

Hierarchical Monitoring

(Ongoing work) (Correction added May 12)

Tim Hua^{1,*} James Baskerville¹ Mia Hopman¹ Henri Lemoine¹ Tyler Tracy^{1,2}

¹MARS

²Redwood Research

*Corresponding author: timhua0@gmail.com

Control Conf March 2025

Have two trusted monitors instead of one

Hierarchical monitoring is a control protocol with two trusted monitors (M, N) instead of one. We typically assume that

- Monitor M is expensive and accurate
- Monitor N is cheaper and less accurate

Two contributions: Theoretical Framework and Empirical Results

We make two contributions to the literature

Two contributions: Theoretical Framework and Empirical Results

We make two contributions to the literature

- ① We develop a simple mathematical framework for monitoring in AI control.
 - We can apply this framework anytime we use trusted monitors, including single-monitor settings.

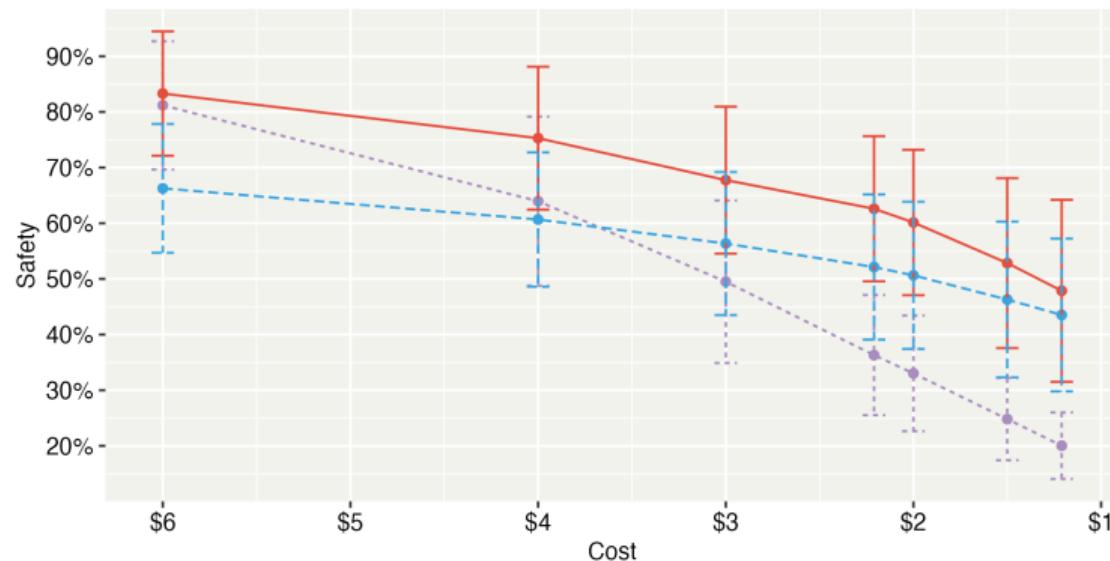
Two contributions: Theoretical Framework and Empirical Results

We make two contributions to the literature

- ① We develop a simple mathematical framework for monitoring in AI control.
 - We can apply this framework anytime we use trusted monitors, including single-monitor settings.
- ② We apply this framework to hierarchical monitoring, and show that it leads to higher safety at the same budget compared to using only one monitor.
 - We use Llama 3.1 8B and Llama 3.1 70B to monitor code outputs from o1-mini.

Hierarchical monitoring tend to have higher safety at the same cost (new ses)

Hierarchical monitoring is safer at the same cost



Policy

Combining both monitors

Llama 3.1 70B only

Llama 3.1 8B only

After fixing the standard errors on the previous slide, it seems as if the three lines are statistically indistinguishable under 95% confidence intervals. This is true in the sense that their means are, but their difference (i.e., how much better the Two monitors system does) are quite a bit more distinguishable, although still not always significantly different from zero.



The old ones were wrong because instead of calculating the bootstrap standard error as

$$\hat{SE}_{boot}(\hat{\theta}) = \sqrt{\frac{1}{B} \sum_{i=1}^B (\theta_i - \bar{\theta})}$$

I was treating each bootstrap as an independent observation (OH MAN THAT WAS SILLY) and calculated

$$\hat{SE}_{stupid}(\hat{\theta}) = B^{-0.5} \sqrt{\frac{1}{B} \sum_{i=1}^B (\theta_i - \bar{\theta})}$$

It doesn't really change the final conclusion though, which is that you should use two monitors because it will do no worse than using only one and sometimes can do better.

Framework sketch: Maximize safety subject to a cost constraint

- **Safety:** We want to maximize the chance of auditing/“catching” a misaligned output from an untrusted model.
 - Following the original control paper, we’ll assume access to human auditors to catch misaligned models.

Framework sketch: Maximize safety subject to a cost constraint

- **Safety:** We want to maximize the chance of auditing/“catching” a misaligned output from an untrusted model.
 - Following the original control paper, we’ll assume access to human auditors to catch misaligned models.
- **Cost Constraint:** Spend no more than Y dollars per aligned output.

Framework sketch: Maximize safety subject to a cost constraint

- **Safety:** We want to maximize the chance of auditing/“catching” a misaligned output from an untrusted model.
 - Following the original control paper, we’ll assume access to human auditors to catch misaligned models.
- **Cost Constraint:** Spend no more than Y dollars per aligned output.
- We measure **safety** on **misaligned outputs** while enforcing **budget constraints** on **aligned** outputs
 - Think of the cost as how much extra you’ll pay if nothing bad happens, while safety is how often you’d be able to catch a misaligned output from your model (if it is indeed misaligned).

