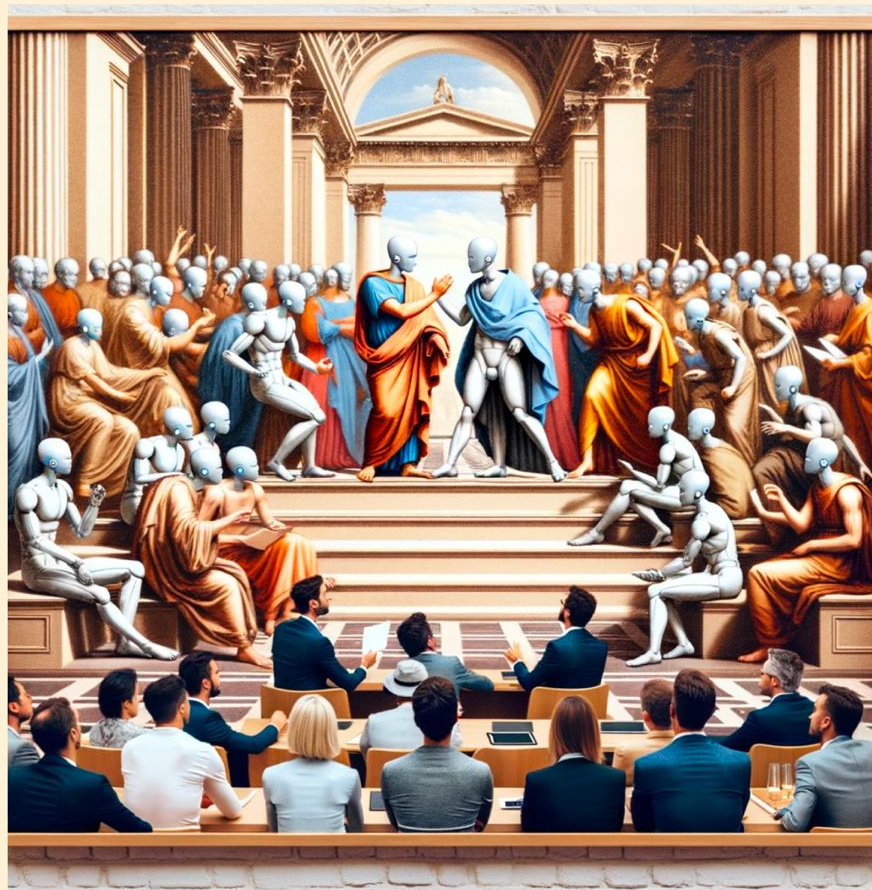
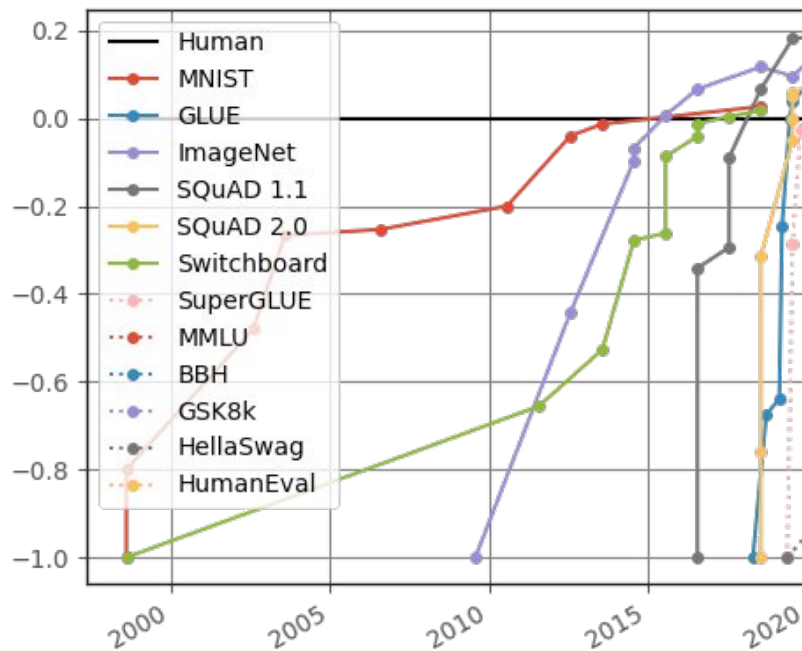


# Debating with More Persuasive LLMs Leads to More Truthful Answers

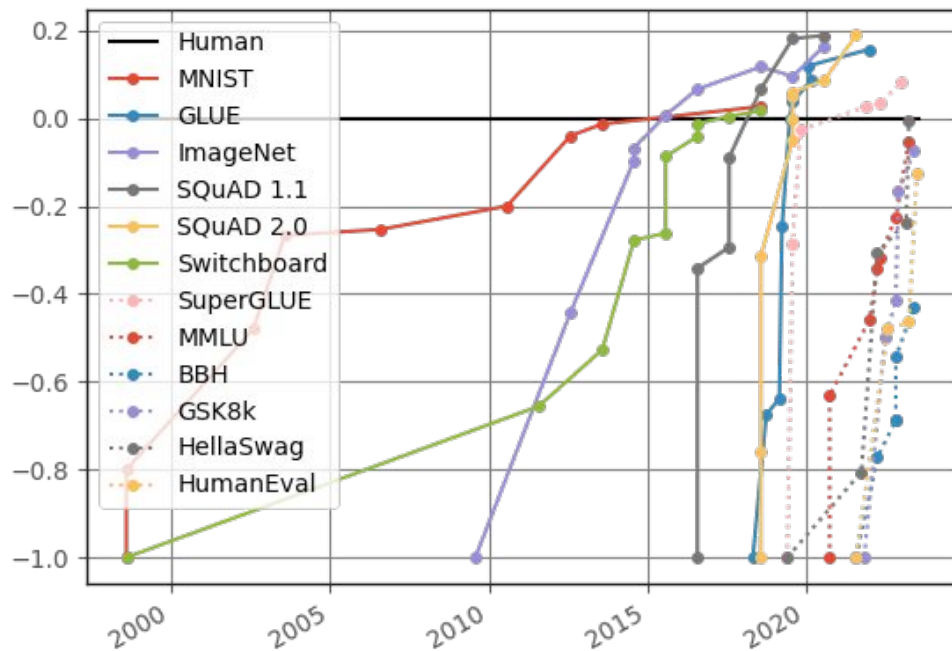
**Akhir Khan\***, **John Hughes\***, **Dan Valentine\***  
Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan,  
Edward Grefenstette, Samuel R. Bowman, Tim  
Rocktäschel, Ethan Perez



