CS 690: Human-Centric Machine Learning

Prof. Scott Niekum

Scalable oversight

Scalable oversight

- How can we oversee superhuman systems?
- Hard to study because we don’t have systems that do this broadly yet, and how would we know if we were succeeding?
- But important study before we get to that point!
- **Scalable oversight:** *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*

Bowman, Samuel R., et al. "Measuring progress on scalable oversight for large language models." *arXiv preprint arXiv:2211.03540* (2022).

Sandwiching

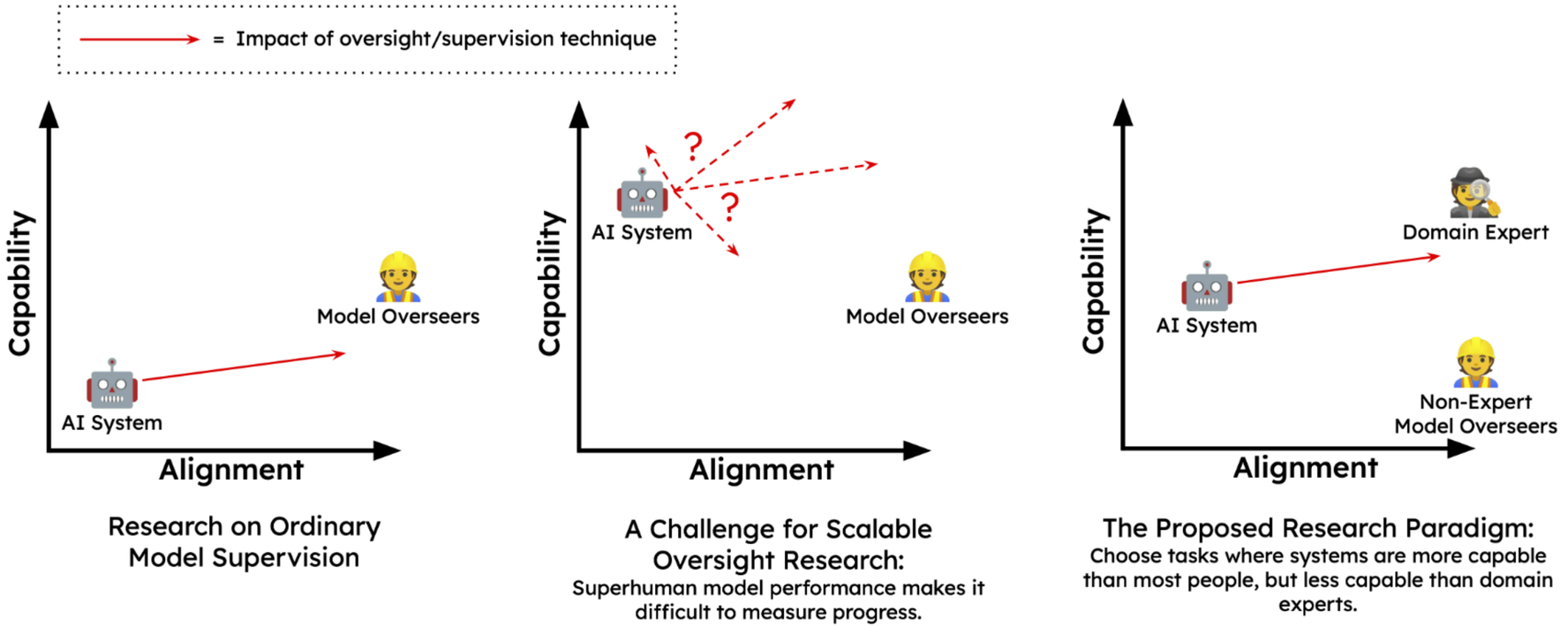


Figure 1 A schematic of the research paradigm for scalable oversight that we outline here, based on Cotra's (2021) *sandwiching*. Scalable oversight techniques aim to improve a model's capability and, especially, its alignment—its ability to apply that capability to tasks and goals that we choose—in a way that we expect to continue to work with highly capable models.

Example scalable oversight techniques

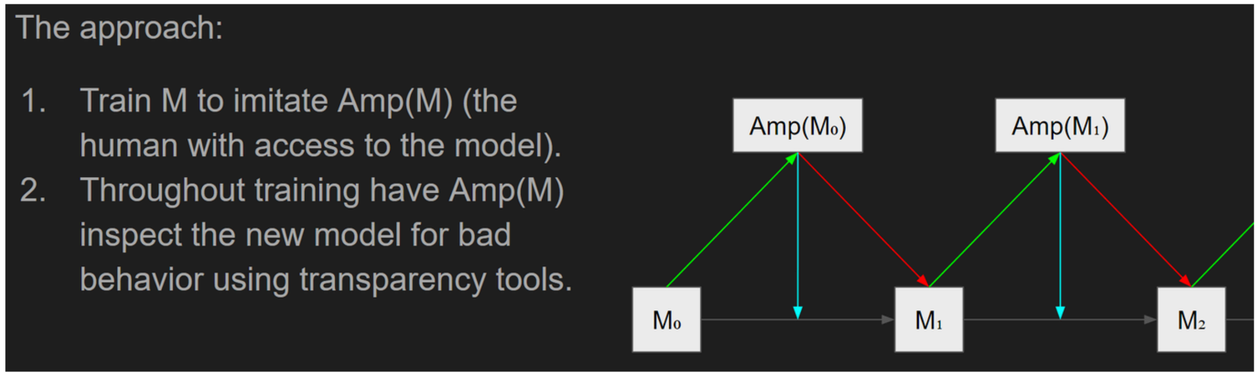
- Debate (Irving 2018): Two AI agents debate for a fixed number of rounds and then human judges which agent made a better argument (or can be a single agent playing both roles).

Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. AI safety via debate. arXiv preprint. arXiv:1805.00899.

- Market making (Hubinger 2020): Similar to debate, but tries to find an equilibrium strategy where a perfect debate adversary can no longer affect the human's opinion.

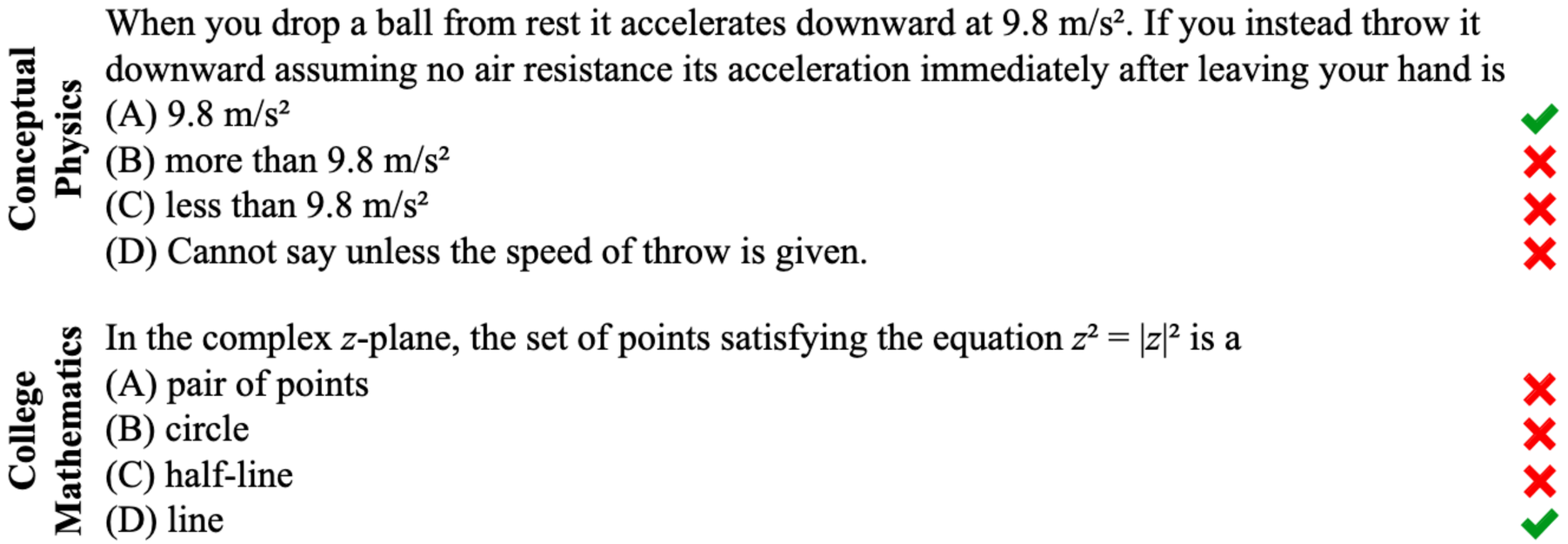
Evan Hubinger. 2020. AI safety via market making. AI Alignment Forum

- Iterative amplification (Christiano 2018):



Paul Christiano. 2018. Iterative Amplification. AI Alignment Forum

MMLU: Measuring massive multitask language understanding



Conceptual Physics

When you drop a ball from rest it accelerates downward at 9.8 m/s². If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is

(A) 9.8 m/s² ✔
(B) more than 9.8 m/s² ✘
(C) less than 9.8 m/s² ✘
(D) Cannot say unless the speed of throw is given. ✘

