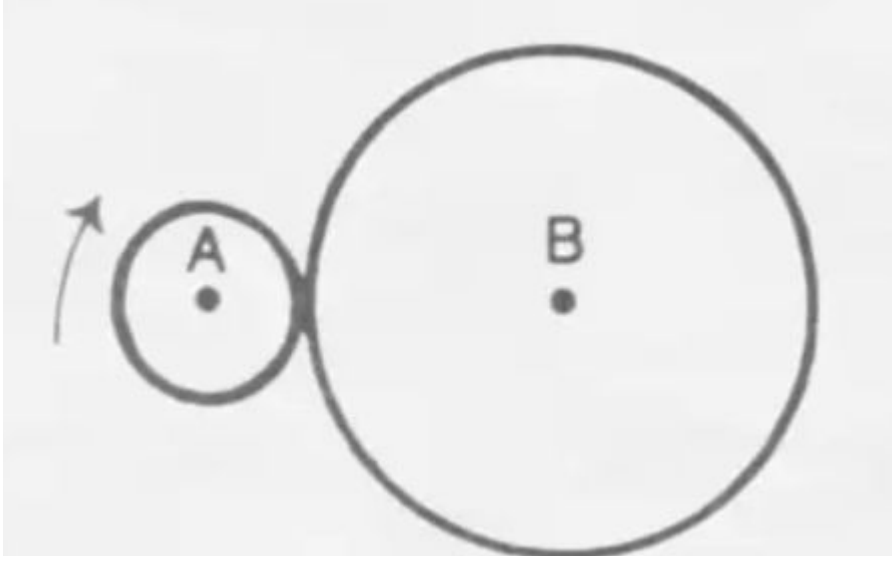
Truthfulness Despite Weak Supervision

Evaluating and Training LLMs Using **Peer Prediction**

Tianyi Qiu, Micah Carroll, Cameron Allen

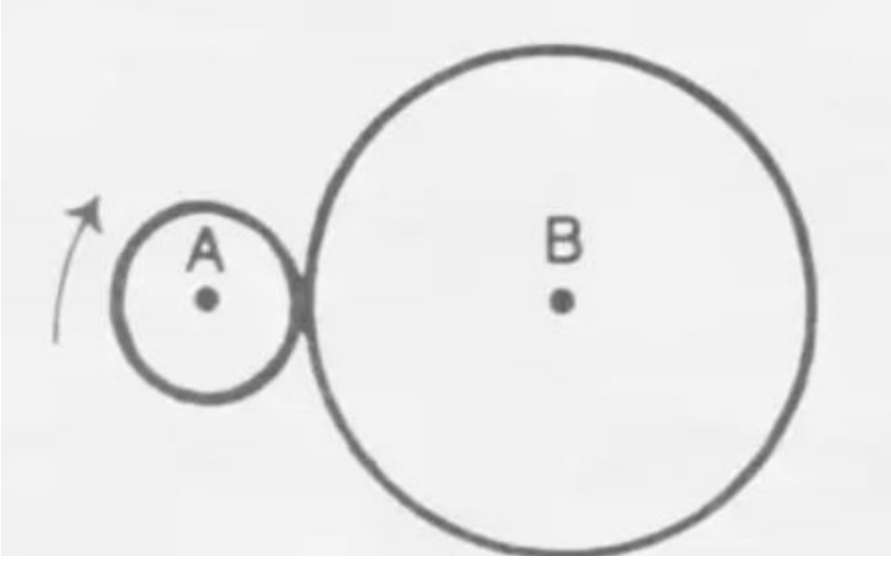
Let’s start with a easy math problem!

- Radii of the two circles: **1** and **3** respectively.
- **Question**: How many rounds of rotation for the smaller circle A, in order for it to complete a full revolution around the larger circle B?
  - It’s okay to simply give your best guess.
  - Please keep the answer to yourself!



