Scalable Reward Learning

Peter Hase
Postdoctoral Researcher, Stanford University
AI Institute Fellow, Schmidt Sciences

Scalable Reward Learning

- Original problem formulation
- Current frameworks
- Empirical work

Original problem formulation

Concrete Problems in AI Safety

Dario Amodei* Google Brain

Chris Olah* Google Brain

Jacob Steinhardt Stanford University

Paul Christiano UC Berkeley

John Schulman OpenAI

Dan Mané Google Brain

Scalable Oversight

5 Scalable Oversight

Consider an autonomous agent performing some complex task, such as cleaning an office in the case of our recurring robot example. We may want the agent to maximize a complex objective like "if the user spent a few hours looking at the result in detail, how happy would they be with the agent's performance?" But we don't have enough time to provide such oversight for every training example; in order to actually train the agent, we need to rely on cheaper approximations, like "does the user seem happy when they see the office?" or "is there any visible dirt on the floor?" These cheaper signals can be efficiently evaluated during training, but they don't perfectly track what we care about. This divergence exacerbates problems like unintended side effects (which may be appropriately penalized by the complex objective but omitted from the cheap approximation) and reward hacking (which thorough oversight might recognize as undesirable). We may be able to

Current Frameworks

Combining weak-to-strong generalization with scalable oversight

A high-level view on how this new approach fits into our alignment plans

JAN LEIKE

DEC 20, 2023

28 6 2 Share

Current Frameworks

Scalable Oversight - improve our evaluation ability

W2S Generalization - just generalize correctly

Empirical Work

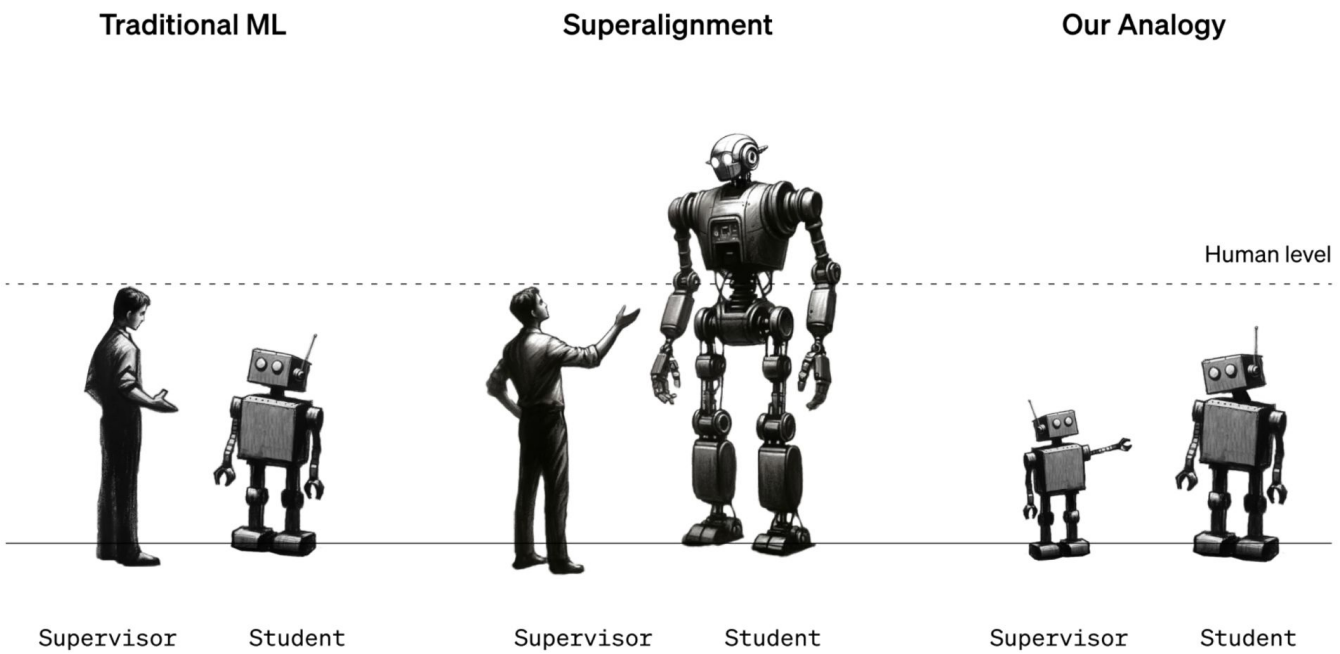
Weak-to-Strong Generalization

- Burns et al. (2023) — W2SG
- Hase et al. (2024) — Easy-to-hard
- Medvedev et al. (2025) — theory

