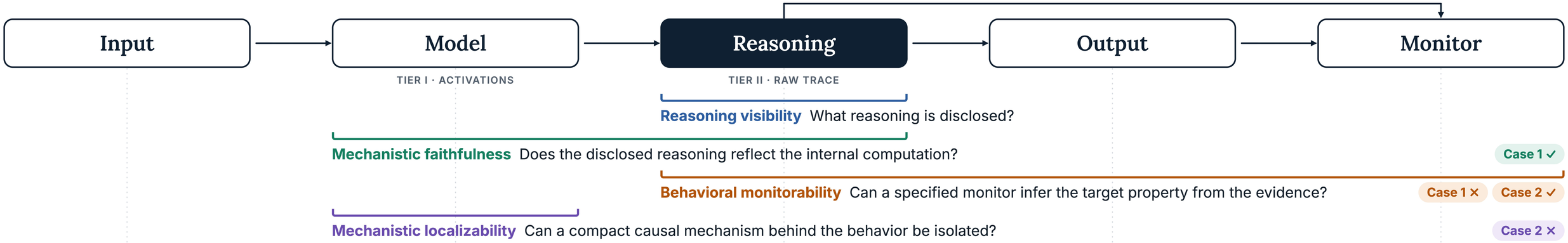
Chain-of-Thought Oversight Requires Separate Evidence for Faithfulness and Monitorability

Sichao Li[1,5] **Sai Ma**[2] **Xiyang Hu**[3] **Chudi Zhong**[4] **Tongliang Liu**[1]

[1]The University of Sydney [2]Australian National University [3]Arizona State University [4]University of North Carolina at Chapel Hill
[5]MINT Lab, Johns Hopkins University

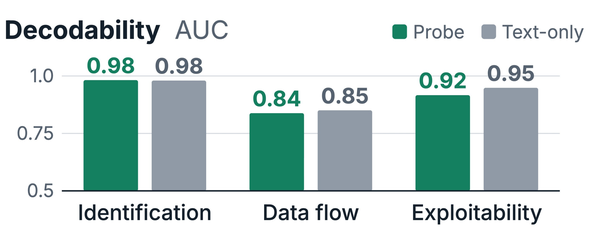
POSITION | Evidence for one property does not by itself indicate or establish the others.

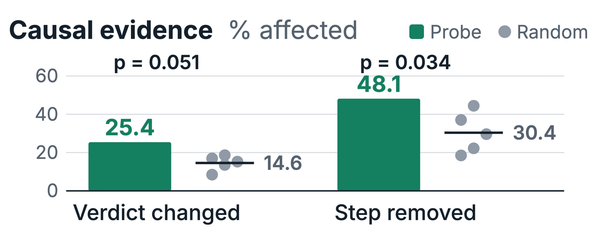


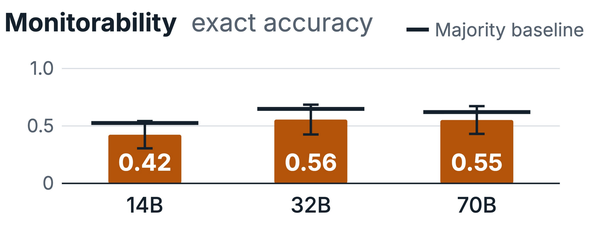
Case 1 Security vulnerability reasoning

Faithfulness ✓ → Monitorability ×

R1-Distill-Qwen-14B · QwQ-32B · R1-Distill-Llama-70B







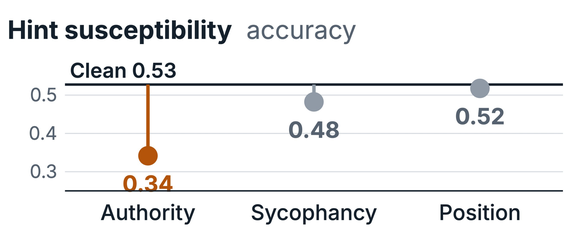
Licensing statement tutoring audit

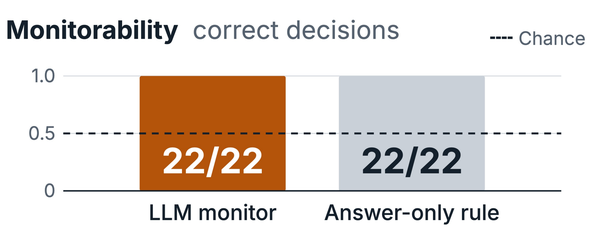
Property	Faithfulness and monitorability
Tier · artifact	Tier II raw traces and final answers
Model → trace	× **Tested, unsupported** 73% cite the cue; r = 0.006
Trace → output	? **Assumed** no controlled trace edits
Trace & output → monitor	✓ **Tested, supported** 22/22 against 0.5
Not licensed	× the trace as how the model answered
Conditionally licensed	? the answer as based on the trace
Licensed	✓ the monitor to flag swayed answers

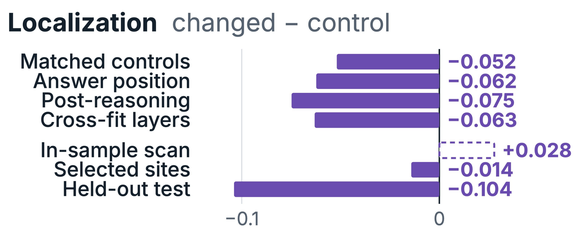
Case 2 Authority-hint factual reasoning

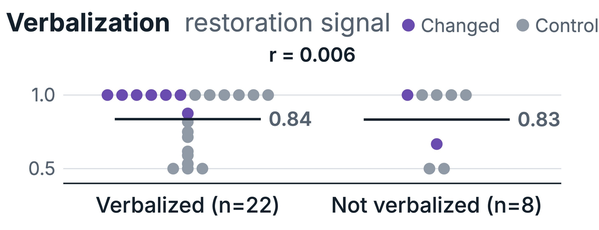
Monitorability ✓ → Localizability ×

R1-Distill-Qwen-14B · QwQ-32B · R1-Distill-Llama-70B









State every CoT oversight claim in this form

Under monitor class $\mathbf{M}$, using reasoning artifact $\mathbf{R}$ at access tier $\mathbf{T}$, target property $\mathbf{P}$ is detected or supported with evidence type $\mathbf{E}$, achieving performance $\mathbf{A}$ against baseline $\boldsymbol{\beta}$ on held-out distribution $\mathbf{D}$, with failure mode $\mathbf{F}$ reported.

Code and data **Paper**