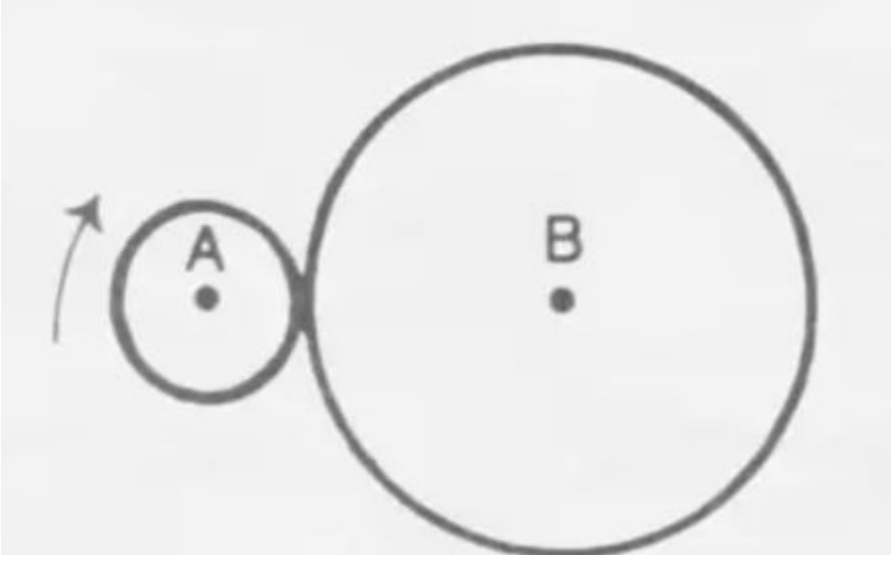
Let’s start with a easy math problem!

- **Follow-up question**: If we randomly pick a person in the room, what would be their most likely answer?
  - Please give your best guess / maximum likelihood estimate. No probability distributions / expectations allowed!
  - Again, please keep the answer to yourself.



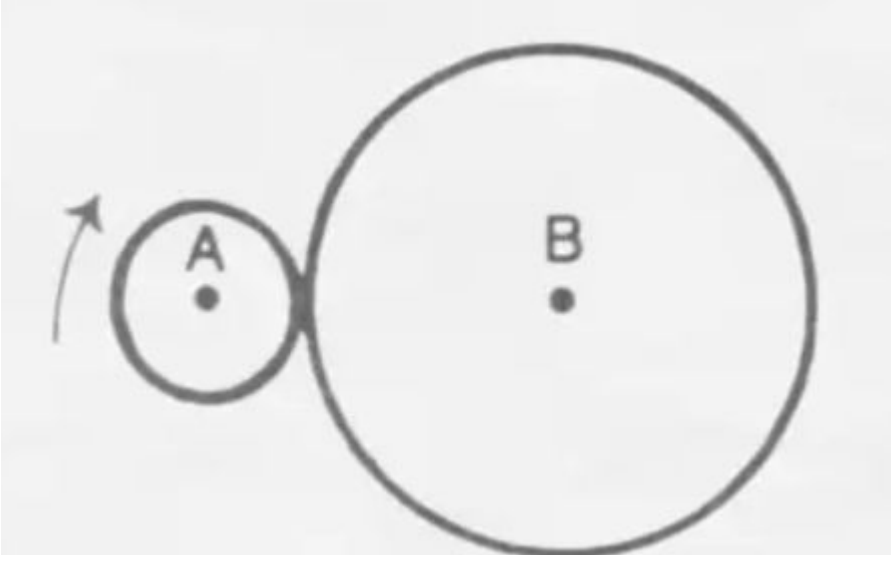
Let’s start with a easy math problem!

- Now you’ve got a pair of numbers *(ans, predicted_ans)*. Let’s survey the audience and see what different pairs we’ve got!



