ROBUST INFORMATION-GAIN CONTROL ACTIVE AGENTIC REASONING UNDER APPROXIMATE BELIEFS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM) agents in multi-turn reasoning often fail to make epistemic progress under approximate belief tracking. We formalize this failure through *Belief Trap Regions* (BTRs), forward-invariant regions of belief space in which expected progress stagnates. We show that even information-gain maximization (IG-max) can induce such traps under belief approximation error: small perturbations in the agent's internal belief may invert action rankings, causing repeated selection of uninformative actions despite the existence of informative alternatives. To mitigate this failure mode, we propose *Distributionally Robust Information Gain* (DR-IG), a decision-time objective that optimizes worst-case information gain over an ambiguity set around the current belief estimate. We show theoretically that robust worst-case control restores the negative drift of a progress potential under bounded belief error whenever uniformly informative actions exist. We further introduce a lightweight verbalized belief proxy that extracts approximate belief states from confidence-annotated model outputs, enabling belief-aware planning without additional training. Across five LLM families, naive IG-based agents collapse into deterministic self-locking loops with near-zero solve rates; in all 250 trapped episodes, the belief is frozen after entry, and the chosen action's oracle-IG rank falls to the bottom of a 1024-action space. DR-IG suppresses these traps to 0% incidence at every shortlist size $K \in \{1, \ldots, 20\}$ and elicitation width $M \in \{3, \ldots, 20\}$, and recovers non-trivial solve rates where IG-max solves nothing. The effect is not an artifact of simulation. In a real browser document object model (DOM), replacing the exact posterior with LLM's verbalized proxy costs 26.5% of solve rate, while DR-IG recovers a significant share. Together, we identify a concrete mechanism underlying active-reasoning failure in LLM agents and demonstrate that robust belief-aware control mitigates these pathologies in a training-free setting.

1 INTRODUCTION

Large language models (LLMs) are increasingly deployed as interactive agents that must reason, plan, and acquire information over multiple turns (Yao et al., 2022; Liu et al., 2023; Wang et al., 2024; Xi et al., 2025). In such settings, effective behavior depends not only on local reasoning ability, but also on the agent's ability to maintain and update a coherent belief over latent hypotheses and to select actions that drive epistemic progress (Yuan et al., 2025; Zhang et al., 2025). However, contemporary LLM agents frequently exhibit pathological behaviors: they repeat redundant queries, revisit uninformative actions, or enter persistent reasoning loops despite the existence of informative alternatives (Huang et al., 2023; Laban et al., 2025; Zhou et al., 2025). These failures persist even after reinforcement learning (RL) or tool-use fine-tuning, suggesting a deeper limitation in how current agents handle uncertainty during multi-step interaction (Wang et al., 2025a; Zou et al., 2026b).

In this work, we study these failures through the lens of approximate belief dynamics. We argue that a central difficulty in active reasoning is that the agent does not plan using the true posterior belief, but rather an imperfect internal approximation induced by amortized inference, heuristic updates, or implicit representations. Under such approximation error, even small perturbations in agent's belief can alter the ranking of candidate actions, causing the agent to repeatedly select actions that appear informative internally but yield no actual epistemic progress.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We formalize this phenomenon as *Belief Trap Regions* (BTRs): forward-invariant regions of belief space in which the expected progress potential fails to decrease (Zou et al., 2026b;a). We show that naive information-gain maximization (IG-max), a common strategy for active reasoning, can induce BTRs under approximate beliefs. In particular, when agent's belief deviates sufficiently from the Bayes-optimal posterior, action rankings under estimated information gain may invert, producing deterministic self-locking loops despite the existence of uniformly informative actions.

To mitigate this failure mode, we propose *Distributionally Robust Information Gain* (DR-IG), a decision-time objective that optimizes worst-case information gain over an ambiguity set around the current belief estimate. Rather than favoring actions that are informative only under a single estimated belief, DR-IG prefers actions whose informativeness remains stable under perturbations. We show theoretically that robust worst-case control restores the negative drift of a progress potential under bounded belief error whenever uniformly informative actions exist.

To operationalize belief-aware planning in LLM agents, we further derive a lightweight verbalized belief proxy that extracts approximate belief states from confidence-annotated model outputs (Kumaran et al., 2026). This provides a practical interface for integrating belief-space control into existing agents without additional training or explicit posterior modeling.

Empirically, across five LLM families on two finite-Θ benchmarks, naive IG-based agents collapse into deterministic self-locking loops with near-zero solve rates. The failure is mechanistically specific: in all 250 trapped episodes the guidance belief is frozen after entry, the top-K shortlist regenerates identically at every turn, and the chosen action's oracle-IG rank falls to $1{,}024$ of $1{,}024$ — while a single-step analysis shows individual ranking errors are nearly harmless in isolation, so the damage accrues through multi-turn compounding rather than catastrophic single-step information loss. DR-IG suppresses these traps to 0% at every shortlist size and elicitation width we test, and recovers non-trivial solve rates where IG-max solves nothing. Driving a real browser document object model (DOM), it also repairs a significant share of the solve rate lost to belief approximation.

2 METHOD

2.1 DRIFT CHARACTERIZATION OF BELIEF DYNAMICS

The POMDP setup, the exact Bayesian update and the amortized-updater model are deferred to Appx. B.1.

Progress Potential. Let $\Psi : \Delta(\Theta) \to \mathbb{R}_{\geq 0}$ be a progress potential, where smaller values indicate better alignment with the ground-truth hypothesis $\theta^\star$. Typical examples include $\Psi(b) = -\log b(\theta^\star)$, which penalizes low posterior mass on the truth and corresponds to log-loss; $\Psi(b) = \mathrm{KL}(\delta_{\theta^\star} \,\|\, b)$, the KL divergence from the degenerate true distribution $\delta_{\theta^\star}$ to b. Define the expected one-step drift

$$\mathcal{D}(b, a) := \mathbb{E}_{x \sim \mathbb{P}(\cdot|b,a)} \left[\Psi(\mathrm{T}(b, a, x)) - \Psi(b)\right]. \tag{1}$$

A Belief Trap Region (BTR) (Zou et al., 2026b;a) is a forward-invariant subset of belief space where the expected progress potential fails to decrease.

Definition 2.1 (Belief Trap Region) *A measurable set $R \subseteq \Delta(\Theta)$ is called a* Belief Trap Region *for the update operator T w.r.t the potential Ψ if for all $b \in R$ and for all $a \in \mathcal{A}$,:*

1. ***(Absorbingness)*** $\mathbb{P}\big(\mathrm{T}(b, a, x) \in R\big) = 1$, *where* $x \sim \mathbb{P}(\cdot \mid b, a)$.
2. ***(Non-negative Drift)*** $\mathcal{D}(b, a) \geq 0$.

Thus, once the belief enters R, it remains in R almost surely, and the expected progress potential does not decrease. Equivalently, $\{\Psi(b_t)\}_{t \geq 0}$ is a submartingale while $b_t \in R$. Beyond just stagnation, we give an escalated version to capture irreversible, linearly accumulating epistemic failure.

Definition 2.2 (Strong Belief Trap Region) *A Belief Trap Region R is called* strong *if there exists $\gamma > 0$ such that for all $b \in R$, $\mathcal{D}(b, a) \geq \gamma$ under any admissible action sequence.*

In this case, $\mathbb{E}[\Psi(b_T)] \geq \Psi(b_0) + \gamma T$, where the potential grows at least linearly in expectation. For instance, if $\Psi(b) = -\log b(\theta^\star)$, then $\mathbb{E}[\log b_T(\theta^\star)] \leq \log b_0(\theta^\star) - \gamma T$, which implies at least a linear decay of log-posterior mass on the true hypothesis.

Failure Modes of Belief Dynamics. After formulating BTRs, we characterize how they arise under Bayesian and amortized belief updates in the controlled setting in App. B.2. These mechanisms differ, but share a unifying structure which induces forward-invariant regions with non-negative drift of Ψ.

2.2 IG-MAX FOR ESCAPING BELIEF TRAPS

We consider the controlled belief dynamics on $\Delta(\Theta)$ with update operator $\mathrm{T}(b_t, a_t, x_t)$. For any belief b and action a, we define the *oracle* expected progress, i.e., information gain (IG), as

$$\mathrm{IG}(b, a) := \mathbb{E}_{\theta\sim b}\mathbb{E}_{x\sim\mathbb{P}(\cdot|\theta,a)}\Big[\Psi(b) - \Psi\big(\mathrm{T}(b, a, x)\big)\Big]. \tag{2}$$

By the law of total expectation, as $\mathbb{P}(x \mid b, a) = \sum_{\theta\in\Theta} b(\theta)\mathbb{P}(x \mid \theta, a)$, we can derive $\mathrm{IG}(b, a) = \mathbb{E}_{x\sim\mathbb{P}(\cdot|b,a)}\big[\Psi(b) - \Psi(\mathrm{T}(b, a, x))\big]$. The *IG-max* policy selects $a_t \in \arg\max_{a\in\mathcal{A}} \mathrm{IG}(b_t, a)$. Suppose $R \subseteq \Delta(\Theta)$ is a candidate region and there exists $\eta > 0$ s.t. for all $b \in R$, $\max_{a\in\mathcal{A}} \mathrm{IG}(b, a) \geq \eta$. Then for all $b \in R$ under IG-max, $\mathbb{E}_{x_t\sim\mathbb{P}(\cdot|b_t,a_t)}[\Psi(b_{t+1}) \mid b_t = b, a_t] \leq \Psi(b) - \eta$. Consequently, $\Psi(b_t)$ decreases at least linearly in expectation while $b_t \in R$, and therefore R *cannot* be a BTR. In addition, if R is forward-invariant, then it is *trap-escaping*: the process cannot stagnate inside R.

2.2.1 APPROXIMATE BELIEFS AND AGENT IG

IG-max is guaranteed to eliminate BTRs whenever sufficiently informative actions exist. However, this guarantee is stated w.r.t the ideal posterior belief b, namely the oracle belief obtained by T_{B} under the true observation model $\mathbb{P}(x \mid \theta, a)$. In practice, the agent does not have access to b. Instead, the ground-truth hypothesis $\theta^\star$ is hidden, and the agent only maintains an approximate belief state $\hat{b}$, together with an approximate predictive updater $\widehat{\mathrm{T}}$. We therefore replace Eq. (2) with an *agent-side* informativeness criterion computed from the agent's current belief proxy $\hat{b}$ and predictive updater $\widehat{\mathrm{T}}$:

$$\widehat{\mathrm{IG}}\left(\hat{b}, a\right) := \mathbb{E}_{\theta\sim\hat{b}}\mathbb{E}_{x\sim\mathbb{P}(\cdot|\theta,a)}\left[\Psi\left(\hat{b}\right) - \Psi\left(\widehat{\mathrm{T}}\left(\hat{b}, a, x\right)\right)\right]. \tag{3}$$

When deviation happens, actions that appear optimal under $\hat{b}$ may not maximize oracle IG under b. We define *belief error* as the discrepancy between $\hat{b}$ and b, measured by a divergence $d(b, \hat{b})$. In particular, if IG is sensitive to belief perturbations, even small estimation errors can invert action preferences. We measure belief approximation error using the L_1 distance between the exact posterior b and the agent's approximate posterior $\hat{b}$, i.e., $d(b, \hat{b}) = \|b - \hat{b}\|_1 = \sum_{\theta\in\Theta} |b(\theta) - \hat{b}(\theta)|$.

