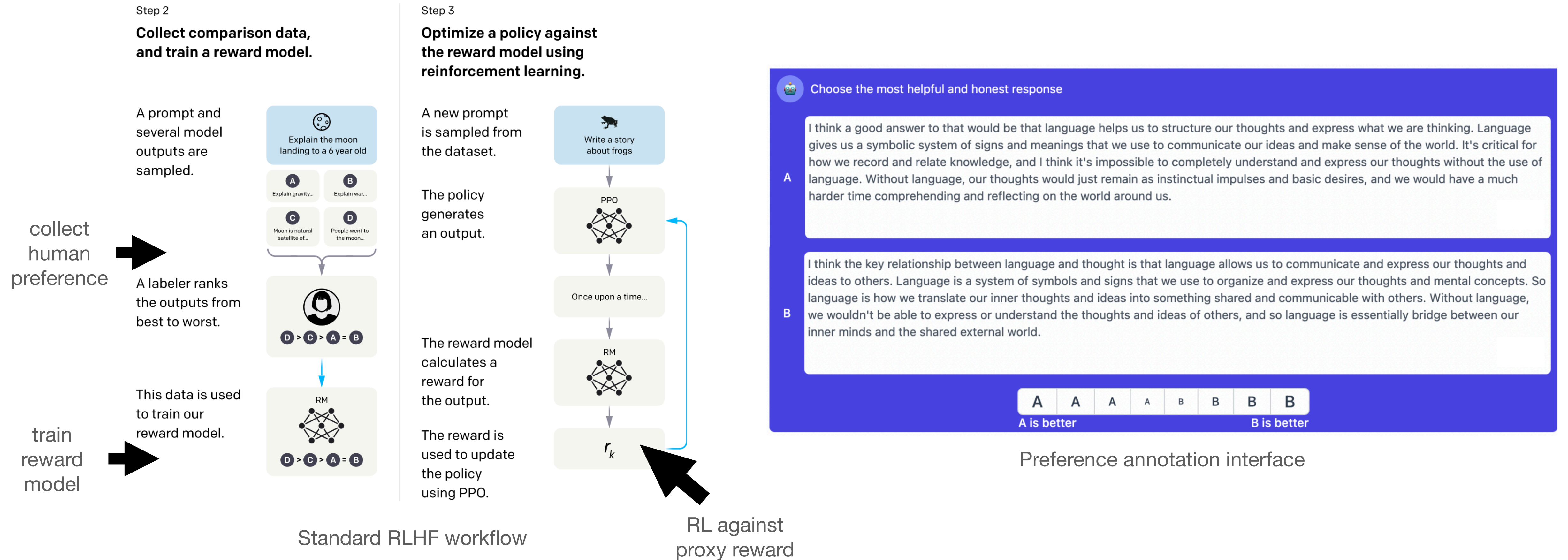
SFT workflow

A screenshot of a chat interface. At the top, it says "Reply as Assistant" with a help icon. Below that, a system message says "Given the following conversation, provide an adequate reply". A user message bubble contains the text: "In the intro scene to the movie 'The Dark Knight Rises', when the character Bane, as played by actor Tom Hardy, replied 'For You.' in response to the character of the CIA agent stating 'You are a big guy.' Please explain the meaning of Bane's reply of 'For you.' in detail." Below the user message are icons for language (en), like, reply, and more options. At the bottom, there is a tip: "Tip: You can use a keyboard shortcut to Review and Submit responses: cmd + Enter". Below the tip are tabs for "Write" and "Preview". A rich text editor toolbar is visible with icons for bold, italic, quote, link, code, list, link, unlink, insert, table, and sum. The text area contains the placeholder "Write your reply here...".