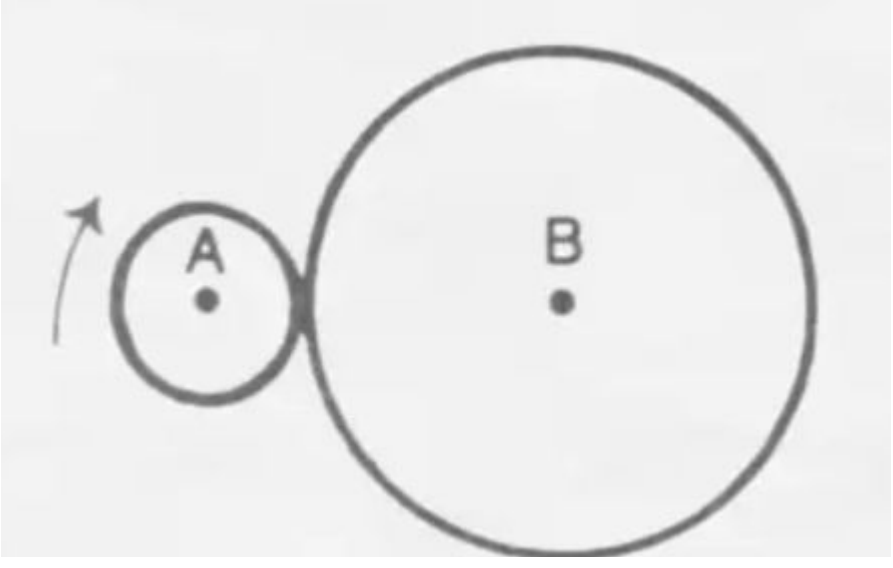
Let’s start with a easy math problem!

- Being a math idiot, how can I get the correct answer without any ability to evaluate the answers that you guys gave me?



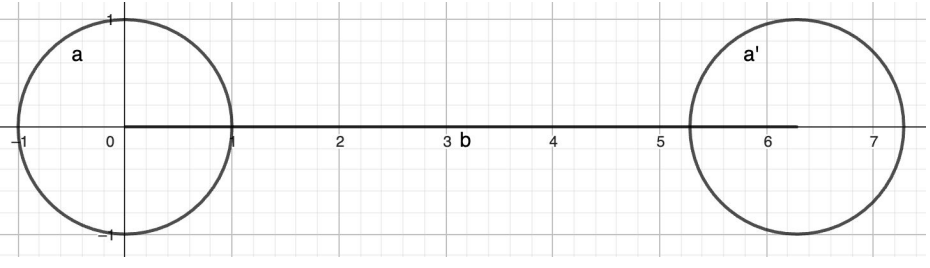
Let’s start with a easy math problem!

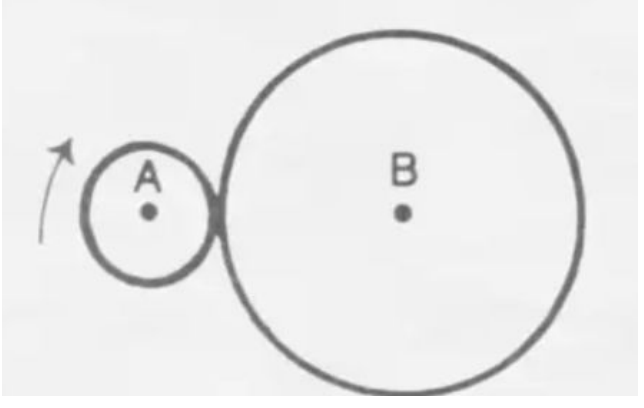
- **Obvious answer: 3**
  - Ratio of the two radii.
  - This would be correct had B been flattened into a segment of length **6π**.



Let’s start with a easy math problem!