Theorem 2.3 (Trap Under Belief Error) *Assume (i) for every $a \in \mathcal{A}$, the map $b \mapsto \mathrm{IG}(b, a)$ is L-Lipschitz w.r.t the metric d, i.e., $|\mathrm{IG}(b, a) - \mathrm{IG}(b', a)| \leq L\, d(b, b')$ for all $b, b' \in \Delta(\Theta)$; (ii) there exists a measurable set $S \subseteq \Delta(\Theta)$ and an action $a_0 \in \mathcal{A}$ such that $\mathrm{IG}(b, a_0) = 0$ and $\mathbb{P}\big(\mathrm{T}(b, a_0, x) \in S \mid b, a_0\big) = 1$, where $x \sim \mathbb{P}(\cdot \mid b, a_0)$, for all $b \in S$; (iii) there exists an action $a_1 \in \mathcal{A}$ and $\eta > 0$ such that $\mathrm{IG}(b, a_1) \geq \eta$ for all $b \in S$. If $\delta \geq \eta/L$, then there exists $(b_0, \hat{b}_0)$ with $b_0 \in S$ and $d(b_0, \hat{b}_0) \leq \delta$ such that the naive IG-max policy $a_t \in \arg\max_a \mathrm{IG}(\hat{b}_t, a)$ selects $a_t = a_0$ for all t, the resulting belief trajectory satisfies $b_t \in S$ almost surely for all t, and $\mathbb{E}[\Psi(b_{t+1}) - \Psi(b_t) \mid b_t, a_t] = 0$. Hence, S is a BTR under naive IG-max.*

Thm 2.3 identifies a sharp threshold phenomenon: *when the belief error δ exceeds η/L, the apparent ranking of actions under $\hat{b}$ can disagree with the oracle ranking under b.* In this regime, the agent may repeatedly select an action that yields zero progress while ignoring an η-informative alternative. Consequently, S becomes a BTR induced purely by belief approximation error. This formalizes how epistemic miscalibration, rather than environmental structure, can create artificial traps.

DISTRIBUTIONALLY ROBUST IG-MAX. To hedge against belief error, assume the Bayes-optimal belief b lies in an ambiguity set $\mathcal{U}(\hat{b}) = \{b \in \Delta(\Theta) : d(b, \hat{b}) \leq \varepsilon\}$, where $d(\cdot, \cdot)$ is the predefined distance measure and ε is the *ambiguity radius*, chosen by the practitioner (unlike the belief error δ of Thm. 2.3, which is a property of the proxy and is not observable at decision time). $\mathcal{U}(\hat{b})$ contains the target belief exactly when $\varepsilon \geq \delta$. Define distributionally-robust information gain (DR-IG)

$$\mathrm{IG}_{\mathrm{DR}}\left(\hat{b}, a\right) := \alpha\mathrm{IG}(\hat{b}, a) + (1 - \alpha) \inf_{b'\in\mathcal{U}(\hat{b})} \mathrm{IG}(b', a). \tag{4}$$

which interpolates between the nominal criterion $\mathrm{IG}(\hat{b}, a)$ and the worst-case IG over the ambiguity set. The weight $\alpha \in [0, 1]$ controls conservatism: $\alpha = 1$ recovers naive IG-max, while $\alpha = 0$ yields the fully robust objective. The infimum is well-defined since $\mathcal{U}(\hat{b})$ is a compact subset of $\Delta(\Theta)$ and $\mathrm{IG}(\cdot, a)$ is continuous. The agent selects $a_t \in \arg\max_{a \in \mathcal{A}} \mathrm{IG}_{\mathrm{DR}}(\hat{b}_t, a)$.

2.2.2 Boundary-Safe IG Maximization

Since $\theta^\star$ is inaccessible at decision time, the oracle IG cannot be computed exactly. Recent studies (Choudhury et al., 2025; Wan et al., 2025) instead optimize an approximate objective $\widehat{\mathrm{IG}}$ with the Shannon entropy $\Psi_{\mathrm{ent}}(b) = H(b)$, which measures overall uncertainty in b. This criterion favors any action that reduces posterior uncertainty, regardless of whether the induced belief shift moves probability mass toward the ground true hypothesis $\theta^\star$ (Lindley, 1956). However, BTRs often occur at the simplex boundary: once the probability of a hypothesis approaches zero, amortized updates or numerical floors prevent recovery.

Theorem 2.4 (Entropy Reduction Is Insufficient for κ-Safety) *There exist a finite hypothesis space Θ, an action set $\mathcal{A}$, an observation model $\mathbb{P}(x \mid \theta, a)$, and an approximate belief updater $\widehat{\mathrm{T}}$ such that: (1) There exists an initial belief $b_0 \in \Delta(\Theta)$ and $\kappa > 0$ for which the IG-max policy with Ψ_{ent} produces a belief b_1 satisfying $\min_{\theta \in \Theta} b_1(\theta) < \kappa$. (2) There exists an action a_{safe} such that the belief after updating with a_{safe} remains in the interior region $\Delta_\kappa(\Theta) := \{b \in \Delta(\Theta) : \min_\theta b(\theta) \geq \kappa\}$.*

Thm. 2.4 demonstrates that maximizing entropy reduction does not guarantee κ-safety, namely, the belief b may still approach the boundary of $\Delta(\Theta)$ arbitrarily closely. This suggests that safety requires controlling not only uncertainty reduction but also proximity to the boundary.

Potential with Boundary Repulsion. Support-preserving update (see Appx. B.3) is an option to remove exact boundary collapse by ensuring that posterior masses never become zero. However, this only guarantees positivity, not safety: beliefs can still approach the boundary arbitrarily closely, where recovery becomes fragile. To enforce a stronger interior-safety property, we introduce an anti-collapse term to prevent elimination traps, and define the *boundary repulsion potential* as

$$\Psi_{\text{repel}}(b) = H(b) - \lambda \sum_{\theta \in \Theta} \log(b(\theta) + \varepsilon_{\mathrm{repel}}), \tag{5}$$

where $\varepsilon_{\mathrm{repel}} > 0$ is a small floor (set to 10^{-8} in the experiments of Appx. D.2). Accordingly, the expected IG is estimated as $\widehat{\mathrm{IG}}(b_t, a^{(k)}) = \mathbb{E}_{\theta \sim b_t} \mathbb{E}_{x \sim \mathbb{P}(\cdot \mid \theta, a)} \left[\Psi_{\text{repel}}(b_t) - \mathbb{E}_m \left[\Psi_{\text{repel}}\big(b^{(k,m)}\big) \right] \right]$. This explicitly penalizes near-boundary belief states to maintain an κ-safe buffer in the simplex interior.

In practice, Ψ_{repel} should be applied in a phase-aware manner. While boundary repulsion prevents premature collapse during early exploration, applying it throughout can hinder convergence and even degrade solve rate. We therefore use Ψ_{repel} only when the belief is still diffuse, and switch back to Ψ_{ent} once the belief has sufficiently converged (e.g., low entropy or high top-1 mass). This treats repulsion as a transient safety mechanism rather than a global objective.

2.3 LLM-Internal Verbalized Belief Proxy

The IG-max planner requires a current belief state in $\Delta(\Theta)$, which is b_t for the oracle setting. In LLM-based agents, however, this posterior is typically *implicit*: the model may maintain graded uncertainty over hypotheses without exposing a normalized distribution. A common proxy $\hat{b}_t$ relies on repeated sampling or answer-frequency aggregation (Choudhury et al., 2025), but such approaches are costly and high-variance when used inside lookahead planning.

Recent mechanistic interpretability work (Kumaran et al., 2026) shows that verbalized confidence reflects internal uncertainty signals informative of answer correctness beyond token-level probabilities. Motivated by this, we introduce a *verbalized belief proxy*. It treats confidence-weighted verbal outputs as samples from an implicit belief distribution and aggregates them into a normalized proxy—a low-cost alternative to Monte Carlo estimation with a principled link to internal uncertainty.

Belief Proxy via Verbalization. Rather than eliciting scores for all hypotheses in Θ, we elicit a belief support of bounded width: a single structured call asks the model for its top-M candidate hypotheses with verbalized confidences, returning pairs $(\hat{\theta}_t^{(m)}, c_t^{(m)}) \in \Theta \times \mathbb{R}_{\geq 0}$ for $m \in \{1, \ldots, M\}$. We aggregate these confidences into an empirical proxy belief as $\hat{b}_t(\theta) = \frac{\sum_{m=1}^{M} c_t^{(m)} \mathbf{1}\left[\hat{\theta}_t^{(m)} = \theta\right]}{\sum_{m=1}^{M} c_t^{(m)}}$, $\theta \in \Theta$. Thus, $\hat{b}_t \in \Delta(\Theta)$ is supported only on the hypotheses the model names, and any hypothesis it does not name receives zero mass. We call M the *elicitation width*: it bounds the size of the *active* support at the current step, not the dimensionality of Θ, so M need not be scaled to the (possibly unknown) hypothesis-space size — Eq. (6) absorbs the induced truncation into the belief error and Prop. 2.5 controls its effect through $\|\hat{b}_t - b_t\|_1$ alone. A repeated-sampling variant (M independent decodes aggregated by confidence-weighted frequency) is covered by the same estimator and ships with the released code; all reported results use the single-call form. Optionally, to mitigate overconfident verbalized scores (Xiong et al., 2023; Wang et al., 2026), we can apply a calibration map ϕ before aggregation, e.g., $\phi(c) = c^{1/\tau}$ with $\tau > 1$. In our implementation, we use a small fixed width ($M = 10$, i.e. one elicitation call per step), to provide a lightweight approximation of $\hat{b}_t$.

Proxy-induced Planning Error. Let b_t^{int} denote the LLM's internal belief over Θ. Then the belief estimation error w.r.t the target belief b_t admits the decomposition

$$d(\hat{b}_t, b_t) := \|\hat{b}_t - b_t\|_1 \;\leq\; \|\hat{b}_t - b_t^{\text{int}}\|_1 + \|b_t^{\text{int}} - b_t\|_1. \tag{6}$$

The first term is a *verbalization error*, capturing the gap between agentic internal uncertainty and the sparse confidence-weighted proxy extracted from its outputs. It is mitigated by evidence that verbalized confidence reflects structured internal representations rather than arbitrary outputs (Kumaran et al., 2026). The second is a *model-belief mismatch*, capturing the deviation of the internal LLM belief from b_t. This decomposition provides a concrete methodological origin for the belief error terms.

Proposition 2.5 (IG Error bound) *Assume that for every action $a \in \mathcal{A}$, the oracle IG $\mathrm{IG}(b, a)$ is L-Lipschitz in b with respect to ℓ_1. Then for any proxy belief $\hat{b}_t$,*

$$\Delta\mathrm{IG} := |\mathrm{IG}(b_t, a) - \mathrm{IG}(\hat{b}_t, a)| \leq L\|\hat{b}_t - b_t\|_1.$$

Combining with Eq. (6)*, we further obtain* $\Delta\mathrm{IG} \leq L\Big(\|\hat{b}_t - b_t^{\text{int}}\|_1 + \|b_t^{\text{int}} - b_t\|_1\Big)$.