Human demonstration data collection

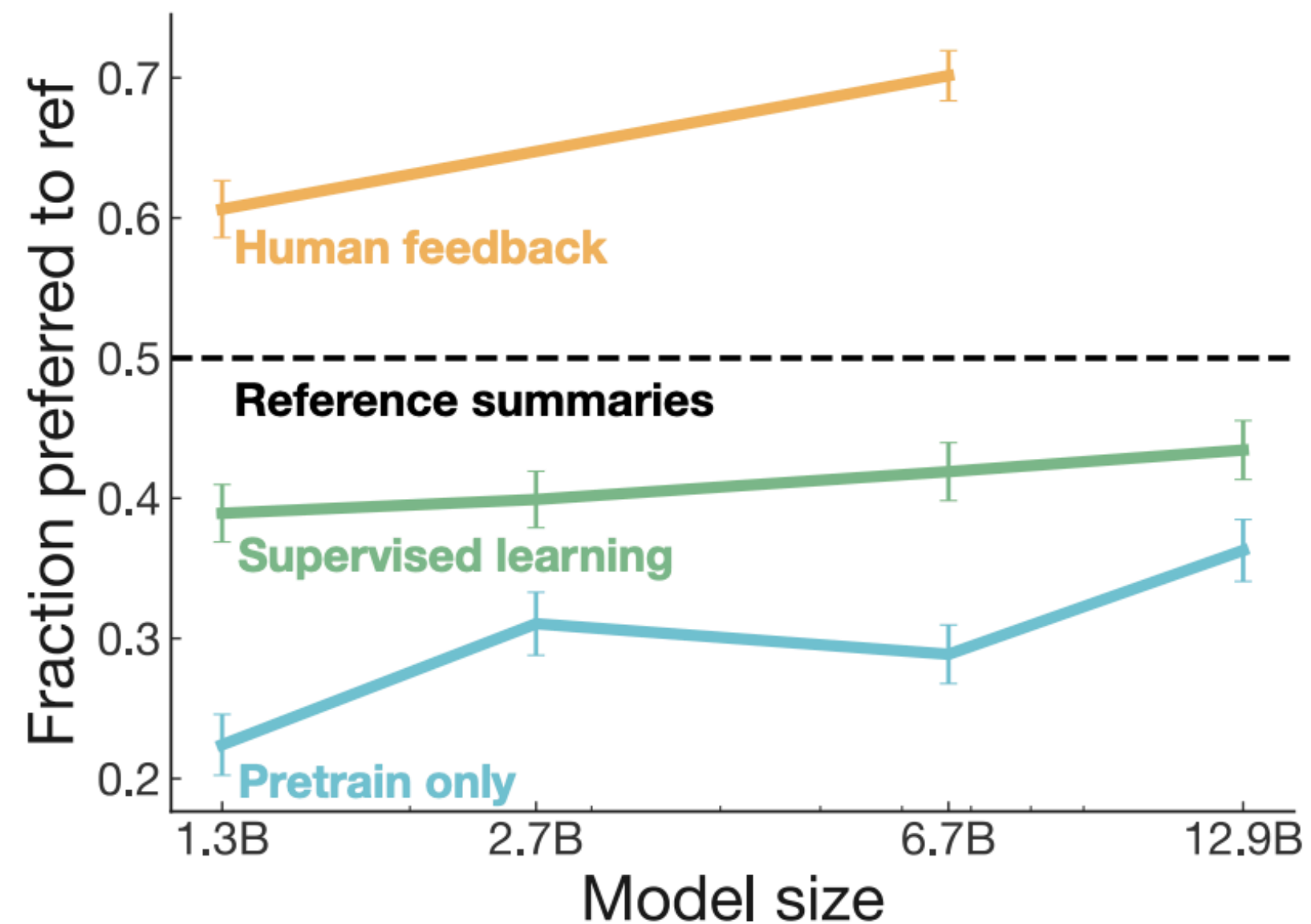
Building instruction-following models

Step 2: Learning from human (preference) feedback



Building instruction-following models

Step 3: Evaluation with human interaction



win rate to ref: fraction of times preferred

Chatbot Arena: Benchmarking LLMs in the Wild with Elo Ratings

by: Lianmin Zheng*, Ying Sheng*, Wei-Lin Chiang, Hao Zhang, Joseph E. Gonzalez, Ion Stoica, May 03, 2023

We present Chatbot Arena, a benchmark platform for large language models (LLMs) that features anonymous, randomized battles in a crowdsourced manner. In this blog post, we are releasing our initial results and a leaderboard based on the Elo rating system, which is a widely-used rating system in chess and other competitive games. We invite the entire community to join this effort by contributing new models and evaluating them by asking questions and voting for your favorite answer.

Rank	Model	Elo Rating	Description
1	vicuna-13b	1169	a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS
2	koala-13b	1082	a dialogue model for academic research by BAIR
3	oasst-pythia-12b	1065	an Open Assistant for everyone by LAION
4	alpaca-13b	1008	a model fine-tuned from LLaMA on instruction-following demonstrations by Stanford
5	chatglm-6b	985	an open bilingual dialogue language model by Tsinghua University
6	fastchat-t5-3b	951	a chat assistant fine-tuned from FLAN-T5 by LMSYS
7	dolly-v2-12b	944	an instruction-tuned open large language model by Databricks

elo rating

Research on instruction-following is hard because...

- for learning from human preference feedback (step 2)
 - A. collecting human preferences is costly: takes days to weeks; has nontrivial dollar cost
 - B. few validated open implementations for common approaches (e.g, hard to get RL to work)
- C. no benchmark for open-ended setting; human evaluation is expensive/slow

**I have a method for learning from preferences.
How do I know if it works for building instruction-following models?**

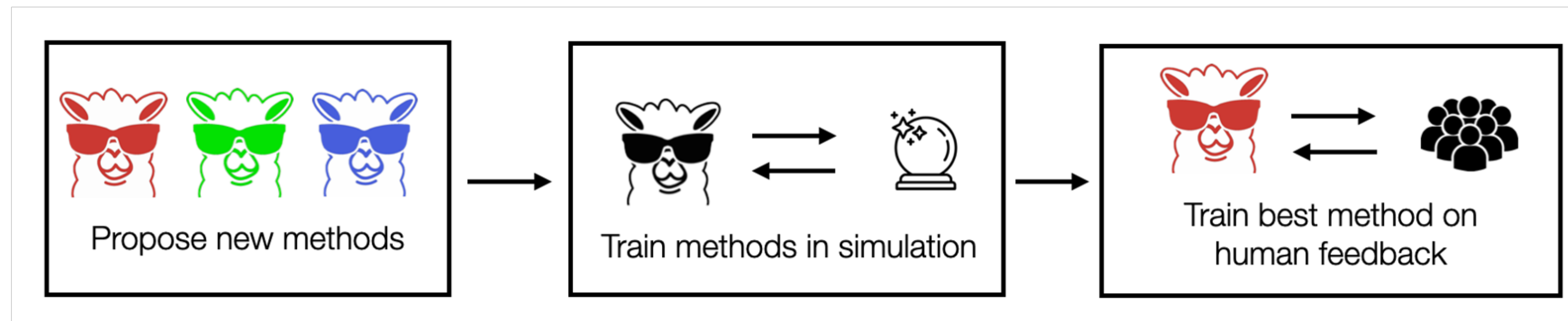
AlpacaFarm addresses the challenges

Helps you answer “if it works” faster and more cheaply

- A. collecting human preference data is costly
 - **AlpacaFarm simulates human feedback with API LLM (GPT4) feedback**
- B. few validated open implementations for common approaches
 - **AlpacaFarm provides reference implementations of popular methods**
- C. no benchmark for open-ended setting; human evaluation is expensive/slow
 - **AlpacaFarm provides a fast and cheap evaluation backed by real usage pattern**