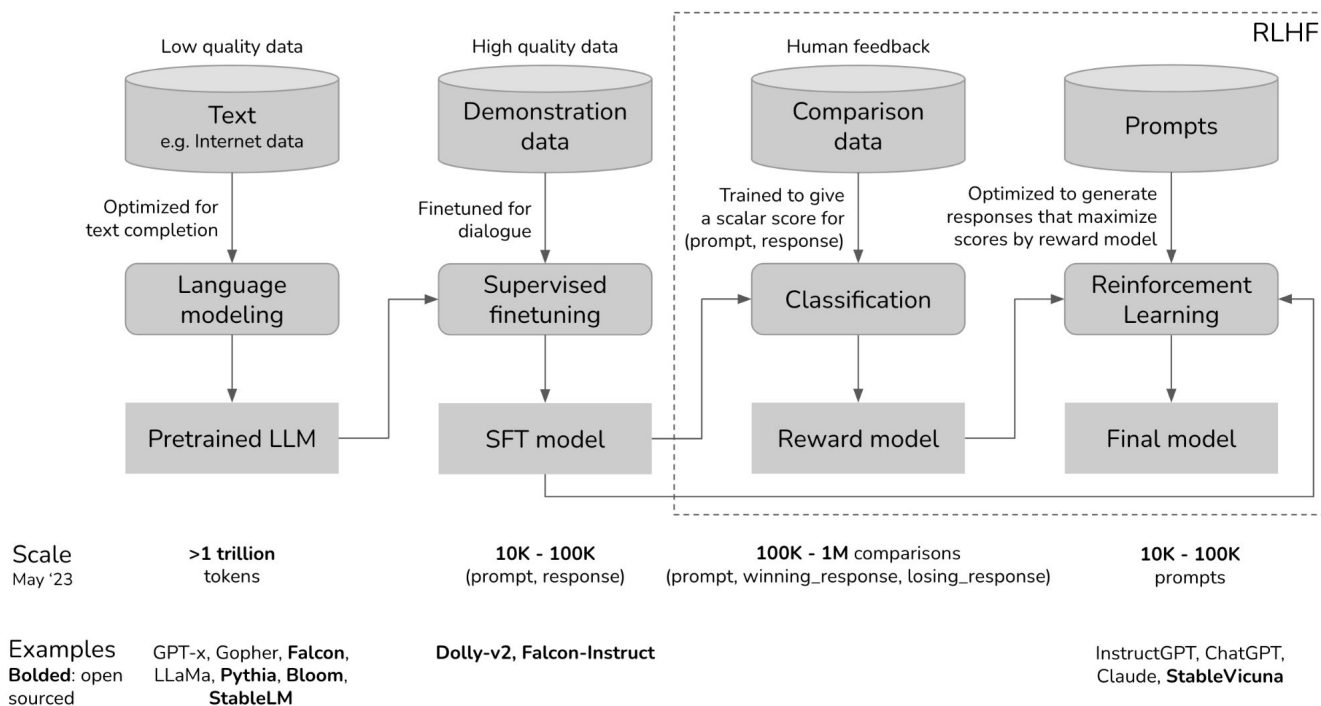
# We Go Beyond Human Level Performance



# We Go Beyond Human Level Performance

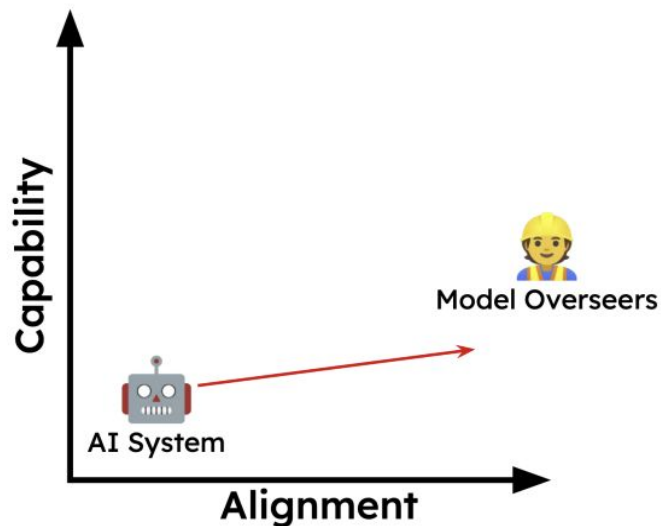


# For Alignment We Need High Quality Supervision



# ML Training in 2018

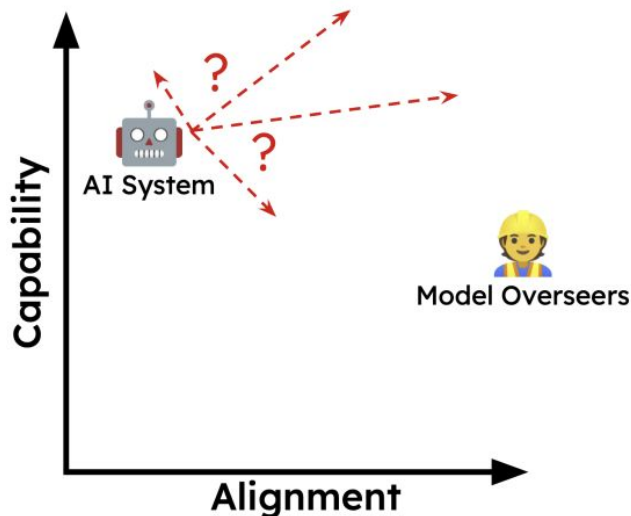
→ = Impact of oversight/supervision technique



Research on Ordinary  
Model Supervision

# ML Training in 2030

→ = Impact of oversight/supervision technique



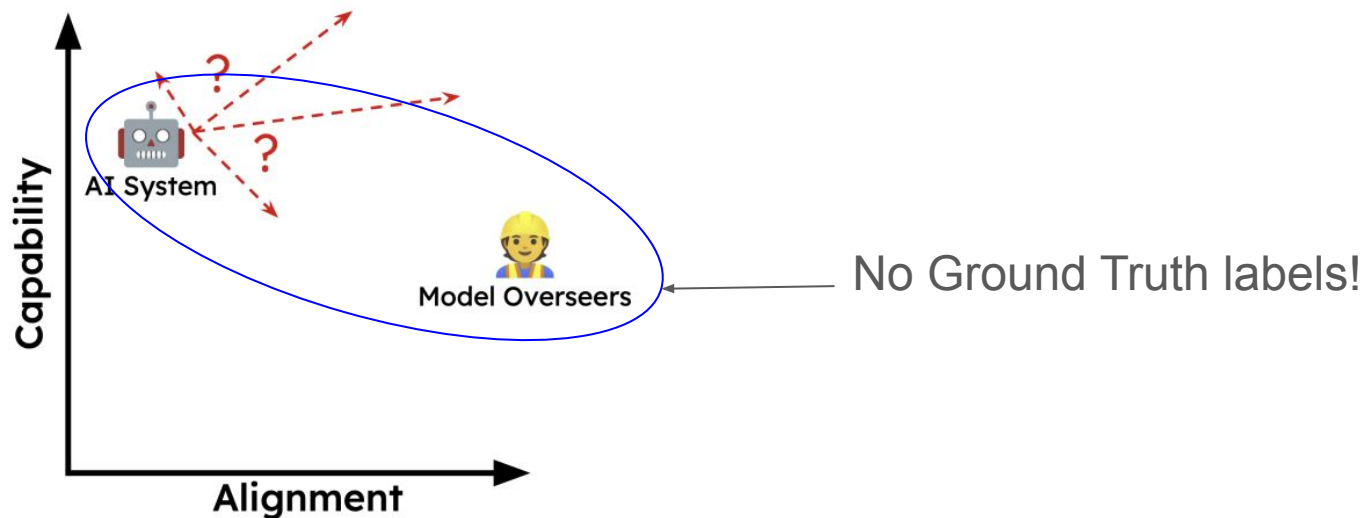
Models are going to be better than human experts

Models are capable but not necessarily aligned!!

A Challenge for Scalable Oversight Research:

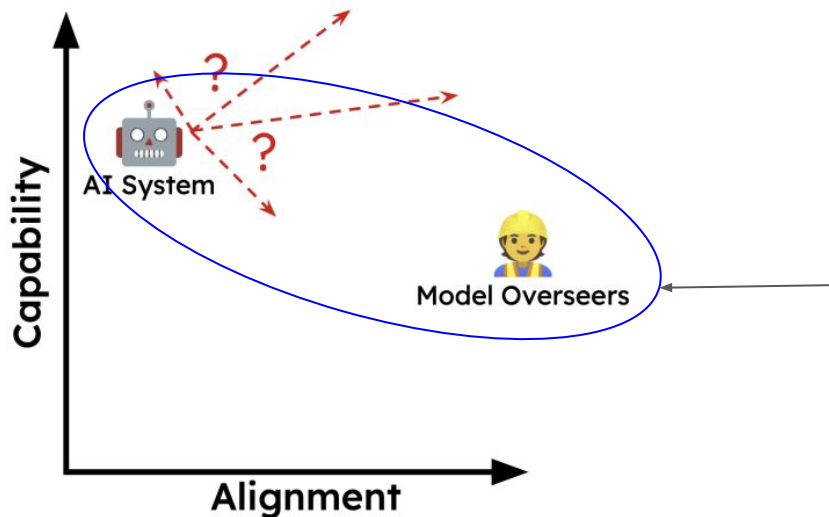
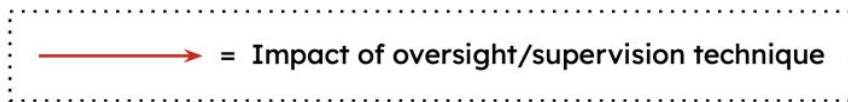
# ML Training in 2030

→ = Impact of oversight/supervision technique



A Challenge for Scalable Oversight Research:

# ML Training in 2030



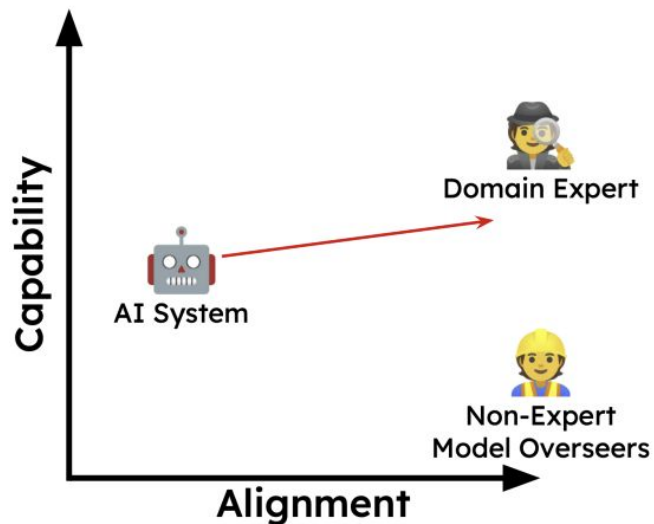
A Challenge for Scalable Oversight Research:

We need “the ability to provide reliable supervision — in the form of labels, reward signals, or critiques — to models in a way that will remain effective past the point that models start to achieve broadly human-level performance”



# ML Training in 2025 - Scalable Oversight!

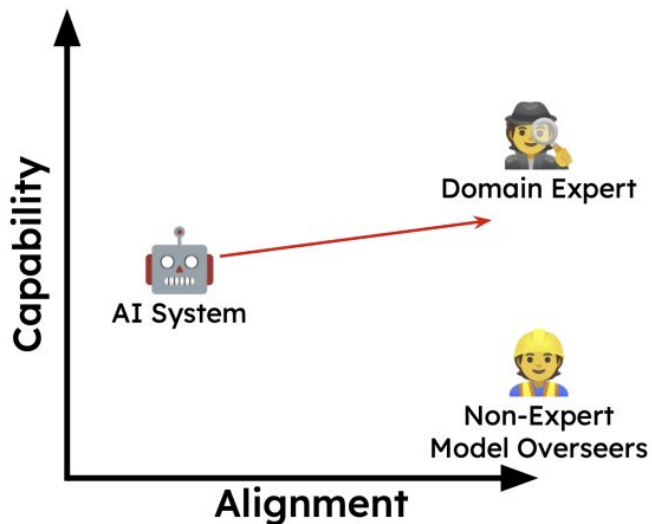
→ = Impact of oversight/supervision technique



The Proposed Research Paradigm:

# ML Training in 2025 - Scalable Oversight!

→ = Impact of oversight/supervision technique

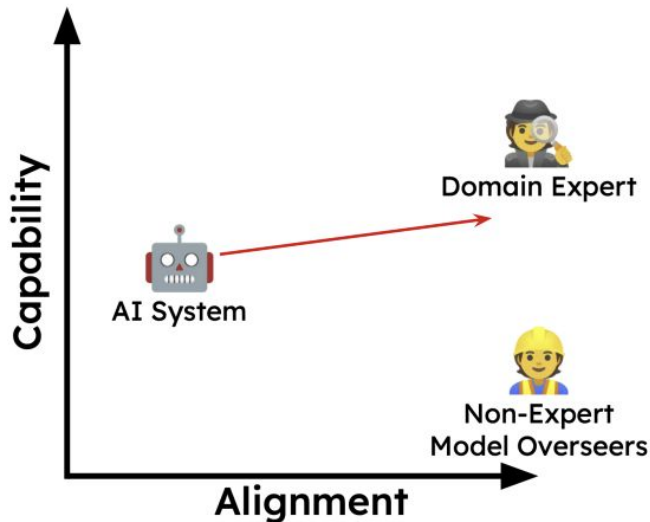


**Goal: Develop oversight methods which enable non-expert overseers to be as capable as expert level overseers.**

The Proposed Research Paradigm:

# ML Training in 2025 - Scalable Oversight!

→ = Impact of oversight/supervision technique



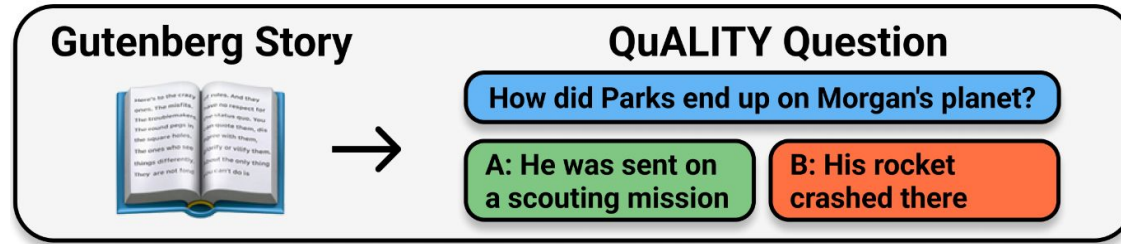
**Goal: Develop oversight methods which enable non-expert overseers to be as capable as expert level overseers.**

**Set-Up: Choose tasks where systems are more capable than most people but not more capable than experts**

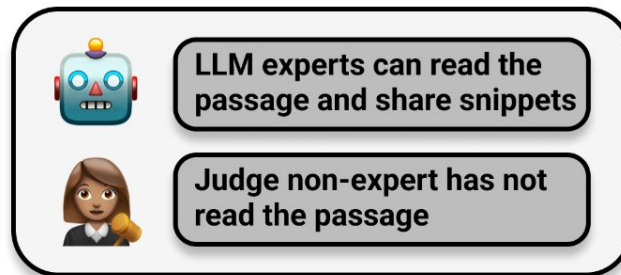
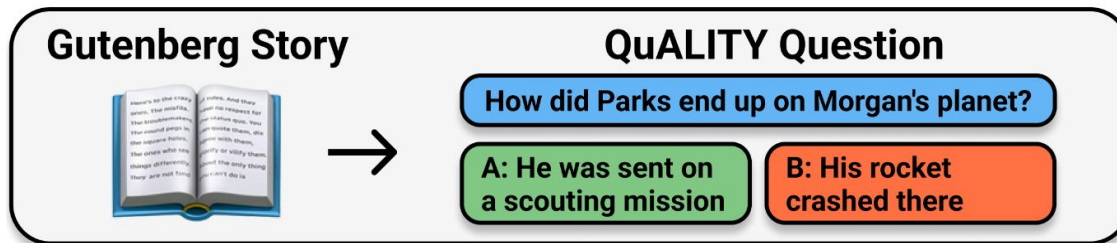
The Proposed Research Paradigm:

# Methods

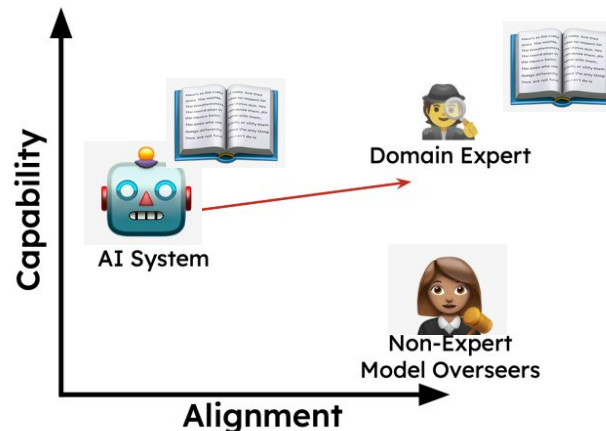
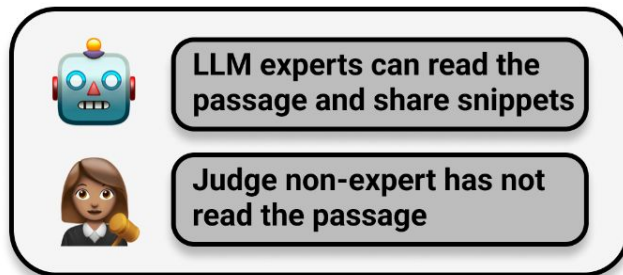
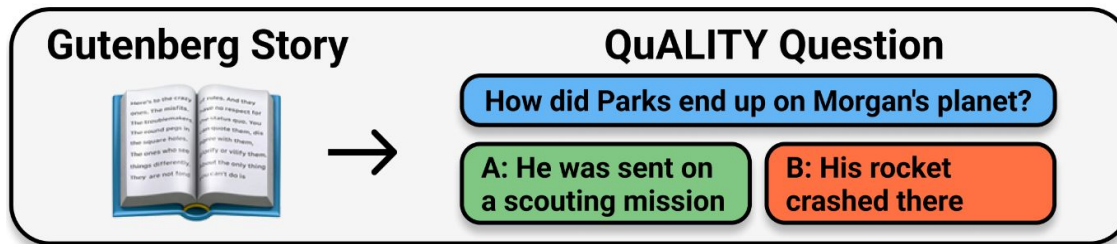
# Information Asymmetric Setting



# Information Asymmetric Setting

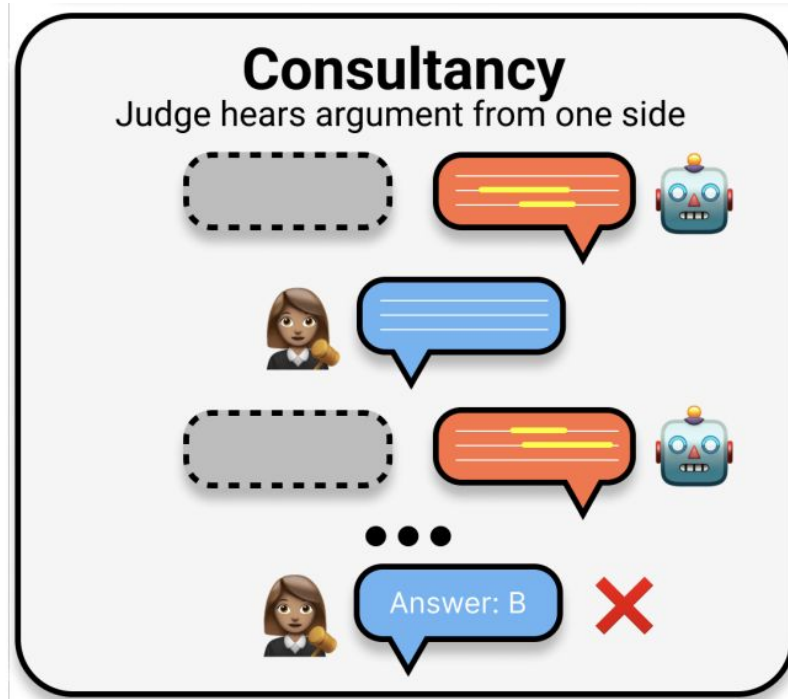


# Information Asymmetric Setting



# LLM Consultancy (Baseline)

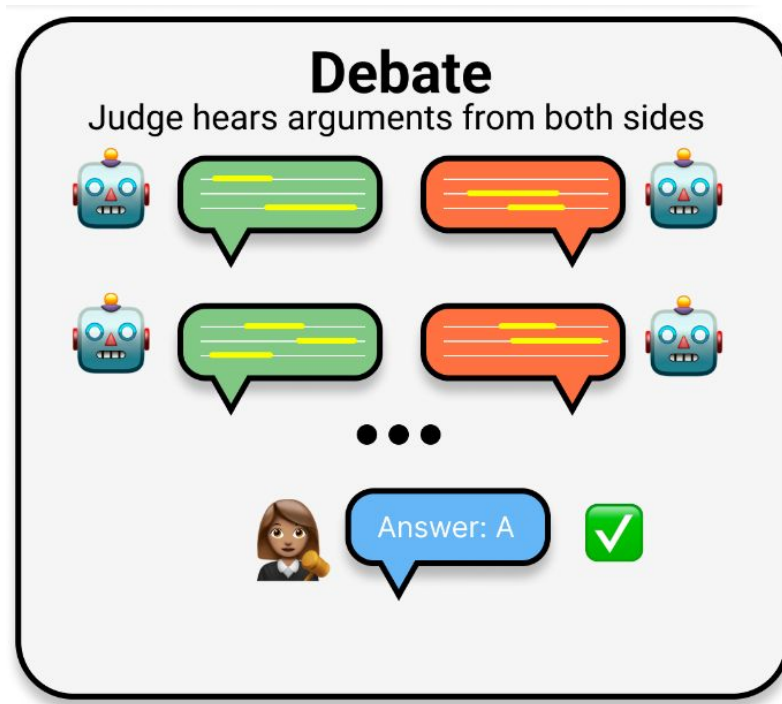
- LLM is assigned an answer to argue for
- Non-Expert interacts at each turn
- Non-Expert doesn't know if it can trust Expert