We then replace the inaccessible b_t by $\hat{b}_t$ when scoring actions, and evaluate the proxy objective in Eq. (3) and (4). In this way, verbalized belief estimation becomes a concrete instance of the approximate-belief setting studied above: the gap from oracle planning is induced by the state-level belief error $\|\hat{b}_t - b_t\|_1$, and, by the Lipschitz continuity assumption on $\mathrm{IG}(b, a)$ in Thm. 2.3, the resulting error in IG estimation is correspondingly controlled.

3 Experiments

We evaluate on several interactive reasoning tasks. **(1) CircuitDecoding (CD)**, adapted from (Zou et al., 2026b), with $|\Theta| = 500$ Boolean circuits, 10-bit queries, and information-theoretic lower bound $\lceil \log_2 500 \rceil = 9$ queries. The agent identifies a hidden Boolean circuit from a large pool by querying with binary inputs and eliminating inconsistent candidates through feedback. **(2) ConceptInduction (CI)**, with $|\Theta| = 500$ subset-threshold classifiers $f_{S,k}(x) = \mathbb{1}[\,|x \cap S| \geq k\,]$, where S is a hidden subset of 12 abstract items and $k \leq |S|$ is a hidden threshold. Each turn, the agent submits a non-empty subset x (action space $|\mathcal{A}| = 4095$) and receives a binary response, with $\max_\text{turns} = 12$. Results on additional tasks are deferred to Appx. D.7–D.8, including LP-500, PP-500, and free-text SituationPuzzles. Experimental details are in Appx. C

3.1 Pervasive Belief Trap Regions across LLM Families and Tasks

To characterize the prevalence of BTRs, we evaluate two contrasting finite-Θ benchmarks under four guidance modes: (i) *Oracle IG* computed from the ground-truth posterior; (ii) *DR-IG*; (iii)

Table 1: Cross-family CD-500 evaluation under four guidance variants ($n = 50$ per cell).

	Solve rate (%)				BTR_{std} (%)			
Model	**Oracle**	**DR-IG**	**IG-max**	**vanilla**	**Oracle**	**DR-IG**	**IG-max**	**vanilla**
GPT-4o-mini	100	30	0	0	0	0	100	0
Claude Haiku 4.5	100	36	0	0	0	2	100	2
Gemini 2.5 FL	100	14	0	0	0	0	100	0
DeepSeek v3.1	100	24	0	0	0	0	100	0
Qwen 2.5 7B	100	2	0	0	0	0	100	0
Average	100	**20.5**	0	0	0	**0.5**	100	0.5

Table 2: Cross-family CI-500 evaluation under four guidance variants ($n = 50$ per cell).

	Solve rate (%)				BTR_{std} (%)			
Model	**Oracle**	**DR-IG**	**IG-max**	**vanilla**	**Oracle**	**DR-IG**	**IG-max**	**vanilla**
GPT-4o-mini	100	8	2	8	0	34	98	76
Gemini 2.5 Flash Lite	100	14	2	0	0	18	98	84
Gemini 2.5 Flash	100	16	2	8	0	18	98	70
Gemini 2.0 Flash	100	16	0	6	0	18	100	80
Gemini Flash Latest	100	18	2	14	0	16	98	54
Average	100	**14.4**	1.6	7.2	0	**20.8**	**98.4**	72.8

LLM-side IG obtained from the verbalized belief proxy; and (iv) *vanilla* committed on verbalized self-confidence without IG at the same budget. Results are reported in Tab. 1 and 2.

Five observations emerge. *(i) LLM-side IG-max collapses universally.* Across all agents, LLM-side IG yields negligible solve rates: exactly 0% on CD and a mean of 1.6% on CI ($4/5$ agents at 2%, Gemini 2.0 Flash at 0%), while BTR_{std} rises to 100% on CD and a mean of 98.4% on CI. This consistently realizes the trap predicted by Thm. 2.3 across two structurally distinct task families. *(ii) Oracle IG approaches task ceilings.* On both CD and CI, Oracle achieves a uniform 100% solve rate across all agents, with 0% BTR. This isolates belief approximation, rather than environmental stochasticity or horizon truncation, as the dominant source of failure. *(iii) DR-IG outperforms both vanilla LLM and IG on CD.* The vanilla baseline yields 0% solve and 0.5% BTR_{std} across 4 agents, indicating neither traps nor solves. DR-IG achieves 20.5% solve at the same 0.5% BTR_{std}: on the harder task, it is the *only* non-oracle method that simultaneously avoids the trap and solves a non-trivial fraction of episodes. *(iv) DR-IG provides recovery.* The $0\% \rightarrow 20.5\%$ CD gain has substantial cross-agent variance (2% for Qwen 2.5 7B; 36% for Claude Haiku 4.5). On CI, DR-IG yields a moderate solve improvement ($1.6\% \rightarrow 14.4\%$ on average, range 8–18%) and consistently reduces catastrophic BTR_{std} (e.g., from 98% to 16% for Gemini Flash Latest, and from 100% to 18% for Gemini 2.0 Flash). The asymmetry reflects task structure: CD's 1024-action binary-input space lets DR-IG eliminate trap-loops with near-zero residual BTR (0.5%), whereas CI's much larger 4095-subset action space leaves residual $\text{BTR}_{\text{std}} \approx 21\%$ even after DR-IG, indicating that the L_1 ambiguity set only partially covers worst-case directions in the verbalized proxy. *(v) IG's* 100% *BTR is a deterministic 5-action loop, not stochastic noise.* Per-trajectory inspection of all 250 BTR-positive CD episodes reveals an identical signature across all 5 agents: entry step $= 5$, stagnation length $= 5$, unique-action ratio $= 0.50$ (only 5 unique actions across 10 turns), and belief max-prob at entry $= 0.06$ (near uniform $1/500$). DR-IG's rare BTR ($1/250$) and vanilla's rare BTR ($1/200$) preserve a unique-action ratio $= 1.00$ and belief max-prob ≈ 0.5 — they are end-of-stagnation transient pauses, qualitatively distinct from IG's framework-induced loop. The remaining gap between DR-IG and Oracle reflects residual belief error not captured by the L_1 ambiguity set at $\varepsilon = 0.5$.

3.2 Pooled BTR as a Near-perfect Negative Predictor of Solve

Aggregating the $1\,950$ episodes of the joint CD+CI pool (5 agents $\times$ 3 guidance modes on CD, 6 agents $\times$ 4 modes on CI), we compute the Haldane-Anscombe-corrected odds ratio of solve given BTR. BTR_{std} has OR $= 0.0055$ $[0.0025, 0.0120]$; the stricter BTR_{int} variant (requires non-recoverable absorption) has OR $= 0.0007$ $[0.0000, 0.0106]$, with $0/673$ BTR-positive episodes solving versus 52.9% of the $1\,277$ non-BTR episodes. Within the CD IG / DR-IG pool, the Spearman correlation between continuous stagnation length and solve is $\rho = -0.833$ ($p < 10^{-4}$, $N = 750$); on CI alone $\rho = -0.728$ ($N = 1\,200$); and on the joint **CD+CI cross-task pool** we obtain $\rho_{\text{stag}} = -0.763$, $\rho_{\text{BTR}_{\text{std}}} = -0.606$, $\rho_{\text{BTR}_{\text{int}}} = -0.529$ (all $p < 10^{-4}$, $N = 1\,950$). These statistics confirm that BTR is not merely a definitional metric but the dominant failure mode of approximate-belief IG agents: any intervention that reduces BTR rate translates directly into solve gains, motivating the worst-case-driven design of DR-IG.

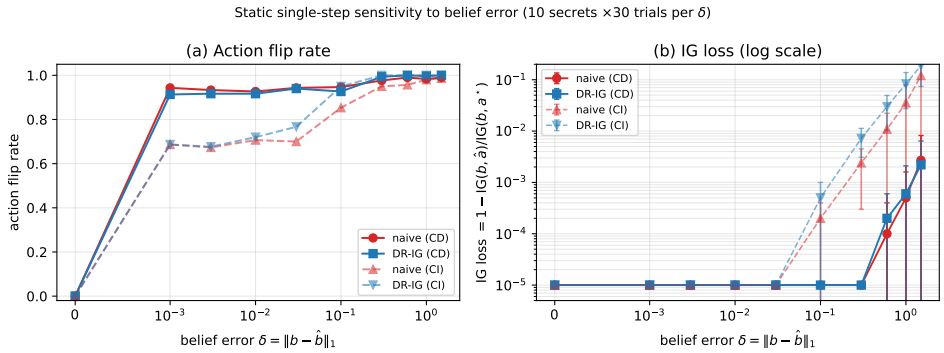


Figure 1: Single-step sensitivity on CI (dashed) and CD (solid). *Left:* the action flip rate rises sharply on both tasks and quickly saturates, indicating that action ranking is highly sensitive to even minimal belief error. *Right:* IG retention remains ≥ 0.965 for $\delta \leq 1.0$ on both tasks, showing that flipped actions are typically near-tied under the oracle IG.

3.3 Action-ranking Sensitivity to Belief Error

Thm. 2.3 predicts a sharp threshold $\delta \geq \eta/L$ beyond which naive IG-max may flip to a non-informative action. To probe this prediction in a controlled single-step setting, we sample 10 ground-truth secrets and, for each, draw 30 random L_1 perturbations $\hat{b} = b + e$ with different values of $\|e\|_1 = \delta$. We report two metrics: (1) ***action flip rate*** $\Pr[\arg\max_a \mathrm{IG}(\hat{b}, a) \neq \arg\max_a \mathrm{IG}(b, a)]$, and (2) ***IG retention*** $\mathrm{IG}(b, \hat{a}) \,/\, \mathrm{IG}(b, a^\star)$, where $\hat{a}$ is selected under $\hat{b}$ and $a^\star$ is the oracle maximiser.

The observed phase transition in flip rate (See Fig. 1) empirically supports the threshold phenomenon predicted by Thm. 2.3: the onset of $\hat{a} \neq a^\star$ occurs below $\delta = 10^{-3}$, far smaller than the $\|\hat{b} - b\|_1$ induced by typical verbalized proxies (e.g., $M = 10$ samples). However, the IG-retention curves show that such single-step flips incur negligible information loss in isolation: when many candidate actions are near-tied in oracle IG, selecting a non-maximiser is informationally inexpensive. The catastrophic effect of ranking instability therefore arises only through *multi-turn compounding*, rather than from any single-step loss. This is consistent with the 0% solve rate of LLM-side IG observed in §3.1, which reflects the cumulative impact of repeated mis-selections. In this sense, the present experiment isolates the sensitivity component of the BTR mechanism, while §3.1 demonstrates its full sequential manifestation as the trap.