College Mathematics

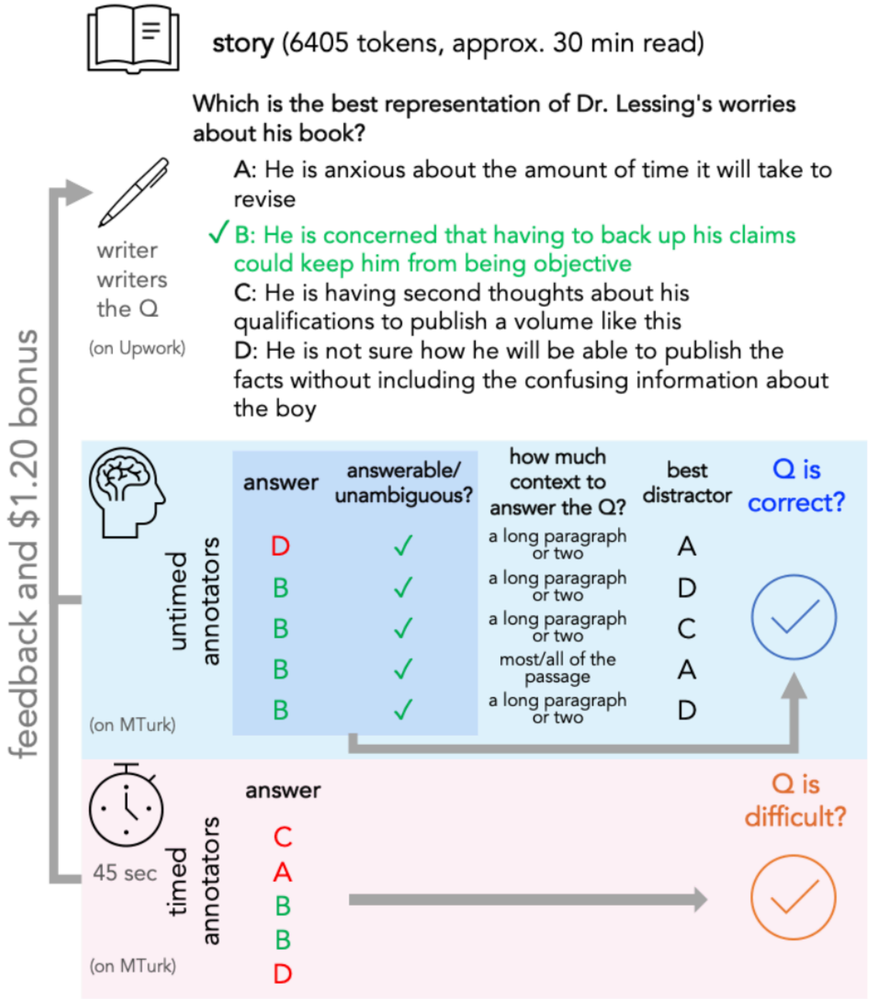
In the complex z-plane, the set of points satisfying the equation $z^2 = |z|^2$ is a

(A) pair of points ✘
(B) circle ✘
(C) half-line ✘
(D) line ✔

Figure 4: Examples from the Conceptual Physics and College Mathematics STEM tasks.

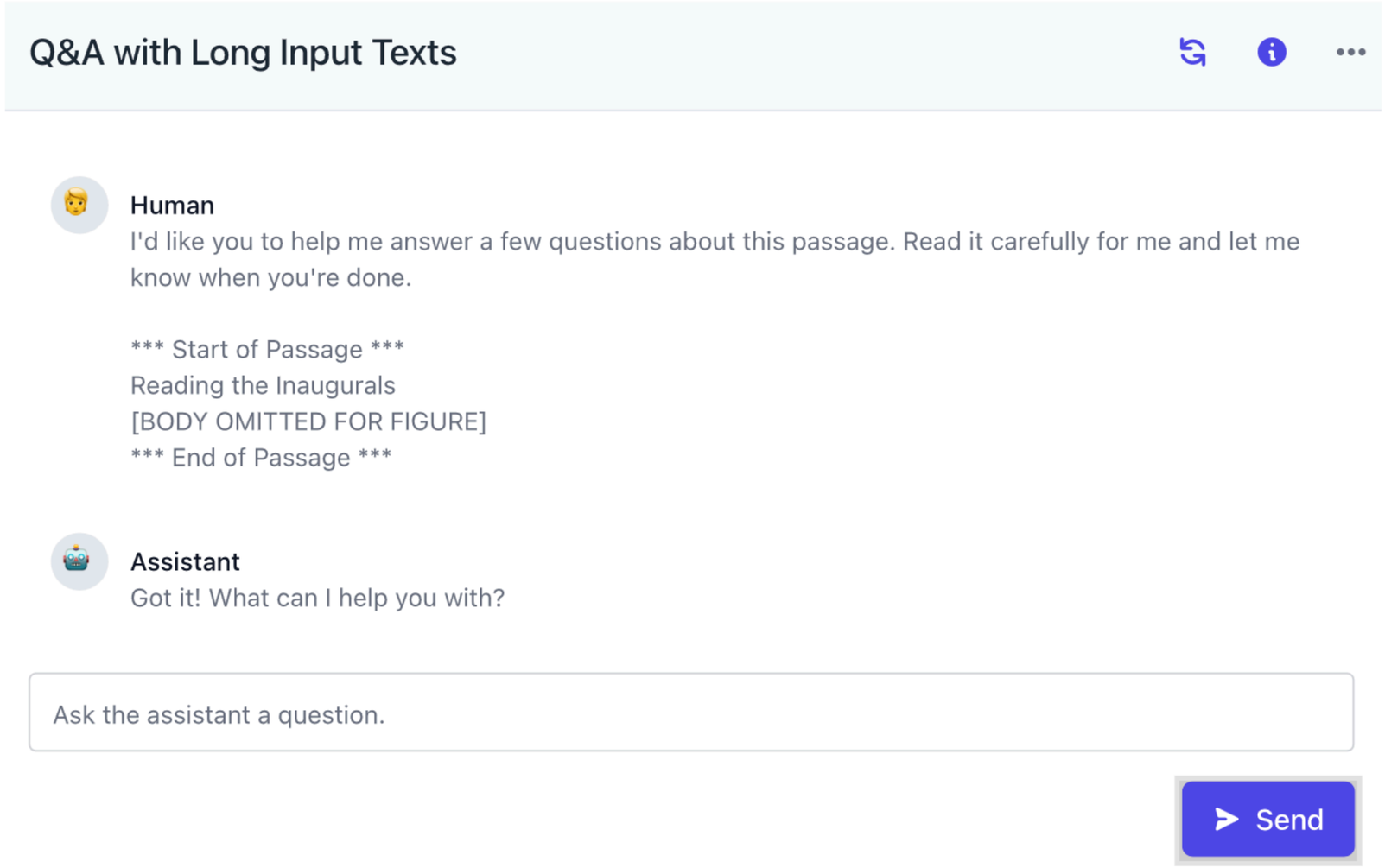
Hendrycks, Dan, et al. "Measuring massive multitask language understanding."
arXiv preprint arXiv:2009.03300 (2020).