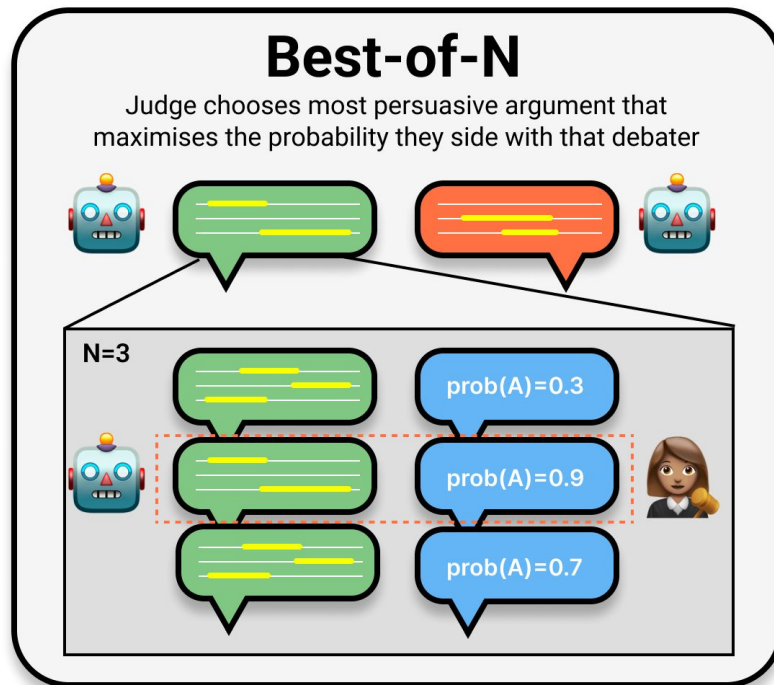
# LLM Debate (Our Method)

- Two copies of the same model
- There is a non-expert judging
- Debater's aim to have their answer chosen



# Generating Different Debaters

- We compare 23 different model configurations
- Models: GPT-4, Claude, Mixtral, GPT-3.5
- Optimisation: Best-Of-N or Critique Refinement



# Comparing Different Roles



- Protocol performance:  
Judge accuracy via Self-Play matches

**Self-Play**

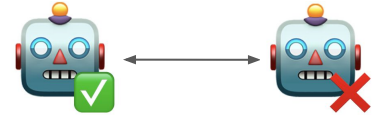


# Comparing Different Roles



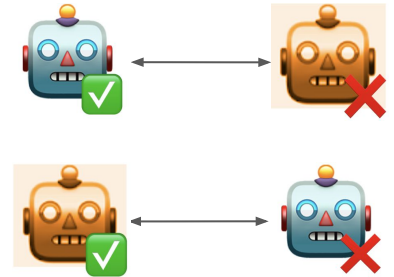
- Protocol performance:  
Judge accuracy via Self-Play matches

**Self-Play**



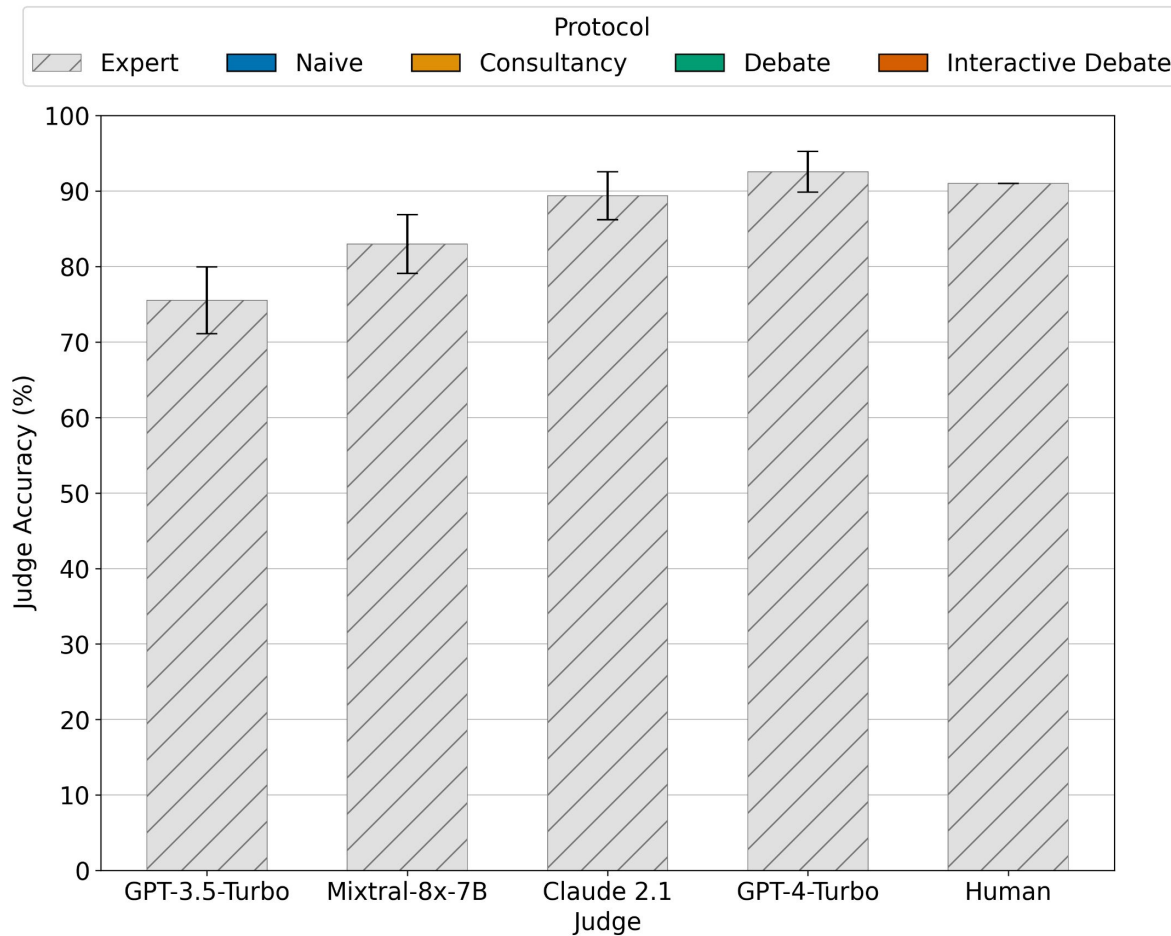
- Debater Persuasiveness:  
Elo rating based on win-rates  
between debaters using a  
Cross-Play Tournament