3.4 Recovery of DR-IG Under Boundary-Induced Belief Shifts

We begin with a single-axis sweep over *naive* IG-max on CD to establish the baseline failure mode. As the mode-shift peak probability p_{peak} increases, naive IG-max retention exhibits a sharp discontinuity: it collapses to 0 at $p_{\text{peak}} = 0.5$, then recovers to 0.99 for larger peaks, where the dominant incorrect hypothesis is sufficiently separated from $\theta^\star$. This cliff pinpoints the regime where boundary-induced ranking inversion is most severe, motivating a two-axis sweep over both p_{peak} and ε. Together, we demonstrate a *recovery* effect. The worst-case term in DR-IG becomes informative when the ambiguity set contains beliefs that drive IG estimates toward zero, i.e., in the boundary-prone configurations identified by Thm. 2.3. In benign regimes, DR-IG is effectively inert, incurring no measurable cost; in contrast, when naive IG-max would stall, even a small ambiguity radius restores substantial progress. This establishes DR-IG as a targeted, training-free safeguard rather than a uniformly conservative modification.

3.5 Ablation and Sensitivity

We show that DR-IG (with $\varepsilon = 0.5$ and *hard_min* aggregation) takes CD-500 from 0% to 20.5% solve while keeping $\mathrm{BTR}_{\text{std}}$ at vanilla levels. We now isolate *which* component of DR-IG is responsible by three axes: (a) ablating the perturbation ($\varepsilon = 0$ reduces to nominal IG), (b) ablating the worst-case aggregation ($\alpha = 1.0$ reduces to mean-over-particles; $\alpha = 0.5$ interpolates between worst-case and mean), and (c) replacing the IG signal entirely (random top-K shortlist).

Table 3: DR-IG ablation on CD. Cells: solve% / BTR_{std}%. DR-IG full and IG / vanilla baselines at $n = 50$; ablation conditions ($\varepsilon = 0$, $\alpha = 0.5$, $\alpha = 1.0$, random top-K) at $n = 30$.

Condition	GPT-4o-mini	Claude Haiku 4.5	Gemini 2.5 FL	Qwen 2.5-7B	Avg.
IG (baseline)	0/100	0/100	0/100	0/100	**0/100**
DR-IG full ($\varepsilon = 0.5$, hard-min)	**30**/0	**36**/2	**14**/0	**2**/0	**20.5/0.5**
$\varepsilon = 0$ / nominal	0/100	0/100	0/100	0/100	0/100
$\alpha = 1.0$ / mean only	0/100	0/100	0/100	0/100	0/100
$\alpha = 0.5$ / interpolated	0/0	0/0	0/0	0/0	0/0
Random top-K	0/0	0/0	0/0	0/0	0/0
Vanilla (reference)	0/0	0/2	0/0	0/0	0/0.5

Table 4: CD-500 top-K sweep under constrained shortlist interfaces, averaged over two model families. Cells report solve rate (%) / BTR_{std} (%). *Loop period* denotes the mean number of unique actions from trap entry to episode end among BTR-positive episodes. At $K = 1$, the policy reduces to the pure $\arg\max$ rule of Thm. 2.3, leaving no shortlist choice to the LLM.

K	1	3	5
IG (point belief)	0.0/**100.0**	0.0/**100.0**	0.0/**100.0**
DR-IG	3.3/**0.0**	**28.4/0.0**	22.0/**0.0**
IG loop period	1.0	2.0	2.0

Ablation Results. Tab. 3 reports CD-500 solve / BTR_{std} for 4 agents. Two patterns emerge cleanly: *(a)* Both $\varepsilon = 0$ (no perturbation) and $\alpha = 1.0$ (mean-only aggregation) *exactly reproduce* IG's 100% BTR_{std} across all 4 agents, and the 5-action loop signature reappears with the same entry/stagnation/diversity/belief profile. Either lacking perturbation or lacking worst-case aggregation reverts the agent to the framework-induced trap. *(b)* The $\alpha = 0.5$ partial-mixing and Random top-K both *break* the trap (BTR $= 0\%$) but do *not* solve (0% across all 4 agents). This separates two independent axes: trap-escape requires only action diversity (which random selection or partial perturbation provides), whereas solve > 0 requires a *correct* IG signal — only pure hard_min worst-case under $\alpha = 0$ supplies an informativeness ranking robust to noisy verbalized belief.

Joint requirement. Combining the ablation and sweeps (See Appx. D.1), solve > 0 on CD requires *both* (i) ε at the right task-specific scale, and (ii) pure hard_min worst-case ($\alpha = 0$); BTR-safety alone is achieved by either ingredient, but only their conjunction recovers solve. This separation between two independent failure axes (deadlock vs. wrong action) is consistent with the trap mechanism in Thm. 2.3: ranking ties under noisy belief produces framework-induced loops (which any perturbation breaks), but recovering an informative *direction* of progress requires worst-case-robust IG.

3.6 Interface-Robust Trap Suppression under Constrained Top-K Interfaces

The constrained decision interface passes a top-K shortlist to the LLM. To determine whether the observed loops arise from the belief dynamics rather than from the specific shortlist width, we evaluate the two guidance modes under the constrained regime $K \in \{1, 3, 5\}$. We use two model families (`gpt-4o-mini` and `gemini-2.5-flash-lite`), with $n = 30$ episodes per cell; the $K = 5$ setting reuses the main-table $n = 50$ cells. Per-model results with Wilson intervals are reported in Appx. D.3.

The shortlist width primarily determines the *period* of the induced loop: IG exhibits 1.0, 2.0, and 2.0 unique post-entry actions at $K = 1$, 3, and 5, respectively. However, the corrupted belief ranking determines whether trapping occurs at all. Across all three constrained shortlist sizes, naive IG traps in 100% of episodes and achieves a 0% solve rate for both model families, including $K = 1$, where no shortlist choice is available to the LLM. In contrast, DR-IG maintains $\text{BTR}_{\text{std}} = 0.0\%$ in all six model-by-K cells, while recovering non-trivial solve rates at $K = 3$ and $K = 5$. The Wilson intervals do not overlap between IG and DR-IG for BTR incidence in this constrained regime. Under the identical interface at $K = 1$, the mean oracle IG of the selected action increases from 0.10 for IG to 0.76 for DR-IG, indicating that the robust objective recovers an informative direction rather than merely breaking action repetition.

Table 5: MiniWoB++ `guess-number`, 200 paired episodes per arm. Block-bootstrap intervals resample latent secrets rather than episodes. Belief error is mean L_1 to the exact posterior (max 2.0).

Controller	Solve% [Wilson 95%]	Block-bootstrap 95%	Belief err (L_1)	vs. IG (McNemar)
Oracle IG (exact posterior)	100.0 [98, 100]	[100, 100]	—	$53/0$, $p < 10^{-3}$
Vanilla (free-form LLM)	85.0 [79, 89]	[69, 95]	0.50	—
DR-IG ($\varepsilon = 0.5$)	**80.0** [74, 85]	[58, 97]	**0.50**	$\mathbf{24/11}$, $\mathbf{p = 0.041}$
IG (point belief)	73.5 [67, 79]	[52, 91]	0.57	—

3.7 Belief Approximation and Its Repair in a Real Browser

All results so far use simulated environments. We now run the same machinery through a real DOM.

The mechanism appears in ordinary web interaction. On 20 standard MiniWoB++ tasks (BrowserGym, ReAct agent, 15-step budget, `gemini-2.5-flash-lite`), latent state is not enumerable and no oracle posterior exists. Vanilla ReAct is strong, succeeding on $15/20$ tasks (75.0%). A web-adapted BTR detector (repeated-action windows plus DOM/URL stagnation, computed online) fires on $5/20$ episodes, and BTR incidence separates success sharply: 20% $(1/5)$ for BTR-positive versus 93% $(14/15)$ for BTR-negative episodes, a Haldane–Anscombe odds ratio for failure given BTR of 29.0 $[2.4, 352.6]$. Every BTR-positive failure ran to the full step budget in a repeated-action loop — the same behavioral signature we formalize in the finite-Θ setting. Each task was run once, so we report this as a task-level association in a pilot.

The DR-IG controller runs end-to-end through a real page. On MiniWoB++ `guess-number` the controller drives real fill-and-submit actions against the live DOM, keeps a local hypothesis set, and elicits the LLM's verbalized posterior over it. We run 10 latent secrets $\times$ 20 replicates $= 200$ paired episodes per arm at temperature 1.0; all arms share seeds, hence the same latent instance and initial page state.

Two results follow. First, belief approximation is the binding constraint on a live page: replacing the exact posterior with the LLM's verbalized proxy under an otherwise identical controller costs 26.5pp of solve rate ($100.0 \to 73.5$; the oracle wins 53 discordant pairs and loses none), and the proxy is genuinely miscalibrated ($L_1 = 0.57$ of a maximum 2.0) — the approximate-belief regime of Thm. 2.3 is not an artifact of simulation. Second, DR-IG repairs a significant share of that loss under the identical controller: 80.0% vs. 73.5%, paired McNemar $24/11$ over 35 discordant pairs ($p = 0.041$), while simultaneously restoring the belief error to the oracle's own level (0.50 vs. 0.57). The free-form vanilla agent reaches 85.0% on this 10-hypothesis task, a difference from DR-IG that the paired test does not resolve ($9/19$, $p = 0.087$).

Robustness to the elicitation width. The verbalized proxy elicits a support of width M (§2.3). Re-running CD-500 DR-IG with $M \in \{3, 5, 20\}$ on both families leaves solve unchanged to the episode (26.7% on `gpt-4o-mini`, 16.7% on `gemini-2.5-flash-lite` at every M; $M = 10$ is the main-table cell at $30.0/14.0\%$) and BTR_{std} at 0% throughout. DR-IG therefore requires no tuning of M to the hypothesis-space size (Appx. D.6).

4 Conclusion

We study active reasoning under approximate beliefs and identify *Belief Trap Regions* as a fundamental failure mode in LLM agents. We show that naive information-gain maximization induces such traps under belief error, and propose *Distributionally Robust Information Gain* (DR-IG) to restore reliable epistemic progress. The supporting evidence spans five LLM families on two finite-Θ benchmarks; 250 trapped trajectories in which the belief is frozen and the shortlist regenerates identically at every post-entry turn; shortlist sizes $K \in \{1, \ldots, 20\}$ and elicitation widths $M \in \{3, \ldots, 20\}$, over which DR-IG's trap suppression never degrades; and a browser-executed task in which the controller drives a real DOM and DR-IG significantly outperforms point-belief IG. Combined with a practical verbalized belief proxy, our approach provides a training-free framework for robust multi-turn reasoning.

AI USE STATEMENT

Large language models are the object of study in this work: every agent evaluated in §3 is an LLM queried through a public API, and the prompts, decoding configuration, model versions and per-cell episode counts are reported in Appx. C. The authors take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies belief-aware control for improving multi-step reasoning in large language model (LLM) agents. By mitigating pathological behaviors such as uninformative action loops and belief collapse, our methods can contribute to more reliable and efficient interactive AI systems. Improved robustness in agentic reasoning can benefit applications that require multi-step information gathering and decision-making, such as scientific discovery, education, and assistive tools. Reducing failure modes like repeated or irrelevant queries may also improve user experience and reduce unnecessary computational cost. More broadly, our framework contributes to understanding how uncertainty should be represented and controlled in sequential decision-making systems.