AlpacaFarm is a gym environment

Build and test methods in simulation; deploy in real world



AlpacaFarm is a gym environment

Build and test methods in simulation; deploy in real world

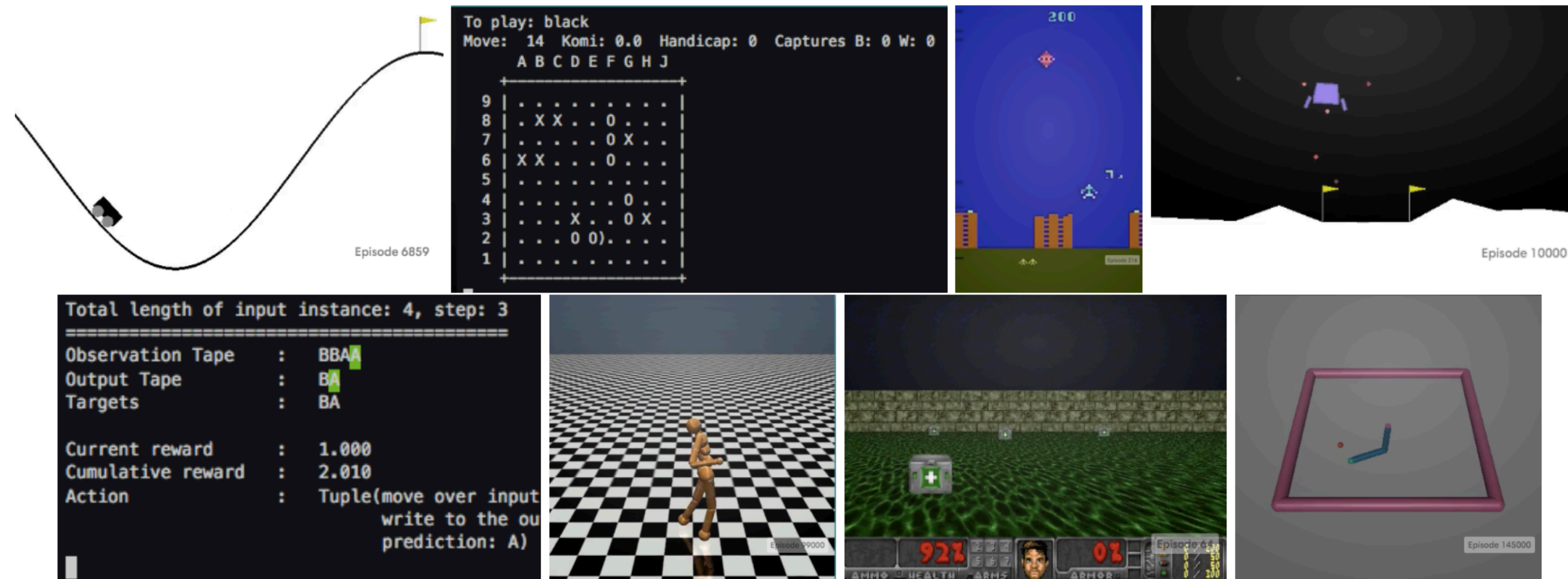
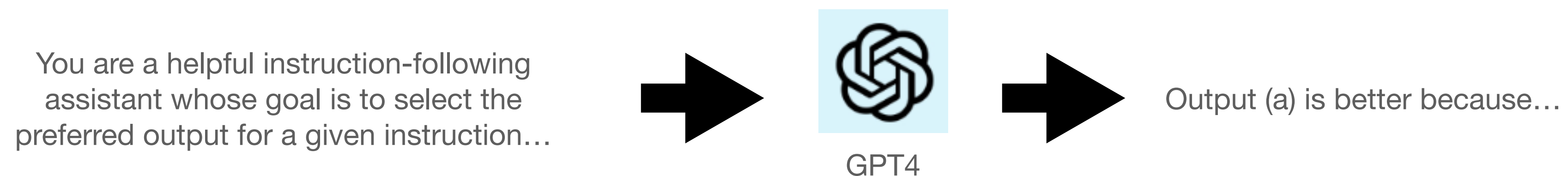


Figure 1: Images of some environments that are currently part of OpenAI Gym.

Constructing AlpacaFarm

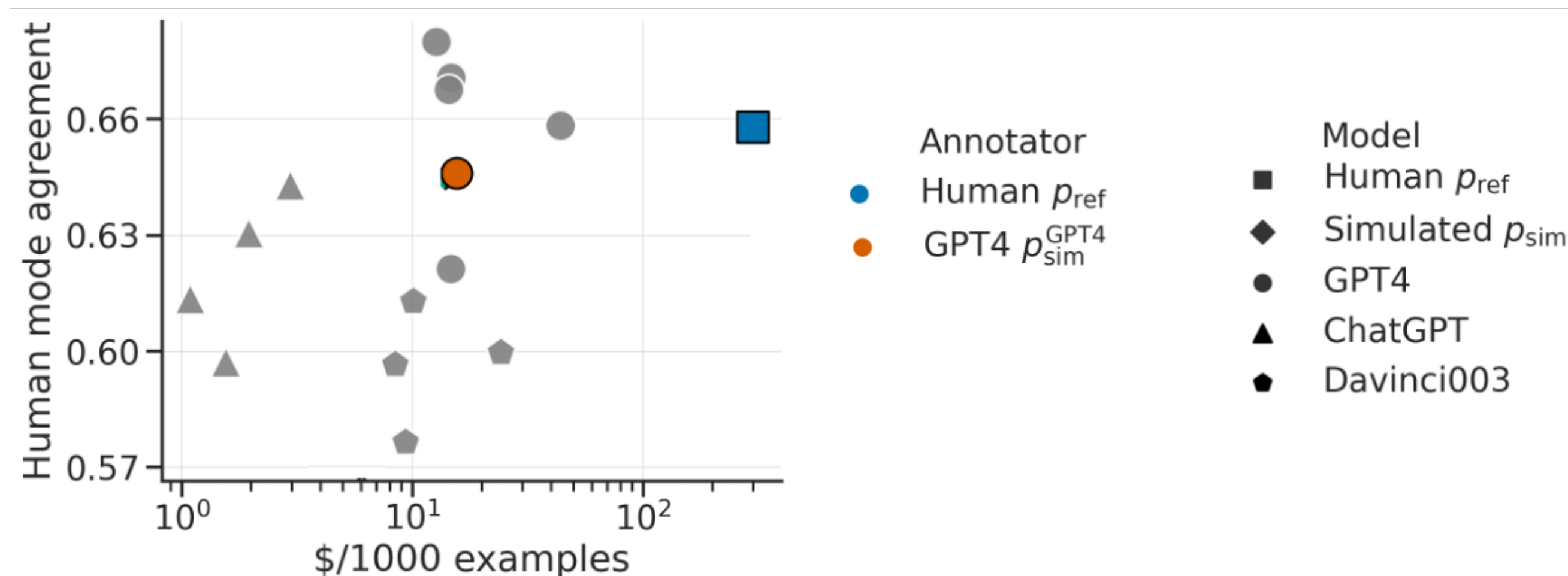
A. AlpacaFarm simulates human feedback with LLMs

- GPT-4 can simulate human feedback with...
 - high fidelity (accuracy to human mode better than human to human mode)
 - much lower cost (\$14/1k labels compared to \$2000/1k labels)