Scalable Oversight

- Irving et al. (2018) — Debate (proposal)
- Bowman et al. (2022) — Sandwiching
- Michael et al. (2023) + Khan et al. (2024) — Debate empirics

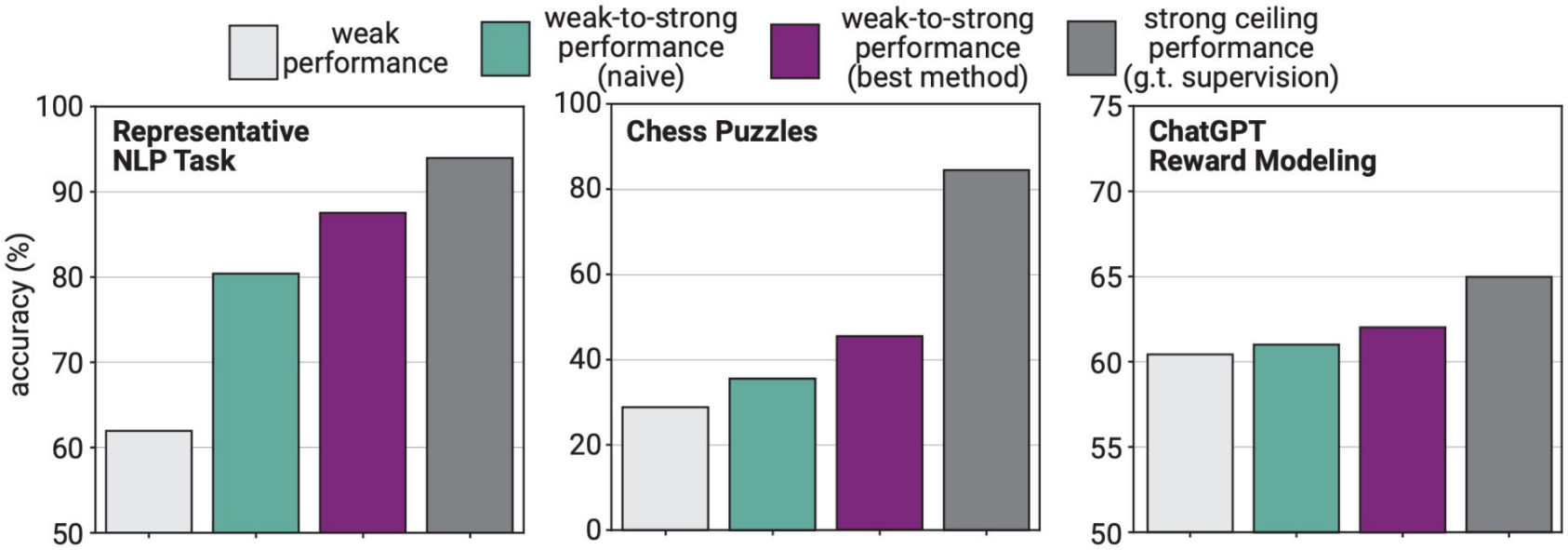
Burns et al. (2023) — W2SG



Burns et al. (2023) — W2SG

Setup - finetune strong student (GPT-4 family) on labels from a weak supervisor; 22 NLP tasks, chess puzzles, ChatGPT RM