REPRODUCIBILITY STATEMENT

Every theoretical claim is stated with its assumptions in §2 and proved in full in Appx. B. The hypothesis spaces, action interfaces, feedback models and turn budgets of CD-500 and CI-500 are specified in §3 and Appx. C, together with the default DR-IG configuration (ambiguity radius $\varepsilon = 0.5$, interpolation weight $\alpha = 0$, elicitation width $M = 10$, shortlist size $K = 5$) and the episode count of every reported cell; Appx. E gives worked interaction traces for each task. All arms within a run share their seeds, so the paired statistics we report (McNemar exact tests, block bootstrap over latent instances) are computed on matched episodes, and the browser-executed runs additionally fix the BrowserGym task seed across controllers so that every arm sees the same latent instance and the same initial page. Per-episode trajectories and per-cell summaries are logged for every number reported in the paper.

REFERENCES

Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. Consistently simulating human personas with multi-turn reinforcement learning. *arXiv preprint arXiv:2511.00222*, 2025.

Kartikeya Badola, Jonathan Simon, Arian Hosseini, Sara Marie Mc Carthy, Tsendsuren Munkhdalai, Abhimanyu Goyal, Tomáš Kočiskỳ, Shyam Upadhyay, Bahare Fatemi, and Mehran Kazemi. Multi-turn puzzles: Evaluating interactive reasoning and strategic dialogue in llms. *arXiv preprint arXiv:2508.10142*, 2025.

Deepro Choudhury, Sinead Williamson, Adam Goliński, Ning Miao, Freddie Bickford Smith, Michael Kirchhof, Yizhe Zhang, and Tom Rainforth. Bed-llm: Intelligent information gathering with llms and bayesian experimental design. *arXiv preprint arXiv:2508.21184*, 2025.

Sanjiban Choudhury. Process reward models for llm agents: Practical framework and directions. *arXiv preprint arXiv:2502.10325*, 2025.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10602–10621, 2023.

Zhirui Deng, Zhicheng Dou, Yutao Zhu, Ji-Rong Wen, Ruibin Xiong, Mang Wang, and Weipeng Chen. From novice to expert: Llm agent policy optimization via step-wise reinforcement learning. *arXiv preprint arXiv:2411.03817*, 2024.

Samuel Gershman and Noah Goodman. Amortized inference in probabilistic reasoning. In *Proceedings of the annual meeting of the cognitive science society*, volume 36, 2014.

Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei W Koh, and Bryan Hooi. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in llms. *Advances in Neural Information Processing Systems*, 37:24181–24215, 2024.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Jonathan Kuck, Shuvam Chakraborty, Hao Tang, Rachel Luo, Jiaming Song, Ashish Sabharwal, and Stefano Ermon. Belief propagation neural networks. *Advances in Neural Information Processing Systems*, 33:667–678, 2020.

Dharshan Kumaran, Arthur Conmy, Federico Barbero, Simon Osindero, Viorica Patraucean, and Petar Velickovic. How do llms compute verbal confidence. *arXiv preprint arXiv:2603.17839*, 2026.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*, 2025.

Wenzhe Li, Shujian Zhang, Wenxuan Zhou, John Lambert, Chi Jin, Andrew Hard, Rajiv Mathews, and Lun Wang. Music: Multi-step instruction contrast for multi-turn reward models. *arXiv preprint arXiv:2512.24693*, 2025.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023.

Shuhaib Mehri, Priyanka Kargupta, Tal August, and Dilek Hakkani-Tür. Multisessioncollab: Learning user preferences with memory to improve long-term collaboration. *arXiv preprint arXiv:2601.02702*, 2026.

Kun Peng, Conghui Tan, Yu Liu, Guohua Tang, Zhongqian Sun, Wei Yang, Zining Zhu, Lei Jiang, Yanbing Liu, and Hao Peng. Dialogue model optimization via agent game and adaptive tree-based grpo. *arXiv preprint arXiv:2602.08533*, 2026.

Yanming Wan, Jiaxing Wu, Marwa Abdulhai, Lior Shani, and Natasha Jaques. Enhancing personalized multi-turn dialogue with curiosity reward. *arXiv preprint arXiv:2504.03206*, 2025.

Hanlin Wang, Chak Tou Leong, Jiashuo Wang, Jian Wang, and Wenjie Li. Spa-rl: Reinforcing llm agents via stepwise progress attribution. *arXiv preprint arXiv:2505.20732*, 2025a.

Jiawei Wang, Yanfei Zhou, Siddartha Devic, and Deqing Fu. Are llm decisions faithful to verbal confidence? *arXiv preprint arXiv:2601.07767*, 2026.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, et al. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*, 2025b.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Siyu Yuan, Zehui Chen, Zhiheng Xi, Junjie Ye, Zhengyin Du, and Jiecao Chen. Agent-r: Training language model agents to reflect via iterative self-training. *arXiv preprint arXiv:2501.11425*, 2025.

Jiaxin Zhang, Prafulla Kumar Choubey, Kung-Hsiang Huang, Caiming Xiong, and Chien-Sheng Wu. Agentic uncertainty quantification. *arXiv preprint arXiv:2601.15703*, 2026.

Zijing Zhang, Ziyang Chen, Mingxiao Li, Zhaopeng Tu, and Xiaolong Li. Rlvmr: Reinforcement learning with verifiable meta-reasoning rewards for robust long-horizon agents. *arXiv preprint arXiv:2507.22844*, 2025.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38, 2024.

Zhanke Zhou, Xiao Feng, Zhaocheng Zhu, Jiangchao Yao, Sanmi Koyejo, and Bo Han. From passive to active reasoning: Can large language models ask the right questions under incomplete information? *arXiv preprint arXiv:2506.08295*, 2025.

Deyu Zou, Yongqiang Chen, Fan Feng, Mufei Li, Pan Li, Yu Gong, and James Cheng. On information self-locking in reinforcement learning for active reasoning of llm agents. *arXiv preprint arXiv:2603.12109*, 2026a.

Deyu Zou, Yongqiang Chen, Jianxiang Wang, Garry YANG, Mufei Li, Qing Da, James Cheng, Pan Li, and Yu Gong. Reducing belief deviation in reinforcement learning for active reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026b.

A Related Works

Active reasoning and information-seeking agents. Recent work has distinguished *active* reasoning from passive reasoning by requiring LLM agents to acquire missing information through interaction rather than reason from a static, sufficient context (Badola et al., 2025). Existing approaches include clarification-oriented prompting, such as ProactiveCoT (Deng et al., 2023), and uncertainty-aware information-seeking methods, such as UoT (Hu et al., 2024), which quantify the value of candidate questions in multi-turn interactions. The rise of this setting has also motivated interactive benchmarks, including AgentBench (Liu et al., 2023), AR-Bench (Zhou et al., 2025), and MultiSessionCollab (Mehri et al., 2026). Our work fits into this broader agenda, but focuses specifically on the reliability of belief-space planning when the agent must act under incomplete and imperfectly represented uncertainty.

Training-time optimization for multi-turn agents. A related line of work improves multi-turn LLM agents through training-time objectives, including process rewards, stepwise supervision, inverse-RL-style learning, and RL-based credit assignment (Deng et al., 2024; Choudhury, 2025; Wang et al., 2025b;a; Wan et al., 2025; Li et al., 2025; Peng et al., 2026; Abdulhai et al., 2025). Recent studies on information self-locking and belief deviation further show that poor action choices and inaccurate latent-state tracking can reinforce each other over multiple turns (Zou et al., 2026a;b). These works are closely related in motivation, as they highlight the coupling between action quality and belief quality. Our perspective is complementary but distinct: rather than improving agents through additional training signals, we study a *training-free*, decision-time planning problem under approximate beliefs, and analyze how robust and boundary-safe objectives can restore epistemic progress without new reward models or policy optimization.

Uncertainty, belief tracking, and information-based control. A central challenge in active reasoning is that action selection depends on a latent belief state that is often only approximately represented. Prior work has explored uncertainty-aware or information-driven control signals for multi-turn agents (Hu et al., 2024; Choudhury et al., 2025; Wan et al., 2025), typically favoring actions that reduce ambiguity or posterior uncertainty. However, these approaches largely assume that the agent's belief or predictive uncertainty estimate is sufficiently accurate for planning. We instead study the failure mode induced by *approximate beliefs*.

Recent work (Kumaran et al., 2026) suggests that LLMs internally compute structured confidence signals during generation, which can be extracted through verbalized outputs. This motivates our use of *verbalized belief proxies* as a practical interface to otherwise implicit belief states.

B Mathematical Analysis

B.1 Preliminary

Belief Dynamics. We model active reasoning as a Partially Observable Markov Decision Process (POMDP) (Kaelbling et al., 1998) with a finite latent hypothesis space Θ. A fixed but unknown ground-truth hypothesis $\theta^\star \in \Theta$ governs observations throughout an episode.

At time t, an agent maintains a belief $b_t \in \Delta(\Theta)$, where $\Delta(\Theta)$ denotes the probability simplex over Θ, and $b_t(\theta)$ denotes the posterior mass assigned to hypothesis θ at time t. At each step, the agent selects an action $a_t \in \mathcal{A}$. The environment produces an interaction signal $x_t \sim \mathbb{P}(\cdot \mid \theta^\star, a_t)$, where $\mathbb{P}(x \mid \theta, a)$ is the observation model conditioned on a latent hypothesis θ and an action a.

Given belief b_t, the corresponding *belief-marginal* predictive distribution over observations is defined as $\mathbb{P}(x \mid b_t, a_t) := \sum_{\theta\in\Theta} b_t(\theta)\mathbb{P}(x \mid \theta, a_t)$. Belief evolution is modeled by a measurable update operator $\mathrm{T} : \Delta(\Theta) \times \mathcal{A} \times \mathcal{X} \to \Delta(\Theta)$, with dynamics given by $b_{t+1} = \mathrm{T}(b_t, a_t, x_t)$.

Bayesian (Multiplicative) Update. Given an observation model $\mathbb{P}(x \mid \theta, a)$, the exact Bayesian update operator $\mathrm{T_B}$ is defined as $\mathrm{T_B}(b, a, x)(\theta) = \frac{\mathbb{P}(x|\theta,a)\, b(\theta)}{\sum_{\theta'\in\Theta} \mathbb{P}(x|\theta',a)\, b(\theta')}$. Equivalently, the log-odds evolve according to $\log \frac{b_{t+1}(\theta)}{b_{t+1}(\theta')} = \log \frac{b_t(\theta)}{b_t(\theta')} + \log \frac{\mathbb{P}(x_t|\theta,a_t)}{\mathbb{P}(x_t|\theta',a_t)}$. This multiplicative structure guarantees normalization and preservation of support whenever $\mathbb{P}(x \mid \theta, a) > 0$ for all $\theta \in \Theta$.

Amortized Belief Network. After receiving observation x_t, the agent employs an internal belief updater T_ϕ, where ϕ denotes learned parameters. The amortized update produces $b_{t+1} = \mathrm{T}_\phi(b_t, a_t, x_t)$. While T_ϕ may be trained to approximate T_B, it generally lacks the algebraic guarantees of exact Bayesian updating. In particular, it may introduce approximation error, spurious fixed points, or unintended support collapse (Gershman & Goodman, 2014; Kuck et al., 2020).