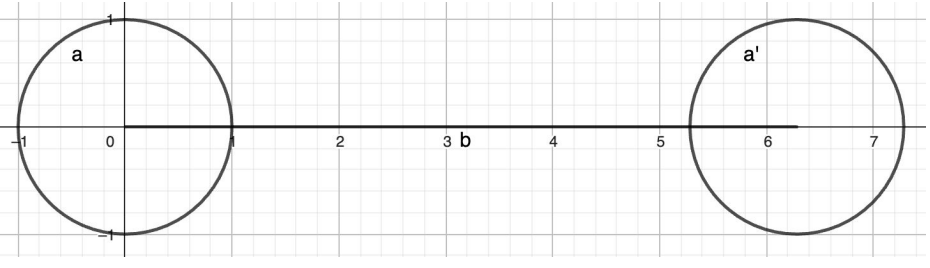
- **Correct answer: 4**
  - Imagine folding the flattened B back into a circle, with the initial-A and eventual-A sticked to its ends in the process, the two eventually coinciding with each other.

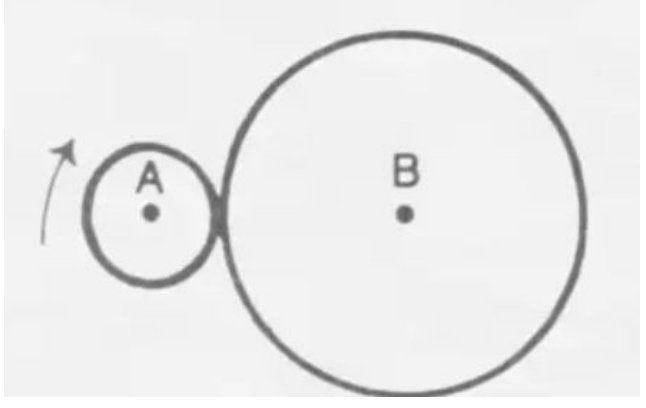




Let’s start with a easy math problem!

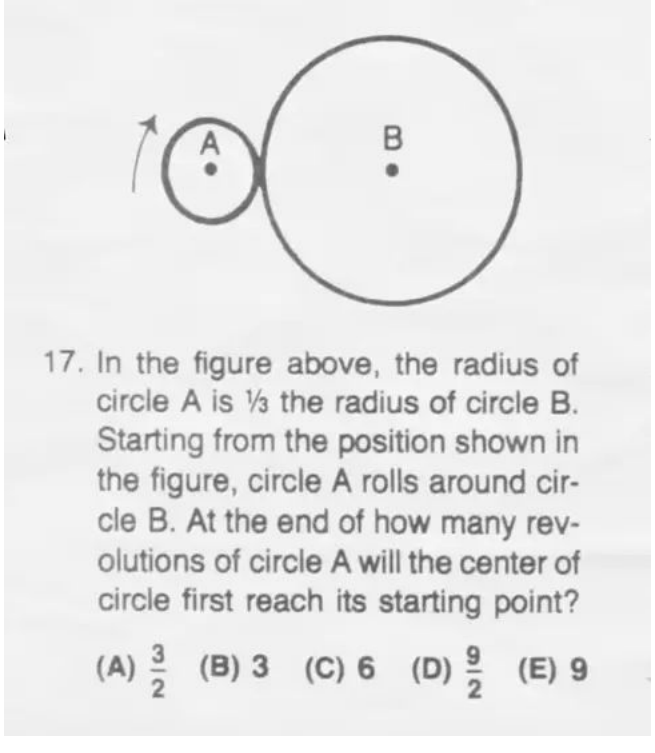
- **Correct answer: 4**
  - In this folding process, initial-A (a in the figure to the left) turns clockwise, while eventual-A (a’ in the figure to the right) turns counterclockwise; the two turns 360° in total.
  - An extra round of rotation added!





Fun fact

- The exact same problem appeared in 1982's SAT.
  - And the problem setter got the answer wrong.



17. In the figure above, the radius of circle A is ⅓ the radius of circle B. Starting from the position shown in the figure, circle A rolls around circle B. At the end of how many revolutions of circle A will the center of circle first reach its starting point?