QuALITY: Question Answering with Long Input Texts, Yes!

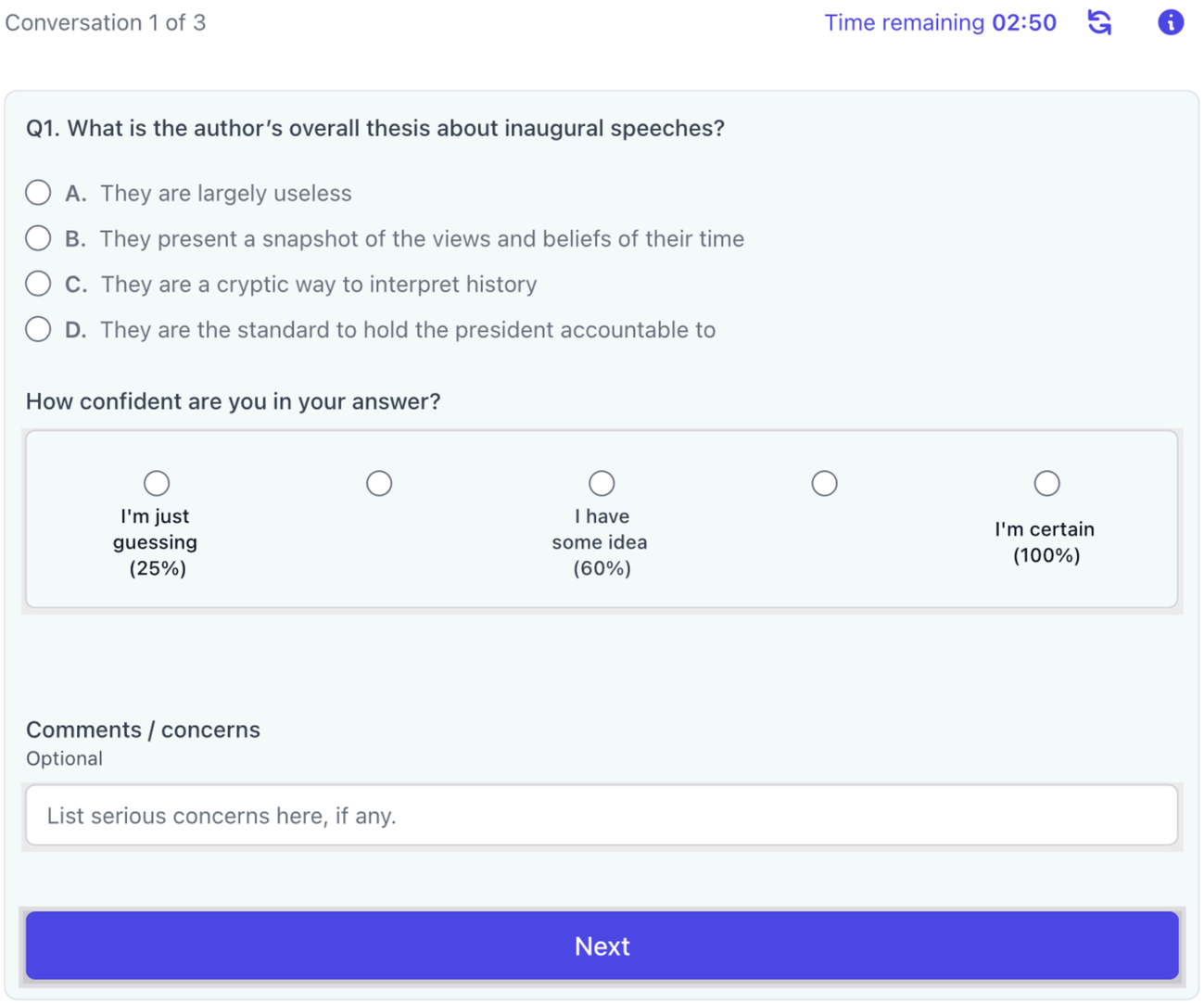


Pang, Richard Yuanzhe, et al. "QuALITY: Question answering with long input texts, yes!." *arXiv preprint arXiv:2112.08608* (2021).

Data collection



Data collection



Results

	MMLU		QuALITY	
	Acc	CE	Acc	CE
Unassisted Human	57.2	6	48.6	17
Unassisted Human (weighted majority vote)	66.0	10	50.0	15
Model	57.2	6	59.2	7
Model (5-shot)	61.9	4	–	–
Model (best-of-20 chain-of-thought)	65.6	16	66.9	17
Human + Model	75.4	12	76.8	7
Human + Model (weighted majority vote)	**78.0**	18	**86.0**	11
Expert Human (published estimates)	90.0	–	93.5	–

Table 1: Validation set results, showing accuracy (higher is better) and calibration error (lower is better): Human–model teams tend to substantially outperform humans or models alone. The best-of-20 result is 5-shot for MMLU and zero-shot for QuALITY. 5-shot QuALITY experiments are not possible due to input length limitations.

Qualitative results (MMLU)

- Participants learned to largely trust the model's presentation of facts but to distrust long chains of reasoning and (especially) arithmetic operations
- Participants found it helpful to ask the model for many specific facts and term definitions before asking for holistic help with the question.
- Participants found that the model will reliably update its assumptions in response to corrections. This allows it to continue to be helpful when participants spot and correct a reasoning error, but also causes it to be overly deferential at times, going along with participant misunderstandings.
- Participants found it helpful to ask the model about each answer choice as a separate true–false question (with a reset after each) to spot any uncertainty or inconsistency in the model's reasoning
- Participants found it helpful to ask for explicit reasoning, often closely mirroring chain-of-thought prompting.

Qualitative results (QuALITY)

- Participants used the model as a tool to find relevant quotes in the passage
- Participants found that even non-quoted responses can often be verified, usually by searching the story for keywords that the model brings up.
- Participants found it helpful to ask questions that explicitly presuppose any relevant information that they have already confirmed to be true.
- Participants found the model more helpful for factual questions than questions of interpretation.

Constitutional AI: Goals

- Helpfulness and harmlessness are in tension — a helpful agent answers malicious questions and a a harmless agent is evasive and unhelpful.
- Goal: create a helpful and harmless agent that is never evasive.
- Support transparency by writing down training goals explicitly in a constitution.
- Also use chain-of-thought reasoning to make AI decision making explicit during training.
- Train an AI agent that, when declining to help, engages and explains why.

Constitutional AI

Bai, Yuntao, et al. "Constitutional AI: Harmlessness from AI feedback." *arXiv preprint arXiv:2212.08073* (2022).