B.2 Failure Modes of Belief Dynamics

B.2.1 Bayesian Update

(B1) Stalling under Uninformative Evidence. Suppose that within a region $R \subseteq \Delta(\Theta)$ and for some action a, all reachable observations are uninformative, i.e., $\mathbb{P}(x \mid \theta, a) = c(x, a)$ for all $\theta \in \Theta$. The observation model becomes hypothesis-independent, so the action provides no discriminative signal. Then for all $b \in R$ and all such x, we have $\mathrm{T}_\mathrm{B}(b, a, x) = b$. Thus R is forward-invariant under action a, and the belief does not change. The progress potential remains constant: $\Psi(b_{t+1}) = \Psi(b_t)$. Equivalently, in log-odds form, for any $\theta \neq \theta'$, $\log \frac{b_{t+1}(\theta)}{b_{t+1}(\theta')} = \log \frac{b_t(\theta)}{b_t(\theta')} + \log \frac{\mathbb{P}(x_t|\theta,a_t)}{\mathbb{P}(x_t|\theta',a_t)}$. Since the likelihood ratio equals 1, the increment is zero and learning stalls completely.

(B2) Irreversible Elimination (Absorbing Error Set). Suppose that for some reachable (a, x) we have $\mathbb{P}(x \mid \theta, a) = 0$. Then for all b with $b(\theta) > 0$, we obtain $\mathrm{T}_\mathrm{B}(b, a, x)(\theta) = 0$. Define the set $\Upsilon_\theta := \{b \in \Delta(\Theta) : b(\theta) = 0\}$. Once entered, Υ_θ is absorbing under any future sequence of actions, since Bayesian updating preserves zeros. Thus, Bayesian updating permits irreversible elimination when the likelihood assigns a true zero.

(B3) Wrong-Direction Drift under Misspecified Likelihood. Suppose a misspecified observation model where observations are systematically better explained by an incorrect hypothesis $\theta' \neq \theta^\star$. Formally, there exists $\gamma > 0$ such that for all $b_t \in R$ and actions a_t, $\mathbb{E}_{x_t \sim \mathbb{P}(\cdot|\theta^\star, a_t)}\left[\log \frac{\mathbb{P}(x_t|\theta^\star,a_t)}{\mathbb{P}(x_t|\theta',a_t)}\right] \leq -\gamma$. The log-odds evolve according to $\log \frac{b_T(\theta)}{b_T(\theta')} = \log \frac{b_0(\theta)}{b_0(\theta')} + \sum_{t=0}^{T-1} \log \frac{\mathbb{P}(x_t|\theta,a_t)}{\mathbb{P}(x_t|\theta',a_t)}$. Taking expectation under the true data-generating process yields $\mathbb{E}\left[\log \frac{b_T(\theta^\star)}{b_T(\theta')}\right] \leq \log \frac{b_0(\theta^\star)}{b_0(\theta')} - \gamma T$. Thus, the posterior mass assigned to $\theta^\star$ decays exponentially; for potentials such as $\Psi(b) = -\log b(\theta^\star)$, this induces a linear increase in $\Psi(b_t)$, yielding systematic non-progress and forming a Strong BTR.

B.2.2 Amortized Belief Network

(A1) Spurious Fixed-Point Attractors. A belief $b^\dagger$ is a fixed point if $b^\dagger = \mathbb{E}_{x \sim \mathbb{P}(\cdot|b^\dagger, a)}[\mathrm{T}_\phi(b^\dagger, a, x)]$ for some action a. If the Jacobian w.r.t b is contractive in a neighborhood $\mathcal{N}$, i.e., $\sup_{b \in \mathcal{N}, a, x} \rho\Big(\frac{\partial \mathrm{T}_\phi}{\partial b}(b, a, x)\Big) < 1$, then $b^\dagger$ is locally asymptotically stable: $b_t \to b^\dagger$ for all $b_0 \in \mathcal{N}$ under repeated application of T_ϕ. If $b^\dagger(\theta^\star) < 1$, $\mathcal{N}$ forms a trap around a stable but incorrect belief.

(A2) Architecture-Induced Elimination. Unlike Bayesian updating, T_ϕ may assign exact zero mass even when the likelihood is non-degenerate. If there exists reachable (b, a, x) with $b(\theta) > 0$ but $\mathrm{T}_\phi(b, a, x)(\theta) = 0$, then the absorbing error set $\Upsilon_\theta = \{b : b(\theta) = 0\}$ becomes reachable without any true zero-likelihood event. This failure mode is strictly stronger than (B2).

(A3) Error Amplification under Composition. Let the per-step approximation error be $e_t := \mathrm{T}_\phi(b_t, a_t, x_t) - \mathrm{T}_\mathrm{B}(b_t, a_t, x_t)$. Assuming differentiability, a first-order expansion yields $b_T^\phi - b_T^{\mathrm{Bayes}} \approx \sum_{k=0}^{T-1} \left(\prod_{j=k+1}^{T-1} J_{\mathrm{Bayes}}(b_j, a_j, x_j)\right) e_k$, where $J_{\mathrm{Bayes}} = \partial \mathrm{T}_\mathrm{B}/\partial q$. If along trajectories inside a region R we have $\left\|\prod_{j=k+1}^{T-1} J_{\mathrm{Bayes}}(b_j, a_j, x_j)\right\| \geq \exp(\lambda(T-1-k))$ for some $\lambda > 0$, then bounded errors $\|e_k\| \leq \varepsilon_{\mathrm{approx}}$ imply $\|b_T^\phi - b_T^{\mathrm{Bayes}}\| \gtrsim \varepsilon_{\mathrm{approx}} \frac{e^{\lambda T}-1}{e^\lambda - 1}$, i.e., exponential blow-up of belief deviation.

(A4) Saturation-Induced Stalling. If in a region R we have $\sup_{b \in R, a, x} \left\|\frac{\partial \mathrm{T}_\phi}{\partial x}(b, a, x)\right\| \approx 0$, then updates become nearly insensitive to observations, so $\mathrm{T}_\phi(b, a, x) \approx \mathrm{T}_\phi(b, a, x')$ for all reachable x, x'. Even when the true likelihood $\mathbb{P}(x \mid \theta^\star, a)$ is informative, the internal belief updater behaves as if evidence carries no signal, leading to a plateau of stalled learning.

B.3 Support-Preserving Updates

Both $\mathrm{T_B}$ and T_ϕ can exhibit trap behavior near the boundary of the simplex. For $\mathrm{T_B}$, once a hypothesis receives zero mass, it can never recover. For T_ϕ, the amortized update can drive plausible hypotheses to zero or near-zero mass, after which recovery becomes difficult or impossible. Let $\hat{\mathbb{P}}(x \mid \theta, a)$ denote the agent's internal predictive model as an approximation to $\mathbb{P}(x \mid \theta, a)$. To avoid this absorbing boundary of the simplex, an option is to introduce a small likelihood floor $\varepsilon_{\mathrm{lik}} > 0$ and strictly ensure

$$\hat{b}_{t+1}(\theta) \propto \max\{\varepsilon_{\mathrm{lik}},\, \hat{\mathbb{P}}(x_t \mid \theta, a_t)\}\hat{b}_t(\theta). \tag{7}$$

The constant $\varepsilon_{\mathrm{lik}}$ prevents zero-likelihood updates due to model misspecification or numerical underflow. As a consequence, the posterior mass remains strictly positive, i.e., there exists $\varepsilon_{\min} > 0$ such that $\hat{b}_t(\theta) \geq \varepsilon_{\min}$ for all θ and t. This removes hard absorbing-error traps and converts permanent exclusion into recoverable uncertainty.

B.4 Proof of Theorem 2.3

We prove that when the belief error exceeds the threshold $\delta \geq \eta/L$, there exists an initial approximate belief $\hat{b}_0$ with $d(b_0, \hat{b}_0) \leq \delta$ such that the naive IG-max policy selects the uninformative action a_0 and induces a belief trap.

Step 1: Lipschitz perturbation bound. By assumption (i), for any $b, \hat{b} \in \Delta(\Theta)$ and any action $a \in \mathcal{A}$,

$$|IG(b, a) - IG(\hat{b}, a)| \leq L\, d(b, \hat{b}).$$

Let $d(b_0, \hat{b}_0) \leq \delta$. Then for any action a,

$$IG(\hat{b}_0, a) \geq IG(b_0, a) - L\delta, \qquad IG(\hat{b}_0, a) \leq IG(b_0, a) + L\delta.$$

Step 2: Compare informative and uninformative actions. By assumption (ii), for all $b \in S$,

$$IG(b, a_0) = 0.$$

By assumption (iii), for all $b \in S$,

$$IG(b, a_1) \geq \eta.$$

Applying the Lipschitz bounds from Step 1 at the true belief $b_0 \in S$ gives

$$IG(\hat{b}_0, a_0) \leq L\delta, \qquad IG(\hat{b}_0, a_1) \geq \eta - L\delta.$$

Step 3: Explicit construction of a ranking-inverting perturbation. Since $\delta \geq \eta/L$, we have $\eta - L\delta \leq 0$. We now construct a concrete $\hat{b}_0$ achieving the inversion.

Let $|\Theta| \geq 2$ and fix any two hypotheses $\theta_+, \theta_- \in \Theta$ with $\theta_+ \neq \theta_-$. Define the perturbation vector

$$e \;=\; \tfrac{\delta}{2}\,(\mathbf{e}_{\theta_-} - \mathbf{e}_{\theta_+}),$$

where $\mathbf{e}_\theta$ denotes the standard basis vector for hypothesis θ, and set $\hat{b}_0 = b_0 + e$. Choosing $\delta \leq \min\{2b_0(\theta_+),\, 2(1 - b_0(\theta_-))\}$ ensures $\hat{b}_0 \in \Delta(\Theta)$, and by construction

$$d(b_0, \hat{b}_0) = \|e\|_1 = \delta.$$

Because $IG(\cdot, a)$ is L-Lipschitz and continuous on the compact set $\Delta(\Theta)$, there exists a direction e of ℓ_1-norm δ along which $IG(\hat{b}_0, a_1)$ is minimised and $IG(\hat{b}_0, a_0)$ is maximised. Evaluating the Lipschitz bounds at this $\hat{b}_0$:

$$IG(\hat{b}_0, a_0) \;\geq\; IG(b_0, a_0) - L\delta \;=\; -L\delta,$$
$$IG(\hat{b}_0, a_1) \;\leq\; IG(b_0, a_1) - L\delta \;\leq\; \eta - L\delta \;\leq\; 0,$$

where the second inequality uses the fact that the perturbation is chosen adversarially to decrease $IG(\cdot, a_1)$ by the full Lipschitz margin $L\delta$, which is achievable since $\|e\|_1 = \delta$. Combining these two bounds,

$$IG(\hat{b}_0, a_0) \;\geq\; -L\delta \;\geq\; \eta - L\delta \;\geq\; IG(\hat{b}_0, a_1),$$

where the middle inequality uses $\eta \geq 0$. Hence, the naive IG-max policy selects a_0:

$$\hat{a} \;=\; \arg\max_{a \in \mathcal{A}} IG(\hat{b}_0, a) \;=\; a_0.$$

Step 4: Invariance of S. By assumption (ii), S is forward-invariant under a_0:

$$\mathbb{P}\big(T(b, a_0, x) \in S \mid b, a_0\big) = 1, \quad x \sim P(\cdot \mid b, a_0).$$

Thus if $b_t \in S$ and the policy selects $a_t = a_0$, then $b_{t+1} \in S$ almost surely. By induction, $b_t \in S$ for all $t \geq 0$.