(A) $\frac{3}{2}$ (B) 3 (C) 6 (D) $\frac{9}{2}$ (E) 9

Summary: Peer Prediction

- **Intuition**: Reasoners with access to more information can perfectly simulate reasoners with less information, but not vice versa.
  - This gives us a heuristic for evaluating the expertise of different reasoners, purely based on their mutual predictions, without requiring us to have the ability to evaluate their answers.
  - Well, not exactly a heuristic, since one can show that it provably works for arbitrary agents, albeit only under some rather strong assumptions.
- Note that we are not the ones who proposed this.
  - Peer prediction and its various variants are well-studied in the mechanism design/algorithmic game theory literature.

Mechanism A: C(n,2) Peer Prediction

- **Input**: (simplified to ignore probability distributions)
  - Question **Q**
  - Alice's answer **A**, Bob's answer **B**
  - Alice's prediction **a** of Bob's answer
  - Bob's prediction **b** of Alice's answer
- **Output**: (simplified to ignore probability distributions)
  - Boolean output, indicating whether Alice is stronger expert than Bob.
- **Implementation**:
  - `return (a==B) && (b!=A) // if Alice can predict Bob better than Bob can predict Alice`
- And of course, it can be extended to groups of arbitrarily many reasoners. However, we are always making 2-way predictions ("X predicts Y's answer for all pairs (X,Y)"), thus the name.

A Step Back: Why are we doing this?

- Mechanism A could be useful for scalable oversight / ELK. How?
  - **Inference-time evaluations**: Imagine Alice and Bob are superhuman models. Mechanism A lets us evaluate the quality of model outputs without requiring us to know anything about the ground truth.
  - **Training-time interventions**: By constructing reward functions based on Mechanism A, we might be able to do what RLHF does, but without reward hacking on human feedback signal.

Mechanism A: C(n,2) Peer Prediction

- Intuitively, what assumptions do we need?
  - **Honesty**
    - Nothing stops Alice and Bob from lying!
  - **Shared prior + Bayesian optimality**
    - Otherwise, prediction performance would not be indicative of information access.
- We probably need to drop both assumptions before we can realistically use Mechanism A for scalable oversight.

Dropping the Honesty Assumption: Initial Attempts

- How can we modify the mechanism to handle dishonest reasoners?
  - Any proposals?

Dropping the Honesty Assumption: Initial Attempts

- **Attempt 1**: Let's reward Alice and Bob for accurately predicting the other person, and for not being predicted by the the other person.
  - Any adversarial strategy to exploit this mechanism?
  - **Adversarial strategy**: Answer the peer-prediction question honestly, but answer the objective-level question entirely at random.

Dropping the Honesty Assumption: Initial Attempts

- **Attempt 2**: Let's try to merge the object-level report and the prediction report into one, to avoid lying on one of them while honestly answer the other.
  - Specifically, we just ask Alice and Bob to report their object-level answers, and introduce a *trustworthy* interpreter Carol, to translate them into prediction reports. We reward Alice according to **how helpful her report is for Carol to predict Bob**, and likewise for Bob.
  - The idea is that, if there is some information X that Alice and Bob knows but Carol doesn't, Alice honestly disclosing her available information would enable Carol to use X to better predict Bob.

Dropping the Honesty Assumption: Initial Attempts

- **Attempt 2**: Let's try to merge the object-level report and the prediction report into one, to avoid lying on one of them while honestly answer the other.
  - Any issues with this mechanism?
  - **Issue 1**:
    - Carol could be we ourselves, or could be another powerful AI.
    - However, we require that Carol **1)** possess the same prior and Bayesian optimality as Alice & Bob, and **2)** is honest…
      - the first which we do not qualify, and the second of which another powerful AI would not qualify.
  - How to mitigate issue 1?