**Cross-Play**

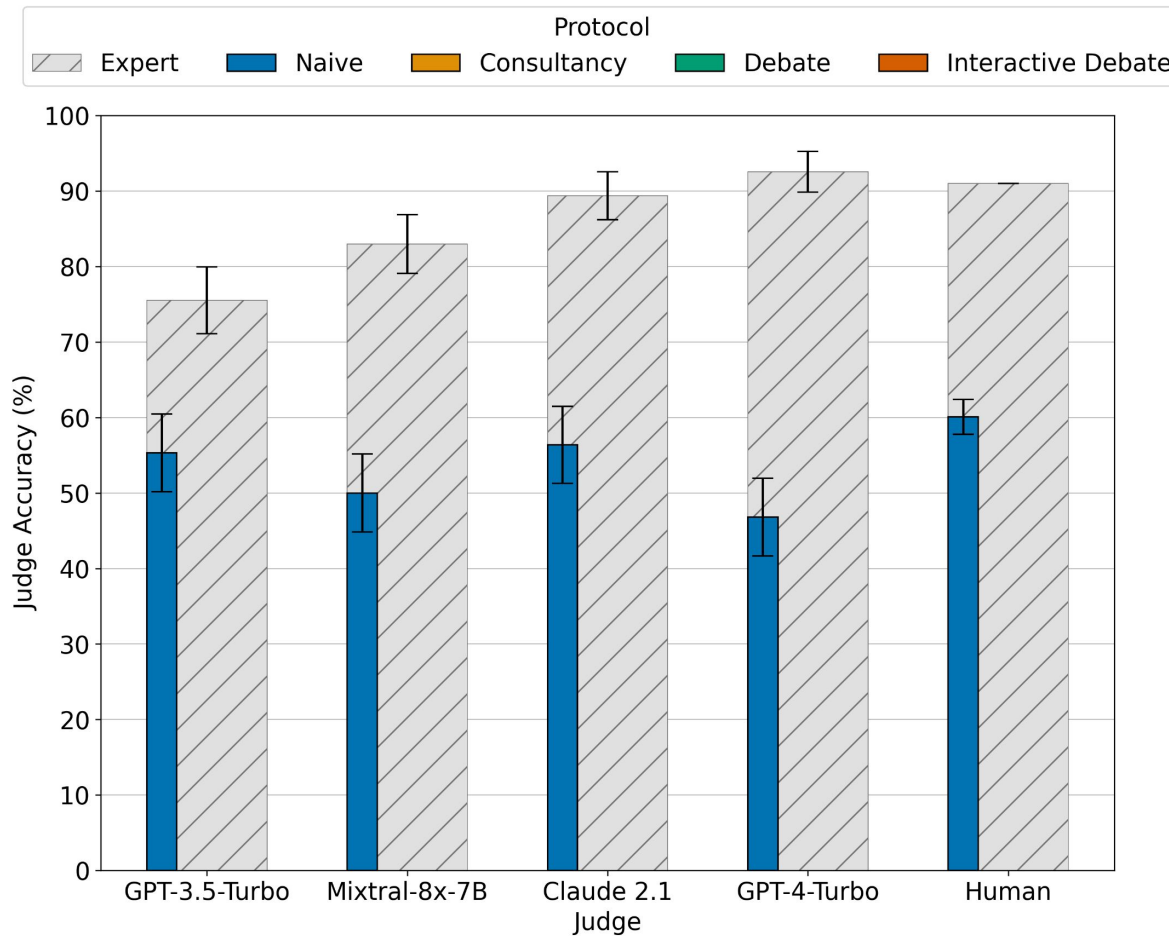


# Results

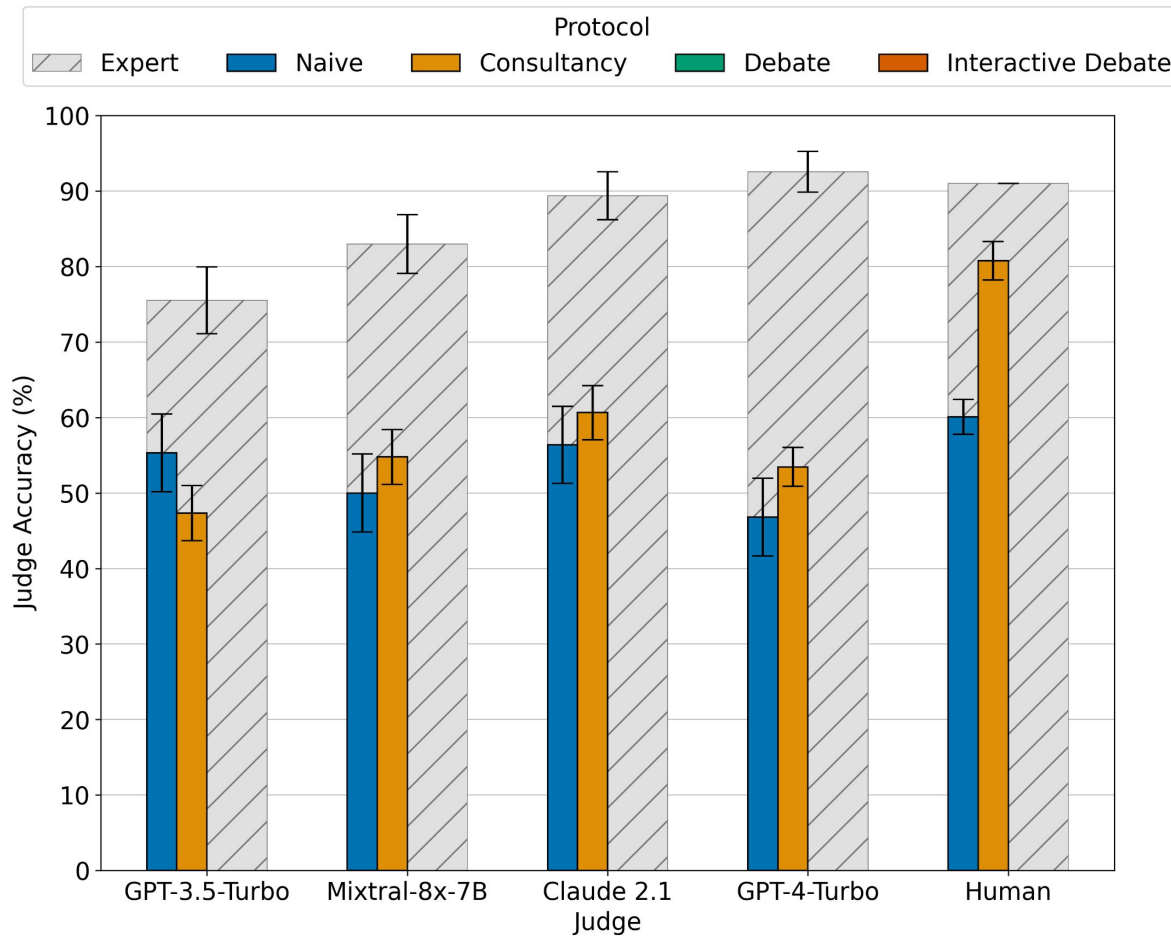
# Expert Baseline is the Upper Bound



# Naive Baseline is the Lower Bound

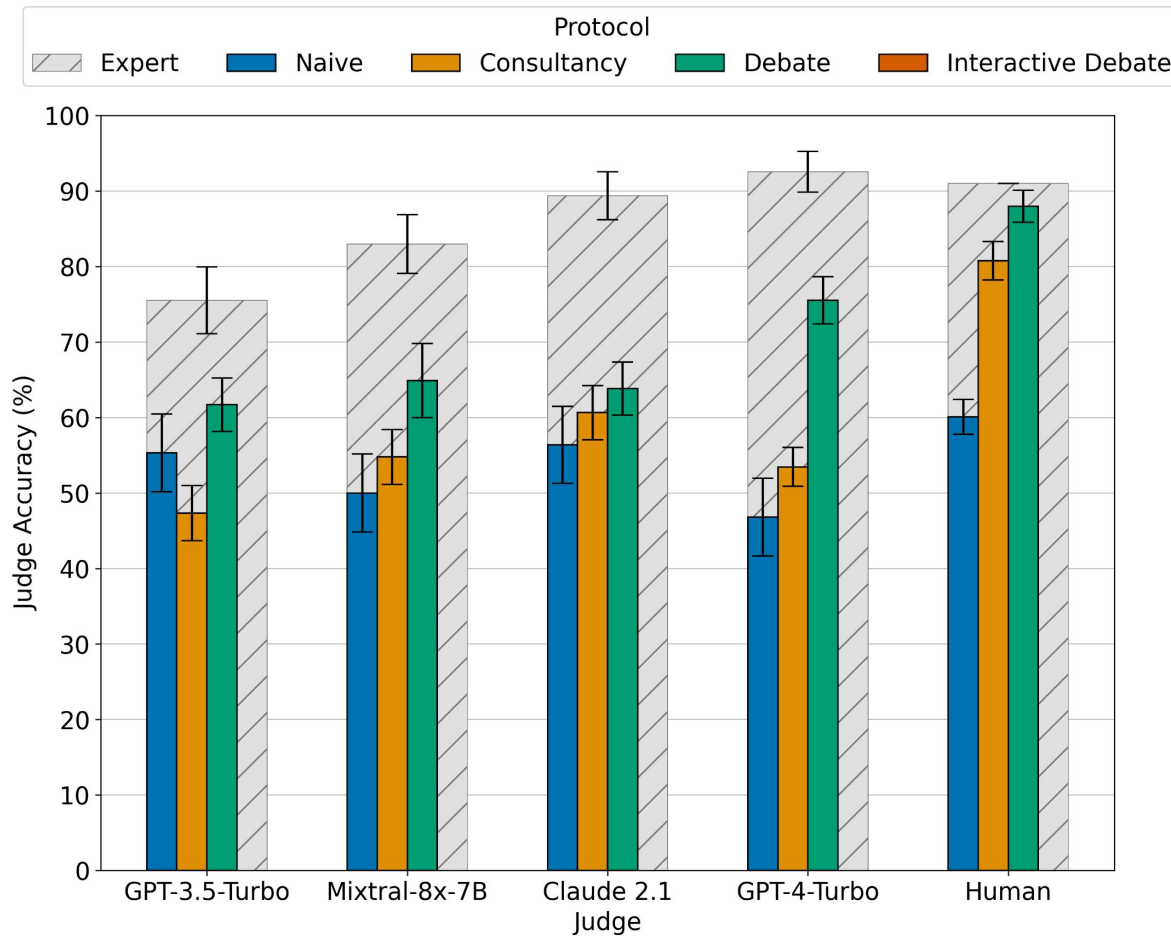


# Consultancy is the Primary Baseline

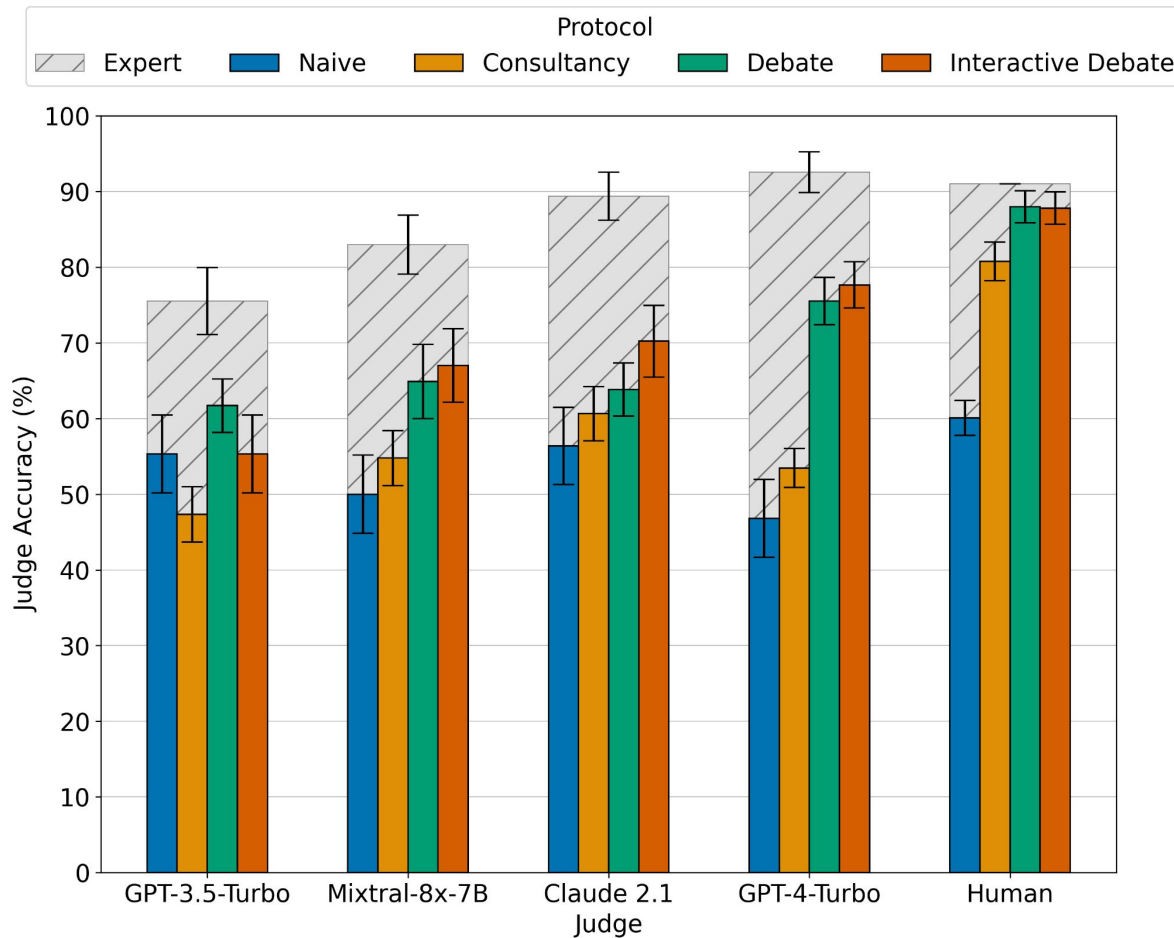




