Learning Structured Trust Policies from Uncertainty, Advisor Signals, and Agreement

Gia Ancone [1] Jaden Chen [1]

[1]Department of Computer Science, Stanford University

Motivation and Background

When should an agent trust external advice? RL agents and humans alike must routinely act on advice of unknown reliability. When an RL agent consults peers, such as LLM advisor ensembles, competence is heterogeneous and not visible at deployment. Existing frameworks either assume reliability away, assume a prior over it, or hand-code single-signal trust rules that fail when correlated/adversarially manipulated:

- Self-uncertainty (Kadavath 2022 [1]) supports introspection, not a downstream trust policy.
- Advisor tone (Mielke 2022 [2], Sharma 2023 [3]) is speaker-side, not listener-side.
- Learning to defer (Madras 2018 [4], Mozannar 2020 [5]) is supervised and binary, with no per-instance advisor signals at inference.
- Action advising RL (Torrey 2013 [6], Subramanian 2023 [7]) decides when to advise; the student's trust is hand-coded.

Instead, we propose a PPO-trained trust policy that combines multiple signals with running observations of advisor behavior to make a per-step trust decision across a multistep episode. We focus on signals that mirror the cues humans use in real-world decision-making: **self-uncertainty, advisor reasoning quality, advisor confidence, and agent-advisor agreement**. We test whether our policy infers reliability r online and whether that inference survives distribution shift.

RL Pipeline

Data. 50,000 synthetic 4-choice questions across 3 reasoning families (arithmetic, probability, and Pearl-style causal/counterfactual [8]) $\times$ 3 difficulties, giving a 40K-train / 10K-test split. Each episode samples 10 tasks under one hidden reliability r.

Per-step pipeline ($T = 10$ steps per episode):

1. **Probe ($\hat{p}$).** A frozen MLP that we trained maps the agent's answer distribution $p_{ag} \in \mathbb{R}^4$ (plus max p_{ag}, $H(p_{ag})$, and task-category one-hot) to a calibrated own-correctness estimate $\hat{p} \in [0, 1]$.
2. **Advisor cues.** A scripted advisor emits an answer, tone (confident vs. hedged), and a reasoning-quality flag.
3. **State assembly (s_t, 8 dims).**
   - *per-task signals (4):* $\hat{p}$, advisor tone, reasoning, agreement
   - *running statistics (4):* step progress, consult rate, observed advisor accuracy, agreement rate
4. **Action.** PPO actor outputs $a_t \in \{\text{trust own}, \text{trust advisor}\}$; critic outputs $V(s_t)$.
5. **Reward.** $R_t = \mathbb{1}[\text{final answer correct}]$; running statistics update for s_{t+1}.

Per-update training (every 16 episodes = 160 transitions):

6. **Rollout buffer.** Steps 1–5 repeat for 10 steps $\times$ 16 episodes under freshly sampled reliabilities r, populating the buffer with $(s_t, a_t, R_t, V(s_t), \log \pi(a_t \mid s_t))$.
7. **PPO update.** See *Technical Specifications*.

Technical Specifications

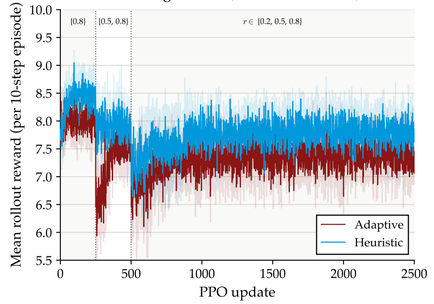
PPO actor–critic. Standard clipped surrogate + GAE(λ=0.95) + entropy bonus. Two-layer tanh MLPs (hidden = 64) for actor and critic. 16 episodes $\times$ 10 steps per rollout, 4 PPO epochs per update, 2500 updates, 3 seeds.

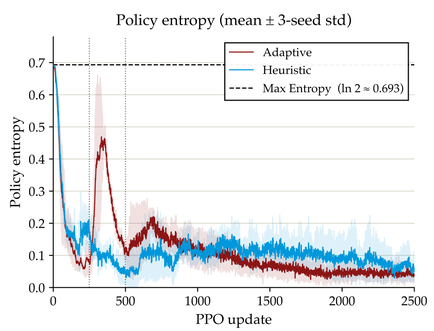
$$L^{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_t\Big[\min(w_t \hat{A}_t, \text{clip}(w_t, 1-\epsilon, 1+\epsilon)\hat{A}_t)\Big]$$

Two policy variants. Both use the same back-loaded curriculum: full reliability diversity introduced only in the final phase. The training progression is 10% / 10% / 80% of updates at $r = 0.8 \rightarrow r \in \{0.5, 0.8\} \rightarrow r \in \{0.2, 0.5, 0.8\}$.

- **Heuristic Variant**: Vanilla PPO, mild early-consult shaping only. Converges to a near always-trust-self rule, which is rewarding due to the statistical layout of the dataset.
- **Adaptive Variant**: Same as heuristic, plus:
  - `force_first_disagree`: override on the first disagreement step so consult outcomes are sampled during training.
  - `advisor_save_bonus`= 0.7: per-step training bonus when consulting flips a wrong answer to a right one.