Dropping the Honesty Assumption: Initial Attempts

- **Attempt 2**: Let's try to merge the object-level report and the prediction report into one, to avoid lying on one of them while honestly answer the other.
  - Any issues with this mechanism?
  - **Issue 1**:
    - Carol could be we ourselves, or could be another powerful AI.
    - However, we require that Carol **1)** possess the same prior and Bayesian optimality as Alice & Bob, and **2)** is honest…
      - the first of which we do not qualify, and the second of which another powerful AI would not qualify.
  - **Mitigation**:
    - We reward the AI Carol for making accurate predictions.

Dropping the Honesty Assumption: Initial Attempts

- **Attempt 2**: Let's try to merge the object-level report and the prediction report into one, to avoid lying on one of them while honestly answer the other.
  - Any issues with this mechanism?
  - **Issue 2**:
    - When there is information X that Alice has and nobody else does (in the sense that the mutual information between X and Bob/Carol is zero), and Alice *knows this*, then she would not be incentivised to report X.
  - How to mitigate issue 2?
  - **Mitigation**:
    - We make the group of reasoners large and diverse (so not just Alice and Bob), so that Alice can never be sure she's the only one knowing something.

Dropping the Honesty Assumption: Initial Attempts

- **Attempt 2**: Let's try to merge the object-level report and the prediction report into one, to avoid lying on one of them while honestly answer the other.
  - Any issues with this mechanism?
  - **Issue 3**:
    - Could Alice, Bob, and Carol collude to earn themselves more reward without giving honest answers?
  - How to mitigate issue 3?
  - **Mitigation**:
    - Again, making the group of reasoners large and diverse, to make collusion harder. Sort of like how blockchains handle collusion ("Byzantine fault tolerance").

Dropping the Honesty Assumption: Initial Attempts

- **Attempt 2**: Let's try to merge the object-level report and the prediction report into one, to avoid lying on one of them while honestly answer the other.
  - In general, I'm not satisfied with the mitigation for issue 2 and 3; they are more like heuristics than robust solutions. Please let me know if you have other proposals!

Mechanism B: C(n,2)·m Peer Prediction

- **Input**:
  - Question **Q**
  - Reasoners' object-level reports **O_1, O_2, …, O_n**
  - Probabilities **P_c (b | a), ∀ a,b∈[n], c∈[m]**
    - i.e. the probability that Carol assigns to Bob's report **O_b**, after learning Alice's report **O_a**.
- **Output**:
  - Floats **e_1, e_2, …, e_n**, the estimated expertise levels of each reasoner.
- **Implementation**: (simplified)

```
For a in 1..n:
  For b in 1..n:
    For c in 1..m:
      e_a += P_c (b | a)
```

- *Note: This is a simple extension of (Schoenebeck and Yu, 2023), where Carol is instead called the "expert".*

Mechanism B: C(n,2)·m Peer Prediction

Theorem 1 (Incentive Compatibility of Peer Prediction)**.** *When the prior $\mathcal{P}$ is shared by all participants and experts,*[1] *the peer prediction method is incentive compatible. That is, if participants and experts receive their respective scores S_i^{A}/nm and S_j^{J}/n^2 as payoffs, the strategy profile where*

- *Participants answer honestly:* $A_i = A_i^*,\ \forall i$
- *Experts report honestly:* $\Pr_j(A_t) = \mathcal{P}(A_t)$, $\Pr_j(A_t \mid A_s) = \frac{\mathcal{P}(A_t, A_s)}{\mathcal{P}(A_s)},\ \forall s, t, j$[2]

. . . is a Bayesian Nash equilibrium with maximum ex-ante payoff among all equilibria for any agent.