Formal Framework

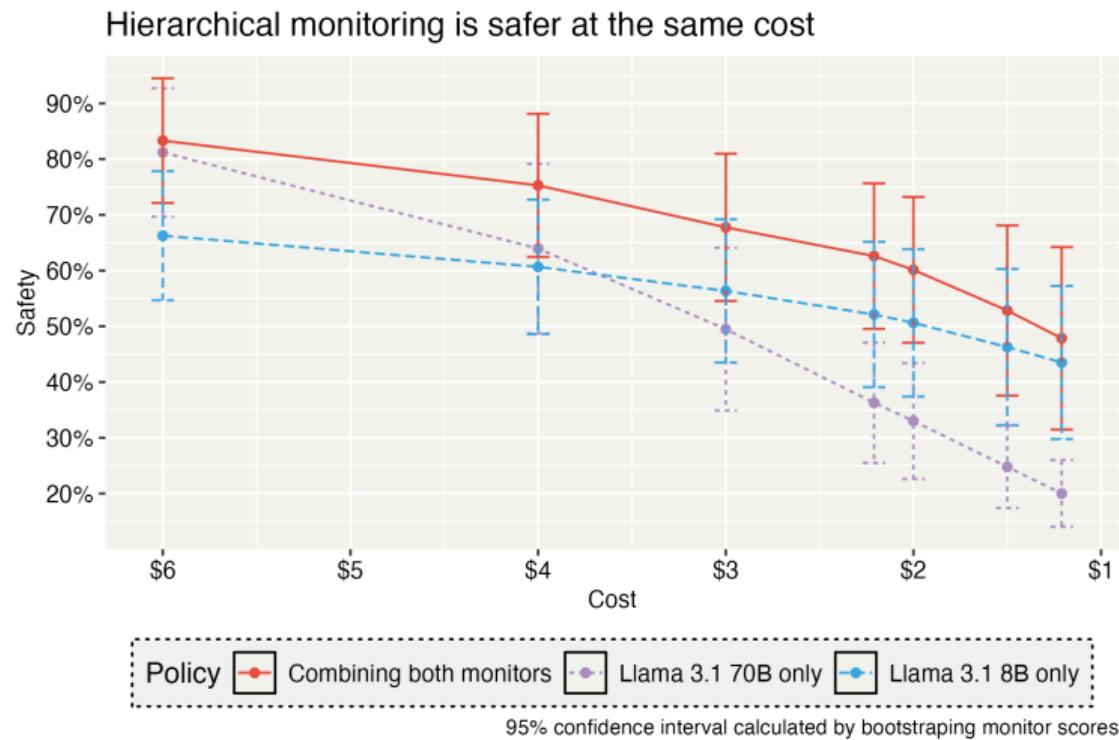
We consider i.i.d. draws from some untrusted model. In each episode, we want to maximize

$$\max P(\text{Audit} | \text{Misaligned Output})$$

such that

$$Y \geq E[\text{Monitor Costs} | \text{Aligned Output}] + \underbrace{P(\text{Audit} | \text{Aligned Output}) \times \text{Cost per Audit}}_{\text{Audit False Positive Rate}}$$

Now I will show you how we applied the framework in this graph



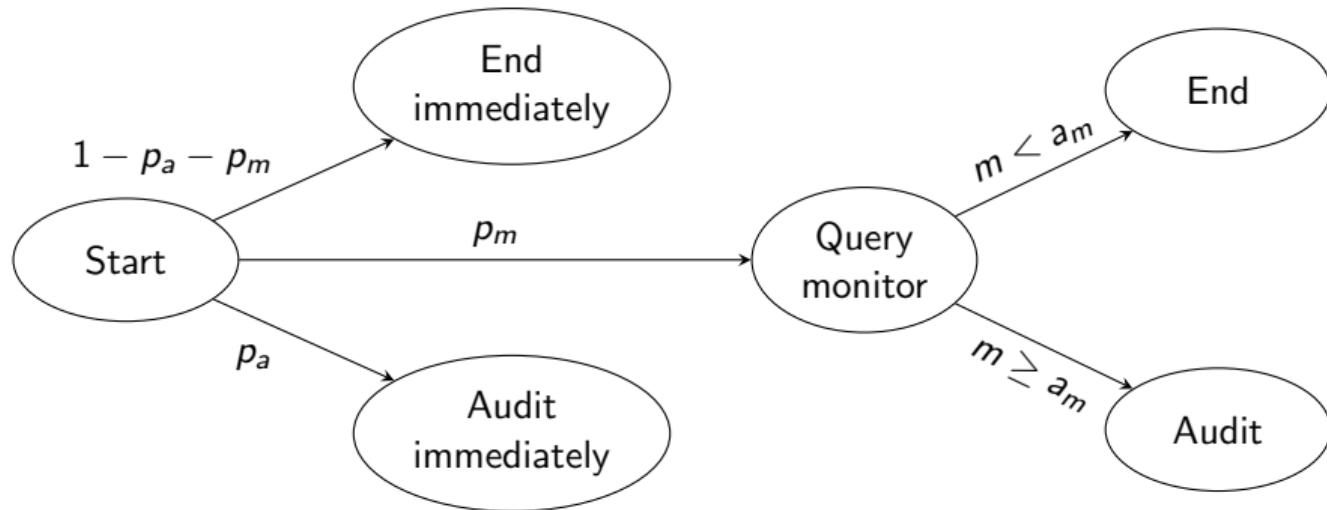
Using our framework, we improve upon baseline trusted monitoring

- In the original trusted monitoring setup in Greenblatt et al. (2023), they call a single monitor and audit if the score is above some threshold.

