CS 690: Human-Centric Machine Learning

Prof. Scott Niekum

Scalable oversight

Presentations next week

- 15 minutes split roughly evenly between group members
- Slightly short is fine, but don’t go long!
- Be prepared for questions
- Even if project failed, give us insight into what you learned

RLAIF

- Learning from constitutional AI preferences outperforms SFT
- But can it outperform RLHF from human preferences?
- And does it need a constitution?
- Are there other ways that LLMs can enable RL?

Lee, Harrison, et al. “RLAIF: Scaling reinforcement learning from human feedback with AI feedback." *arXiv preprint arXiv:2309.00267* (2023).

RLAIF

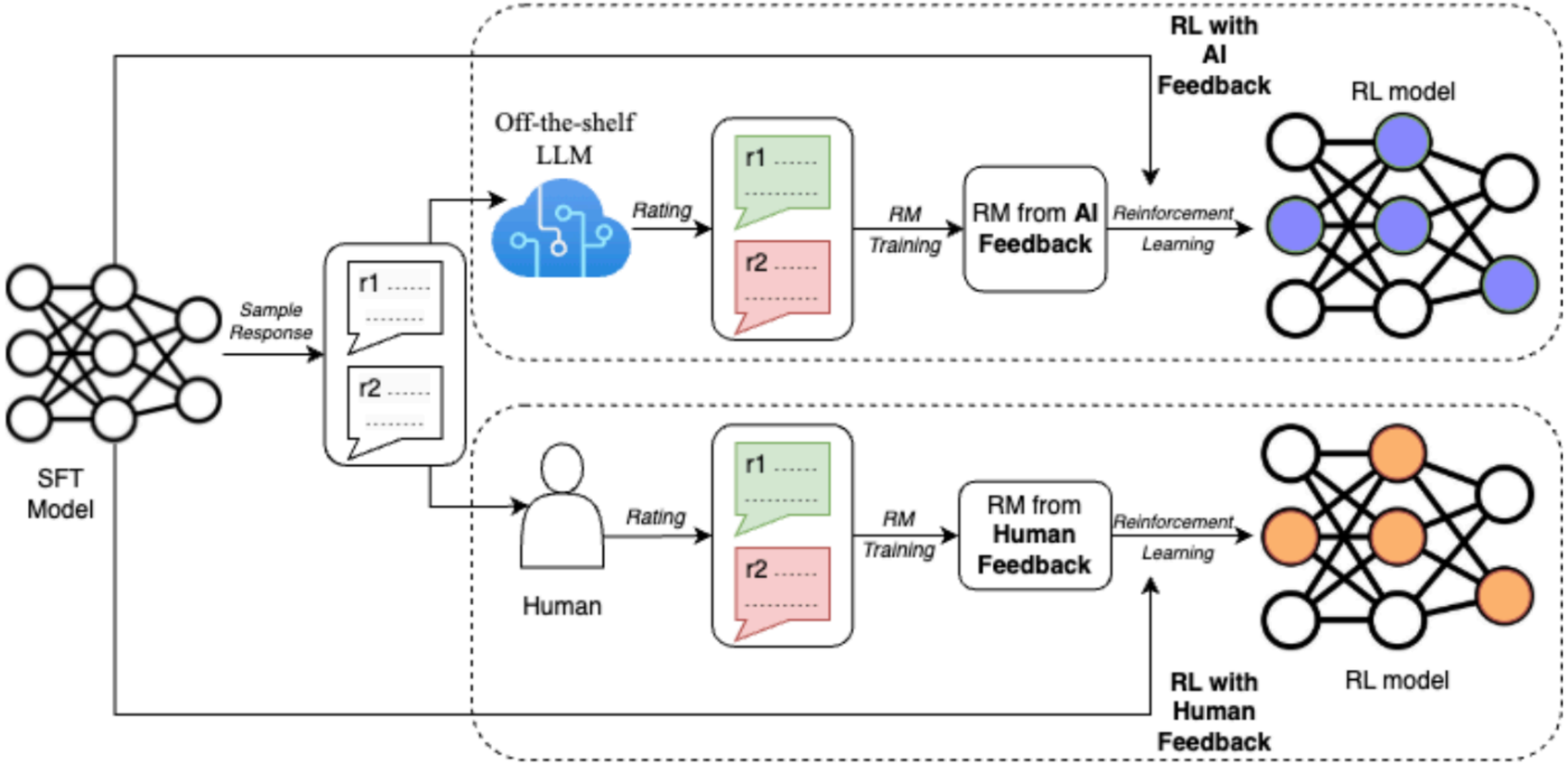


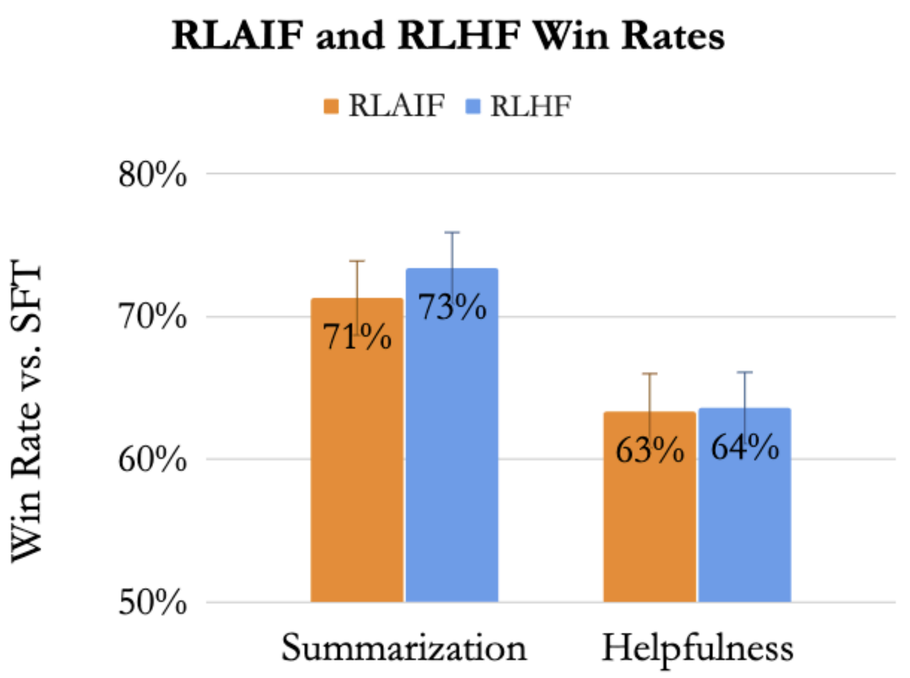
Figure 2: A diagram depicting RLAIF (top) vs. RLHF (bottom)

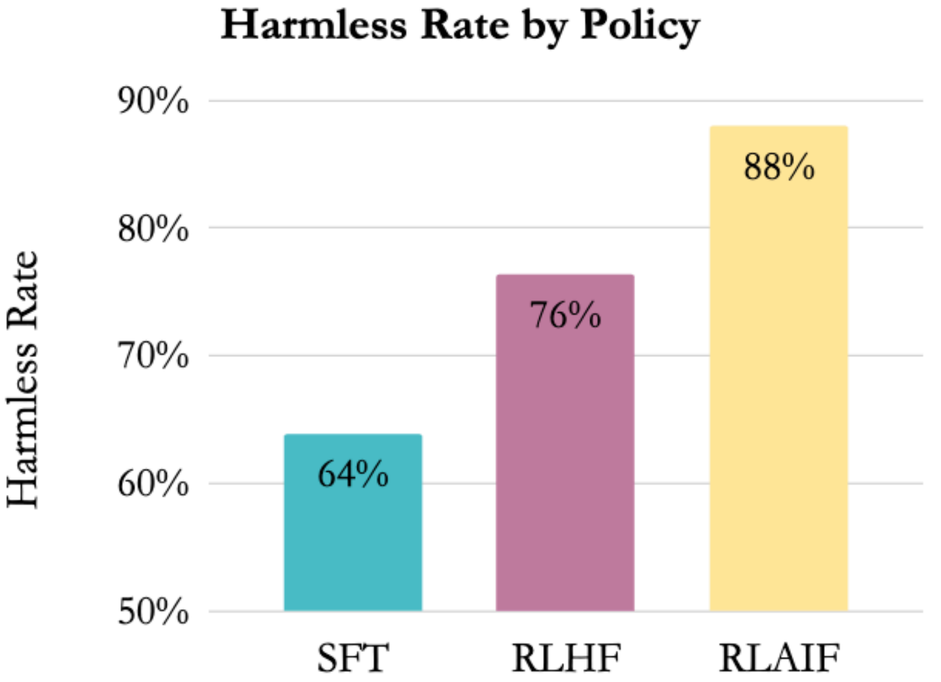
How to extract preferences?