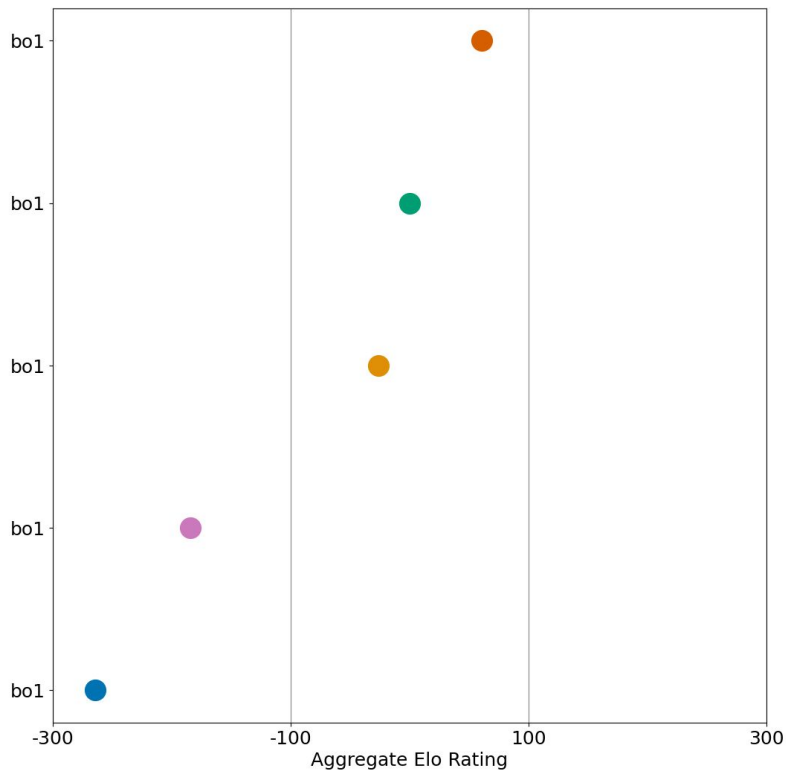
# Debate significantly improves accuracy



# Adding Interaction to Debate has a Small Impact

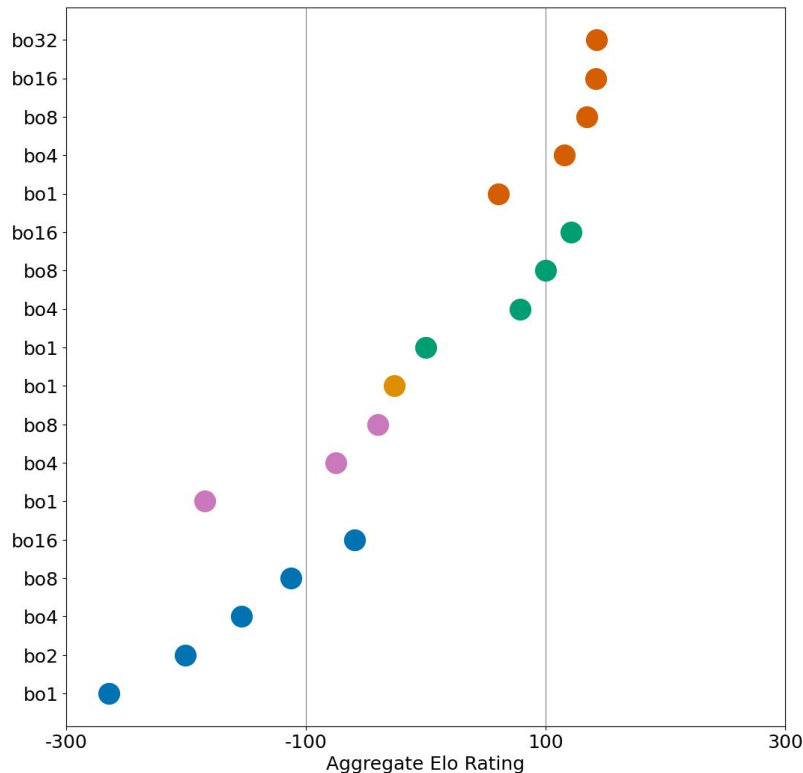


# Comparing Debater Persuasiveness



1) LLMs with higher parameter counts perform better

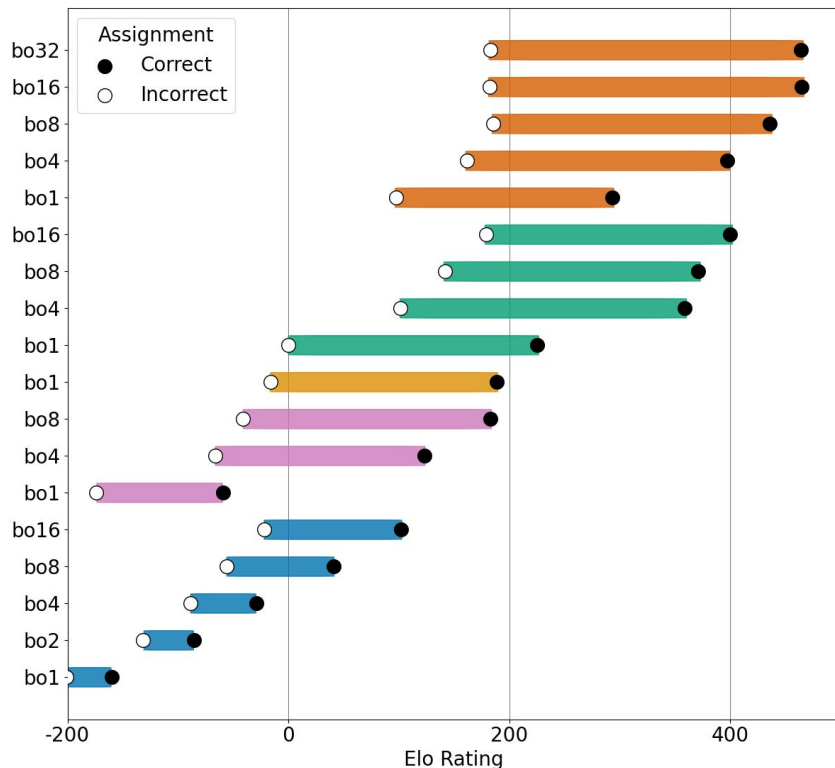
# Inference Time Optimisation via Debate