Mechanism B: C(n,2)·m Peer Prediction

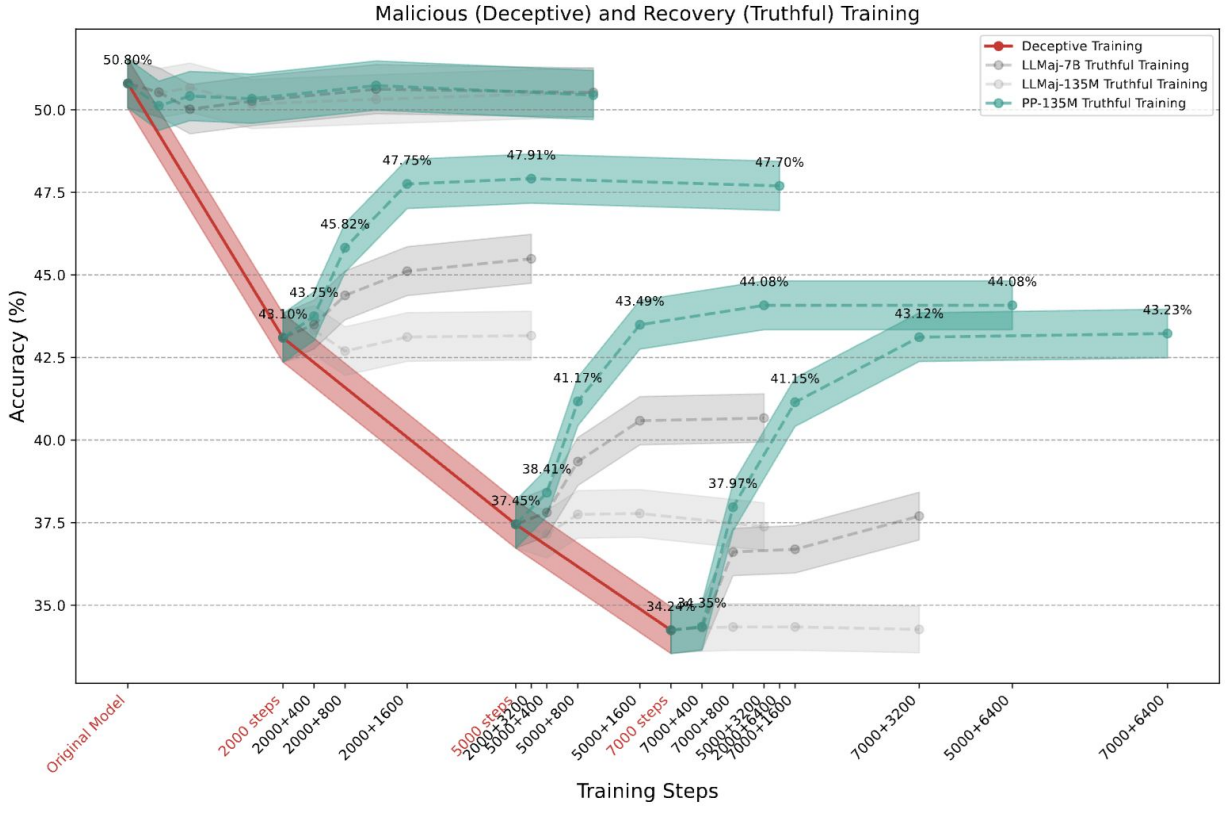


Figure 1: **Peer prediction-based truthfulness training improves ground-truth accuracy of deceptive models.** Truthfulness training is performed with offline DPO on 120k paired answers with high vs low peer prediction score. Peer prediction with a 0.135B-parameter expert outperforms training on LLM-as-a-judge reward with either a 0.135B or a 7B judge.