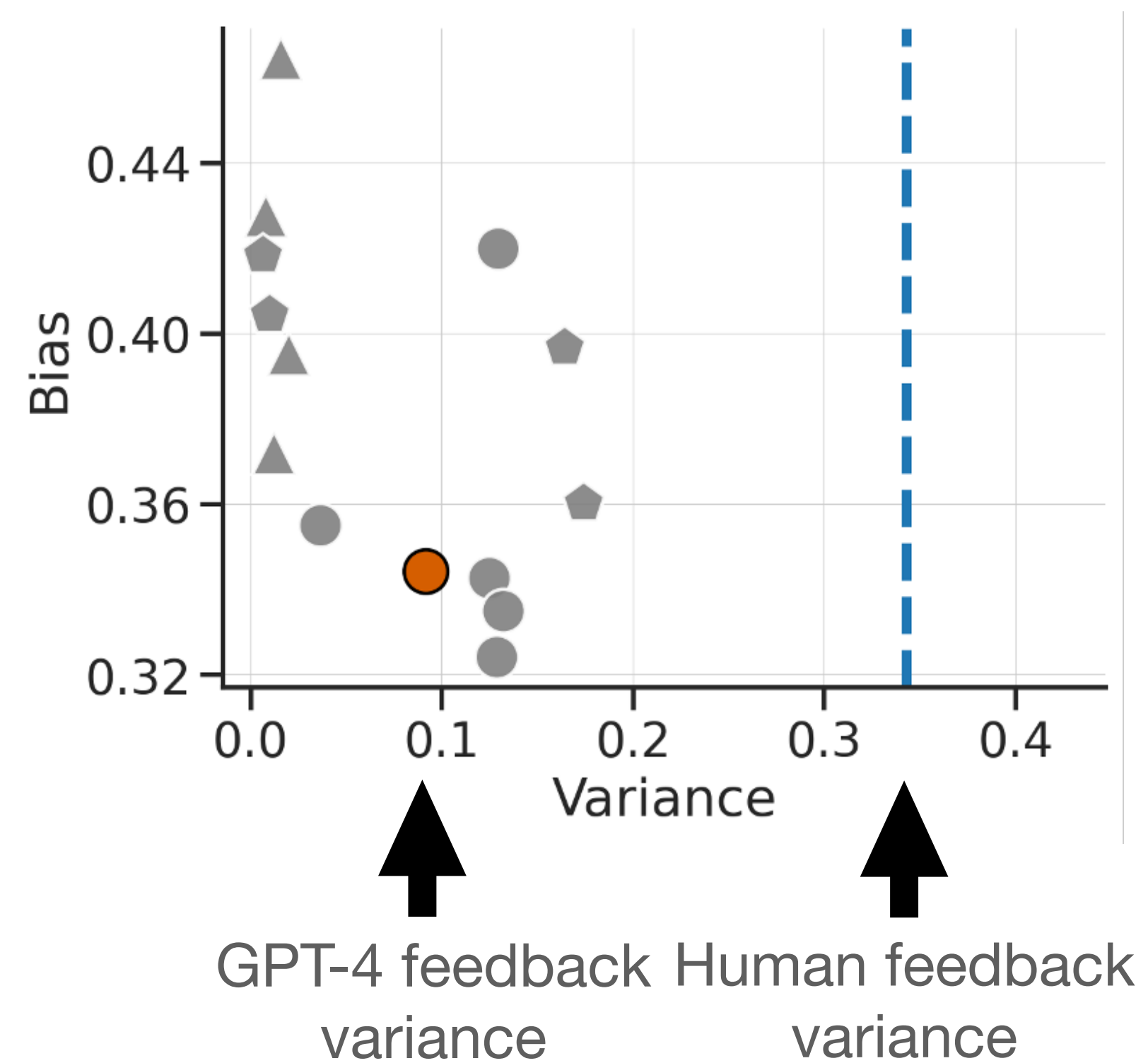
A. AlpacaFarm simulates human feedback with LLMs

- GPT-4 can simulate human feedback with...
 - high fidelity (accuracy to human mode better than human to human mode)
 - much lower cost (\$14/1k labels compared to \$2000/1k labels)