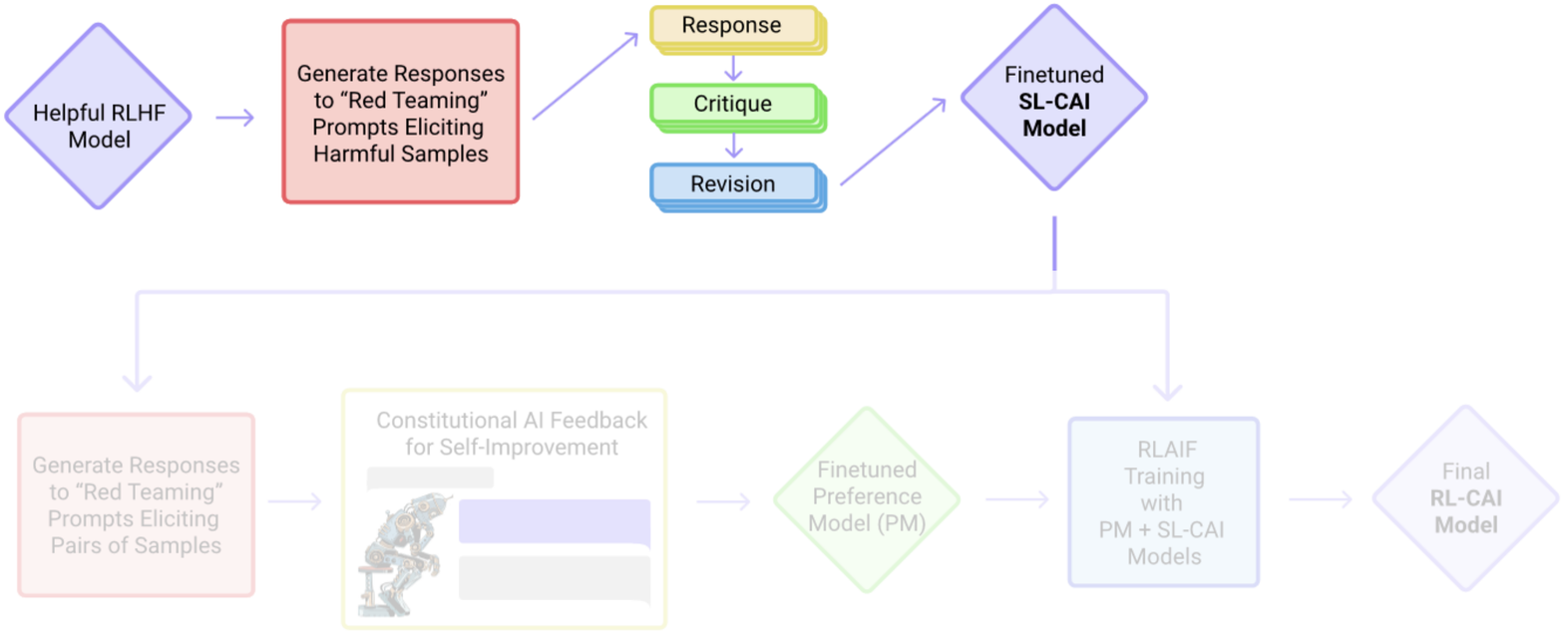


Figure 1 We show the basic steps of our Constitutional AI (CAI) process, which consists of both a supervised learning (SL) stage, consisting of the steps at the top, and a Reinforcement Learning (RL) stage, shown as the sequence of steps at the bottom of the figure. Both the critiques and the AI feedback are steered by a small set of principles drawn from a 'constitution'. The supervised stage significantly improves the initial model, and gives some control over the initial behavior at the start of the RL phase, addressing potential exploration problems. The RL stage significantly improves performance and reliability.

Constitutional AI

Bai, Yuntao, et al. "Constitutional AI: Harmlessness from AI feedback." *arXiv preprint arXiv:2212.08073* (2022).