- Prompt structure:
  - (1) Preamble instructions
  - (2) Few-shot exemplars (optional)
  - (3) Sample to annotate
  - (4) Ending (e.g. “Preferred response =“)
- Answer: logprobs of “1” and “2” tokens

- Positional bias: present both ways and average logits
- Length bias can be an issue as well

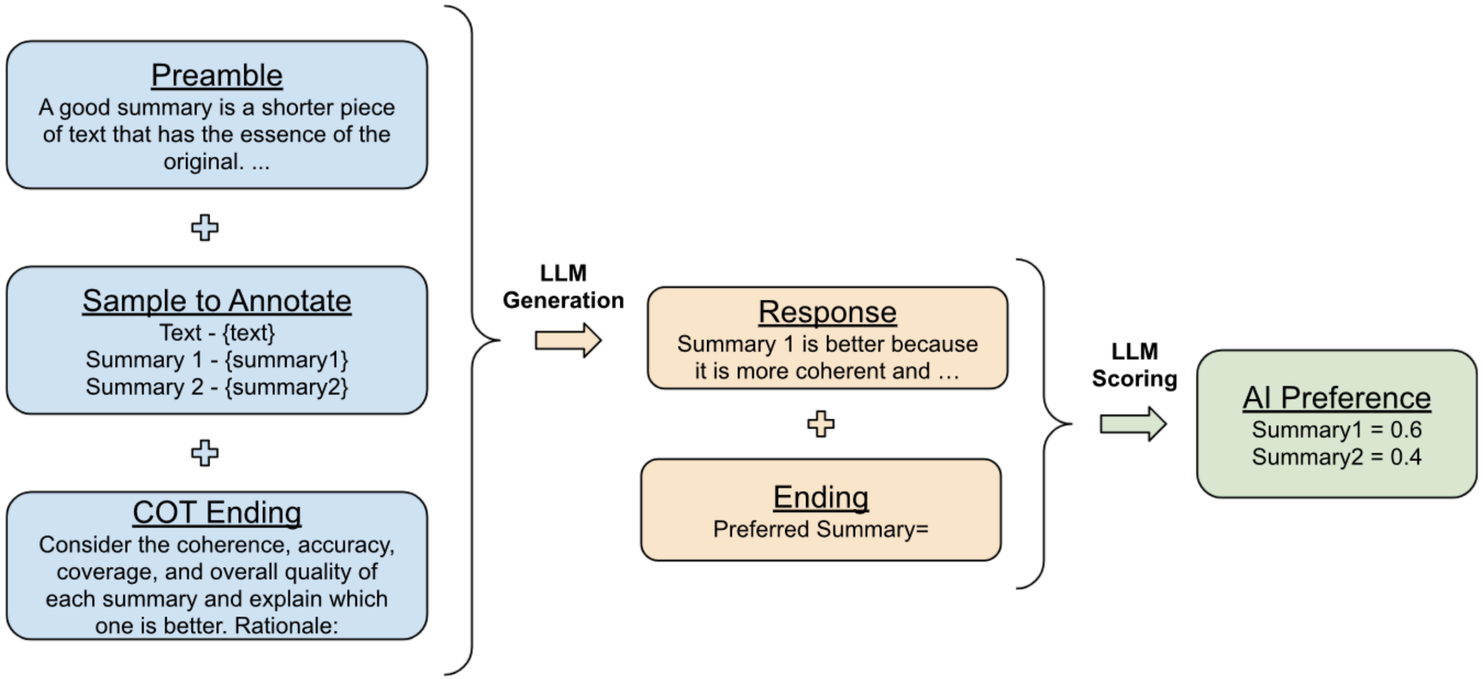
RLAIF


RLAIF and RLHF Win Rates
RLAIF
RLHF
Win Rate vs. SFT
80%
70%
60%
50%
71%
73%
63%
64%
Summarization
Helpfulness


Harmless Rate by Policy
Harmless Rate
90%
80%
70%
60%
50%
64%
76%
88%
SFT
RLHF
RLAIF

Variants

- Chain of thought — slightly helpful



- Averaging multiple chains of thought — worse!
- RLHF + RLAIF — no better than RLHF alone
- Few-shot exemplars — mixed results

Direct RLAIF

- Have LLM output a reward instead of a preference
- Prompt: “You are an expert summary rater. Given a TEXT and a SUMMARY, your role is to provide a SCORE from 1 to 10 that rates the quality of the SUMMARY given the TEXT, with 1 being awful and 10 being a perfect SUMMARY.”, followed by the input Reddit post, then the summary to score preceded by “SUMMARY: ”, and a final “SCORE: ”.