Burns et al. (2023) — W2SG



Burns et al. (2023) — W2SG

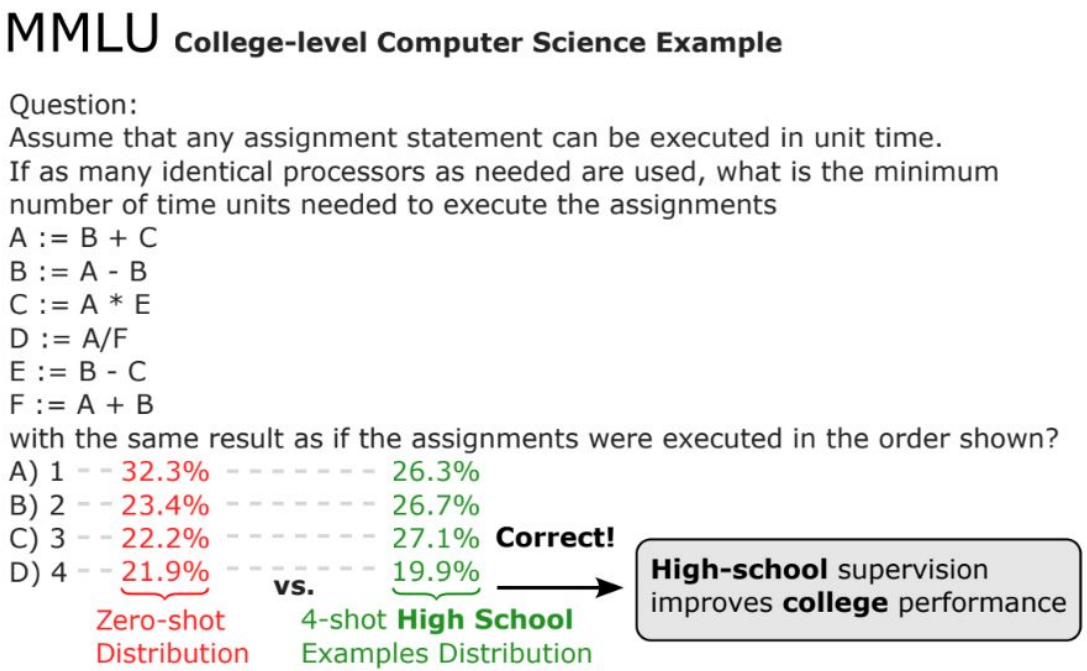
$$\text{PGR} = \frac{\text{weak-to-strong} - \text{weak}}{\text{strong ceiling} - \text{weak}} = \frac{\text{——}}{\text{·····}}$$

Burns et al. (2023) — Results

NLP / chess / RM

- NLP: “the confidence loss increases median PGR from about 25% to nearly 80%”
- chess PGR “above 40%” for small gaps;
- RM “only able to recover roughly 10% of the performance gap”

Hase et al. (2024) — Easy-to-Hard



Hase et al. (2024) — Easy-to-Hard

Q: John hires a driving service to get him to work each day.
His work is 30 miles away and he has to go there and back each day.
He goes to work 5 days a week for 50 weeks a year.
He gets charged $2 per mile driven and he also gives his driver a $150 bonus per month
How much does he pay a year for driving?
A: John goes to work 5 days a week for 50 weeks a year.
John goes to work 5 x 50 = <<5*50=250>>250 times a year.
John pays $2 per mile driven.
John pays $2 x 30 x 2 = <<2*30*2=120>>120 for each trip.
John pays $120 x 250 = <<120*250=30000>>30000 for each year.
John pays $150 bonus per month.
John pays $150 x 12 = <<150*12=1800>>1800 for each year.
John pays $30000 + $1800 = <<30000+1800=31800>>31800 for each year.
So the answer is 31800