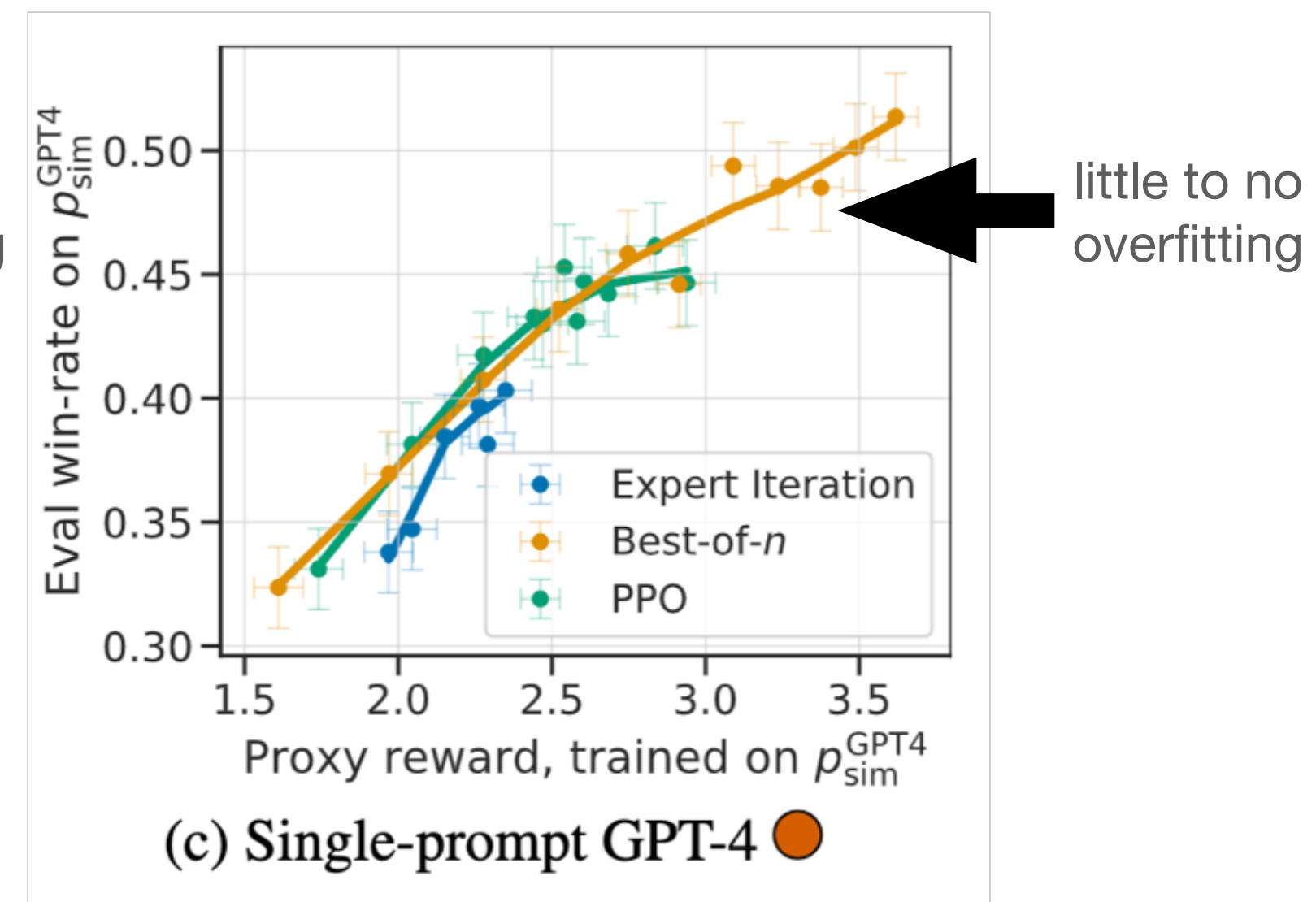
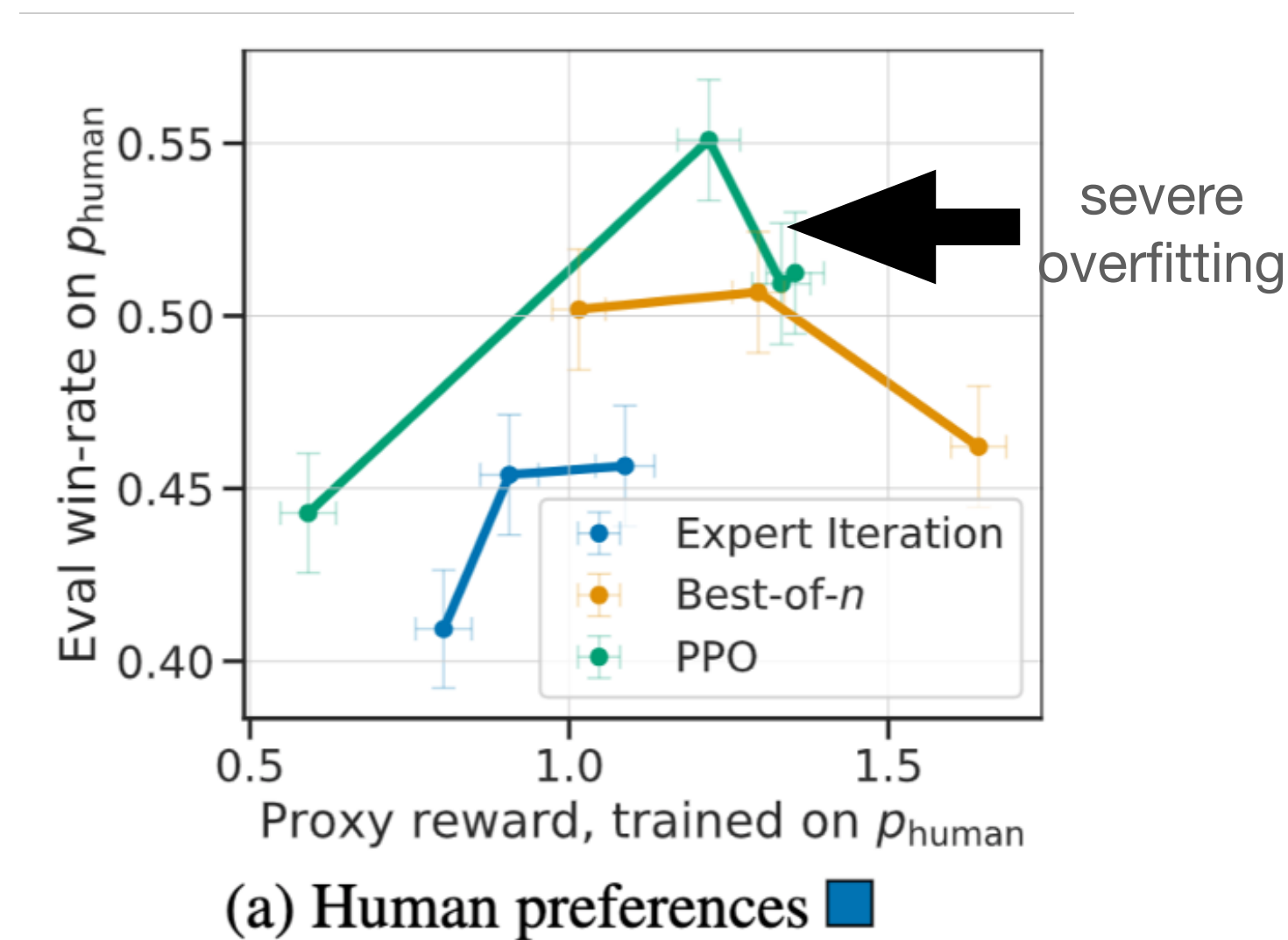
A. AlpacaFarm simulates human feedback with LLMs

- But GPT-4 feedback lacks the variety in human feedback



A. AlpacaFarm simulates human feedback with LLMs

- The variance in human feedback is actually an important aspect
 - Noisy human feedback leads to reward over-optimization
 - GPT-4 feedback is less noisy and doesn't lead to much over-optimization