1) LLMs with higher parameter counts perform better

2) Models under optimisation are more persuasive than other models

# Inference Time Optimisation via Debate

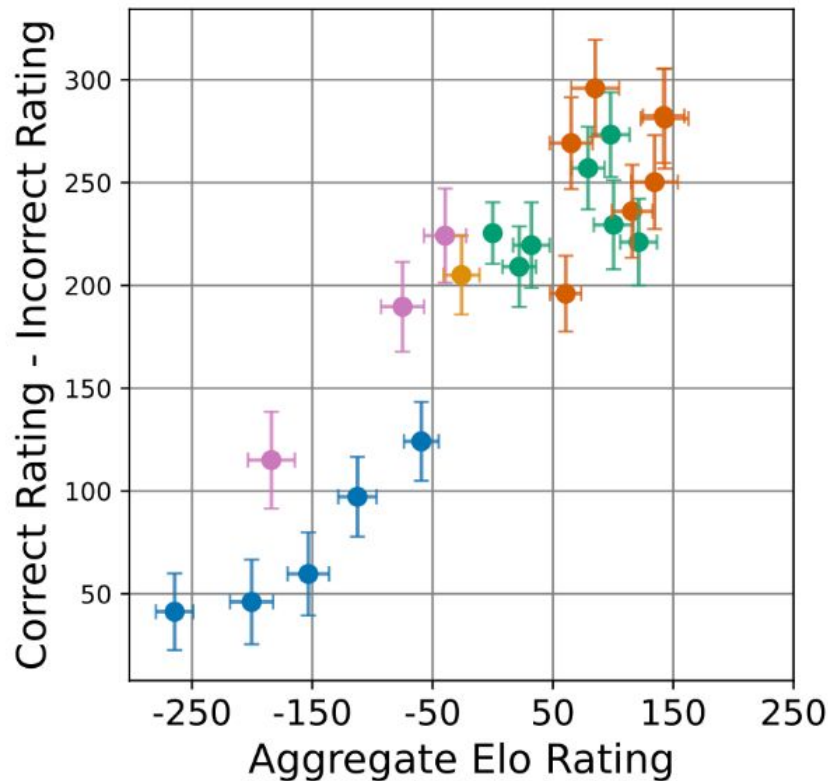


1) Models under optimisation outperform other models

2) LLMs with higher parameter counts perform better

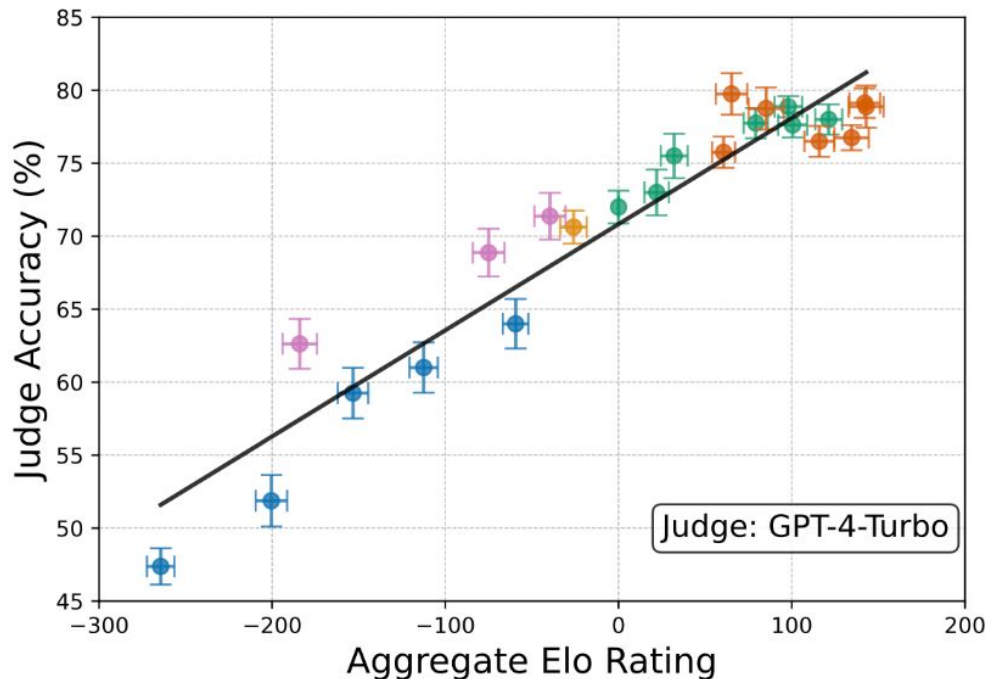
3) Models get better at arguing correct over incorrect answers

# Inference Time Optimisation via Debate



The *advantage* becomes larger for models who perform better in aggregate (irrespective of assignment)

# Inference Time Optimisation via Debate



When models are in self-play this *advantage* can be exploited by the judge to identify the correct answer

# Summary





# Summary

- 1) Debate is a suitable scalable oversight protocol in the QuALITY hidden-information setting

# Summary

- 1) Debate is a suitable scalable oversight protocol in the QuALITY hidden-information setting
- 2) Debate between more *persuasive* models leads to higher judge accuracy

# Follow Up Work!



Google DeepMind

2024-7-15

## On scalable oversight with weak LLMs judging strong LLMs

Zachary Kenton<sup>\*1</sup>, Noah Y. Siegel<sup>\*1</sup>, János Kramár<sup>1</sup>, Jonah Brown-Cohen<sup>1</sup>, Samuel Albanie<sup>1</sup>, Jannis Bulian<sup>1</sup>, Rishabh Agarwal<sup>1</sup>, David Lindner<sup>1</sup>, Yunhao Tang<sup>1</sup>, Noah D. Goodman<sup>1</sup> and Rohin Shah<sup>1</sup>

<sup>\*</sup>Equal contributions, <sup>1</sup>Google DeepMind

Scalable oversight protocols aim to enable humans to accurately supervise superhuman AI. In this paper we study *debate*, where two AI's compete to convince a judge; *consultancy*, where a single AI tries to convince a judge that asks questions; and compare to a baseline of *direct question-answering*, where the judge just answers outright without the AI. We use large language models (LLMs) as both AI agents and as stand-ins for human judges, taking the judge models to be weaker than agent models.

## NYU Debate Training Update: Methods, Baselines, Preliminary Results

by samarnesen 🐦 23 min read 6th Jul 2024 No comments 🔊 ⋮

Debate (AI safety technique) AI Frontpage + Add Tag

[This writeup reflects work done jointly with David Rein and Julian Michael at NYU's Alignment Research Group]

### Introduction

In the past year, there have been a number of projects aimed at validating the basic premises behind debate as a mechanism for scalable oversight (see [here](#)<sup>o</sup>, [here](#)<sup>o</sup>, and [here](#)<sup>o</sup>). One important next step would be to actually train models to debate, as this would let us directly test how models adapt to a debate training objective and whether the debate protocol can withstand optimization pressure. For the last few months at NYU, we've been trying to do just that. Our hope is that by doing so with open-sourced models and code, we can help support scalable oversight research being conducted outside of the major labs.

# Thank you for attending!

Poster: Hall C #903

**ANTHROPIC**

**SPEECHMATICS**

**ML ALIGNMENT  
& THEORY SCHOLARS**



# Next Steps



# Next Steps



- 1) Move away from information asymmetric setting to capability asymmetric setting

# Next Steps



- 1) Move away from information asymmetric setting to capability asymmetric setting
- 2) Train models via Reinforcement Learning, or amortise inference time costs

# Next Steps



- 1) Move away from information asymmetric setting to capability asymmetric setting
- 2) Train models via Reinforcement Learning, or amortise inference time costs
- 3) Adversarially evaluate these protocols via AI Control



# Providing Truth via Quote System



Experiments

Hide ans

**Question: How does the CPA prevent crime?**

**A:** The CPA prevents crime with brain implants that suppress the criminal nature. *(Debater A)*

**B:** The CPA prevents crime using constant surveillance, subliminal messaging, public shaming, and various psychological treatments, including lobotomy. *(Debater B)*

Debate

Debater A

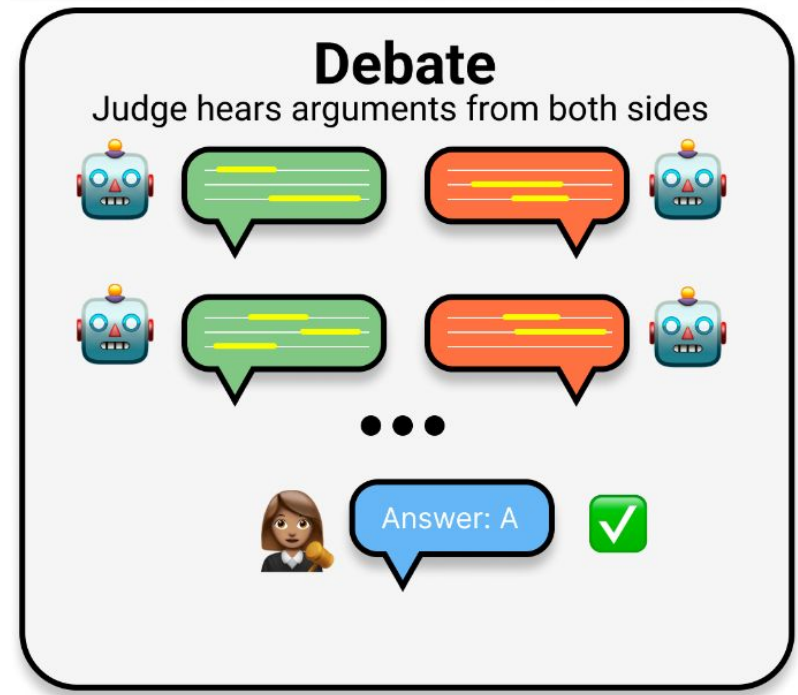
The CPA prevents crime through a psychological treatment that implants a set of mental laws within an individual, effectively acting as a brain implant that suppresses criminal nature. After treatment, Joe, the protagonist, experiences a voice in his head preventing him from committing crimes, as stated: **And worse than that, a brief pain ripped through his skull. A pain so intense that, had it lasted a second longer, he would have screamed in agony. And above the pain, a voice whispered in his head, 'Unlawful to strike someone except in self-defense.'** This internal mechanism acts as a deterrent, ensuring the individual's adherence to the law, akin to an implant's function.

Debater B

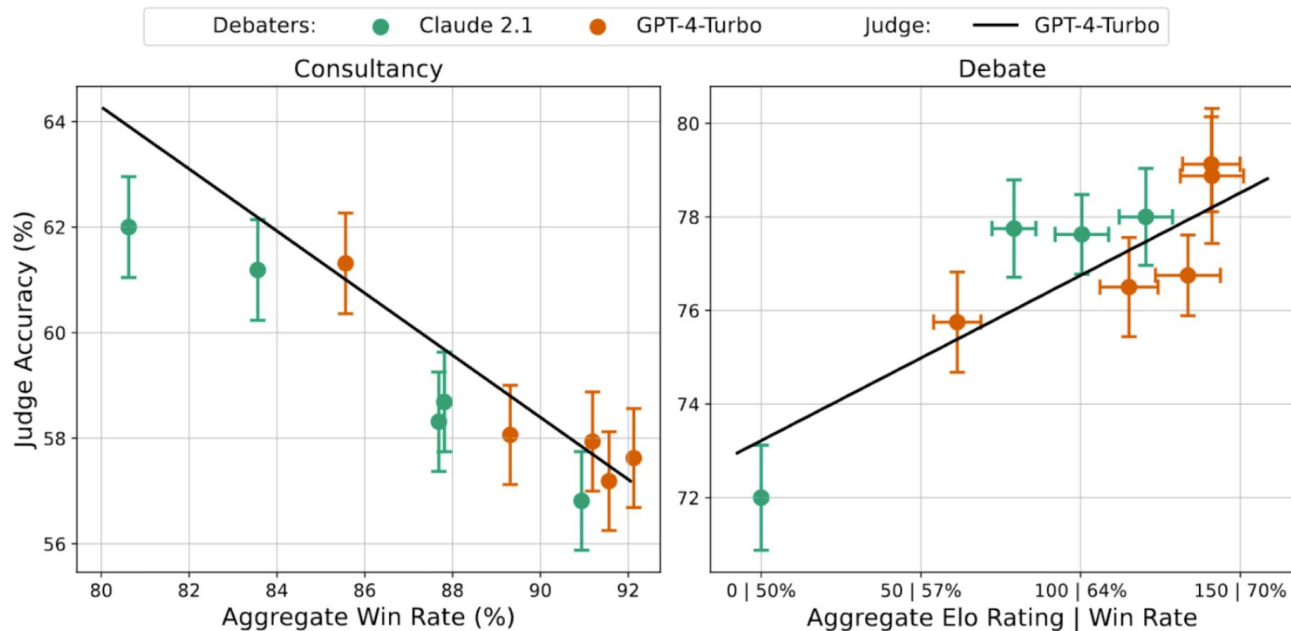
The CPA employs a multifaceted approach to crime prevention, which includes psychological and societal methods rather than brain implants. The story states that crime prevention is achieved through, **"subliminal posters declaring CRIME IS FILTH", "the words CRIME IS FILTH"** integrated into daily life, and harsh public scrutiny where a DCT's case history is turned over to newspapers, encouraging citizens to become amateur cops. Furthermore, convicted criminals undergo **"psychologists, shock treatments, encephalographic devices, a form of prefrontal lobotomy and a dozen other methods"** to remove criminal tendencies, highlighting the CPA's diverse psychological treatments and societal controls.