Using our framework, we improve upon baseline trusted monitoring

- In the original trusted monitoring setup in Greenblatt et al. (2023), they call a single monitor and audit if the score is above some threshold.
- We improve on this baseline since calling the monitor 100% of the time is not always optimal.

Auditing/Ending without calling the monitor at all



Auditing/Ending without calling the monitor at all

- The only thing you need to understand about these equations is that they represent the chart on the previous slide, and we can easily solve them to maximize safety while remaining under our budget constraint.

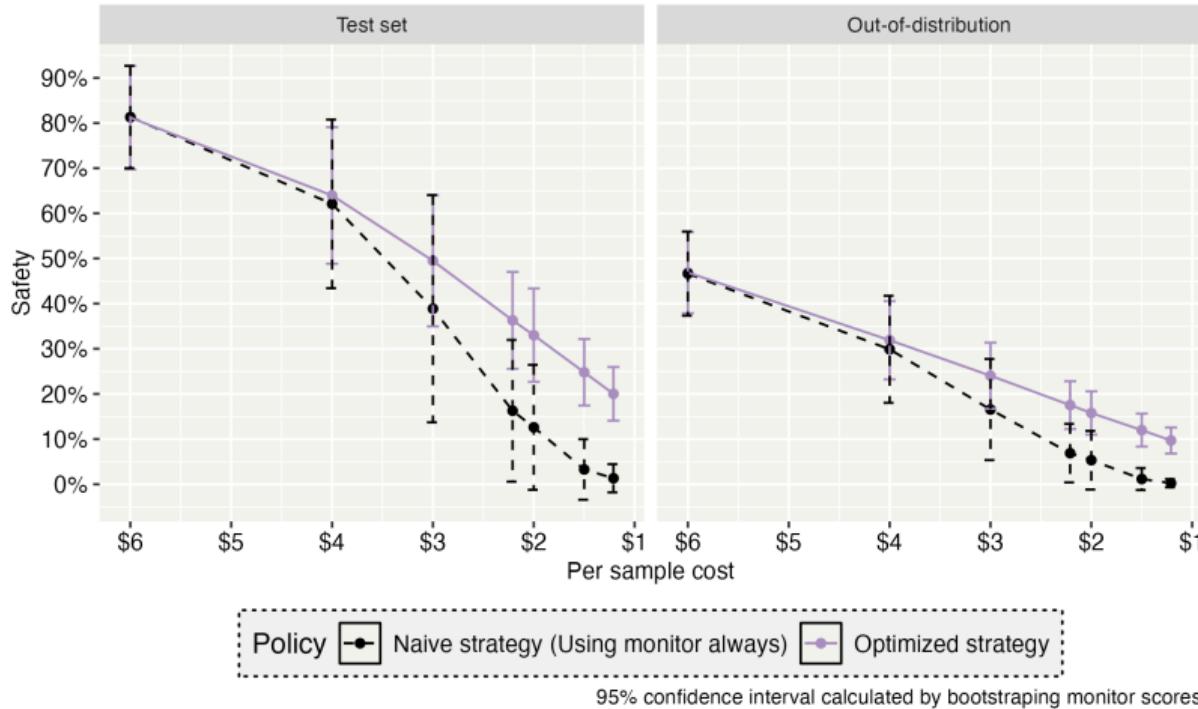
$$\max_{p_a, p_m, a_m} P(\text{Audit} \mid \text{Misaligned Output}) = p_m \cdot P(m \geq a_m \mid \text{Misaligned Output}) + p_a$$

such that

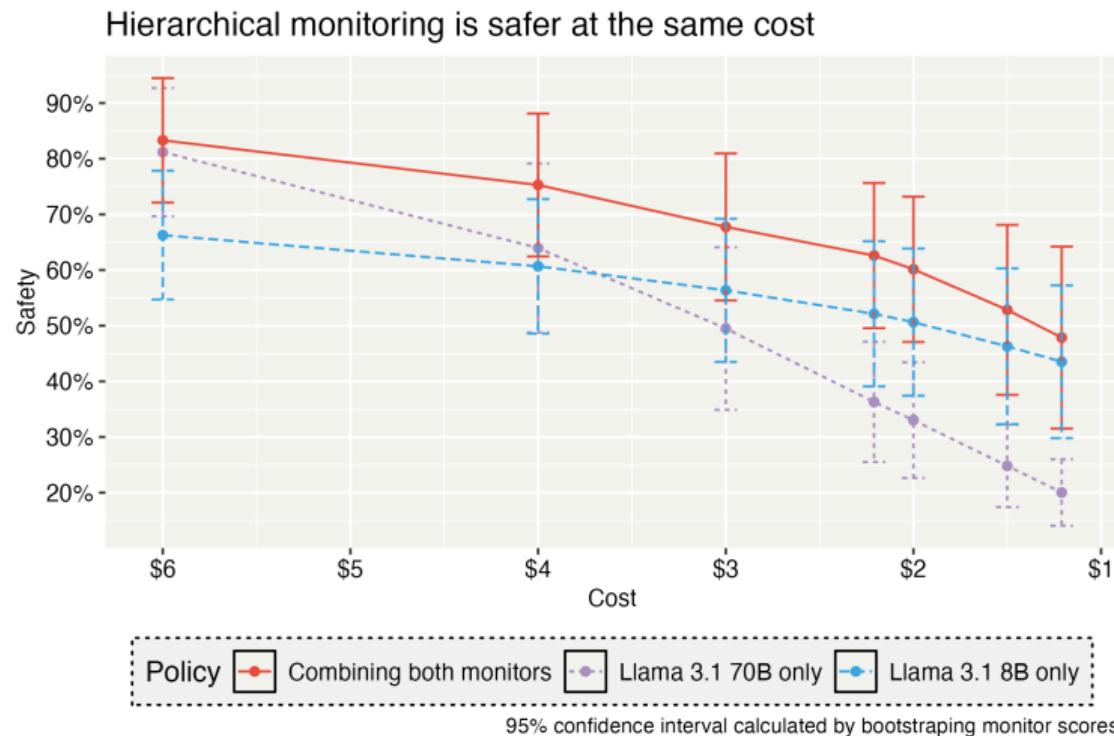
$$Y \geq p_m(c_m + P(m \geq a_m \mid \text{Aligned Output}) \cdot c_a) + p_a \cdot c_a$$