Step 5: Zero drift. From assumption (ii),

$$IG(b, a_0) \;=\; \mathbb{E}[\Psi(b) - \Psi(T(b, a_0, x))] = 0,$$

which rearranges to

$$\mathbb{E}[\Psi(b_{t+1}) - \Psi(b_t) \mid b_t,\ a_t = a_0] = 0.$$

Conclusion. We have exhibited an initial pair $(b_0, \hat{b}_0)$ with $b_0 \in S$ and $d(b_0, \hat{b}_0) \leq \delta$ such that:

- the naive IG-max policy selects $a_t = a_0$ at every step (Step 3),
- the belief trajectory satisfies $b_t \in S$ almost surely for all t (Step 4),
- the expected progress potential has zero drift: $\mathbb{E}[\Psi(b_{t+1}) - \Psi(b_t) \mid b_t] = 0$ (Step 5).

Therefore, S satisfies the BTR definition under naive IG-max.

B.5 Proof of Theorem 2.4

We construct an explicit counterexample satisfying all conditions of the theorem.

Construction. Let the hypothesis space be

$$\Theta = \{\theta_1, \theta_2\},$$

and let the initial belief be uniform:

$$b_0(\theta_1) = b_0(\theta_2) = \tfrac{1}{2}.$$

Fix any $\kappa \in (0, 1/2)$ and choose $\rho \in (0, \kappa)$. Let the action set be

$$\mathcal{A} = \{a_{\text{col}}, a_{\text{safe}}\}.$$

Define the observation space $\mathcal{X} = \{x_0\}$ with a single observation, and set

$$P(x_0 \mid \theta, a) = 1 \qquad \text{for all } \theta \in \Theta,\ a \in \mathcal{A}.$$

This observation model is deliberately hypothesis-independent: it is a valid probability kernel over $\mathcal{X}$, but every observation is completely uninformative about θ. Consequently, the exact Bayesian updater T_B satisfies $T_B(b_0, a, x_0) = b_0$ for both actions, so no exact-Bayesian agent could reduce entropy at all in this environment.

Define the approximate belief updater $\widehat{T}\colon \Delta(\Theta) \times \mathcal{A} \times \mathcal{X} \to \Delta(\Theta)$ by

$$\widehat{T}(b_0, a_{\text{col}}, x_0) = (1 - \rho,\ \rho), \qquad \widehat{T}(b_0, a_{\text{safe}}, x_0) = \big(\tfrac{1}{2},\ \tfrac{1}{2}\big).$$

Both outputs lie in $\Delta(\Theta)$, so $\widehat{T}$ is a well-defined measurable map into the simplex, as required by the definition of an approximate updater in Section B.1. Note that $\widehat{T}(\cdot, a_{\text{col}}, \cdot)$ moves probability mass toward the boundary of $\Delta(\Theta)$ despite the observation being uninformative — precisely the failure mode of amortized updaters identified in (A2) of Appendix B.2.

Let the potential be the Shannon entropy $\Psi_{\text{ent}}(b) = H(b)$.

Entropy-IG-max selects a_{col}. Since $|\mathcal{X}| = 1$, the expectation over observations is trivial. The entropy reduction under each action is:

$$H(b_0) - H\Big(\widehat{T}(b_0, a_{\mathrm{safe}}, x_0)\Big) = H\big(\tfrac{1}{2}, \tfrac{1}{2}\big) - H\big(\tfrac{1}{2}, \tfrac{1}{2}\big) = 0,$$

$$H(b_0) - H\Big(\widehat{T}(b_0, a_{\mathrm{col}}, x_0)\Big) = H\big(\tfrac{1}{2}, \tfrac{1}{2}\big) - H(1-\rho,\, \rho).$$

Since $\rho \in (0, 1/2)$, the binary entropy function is strictly concave and achieves its maximum at the uniform distribution, so

$$H(1-\rho,\, \rho) < H\big(\tfrac{1}{2}, \tfrac{1}{2}\big)\,,$$

and therefore the entropy reduction under a_{col} is strictly positive:

$$H(b_0) - H\Big(\widehat{T}(b_0, a_{\mathrm{col}}, x_0)\Big) > 0 = H(b_0) - H\Big(\widehat{T}(b_0, a_{\mathrm{safe}}, x_0)\Big)\,.$$

Hence entropy-based IG-max strictly prefers a_{col} over a_{safe}.

The resulting belief violates κ-safety. The belief after taking a_{col} is

$$b_1 = \widehat{T}(b_0, a_{\mathrm{col}}, x_0) = (1-\rho,\, \rho).$$

Since $\rho < \kappa$ by construction,

$$\min_{\theta \in \Theta} b_1(\theta) = \rho < \kappa,$$

so $b_1 \notin \Delta_\kappa(\Theta) := \{b \in \Delta(\Theta) : \min_\theta b(\theta) \geq \kappa\}$. Entropy-based IG-max has driven the belief outside the κ-safe interior.

A safe action exists. Under a_{safe}, the updated belief is

$$\widehat{T}(b_0, a_{\mathrm{safe}}, x_0) = \big(\tfrac{1}{2},\, \tfrac{1}{2}\big),$$

so

$$\min_{\theta \in \Theta} \widehat{T}(b_0, a_{\mathrm{safe}}, x_0)(\theta) = \tfrac{1}{2} \geq \kappa,$$

since $\kappa < 1/2$. Hence a_{safe} keeps the belief in $\Delta_\kappa(\Theta)$.

Conclusion. We have exhibited a finite hypothesis space Θ, an action set $\mathcal{A}$, a valid observation model P, and a well-defined approximate updater $\widehat{T}$ such that:

1. entropy-based IG-max selects a_{col}, producing a belief $b_1 \notin \Delta_\kappa(\Theta)$, and
2. a safe action a_{safe} exists whose updated belief remains in $\Delta_\kappa(\Theta)$.

Therefore, maximising entropy reduction does not guarantee κ-safety, even when a safe action exists.

B.6 Proof of Theorem 2.5

Fix an arbitrary action $a \in \mathcal{A}$. By assumption, the oracle information gain $IG(\cdot, a)$ is L-Lipschitz with respect to the ℓ_1 metric. Therefore, for any two beliefs $b, b' \in \Delta(\Theta)$,

$$\big|IG(b, a) - IG(b', a)\big| \leq L\|b - b'\|_1.$$

Taking $b = b_t$ and $b' = \hat{b}_t$ gives

$$\Delta_{IG} := \big|IG(b_t, a) - IG(\hat{b}_t, a)\big| \leq L\|b_t - \hat{b}_t\|_1.$$

It remains to relate the proxy error to the internal-belief and model-belief errors. By the triangle inequality for the ℓ_1 norm,

$$\|\hat{b}_t - b_t\|_1 \leq \|\hat{b}_t - b_t^{\mathrm{int}}\|_1 + \|b_t^{\mathrm{int}} - b_t\|_1.$$

Substituting this bound into the Lipschitz inequality yields

$$\Delta_{IG} \leq L\left(\|\hat{b}_t - b_t^{\mathrm{int}}\|_1 + \|b_t^{\mathrm{int}} - b_t\|_1\right).$$

This proves the claimed bound.

C Experiment Details

Default hyperparameters. DR-IG uses an L_1 ambiguity radius $\varepsilon = 0.5$, interpolation weight $\alpha = 0$ (pure `hard_min` worst-case aggregation; see §3.5 for the ablation establishing $\alpha = 0$ as required for solve), $M = 8$ particle-belief samples for the inner expectation in Eq. (4), and a shortlist of 50 candidate hypotheses for the inner worst-case search. The verbalized belief proxy uses $M = 10$ confidence-annotated samples per round. Each evaluation cell reports $n = 50$ episodes unless explicitly indicated otherwise; cell-level standard deviations across 2–3 random seeds remain below 5 percentage points for all Oracle and LLM-side cells, and the qualitative LLM-side-collapse pattern is invariant across seeds.

CD-500. On the CD task ($|\Theta| = 500$, max_turns $= 10$), we evaluate `gpt-4o-mini`, `claude-haiku-4.5`, `gemini-2.5-flash-lite`, `deepseek-chat-v3.1`, and `qwen-2.5-7b-instruct` ($n = 50$ episodes per cell). The vanilla baseline (no IG, no maintained belief; commitment triggered when verbalized self-reported max-prob ≥ 0.99) is reported on all 5 agents at $n = 50$. The ablation conditions ($\varepsilon = 0$ / nominal, $\alpha = 0.5$ / interpolated, $\alpha = 1.0$ / mean only, random top-K) use $n = 30$ on the 4 shared agents. The ε/α sweep uses $n = 30$ on `gpt-4o-mini` and `claude-haiku-4.5`. All cells use the verbalized belief probe with $M = 10$ confidence-weighted samples per round; the IG / DR-IG variants pass a top-$K = 5$ shortlist to the LLM as forced-choice; the random-top-K ablation replaces the IG ranking with a uniform random permutation while keeping the constrained framework intact.

CI-500. On the CI task ($|\Theta| = 500$, $\max$ _turns $= 12$, action space $|\mathcal{A}| = 4095$ non-empty subsets of 12 items), we evaluate `gpt-4o-mini`, `gemini-2.5-flash-lite`, `gemini-2.5-flash`, `gemini-2.0-flash`, and `gemini-flash-latest` ($n = 50$ episodes per cell, fixed seed 0 for the rule pool so the same 500 subset-threshold rules are used across all agents). Each rule is generated by sampling a uniform non-empty subset $S \subseteq \{A, \ldots, L\}$ and a threshold $k \in \{1, \ldots, |S|\}$ uniformly; the feedback matrix $f_{S,k}(x) \in \{0, 1\}$ is precomputed over all 4095 candidate queries. The vanilla baseline (no IG, no maintained belief; commitment triggered when verbalized self-reported max-prob ≥ 0.99) and the IG / DR-IG / Oracle modes all use the verbalized belief probe with $M = 10$ confidence-weighted samples per round; the IG and DR-IG variants pass a top-$K = 5$ shortlist of candidate subsets to the LLM as forced-choice. Solve is defined as the agent committing to the correct rule ID by $\max$ _turns; BTR_{std} uses the same 5-step stagnation-length detector as CD-500.

Proxy Belief Implementation. At round t, we sample M verbalized pairs $\{(\hat{\theta}_t^{(m)}, c_t^{(m)})\}_{m=1}^{M}$, aggregate them into a normalized proxy $\hat{b}_t$, and use $\hat{b}_t$ as the planning state for k-step lookahead IG-max. This yields a substantially cheaper alternative to large-sample Monte Carlo belief estimation while making the approximation error explicit at the belief-state level. Because sparse verbalized proxies can distort action ranking and underweight weak but still viable hypotheses (Zhao et al., 2024; Huang et al., 2025; Zhang et al., 2026), our two core mechanisms are particularly important here: DR-IG provides robustness to belief misspecification, while the boundary-safe objective discourages trajectories that collapse toward the simplex boundary.