Direct RLAIF

Win Rate			Harmless Rate	
Comparison	**Summa -rization**	**Helpful dialogue**	**Model**	**Harmless dialogue**
RLAIF **vs** SFT	71%	63%	SFT	64%
RLHF **vs** SFT	73%	64%	RLHF	76%
RLAIF **vs** RLHF	50%	52%	RLAIF	88%
Same-size RLAIF **vs** SFT	68%			
Direct RLAIF **vs** SFT	74%			
Direct RLAIF **vs** Same-size RLAIF	60%			

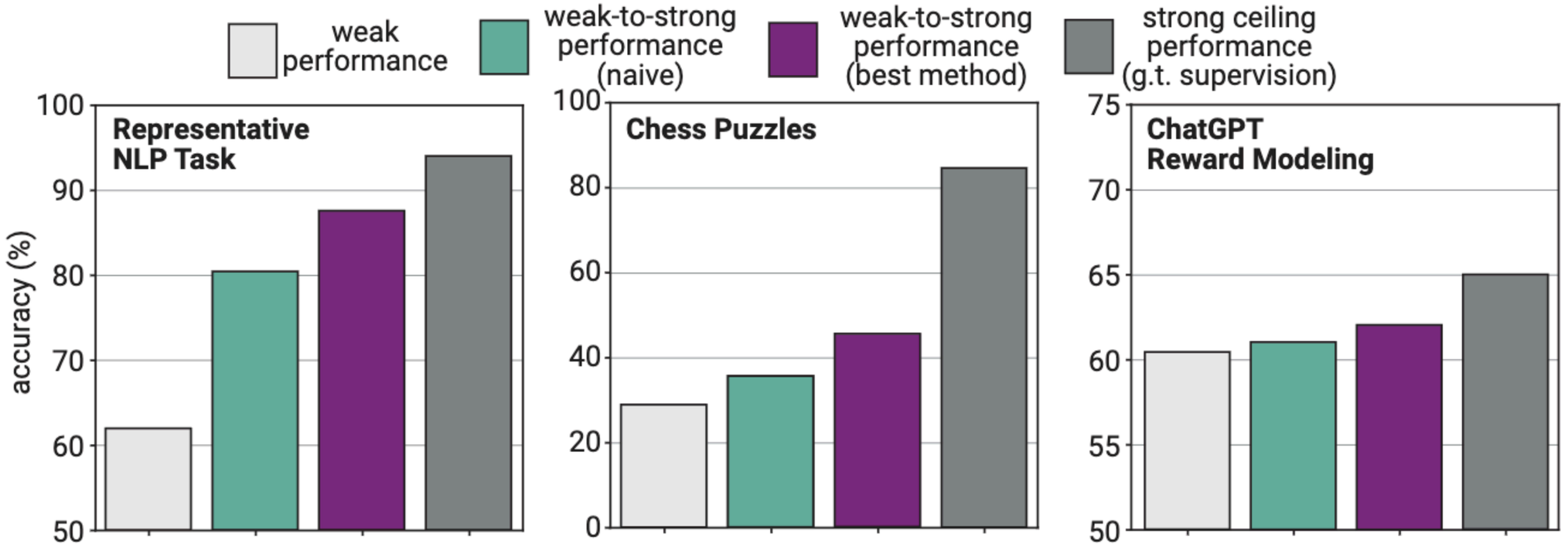
Table 1: **Left side:** Win rates when comparing generations from two different models for the summarization and the helpful dialogue tasks, judged by human evaluators. **Right side:** Harmless rates across policies for the harmless dialogue task, judged by human evaluators.