Optimizing Llama 3.1 70B single monitor performance

Assuming auditing costs \$100 per sample, Llama 3.1 70B costs \$1



Now I'll explain how we calculated performance with two monitors



It's much harder to search through the possible policy space in a two-monitor setting

- There are a lot of choices:
 - (e.g., which monitor to use first, when to defer to second monitor, when to audit immediately)

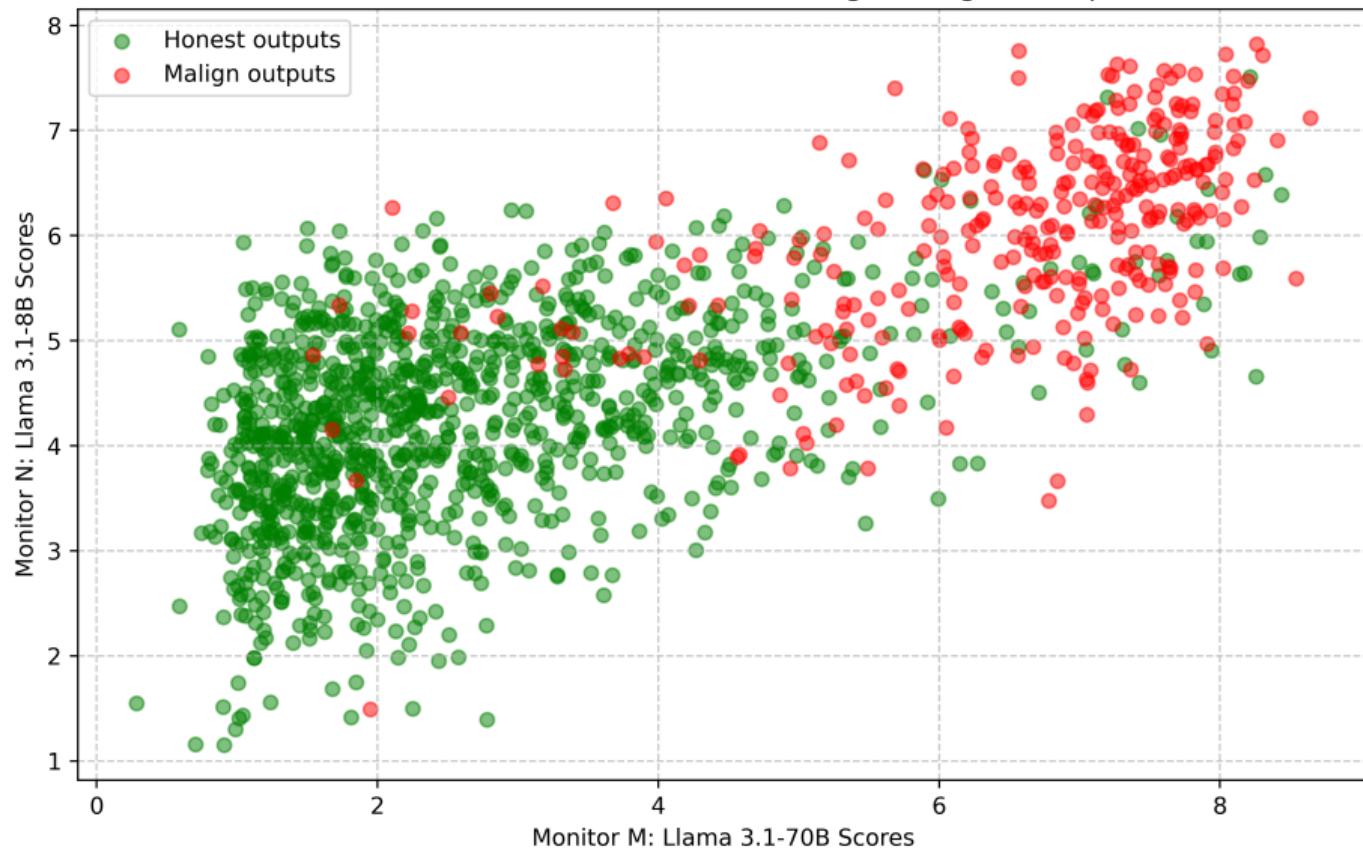
It's much harder to search through the possible policy space in a two-monitor setting

- There are a lot of choices:
 - (e.g., which monitor to use first, when to defer to second monitor, when to audit immediately)
- We tried and failed to solve for the whole system analytically or numerically