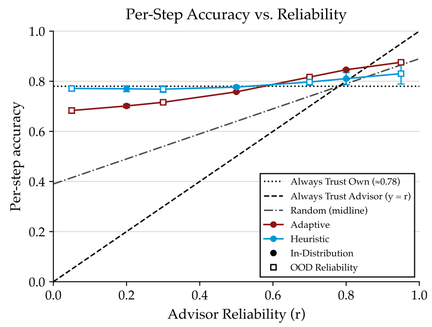
Training



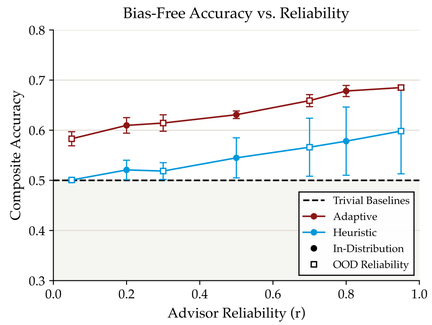


Both policies converge stably across 3 seeds. Heuristic commits to its decision rule earlier while Adaptive takes longer to settle on its multi-signal rule. Note the dips in rewards as new advisor reliability levels are introduced.

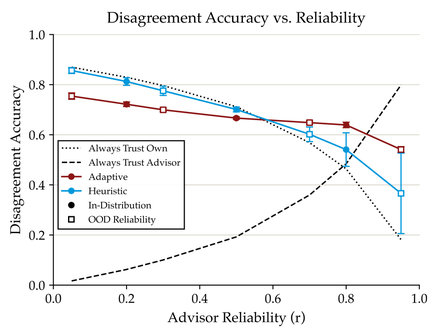
Results



Per-step accuracy. Heuristic hugs the always-trust-own baseline (data statistics hack, not robust). Adaptive's slope crosses below at r=0.05 but *above* the trivial baseline by +9.6 pp at r=0.95, showing non-trivial OOD extrapolation.

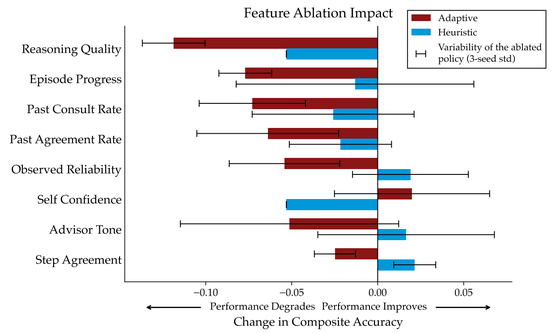


Bias-free accuracy. Composite = (loyalty + deferral)/2, where loyalty = $P(\text{trust own} \mid \text{self right, advisor wrong})$ and deferral = $P(\text{trust advisor} \mid \text{self wrong, advisor right})$. Adaptive consistently clears the trivial 0.5 floor, and Heuristic's seed std balloons to **17**$\times$ Adaptive's at r=0.95.

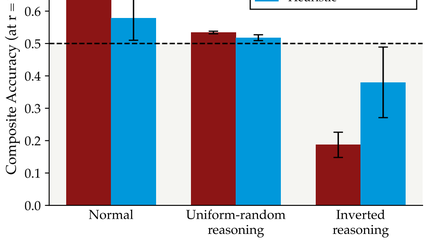


Disagreement accuracy. Fraction of disagreement steps where the actor picked the correct answer. Unlike composite, this is weighted by class frequency: at low r disagreements are mostly "self right, advisor wrong" so blanket-loyalty wins. Heuristic dominates at low r; **Adaptive** dominates at high r.

Feature Examination



- **Self-uncertainty ($\hat{p}$) stabilizes Adaptive training** despite a slightly negative mean contribution: removing it lifts composite by 2 pp but inflates seed std **3.4**$\times$ at r=0.8.
- **Heuristic catastrophically collapses** to fixed-action (0.5) when either $\hat{p}$ or reasoning is removed.



Reasoning signal attack. Inverting the reasoning signal collapses Adaptive (**3.4**$\times$ more damaging than uniform noise) but barely moves Heuristic: Adaptive genuinely depended on reasoning content, while Heuristic largely ignored it.

Takeaways, Limitations, Future Work

We found that combining step-level signals with running episode statistics, forced exploration, and targeted reward shaping produced true learning. One ablation demonstrated that self-uncertainty carried a slightly negative mean contribution but stabilized training across seeds — mean-impact ablation alone would discard a load-bearing feature. Furthermore, policy evaluation must look past aggregate accuracy: bias-free composite, OOD stress tests, and directional signal attacks are what distinguish a robust policy from a shortcut. Our scripted agent and advisor keep findings LLM-agnostic by construction; real-LM advisor behavior is a separate open question, as the probe does not consider raw question text, which could be a future direction. Finally, reliability is sampled from a discrete set, so continuous/drifting reliability remains untested but has potential.

gancone@stanford.edu jaden1@stanford.edu