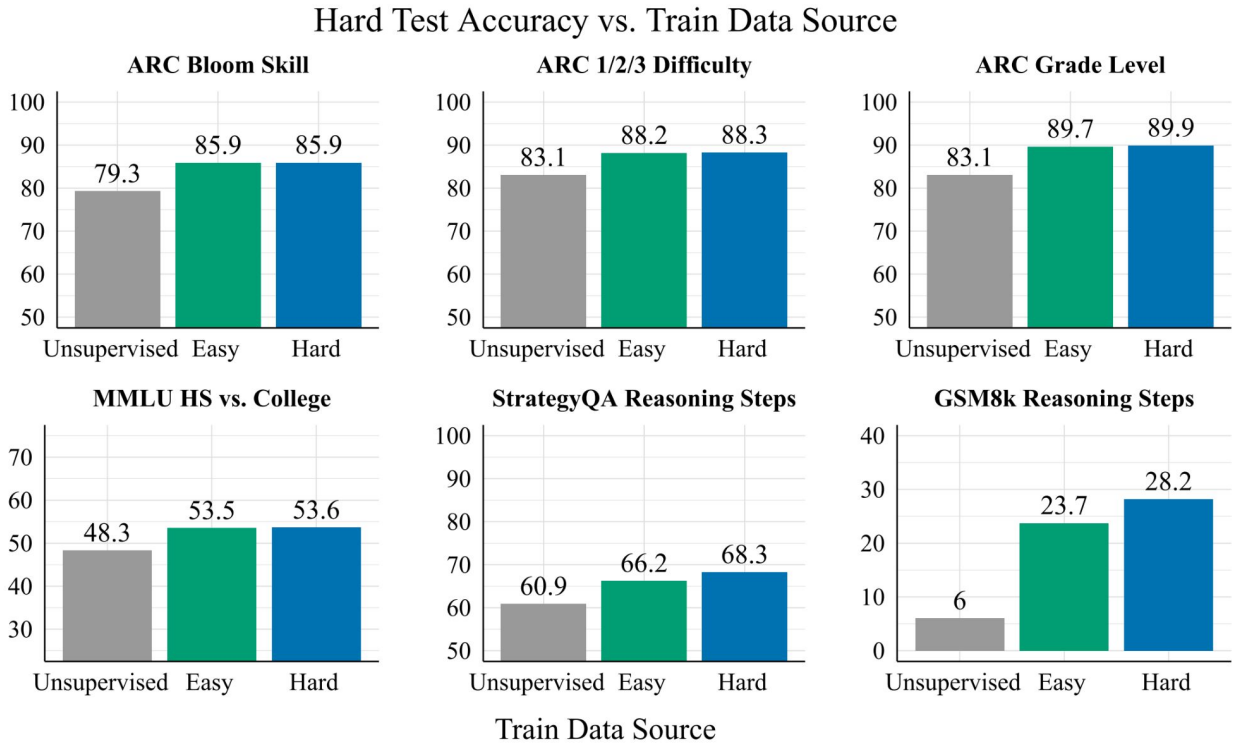
Correct on hard problem given only easy data in prompt

Hase et al. (2024) — Easy-to-Hard

Setup

- 70b Llama 2
- finetune/ICL/probe on easy data only, evaluate on hard;
- seven hardness measures (six human, one model-based);
- QA datasets from 3rd-grade science to college STEM

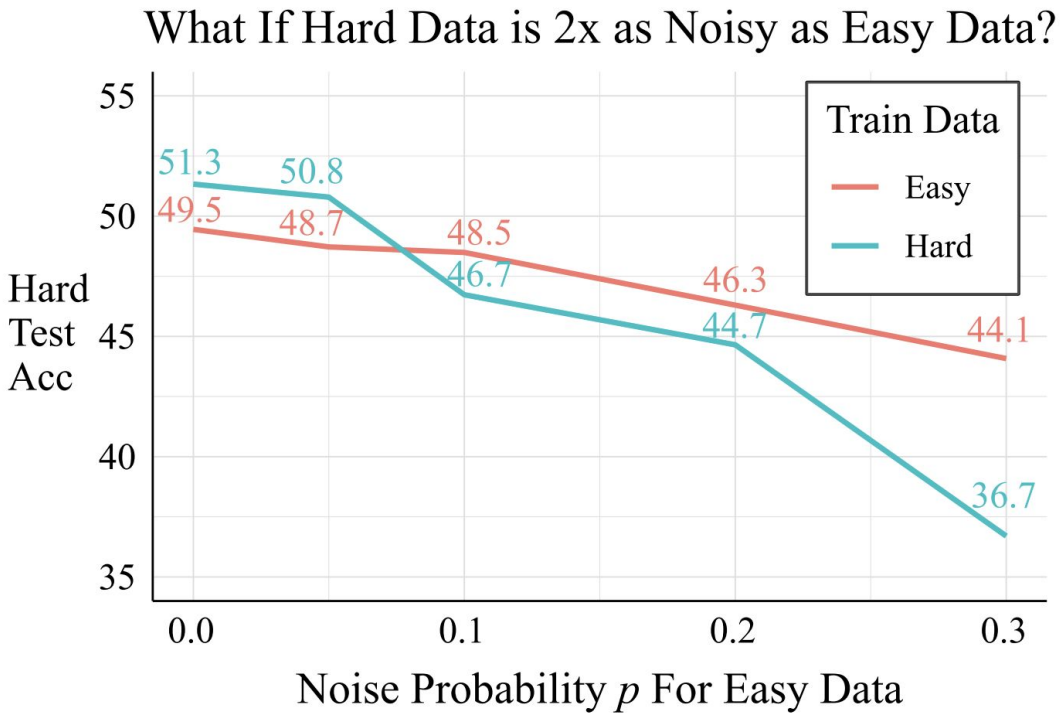
Hase et al. (2024) — Results



Hase et al. (2024) — Results

$$\frac{\text{Easy} - \text{Unsupervised}}{\text{Hard} - \text{Unsupervised}}$$

Hase et al. (2024) — Results



Hase et al. (2024) — Results

Hase et al. (2024) — Results

Takeaways

- SGR is usually between 70% and 100%
- plausibly, collecting easy data better than collecting hard data
- SGR declines as you fix ceiling and weaken the teacher