Weak-to-strong generalization

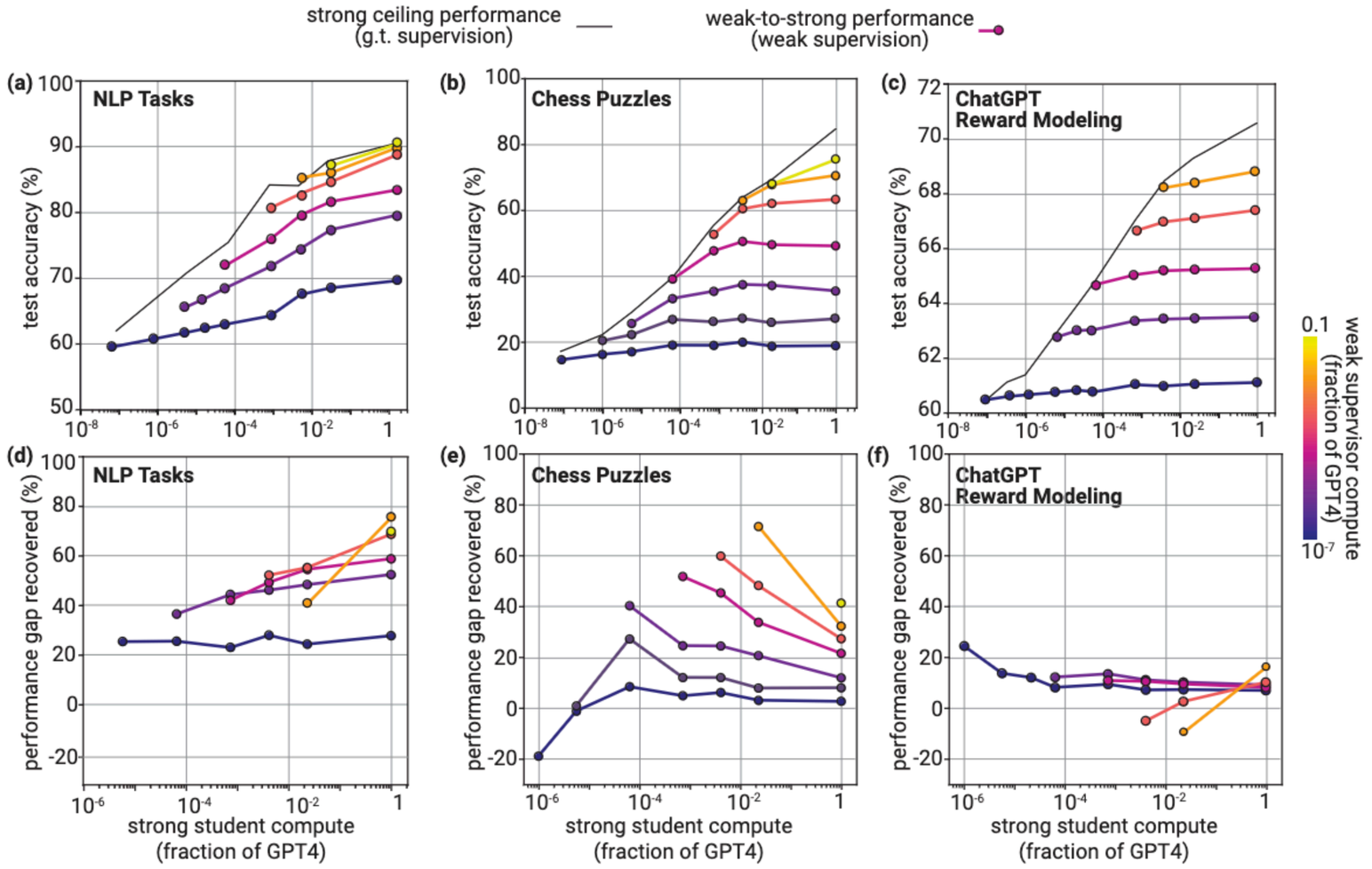
- Similar to scalable oversight, but asks the question: Can similar techniques work without humans at all?
- Strong model directly imitates weak model labels (not preferences in this work)
- Can be thought of as an analogy that lets us study what human supervision of superhuman models might look like
- Or as something more directly useful in settings that humans will not be able to reliably supervise, even with AI assistance
- Question: Why should weaker models be able to supervise stronger models effectively?

Burns, Collin, et al. "Weak-to-strong generalization: Eliciting strong capabilities with weak supervision." *arXiv preprint arXiv:2312.09390* (2023).