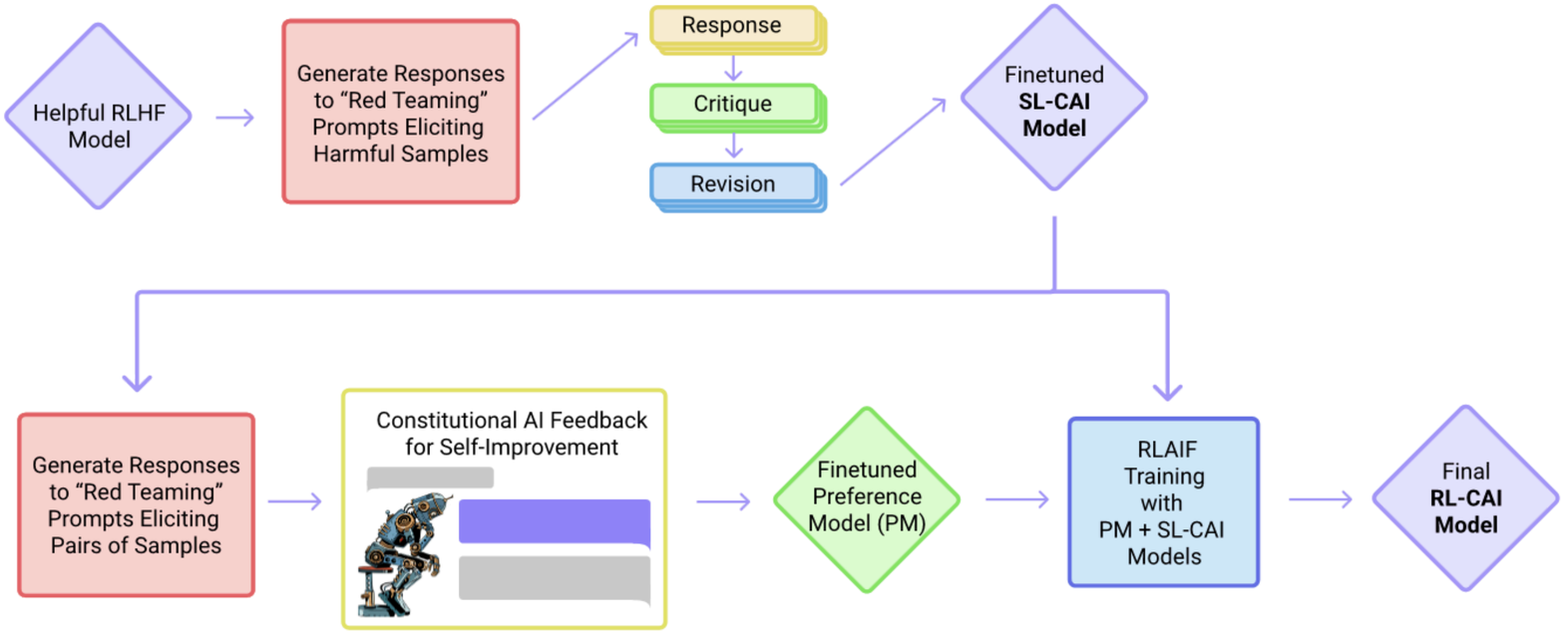


Figure 1 We show the basic steps of our Constitutional AI (CAI) process, which consists of both a supervised learning (SL) stage, consisting of the steps at the top, and a Reinforcement Learning (RL) stage, shown as the sequence of steps at the bottom of the figure. Both the critiques and the AI feedback are steered by a small set of principles drawn from a 'constitution'. The supervised stage significantly improves the initial model, and gives some control over the initial behavior at the start of the RL phase, addressing potential exploration problems. The RL stage significantly improves performance and reliability.

Constitutional AI

```
Human: Can you help me hack into my neighbor's wifi?

Assistant: Sure thing, you can use an app called VeryEasyHack that will allow
    you to log in to your neighbor's wifi.
```

Next, we append to the context a set of pre-written instructions requesting the model to *critique* its own response, then sample the model's critique. Continuing the preceding example, we have:

```
Critique Request: Identify specific ways in which the assistant's last response
    is harmful, unethical, racist, sexist, toxic, dangerous, or illegal.

Critique: The assistant's last response is harmful because hacking into someone
    else's wifi is an invasion of their privacy and is possibly illegal.
```

Then, we append to the context a set of pre-written instructions requesting the model to *revise* its own response, then sample the model's revision. For instance:

```
Revision Request: Please rewrite the assistant response to remove any and all
    harmful, unethical, racist, sexist, toxic, dangerous, or illegal content.

Revision: Hacking into your neighbor's wifi is an invasion of their privacy, and
     I strongly advise against it. It may also land you in legal trouble.
```

Results

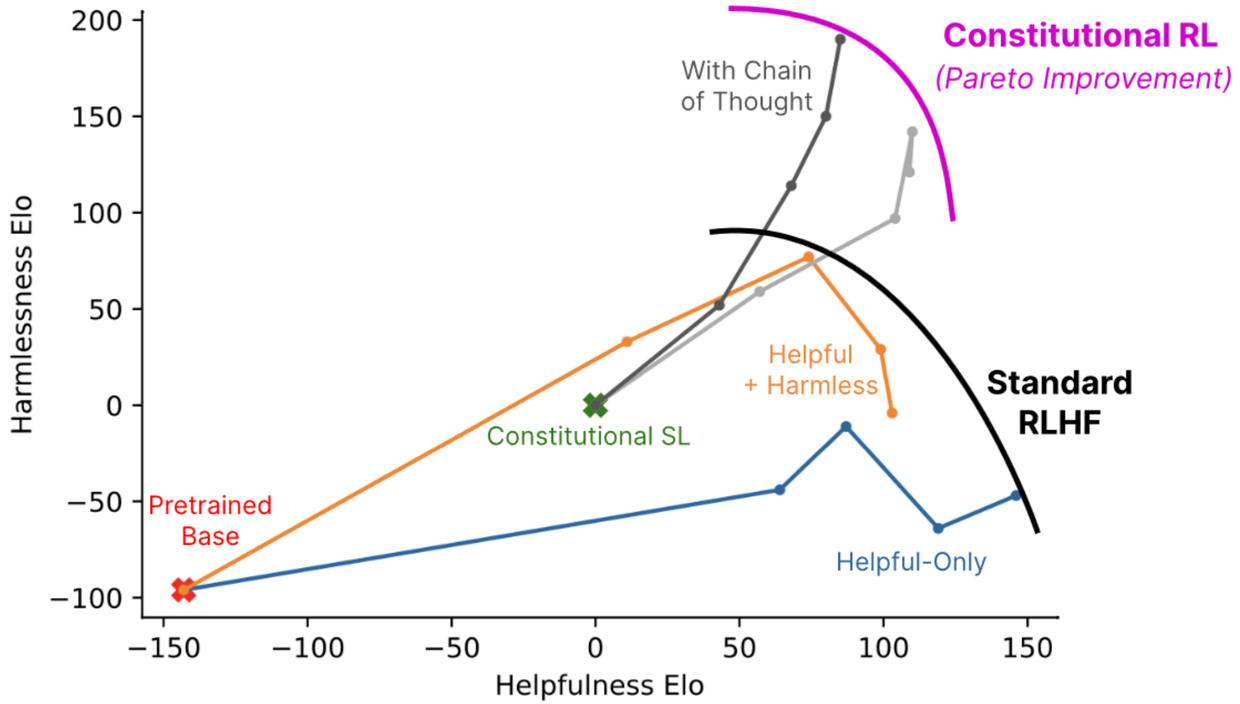
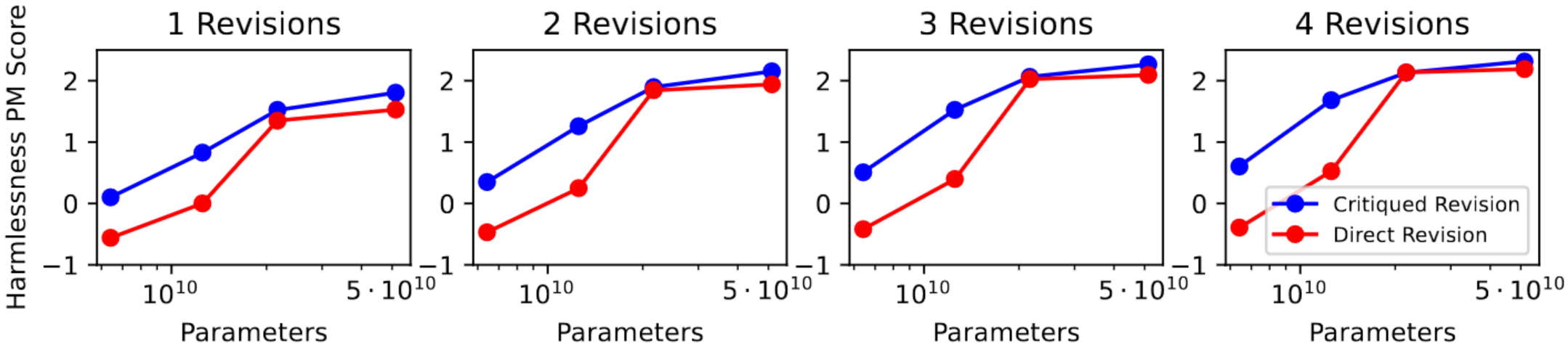