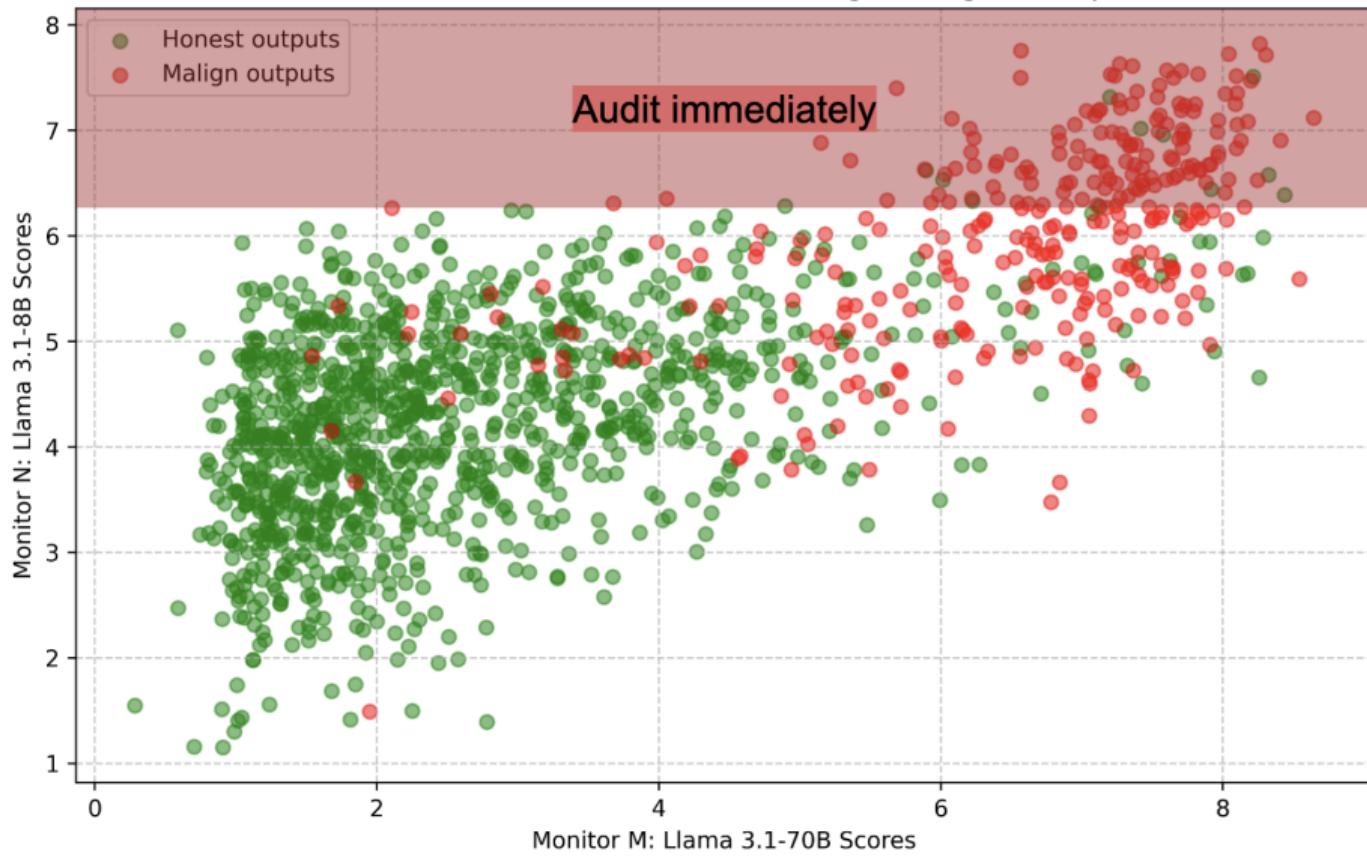
It's much harder to search through the possible policy space in a two-monitor setting

- There are a lot of choices:
 - (e.g., which monitor to use first, when to defer to second monitor, when to audit immediately)
- We tried and failed to solve for the whole system analytically or numerically
- Instead, we restrict ourselves to a subset of policies with fewer free parameters. I will show you the strategy we used for the previous slide; it has only two free parameters which we sweep over.