Weak-to-strong generalization



Weak-to-strong generalization



Weak-to-strong generalization

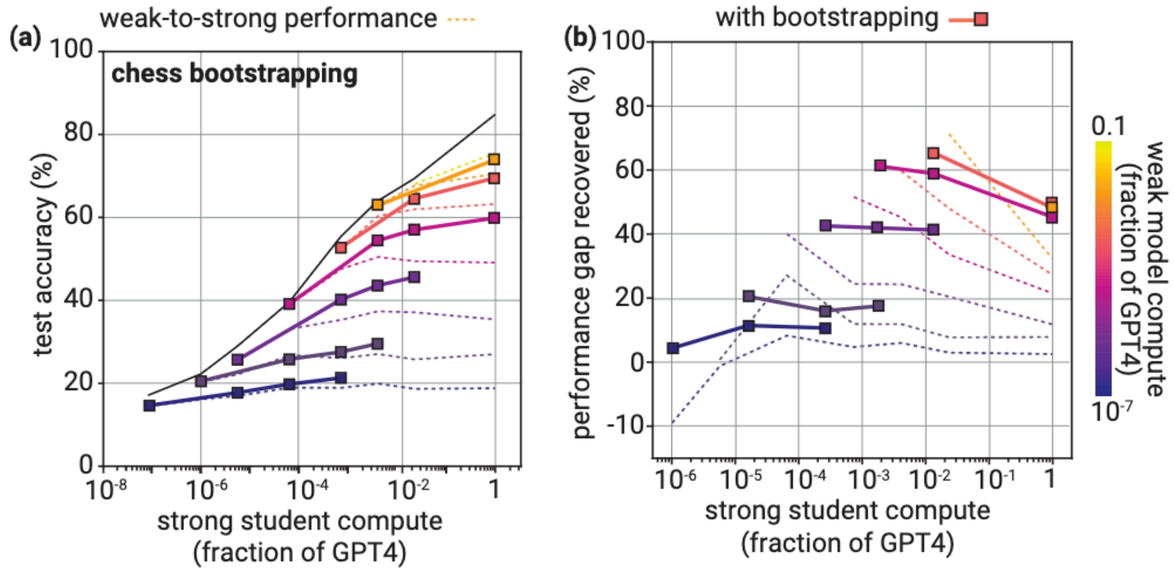


Figure 4: **Bootstrapping improves weak-to-strong generalization on chess puzzles.** **(a)** Test accuracy as a function of strong student size. Accuracy of students trained with ground truth in black, accuracy of students naively trained with weak supervision shown with dotted lines (hue indicates size of weak supervisor). Accuracies of students trained via bootstrapping shown with colored squares (including both the final weak-to-strong performance and the performance of the intermediate models during bootstrapping). **(b)** Same as **a** with PGR. By taking multiple small steps instead of one big step we see substantially improved generalization, especially for larger student models.

Limitations

Limitations. Our setup still has important disanalogies to the ultimate problem of aligning superhuman models. We view our setup as removing one of the main disanalogies in prior work, not as providing a final, perfectly analogous setup. Two remaining disanalogies include:

1. **Imitation saliency.** Future superhuman models will likely have salient representations of human behaviors, but our strong models may not have learned features relevant for imitating weak model predictions; simply imitating the weak supervisor may thus be an easier failure mode to avoid in our setting than it will be in the future. More generally, the types of errors weak models make today may be different from the types of errors humans will make when attempting to supervise superhuman models.
2. **Pretraining leakage.** Our pretraining data implicitly contains supervision from humans. It may thus be artificially easy to elicit strong models' capabilities in our setting, since they were directly pretrained to observe strong (human-level) performance. Superhuman-level performance may not be directly observed in the same way—superhuman knowledge might be more latent, e.g. because it was learned from self-supervised learning—and thus might be harder to elicit from superhuman models in the future.

Open questions

- Does weak-to-strong generalization work with preferences? Or other types of supervision such as process critique?
- Are there fundamental limits to weak-to-strong generalization (e.g. if chain gets long enough, when does alignment get lost?)
- Why is the success of each approach so sensitive to domain?
- How can less naive supervision be performed?

Course review

- Behavior cloning
- RL
- Reward specification
- Interactive RL
- IRL / Bayesian IRL
- Adversarial imitation learning
- RLHF: Reward-based, reward-free, and fine-grained
- Models of human preferences and rationality
- Performance guarantees
- Cooperation and corrigibility
- Multimodal signals and natural language for reward inference / design
- Scalable oversight