Medvedev et al. (2025) — Theory

Weak-to-Strong Generalization Even in Random Feature Networks, Provably

Marko Medvedev[1,*,†], Kaifeng Lyu[2,*], Dingli Yu[3], Sanjeev Arora[4], Zhiyuan Li[5], Nathan Srebro[5]

[1]University of Chicago [2]Simons Institute, UC Berkeley [3]Microsoft Research
[4]Princeton University [5]Toyota Technological Institute at Chicago

Medvedev et al. (2025) — Main Idea

Setup

- Two layer networks
- Random first layer
- Stronger student has wide 2nd layer
- Weaker teacher has narrow 2nd layer
- Train student on teacher labels **with early stopping**

Intuition

- First layer is a random up-projection
- Wide student layer has good inductive bias, learns real features before noise

Medvedev. (2025) — Results

If teacher loss is 0...

- PGR → 1

If teacher loss is non-zero...

- Student *can outperform* teacher
- But PGR < 1

Medvedev. (2025) — Results

W2S not only about knowledge elicitation, also about inductive bias

IMPORTANT

- Early stopping *very important* in W2S
- Early stopping *does not matter* in E2H (with right reg.)

Active directions

Pretty active area of research, a lot of recent work on this

- Weak supervision on reasoning
- Exploiting preference learning for extrapolation
- Completely unsupervised approaches