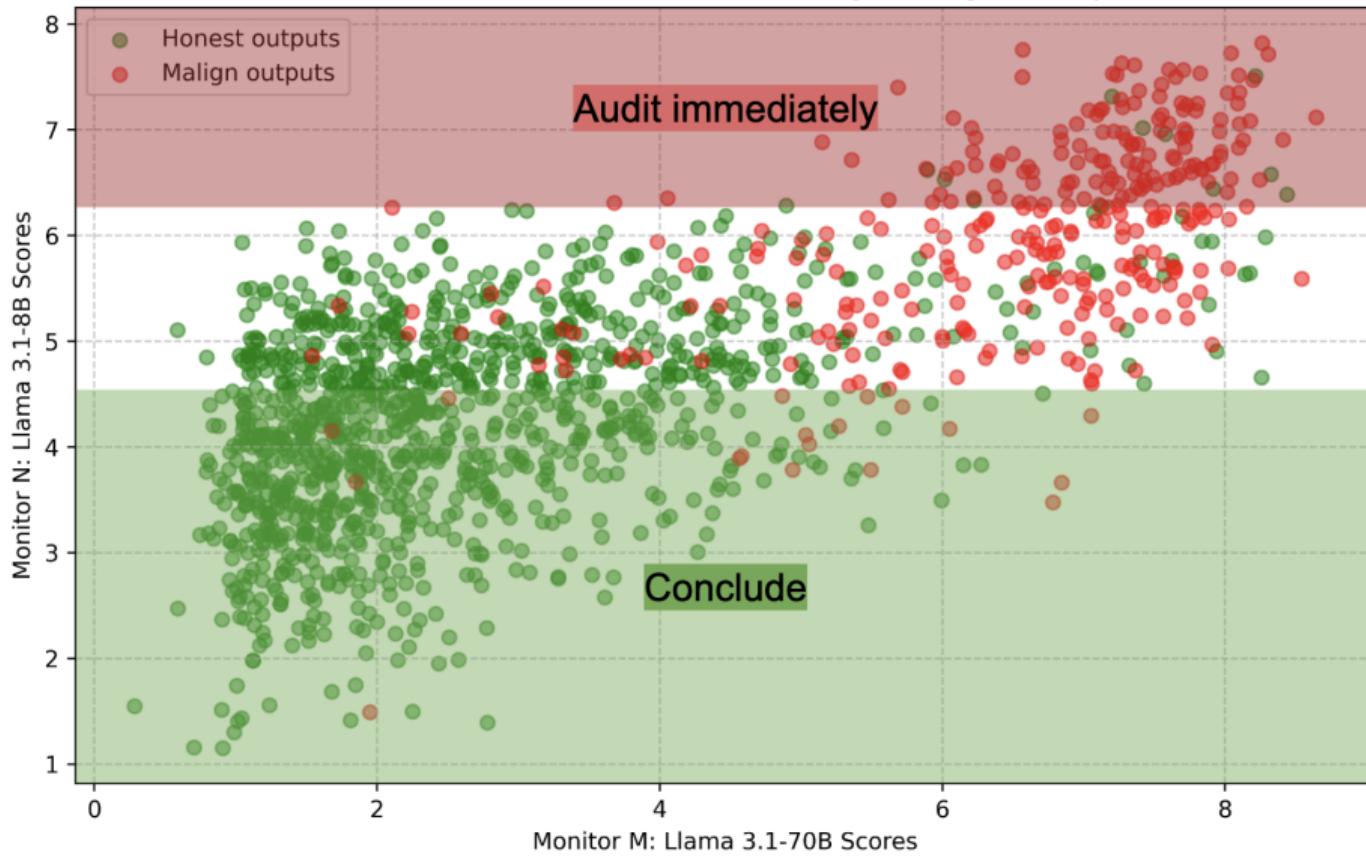
Scatter Plot of Monitor Scores for Misaligned/Aligned Outputs



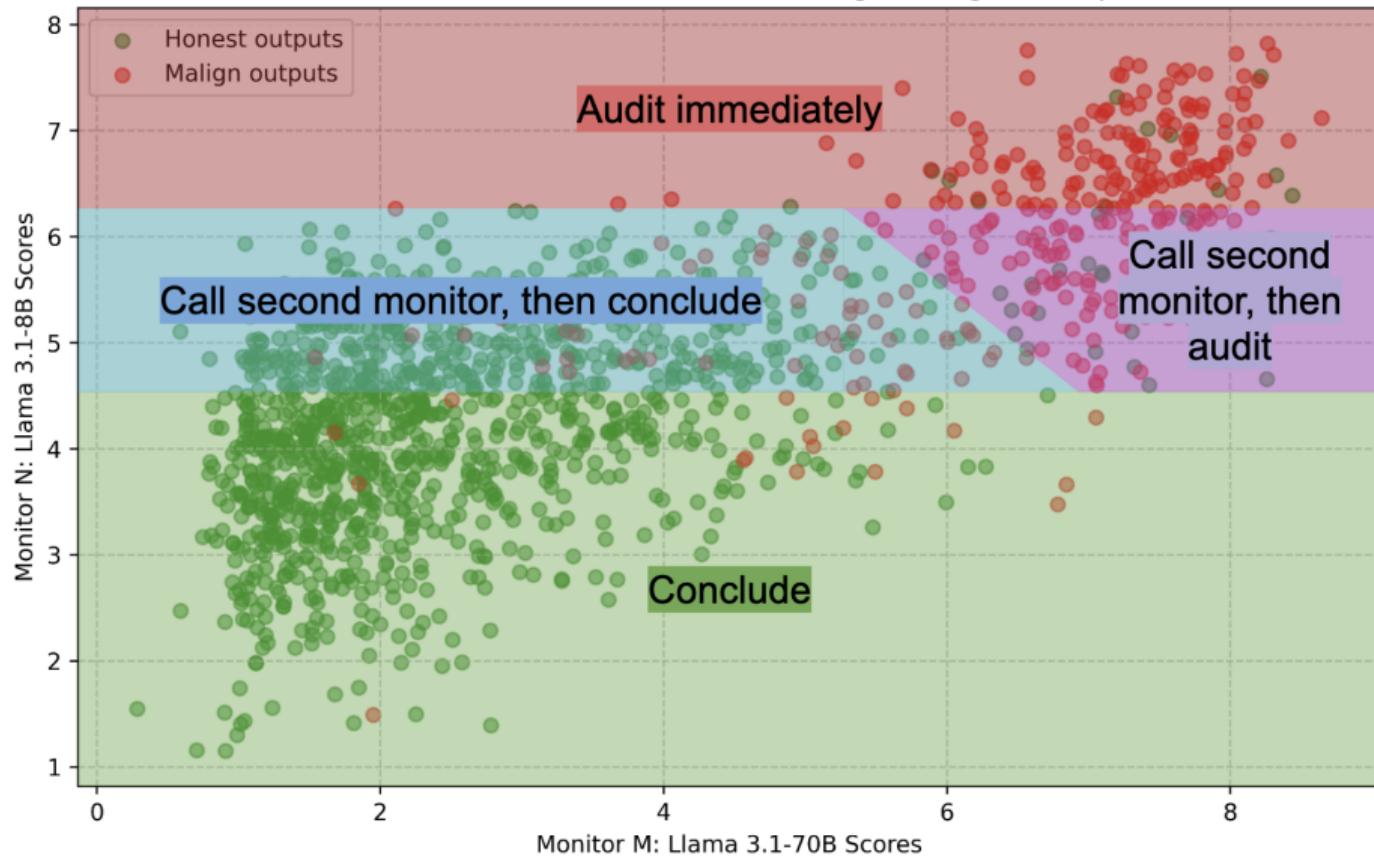
Scatter Plot of Monitor Scores for Misaligned/Aligned Outputs