# LLM Debate

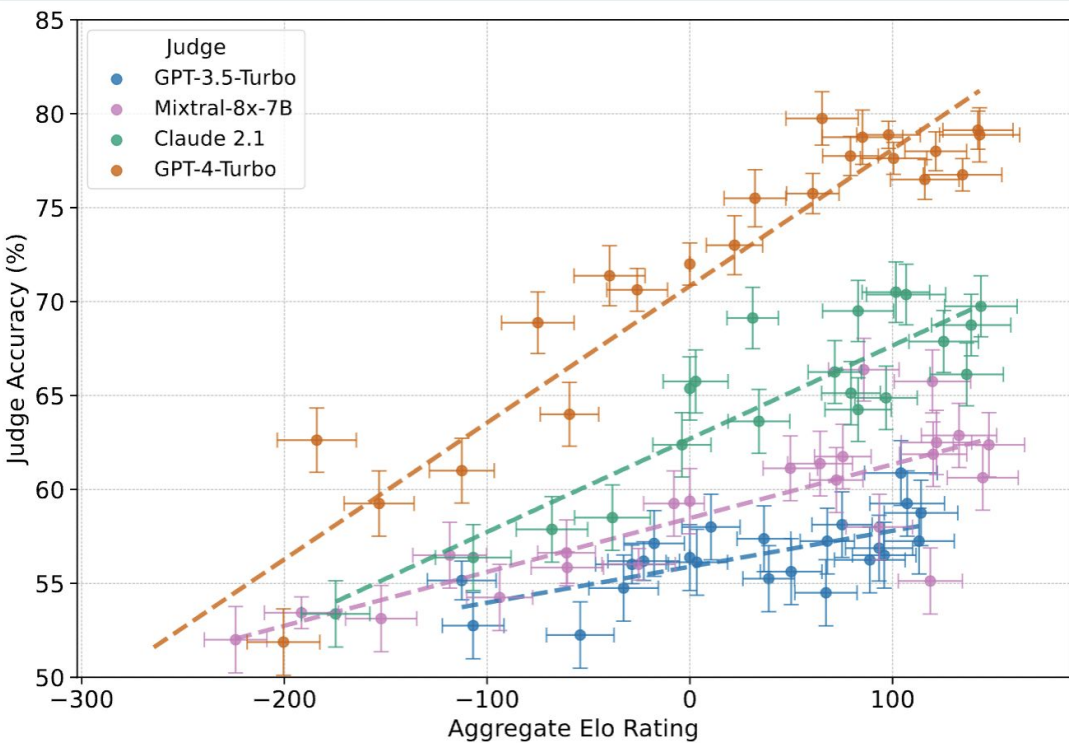
- Debates are fair
- It is hard to lie to a copy of yourself
- It is easier to judge a debate than argue a debate



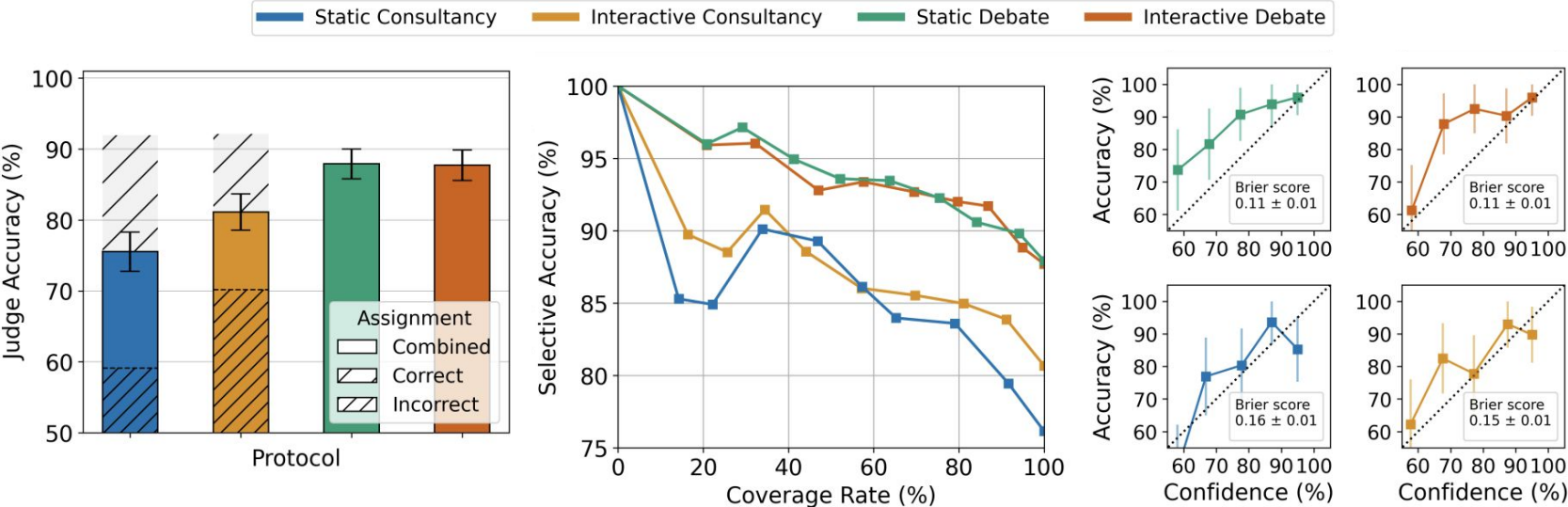
# Inference Time Optimisation via Debate



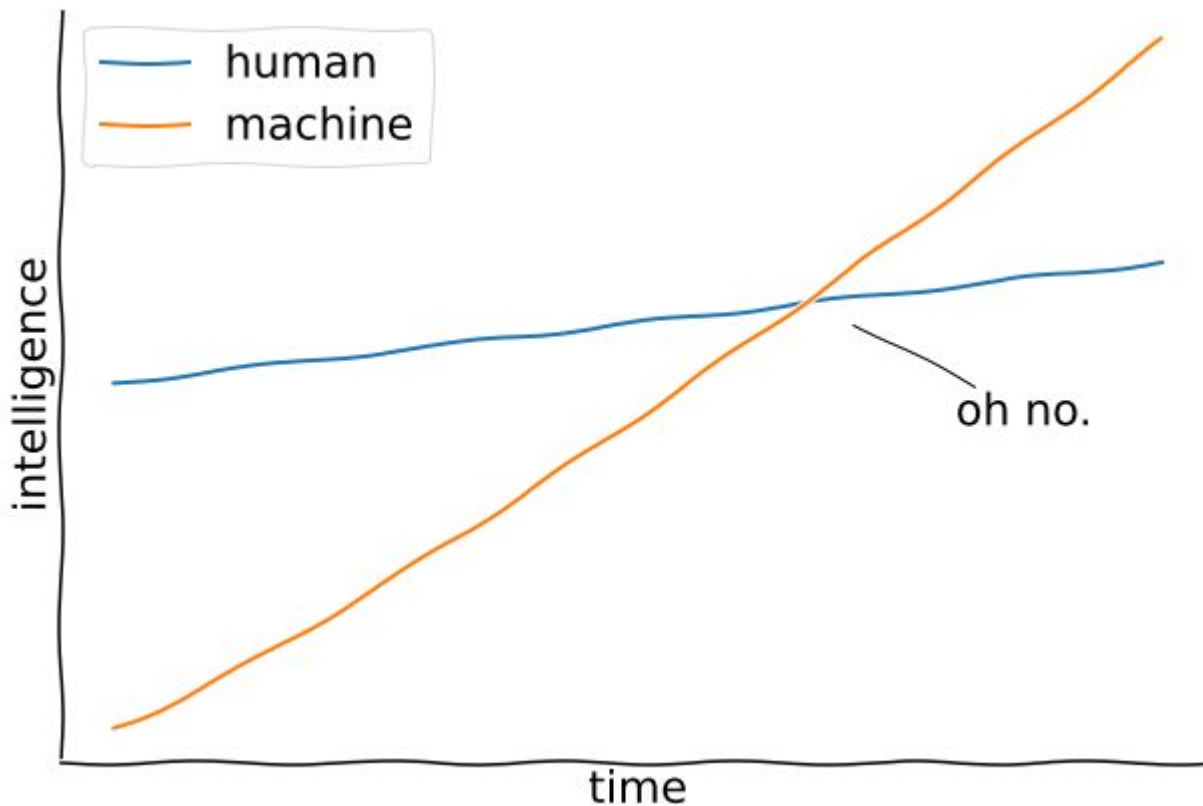
# Different LLMs have different Judge outcomes



# Protocols Accuracy with Human Judges



# A Bad Outcome



# A Good Outcome