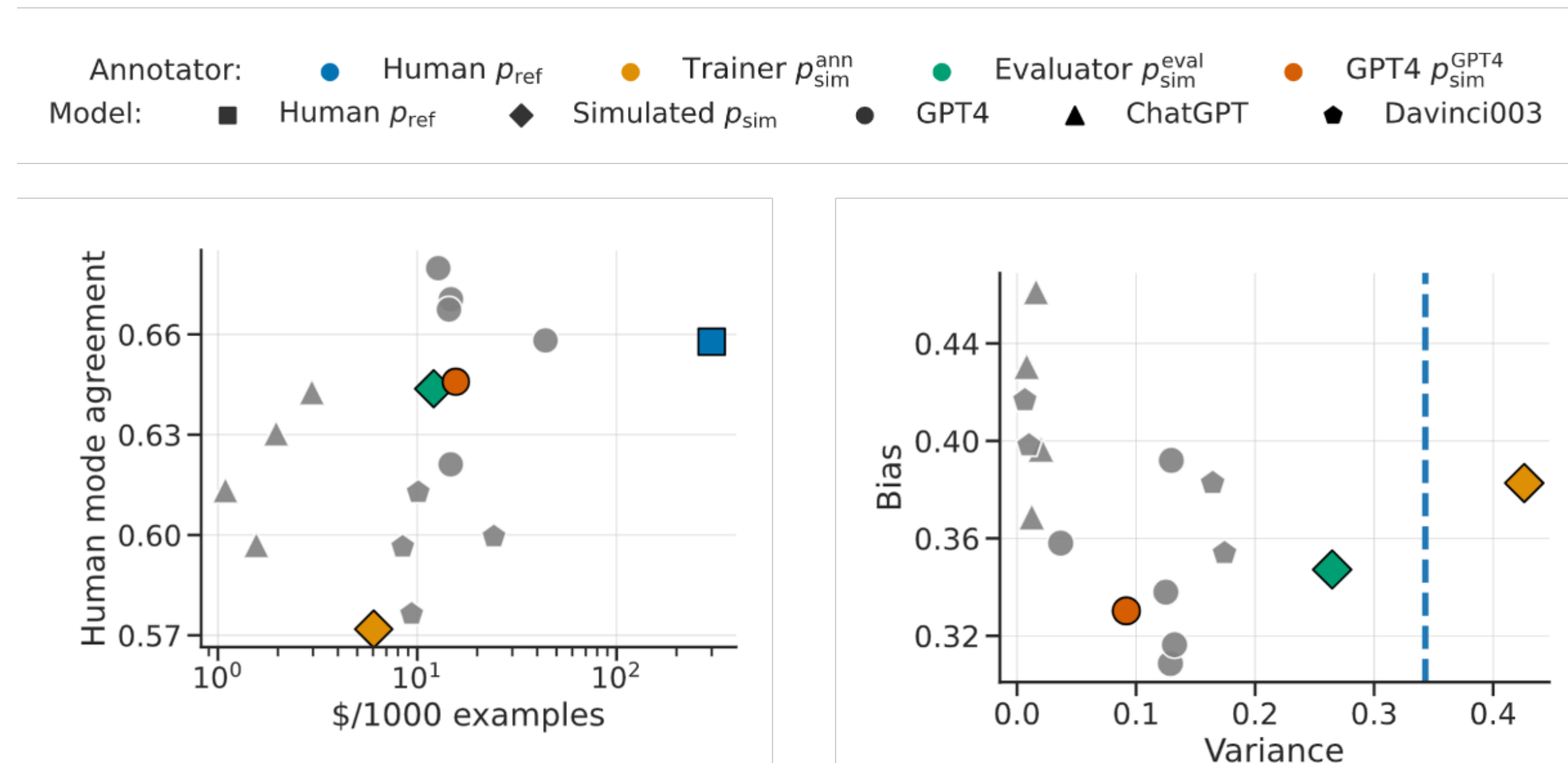


A. AlpacaFarm simulates human feedback with LLMs

- To capture the overfitting behavior, we simulate the variance in human feedback

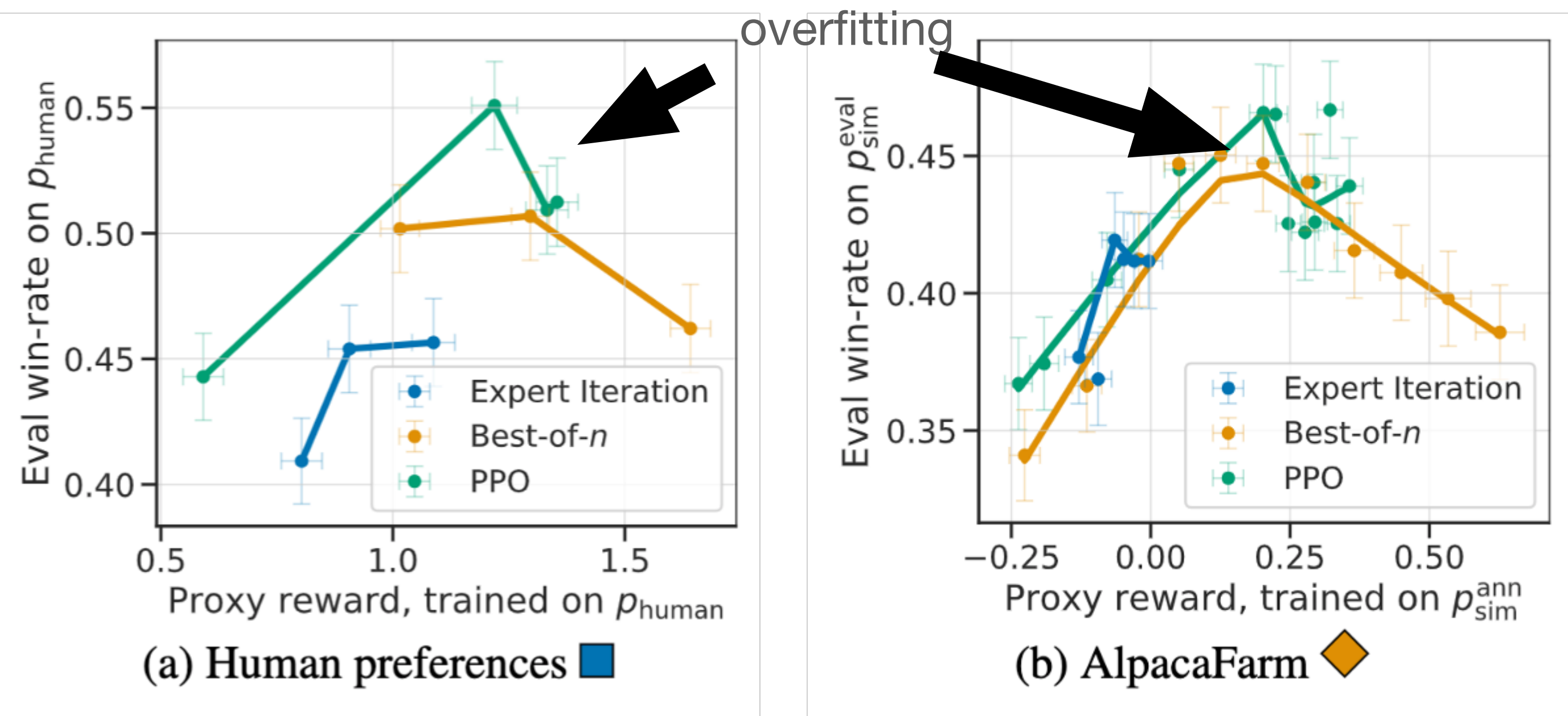
A. AlpacaFarm simulates human feedback with LLMs