Scatter Plot of Monitor Scores for Misaligned/Aligned Outputs



Scatter Plot of Monitor Scores for Misaligned/Aligned Outputs



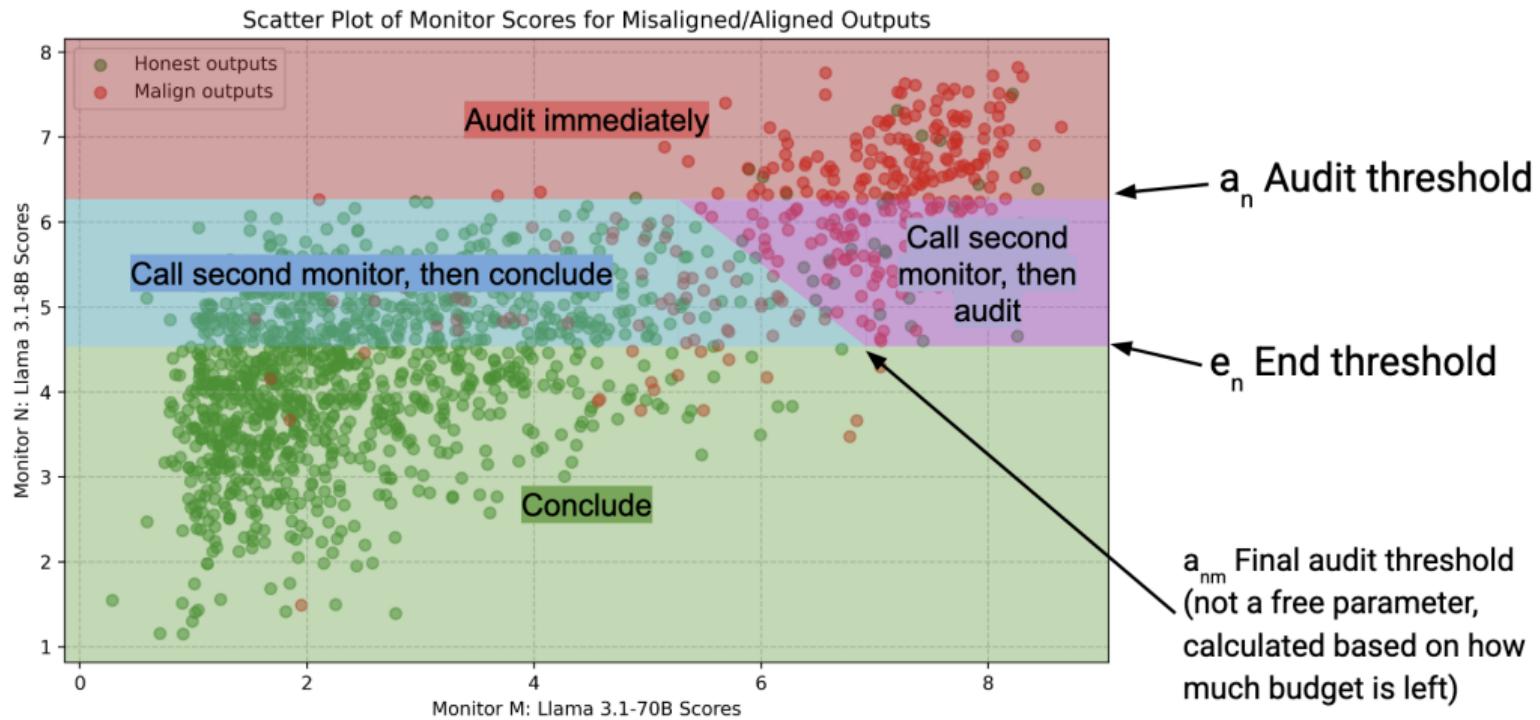
General Procedure for Finding and Evaluating a Policy

- ① We first pick some possible way to use the two monitors.

General Procedure for Finding and Evaluating a Policy

- ① We first pick some possible way to use the two monitors.
- ② We fit our free parameters to optimize safety on some training dataset of monitor scores.