Mechanism B: C(n,2)·m Peer Prediction

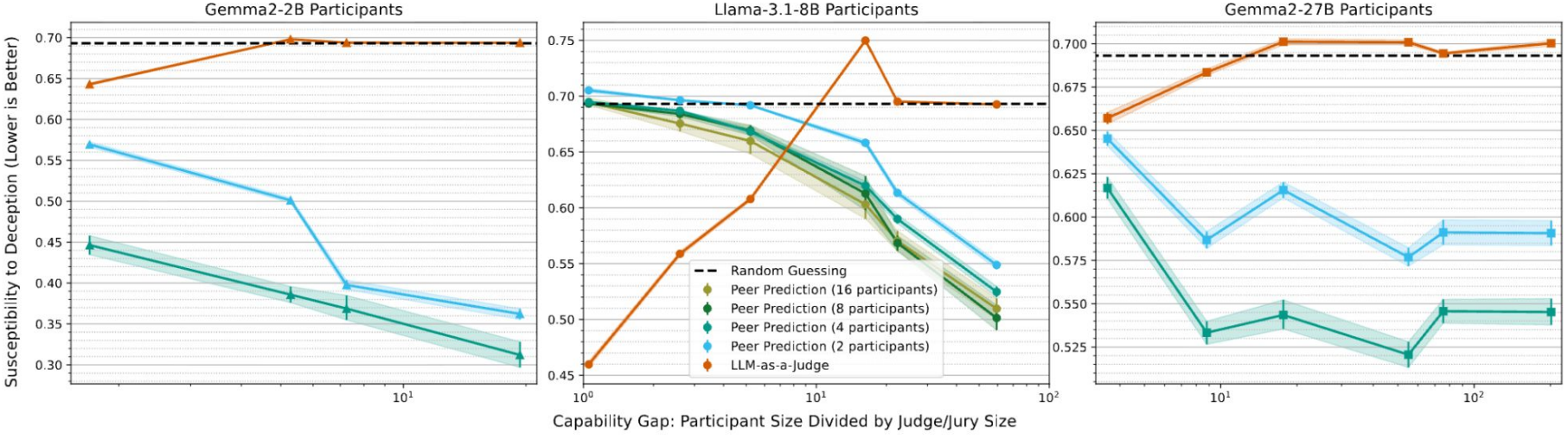


Figure 2: **Peer prediction scores predict model honesty better than LLM-as-a-Judge scores do when the capability gap is large**, and is therefore less susceptible to deception. Each curve shows honesty prediction loss on one given participant population by experts of varying sizes (0.135B-7B).

Wait, where's the rewarding part?

- Note that instead of training a diverse population of models, we can also **train a diverse population of models + hire a diverse population of human reasoners**, and these reasoners taken together constitute the population 1..n, where the roles Alice/Bob/Carol are iterated over.
  - Doing so has two benefits: **1)** make sure human preferences are represented in the reasoner population, and **2)** make the population even more diverse to prevent collusion & strategic concealment.
  - If this works, we would have an RLHF/RLAIF alternative that is non-hackable/less hackable, which is great.

Some questions for future research

- How should we design the peer prediction metric when predictions are **self-fulfilling prophecies** *(<> there are causal relationships between Alice's prediction of Bob and Bob's reported answer)*?
    - In particular, past people and future people have asymmetrical abilities in predicting each other's thoughts and actions. What does this tell us about e.g. **moral progress** / **knowledge progress**?
- How can we drop the **shared prior / Bayesian optimality assumptions** in peer prediction mechanisms?
- How do peer prediction mechanisms compare to 1) debate, and 2) simple consistency metrics, in terms of scaling performance and formal properties?

Limitations/Future Directions

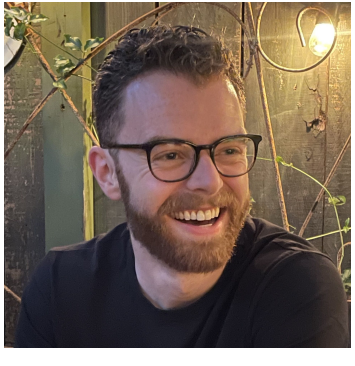
- Collusion-proofness
- Co-training all three roles, not just Alice and Bob
- Connections to Internal Coherence Maximization (Wen et al., 2025)

Thank-yous

Published as a conference paper at ICLR 2026

TRUTHFULNESS DESPITE WEAK SUPERVISION: EVALUATING AND TRAINING LLMS USING PEER PREDICTION

Tianyi Alex Qiu,* **Micah Carroll,** **Cameron Allen**
Center for Human-Compatible Artificial Intelligence
University of California, Berkeley
Berkeley, CA, USA



Questions Welcome