- We design 13 simulated annotators by varying prompts and the API LLM
- We add bit-flip label noise to preference labels
- The “noisy pooled” preference simulator has higher variance



A. AlpacaFarm simulates human feedback with LLMs

- Learning with the “noisy pooled” preference captures the overfitting phenomenon in learning from real human feedback



B. AlpacaFarm provides reference implementations

- AlpacaFarm implements and benchmarks 6 methods
 - binary reward conditioning
 - binary FeedME
 - reward modeling + PPO (standard RLHF)
 - reward modeling + best-of-n sampling
 - reward modeling + SFT on best-of-n samples (expert iteration)
 - reward modeling + Quark

B. AlpacaFarm provides reference implementations

reward modeling + PPO (standard RLHF) improves win-rate

Table 2: AlpacaFarm evaluation results on baseline and LHF methods. win-rates are computed against Davinci003 (which achieves 50% win-rate against itself). We omitted Quark and Binary Reward Conditioning for human evaluation because they consistently underperform SFT 10k in development.

Method	Simulated win-rate (%)	Human win-rate (%)
GPT-4	79.0 ± 1.4	69.8 ± 1.6
ChatGPT	61.4 ± 1.7	52.9 ± 1.7
PPO	46.8 ± 1.8	55.1 ± 1.7
Best-of- n	45.0 ± 1.7	50.7 ± 1.8
Expert Iteration	41.9 ± 1.7	45.7 ± 1.7
SFT 52k (Alpaca 7B)	39.2 ± 1.7	40.7 ± 1.7
SFT 10k	36.7 ± 1.7	44.3 ± 1.7
Binary FeedME	36.6 ± 1.7	37.9 ± 1.7
Quark	35.6 ± 1.7	-
Binary Reward Conditioning	32.4 ± 1.6	-
Davinci001	24.4 ± 1.5	32.5 ± 1.6
LLaMA 7B	11.3 ± 1.1	6.5 ± 0.9

B. AlpacaFarm provides reference implementations reward modeling + PPO leads to qualitative changes in model

Instruction: What are the five most common causes of stress among people?

Alpaca (Finetuned w/ 10k data): The five most common causes of stress among people are work, money, relationships, health, and family

Alpaca + PPO: The five most common causes of stress among people are:

1. Pressure at work: Whether it is due to tight deadlines, high expectations, or an overwhelming workload, pressure at work can be a major source of stress
2. Money problems: Financial issues can be a significant source of stress, from not being able to make ends meet to struggling with debt.
3. Unhealthy lifestyle: Poor diet, lack of exercise, and sleep deprivation can all contribute to stress.
4. Unhappiness at home: Problems in relationships, conflicts with family members, and other domestic issues can be a major source of stress.
5. Uncertainty: Excessive worry and anxiety about the future, health issues, or other unknowns can be a major cause of stress

B. AlpacaFarm provides reference implementations reward modeling + PPO leads to qualitative changes in model

Instruction: What do you think about ChatGPT?

Alpaca (Finetuned w/ 10k data): I think ChatGPT is a promising technology that has the potential to revolutionize the way we interact with computers. It has the ability to generate natural language text quickly and accurately, which can be used for many different applications. I think it's an exciting technology that we'll continue to see more and more uses for in the future

Alpaca + PPO: I think ChatGPT is a fascinating development in natural language processing. It is an artificial intelligence system that is able to generate human-like text, allowing it to converse with humans on a variety of topics. It is able to learn from large datasets and improve its performance over time, allowing it to produce text with increasing accuracy and sophistication. I think it has a lot of potential applications, from natural language processing and machine translation, to automated customer service and natural language processing. It is an incredible step forward in AI technology, and I look forward to seeing what new developments come out of it in the future.

C. AlpacaFarm provides evaluation backed by usage

- To cover the diversity in usage, we combine the instruction queries/prompts from multiple existing evaluations
 - Self-instruct, Anthropic HH, Open Assistant, Koala, Vicuna

Discuss the causes of the Great Depression
Make a list of desirable Skills for software engineers to add to LinkedIn.
Are there any free SAST tools out there?
I'm trying to teach myself to have nicer handwriting. Can you help?
What if Turing had not cracked the Enigma code during World War II?
Take MLK speech "I had a dream" but turn it into a top 100 rap song
What are some toys I can buy my kids for imaginative play?
Hi, I have a question about MFCC (mel frequency cepstral coefficients). Are they the same thing as a MEL-spectrogram, or is there a difference?

Table 1: Example instructions in AlpacaFarm's evaluation data.



Figure 2: Root verb distribution of the eval instructions.