Previous strategy has two free parameters a_n and e_n

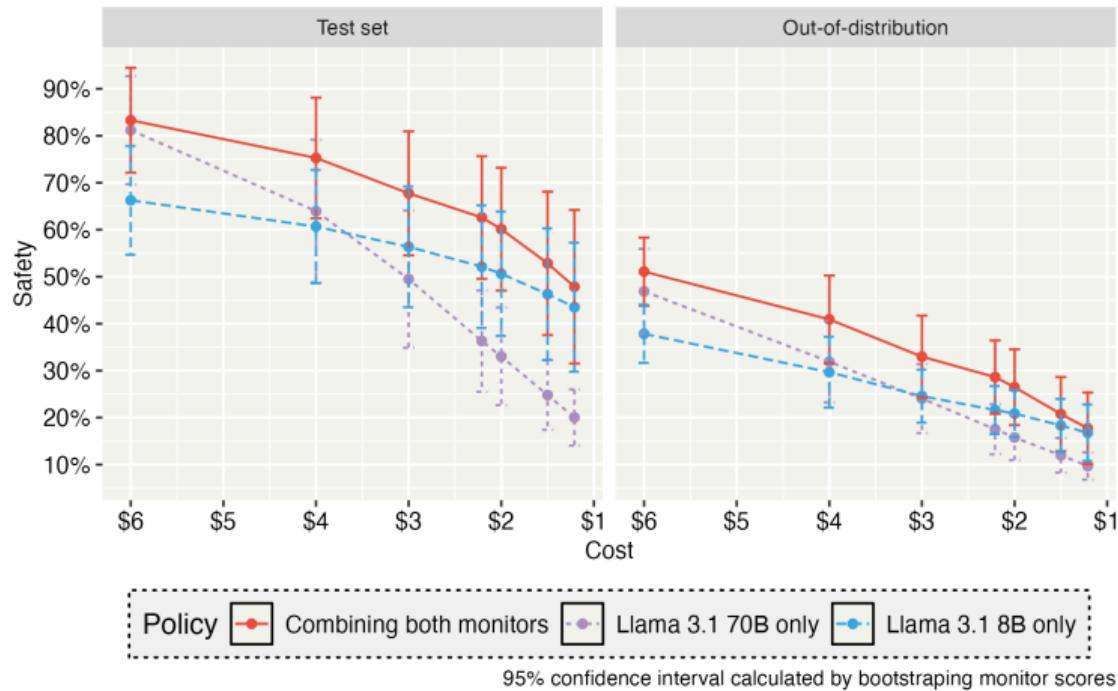


General Procedure for Finding and Evaluating a Policy

- ① We first pick some possible way to use the two monitors.
- ② We fit our free parameters to optimize safety on some training dataset of monitor scores.
- ③ We evaluate the strategy's safety on some test dataset of misaligned outputs and an out of distribution set of misaligned outputs.

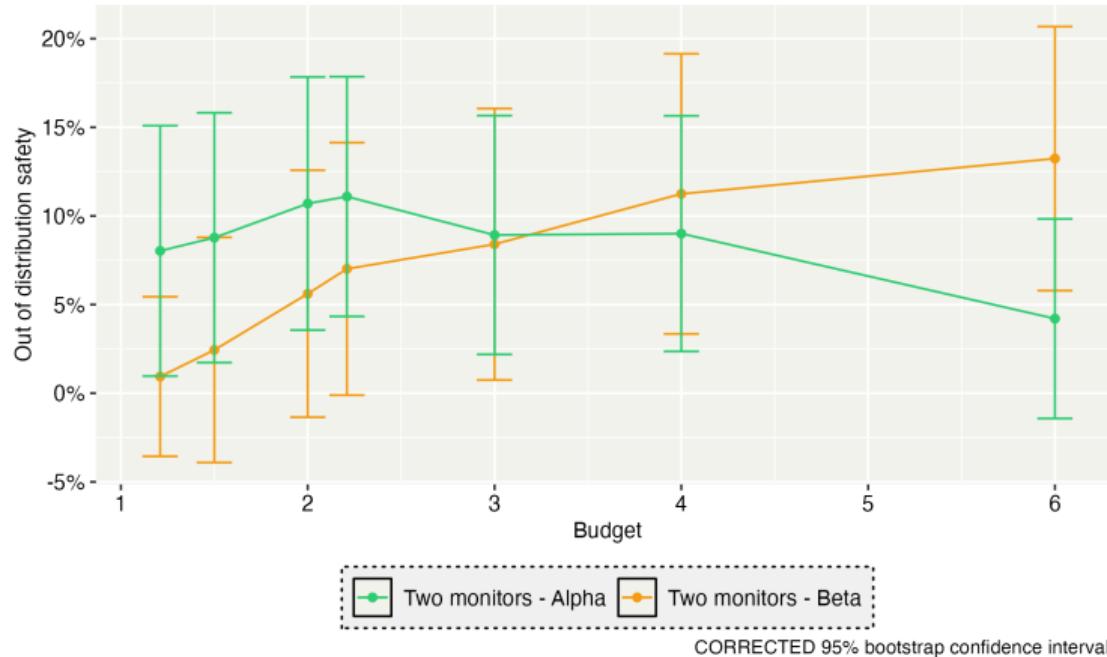
Hierarchical monitoring generally works better out of distribution

Testing our policy on i.i.d. and out of distribution set



Here's the version with differences instead of raw means

Percentage point improvement in safety in using two monitors (OOD)
Relative to using only Alpha/Beta monitor. Higher is better



Come to our poster session later today!

Chat about open questions

- What costs parameters should we experiment on?
- What other real/simulated monitor scores should we look at?

Learn more results

- How we think about extension to defer to trusted, not just auditing
- Binary Monitor Strategy
- Sweeping over audit costs

Thanks for listening

- Slides can be found at timhua.me/control_slides.pdf (link also in my swapcard bio).

Thanks for listening

- Slides can be found at timhua.me/control_slides.pdf (link also in my swapcard bio).
- Hey I just gave a talk, and this is crazy, but I don't have a job, so hire me maybe.