Scalable Oversight

Now we’re talking about humans figuring out the right answer

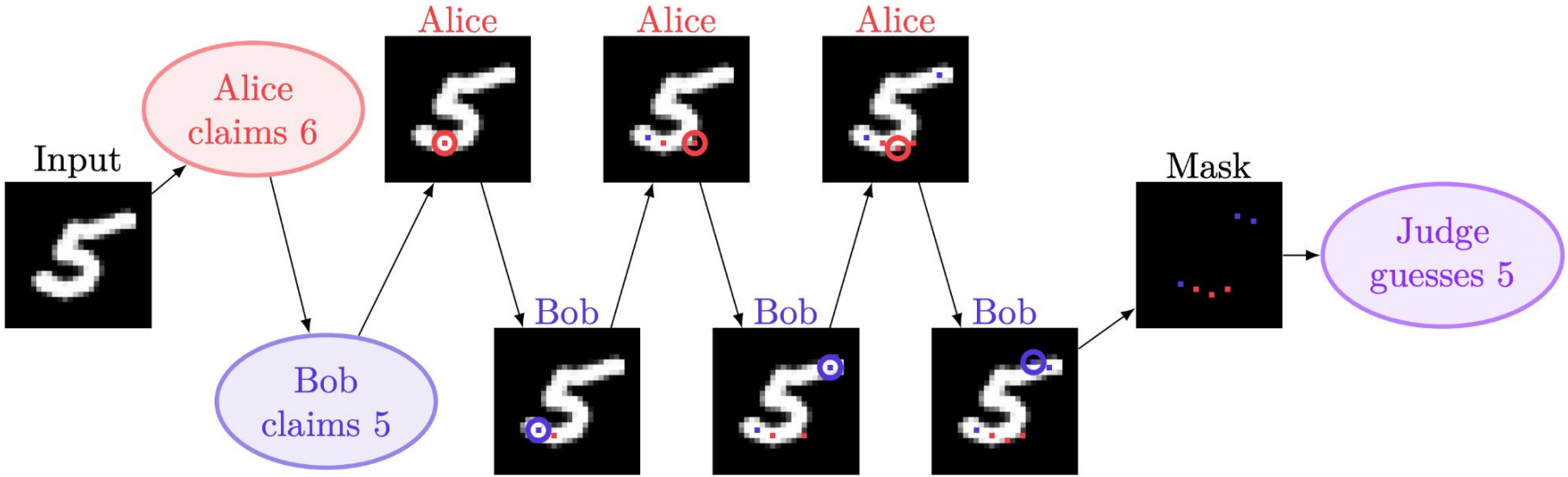
Irving et al. (2018) — Debate

AI safety via debate

Geoffrey Irving* Paul Christiano Dario Amodei

OpenAI

Irving et al. (2018) — Debate



Irving et al. (2018) — Debate

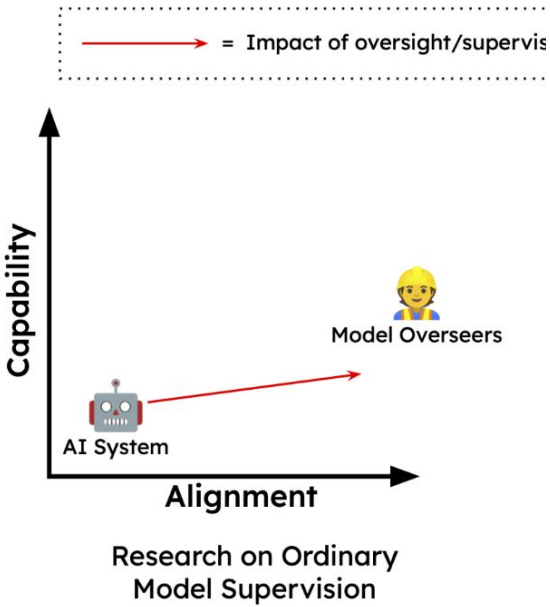
Central hope:
“In the debate game, it is harder to lie than to refute a lie”

Irving et al. (2018) — Results

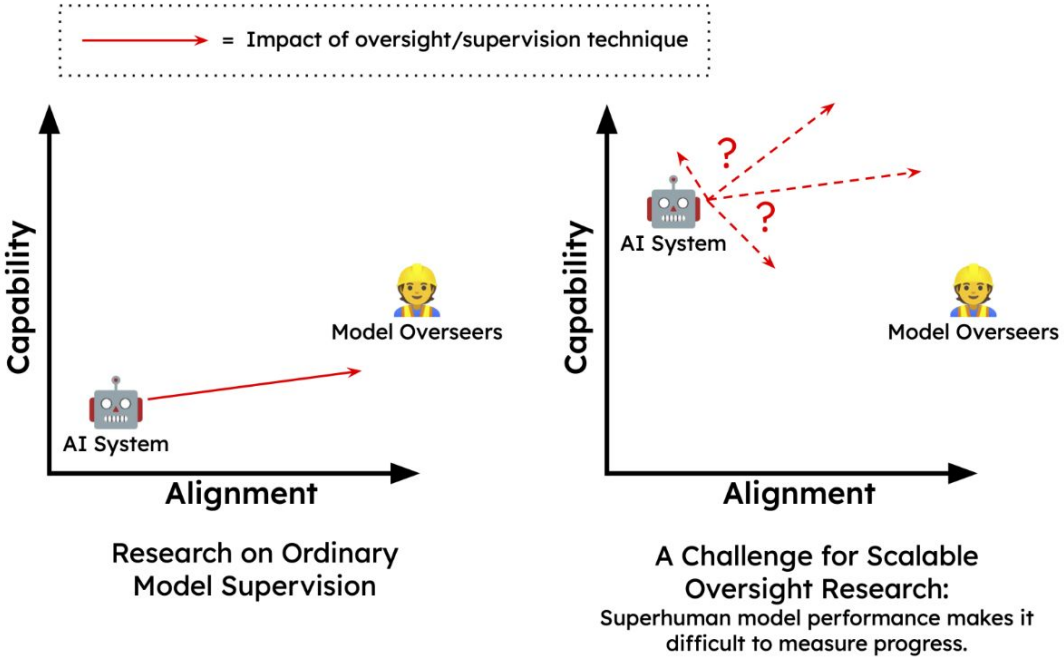
MNIST result - “boosting the classifier’s accuracy from 59.4% to 88.9% given 6 pixels and from 48.2% to 85.2% given 4 pixels”

Complexity - “debate with optimal play can answer any question in PSPACE given polynomial time judges (direct judging answers only NP questions)”