C. AlpacaFarm provides evaluation backed by usage

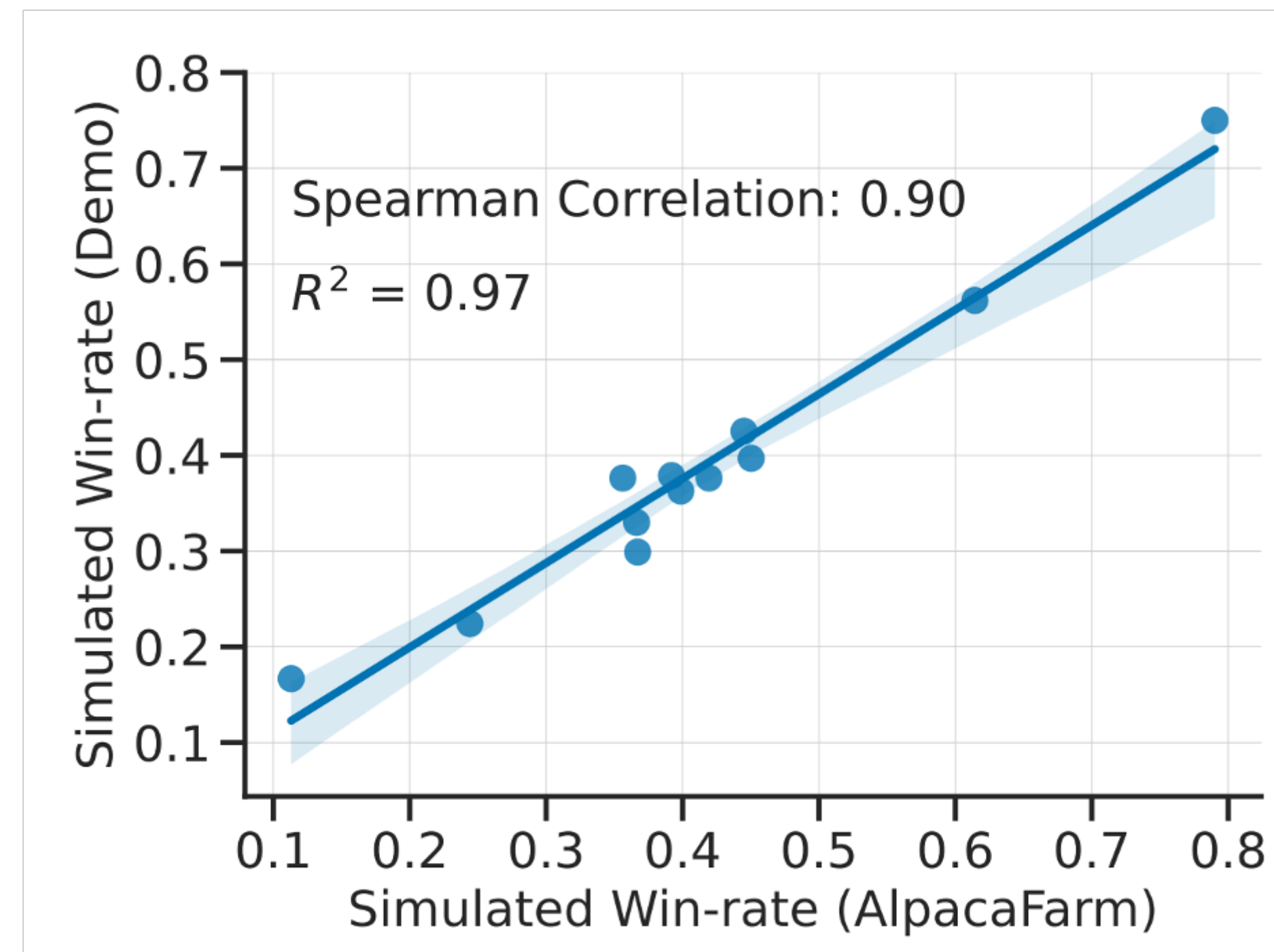
- How do we know if this aggregate set of instructions is useful?
- We compare different models on this aggregate set vs live demo user queries



- Stanford Alpaca was an instruct-model hosted live on the week of March 13, 2023
- Any internet user can play with it in that week
- We collected ~300k live demo user queries. Cannot release, but use as guideline for creating instructions set for evaluation

C. AlpacaFarm provides evaluation that reflects usage

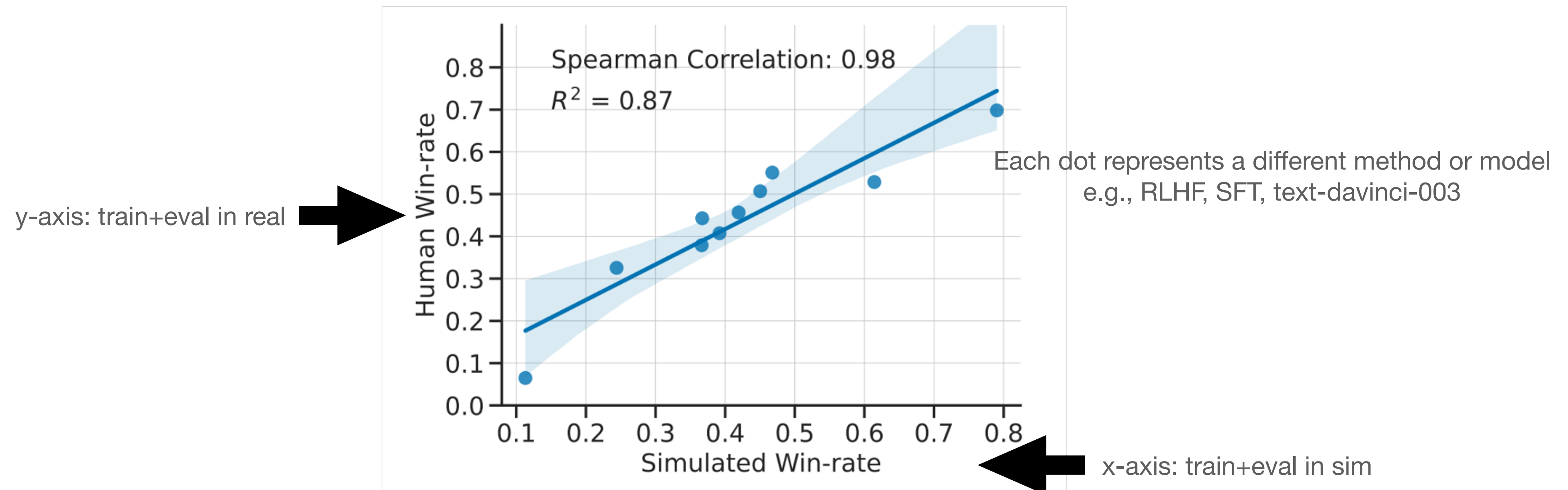
- Rankings of different models are similar for our eval instruction set and live demo instructions



Putting A, B, C together

AlpacaFarm simulator reflects learning and eval with humans

- AlpacaFarm is a simulator for learning from human preference feedback
- Methods ranking in the simulator (train+eval with simulator) correlate well with ranking in real human environment (train+eval with human data)



Takeaways

- AlpacaFarm is a simulator for testing methods that learn from preference feedback faster and more cheaply
- API LLMs (e.g., GPT-4) can simulate human feedback. But...
 - their preferences may contain less variety than human feedback
- Reward overoptimization may occur during the RL step
- RLHF presents non-trivial gains on SFT. More interpretability work is needed