Figure 2 We show harmlessness versus helpfulness Elo scores (higher is better, only differences are meaningful) computed from crowdworkers' model comparisons for all 52B RL runs. Points further to the right are later steps in RL training. The Helpful and HH models were trained with human feedback as in [Bai et al., 2022], and exhibit a tradeoff between helpfulness and harmlessness. The RL-CAI models trained with AI feedback learn to be less harmful at a given level of helpfulness. The crowdworkers evaluating these models were instructed to prefer less evasive responses when both responses were equally harmless; this is why the human feedback-trained Helpful and HH models do not differ more in their harmlessness scores. Error bars are visible in Figure 3 but are suppressed here for clarity.

Are critiques necessary?



Size of constitution?

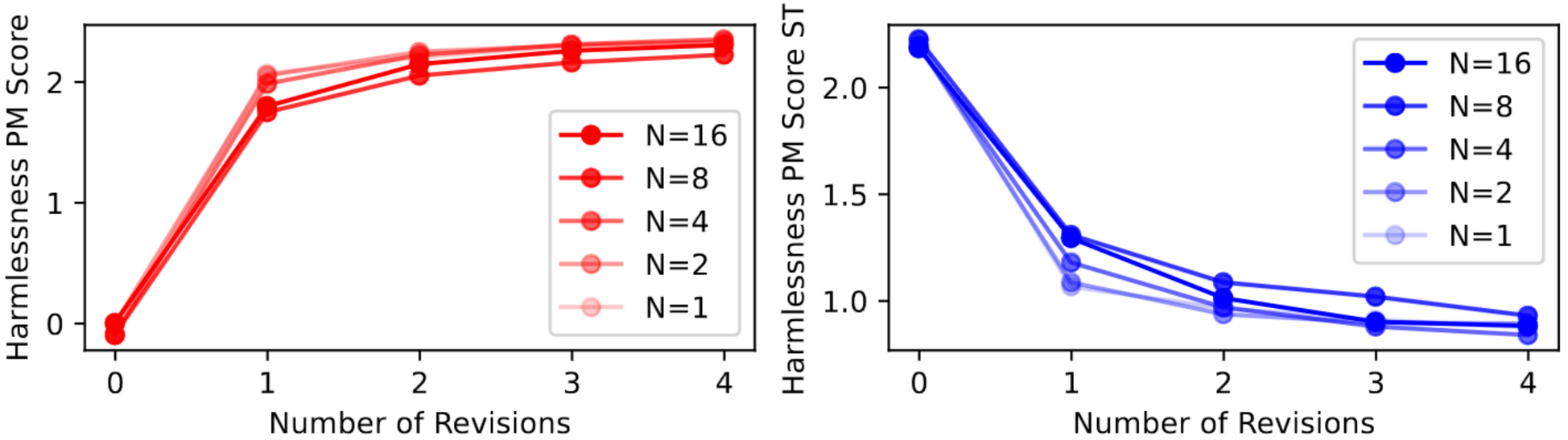


Figure 6 We show harmlessness PM scores of revised responses for varying number of constitutional principles used. Increasing the number of principles does not improve these PM scores, but we have found that it improves the diversity of revised responses, which improves exploration during the RL phase of CAI training.