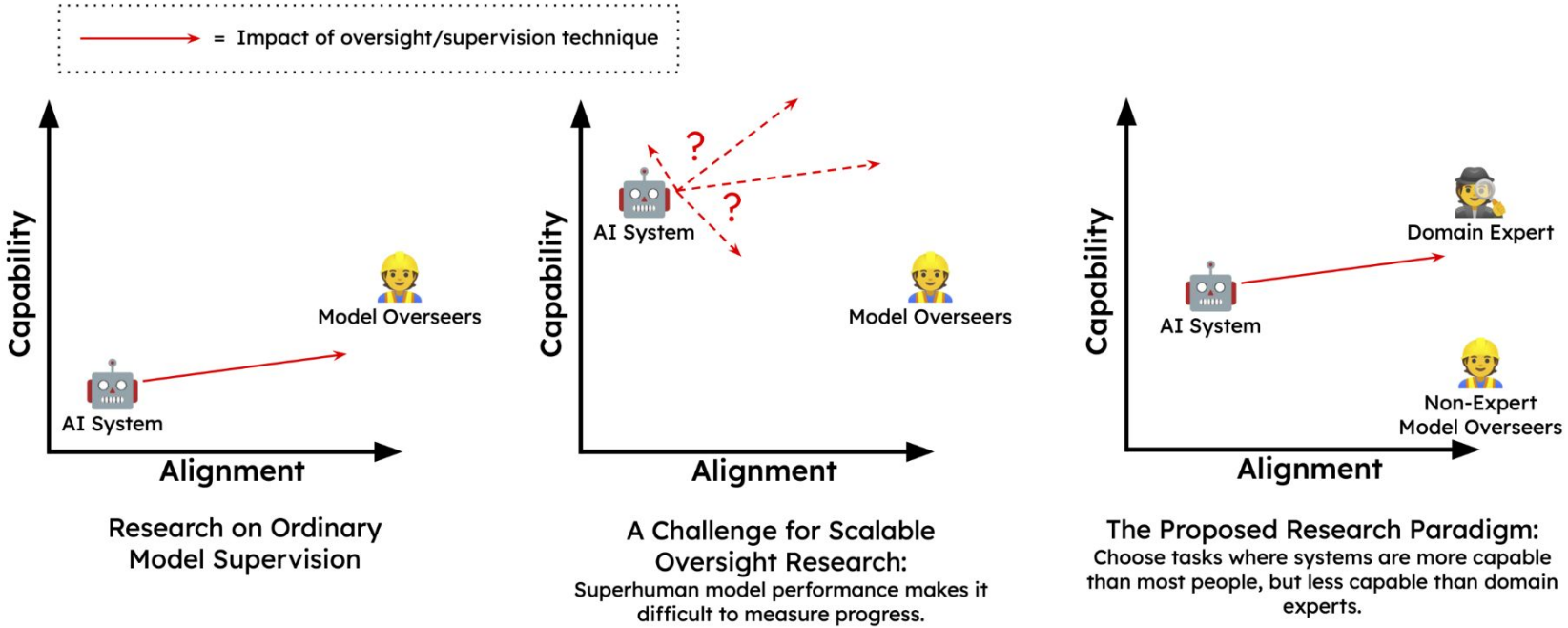
Bowman et al. (2022) — Sandwiching



Bowman et al. (2022) — Sandwiching



Bowman et al. (2022) — Sandwiching



Bowman et al. (2022) — Results

Setup - LLMs help people answer MMLU and QuaLITY questions

Finding - “human participants who interact with an unreliable large-language-model dialog assistant through chat — a trivial baseline strategy for scalable oversight — substantially outperform both the model alone and their own unaided performance”

Michael (2023) + Khan (2024) — Setup

Michael et al. - human debaters on QuALITY; “debate between two unreliable experts can help a non-expert judge more reliably identify the truth”

Khan et al. (ICML 2024 Best Paper) - LLM debaters on QuALITY; stronger experts debate, non-expert judge (model or human) chooses

Michael + Khan — Results

Michael et al.
- “debate performs significantly better, with 84% judge accuracy compared to consultancy’s 74%”
- but what will happen as models get more persuasive?

Khan et al.
- “76% and 88% accuracy respectively
- (naive baselines obtain 48% and 60%)”
- “**optimising expert debaters for persuasiveness in an unsupervised manner** improves non-expert ability”

Michael + Khan — Results

The truth wins out (for time-bounded, information-asymmetric task)

Active directions

Somewhat active area of research, less recent work on this

- Applications in social/political topics with humans
- Making models better debaters, e.g. more legible arguments
- Tasks where judge has as much evidence as debaters

Thank You!

Contact Info:
Peter Hase
phase@stanford.edu
https://peterbhase.github.io