D Additional Results

D.1 ε / α Sweeps

To isolate when the robust term in Eq. (4) becomes operationally relevant, we vary $p_{\text{peak}} \in \{0.3, 0.5, 0.7\}$ together with different ambiguity radii. Each $(\text{task}, p_{\text{peak}}, \varepsilon)$ configuration aggregates 200 trials across 10 secrets. In Fig. 2, the effect on CD is highly localized: for $p_{\text{peak}} = 0.5$, naive IG ($\varepsilon = 0$) collapses to zero retention, whereas any positive ambiguity radius restores performance and yields monotonic gains, reaching 0.93 at $\varepsilon = 1.5$. For other values of p_{peak}, CD remains largely insensitive to ε. This behaviour is quantified in Tab. 6 with the most adversarial configuration (CD with $p_{\text{peak}} = 0.5$). Here, the robust term is essential: increasing ε lifts retention from 0.00 to 0.51, and further to 0.93 at $\varepsilon = 1.5$, with a consistent

Table 6: IG retention vs. DR-IG radius ε (CD, $p_{\text{peak}} = 0.5$).

ε	Retention
0.0	0.00
0.1	0.51
0.3	0.73
0.5	0.75
0.8	0.82
1.0	0.84
1.5	**0.93**

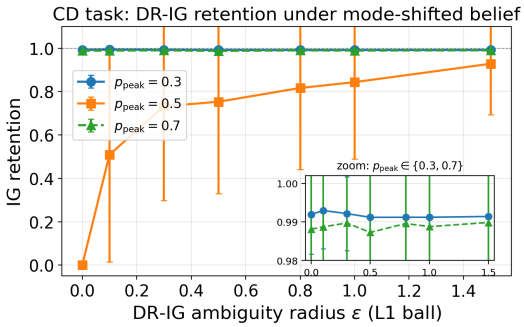


Figure 2: IG retention under mode-shifted beliefs as a function of ε. $p_{\text{peak}} = 0.5$ (orange) shows naive IG ($\varepsilon = 0$) collapsing to retention 0, while any positive ε recovers retention monotonically, reaching 0.93 at $\varepsilon = 1.5$. $p_{\text{peak}} \in \{0.3, 0.7\}$ stay near 1.0; the lower-right inset zooms on $[0.98, 1.0]$ to separate them (0.992 vs 0.989).

Table 7: CD-500 DR-IG sweep over ε and α.

ε sweep ($\alpha = 0$, hard_min, fixed)				
ε	0	0.1 / 0.3	0.5 (full)	0.8 / 1.0 / 1.5
GPT-4o-mini	0/100	0/0 / 0/0	**30/0**	0/0 / 0/0 / 0/0
Claude Haiku 4.5	0/100	0/0 / 0/0	**36/2**	0/0 / 0/0 / 0/0
α sweep ($\varepsilon = 0.5$, interpolated, fixed)				
α	0 (hard_min)	0.25	0.5	0.75 / 1.0 (mean)
GPT-4o-mini	**30/0**	0/0	0/0	0/0 / 0/100
Claude Haiku 4.5	**36/2**	0/0	0/0	0/0 / 0/100

monotonic trend. Outside this regime, all other $(\text{task}, p_{\text{peak}})$ combinations exhibit variation below 0.5%.

To check whether $\varepsilon = 0.5$ and $\alpha = 0$ form a robust plateau or a single-point sweet spot, we sweep along each axis separately in Tab. 7. Two findings: *(a) ε sweep ($\alpha = 0$ fixed)*: only $\varepsilon = 0.5$ produces non-zero solve; $\varepsilon \in \{0.1, 0.3, 0.8, 1.0, 1.5\}$ all yield 0% solve while keeping BTR_{std} at 0%. The BTR-safety property is robust under any $\varepsilon > 0$, but the solve recovery has a single peak at $\varepsilon = 0.5$ on CD-500 — reflecting that this scale matches the typical $\|\hat{b} - b\|_1$ induced by the verbalized proxy. *(b) α sweep ($\varepsilon = 0.5$ fixed)*: solve degrades sharply with any mean-mixing. Even $\alpha = 0.25$ (75% worst-case $+25\%$ mean) collapses solve to 0%, while $\alpha = 1.0$ recovers the full 100% BTR. Pure worst-case is therefore strictly required, not merely encouraged.

D.2 Boundary-Safe Potential

Thm. 2.4 motivates replacing the entropy potential Ψ_{ent} with the boundary-safe variant Ψ_{repel}. To evaluate its empirical behavior, we sweep λ under *Oracle* + Ψ_{repel}, while fixing $\varepsilon_{\text{repel}} = 10^{-8}$. The smallest value corresponds to the default $1/|\Theta|$ prior and effectively disables repulsion. In parallel, we sweep ε to contrast with a regime where the verbalized belief proxy remains well within the interior of the simplex, so that the inner $\inf$ in Eq. (4) plays only a limited role.

As shown in Fig. 3 (left), the effect of λ is highly non-linear. Increasing λ from the prior level ($\approx 2{\cdot}10^{-6}$) to 10^{-2} causes the Oracle solve rate to collapse from 96% to 2%, while the BTR rate simultaneously rises from 0% to 98%. This sharp transition is further detailed in Tab. 8: performance remains stable up to $\lambda = 10^{-4}$, but deteriorates rapidly beyond this point, with a two-thirds drop in solve rate already at $\lambda = 10^{-3}$. In contrast, the DR-IG ε sweep in Fig. 3 (right) exhibits a much milder profile, with only marginal variation in solve rate (0–4%) and BTR_{std} (8–16%), indicating that in the absence of explicit boundary pressure the belief trajectory remains largely interior.

Table 8: Oracle + Ψ_{repel} ($n = 50$ per cell). The repulsion weight λ exhibits a sharp threshold: at $\lambda = 10^{-3}$ the solve rate drops by two-thirds, and at $\lambda = 10^{-2}$ the agent is trapped.

λ	$1.98{\cdot}10^{-6}$ (prior)	10^{-4}	10^{-3}	10^{-2}
Solve (%)	**96**	92	34	2
BTR_{std} (%)	**0**	0	58	98

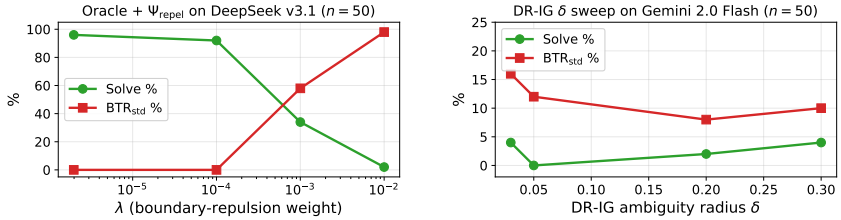


Figure 3: *Left:* As the boundary-repulsion weight λ grows from $1/|\Theta| \approx 2{\cdot}10^{-6}$ to 10^{-2}, the Oracle solve rate collapses from 96% to 2% with BTR rate exploding from 0% to 98%. *Right:* the DR-IG ε sweep exhibits a much milder profile (0–4% solve, 8–16% BTR_{std}) across the ε range, indicating that without explicit boundary-repulsion the verbalized proxy stays in the interior of $\Delta(\Theta)$.

These results highlight a trade-off. When λ is too small, boundary repulsion is ineffective and provides no additional safety; when it is too large, it actively impedes posterior concentration on $\theta^\star$, reintroducing a BTR. The useful operating regime is therefore: $\lambda \leq 10^{-4}$ preserves Oracle-level performance while mitigating the entropy-only failure modes identified in Thm. 2.4. This supports a phase-aware interpretation of Ψ_{repel}: boundary repulsion should act as a transient safeguard during early, diffuse belief states, and be attenuated or removed once the posterior has sufficiently concentrated. The contrast with the comparatively mild ε sweep further clarifies that boundary control and ambiguity control address distinct failure modes. Specifically, ε mitigates ranking inversion under *interior* belief error (§3.3), whereas λ prevents support collapse at the simplex *boundary*.

D.3 Top-K Sweep: Per-model Cells with Wilson Intervals

Per-model detail for Tab. 4. Each cell is an independent run; solve and BTR_{std} are binomial proportions with Wilson score 95% intervals. $K = 5$ rows reuse the main-table cells ($n = 50$); all other rows are sweep cells ($n = 30$).

Table 9: CD-500 top-K sweep, per model. Cells: solve% [95% CI] / BTR_{std}% [95% CI].

Mode	Model	$K = 1$	$K = 3$	$K = 5$	$K = 10$
IG	`gpt-4o-mini`	0.0 [0,11] / 100 [89,100]	0.0 [0,11] / 100 [89,100]	0.0 [0,7] / 100 [93,100]	43.3 [27,61] / 3.3 [1,17]
	`gemini-2.5-flash-lite`	0.0 [0,11] / 100 [89,100]	0.0 [0,11] / 100 [89,100]	0.0 [0,7] / 100 [93,100]	43.3 [27,61] / 3.3 [1,17]
DR-IG	`gpt-4o-mini`	3.3 [1,17] / 0.0 [0,11]	36.7 [22,55] / 0.0 [0,11]	30.0 [19,44] / 0.0 [0,7]	13.3 [5,30] / 0.0 [0,11]
	`gemini-2.5-flash-lite`	3.3 [1,17] / 0.0 [0,11]	20.0 [10,37] / 0.0 [0,11]	14.0 [7,26] / 0.0 [0,7]	6.7 [2,21] / 0.0 [0,11]

At $K = 20$ (omitted from the table for width) IG reaches 50.0% [33,67] on `gpt-4o-mini` and 43.3% [27,61] on `gemini-2.5-flash-lite` with $\text{BTR}_{\text{std}} = 3.3\%$ [1,17] on both, while DR-IG is at 10.0% [4,26] / 0.0% [0,11] on both. Adjacent-K solve intervals overlap within a source, so it is the qualitative pattern — IG collapse at small K versus K-invariant DR-IG trap suppression — rather than any single point estimate that the sweep establishes. At $K \geq 10$ the shortlist stops binding and the LLM's own judgment can override the corrupted ranking (mean oracle IG of the selected action rises from 0.10–0.50 to 0.86).

D.4 Trajectory-level Invariance of Trapped Episodes

We re-analyze every BTR_{std}-positive episode in the CD-500 IG-constrain cells (5 model families $\times$ 50 episodes $= 250$ trapped episodes) from the detected entry step onward.

Table 10: Post-entry invariance in trapped CD-500 episodes. *Identical shortlist*: the regenerated top-K IG shortlist is unchanged at every subsequent turn. *Actions confined*: every subsequent action lies inside the entry-step shortlist. L_1 columns report the maximum consecutive change.

Model family	BTR-positive	Identical shortlist	Actions confined	max L_1 (proxy / oracle)
`gpt-4o-mini`	50/50	50/50	50/50	$0.0 / 3.6 \times 10^{-10}$
`claude-haiku-4.5`	50/50	50/50	50/50	$0.0 / 3.6 \times 10^{-10}$
`deepseek-chat-v3.1`	50/50	50/50	50/50	$0.0 / 3.6 \times 10^{-10}$
`gemini-2.5-flash-lite`	50/50	50/50	50/50	$0.0 / 3.6 \times 10^{-10}$
`qwen-2.5-7b-instruct`	50/50	50/50	50/50	$0.0 / 3.6 \times 10^{-10}$
Pooled	**250/250**	**250/250**	**250/250**	median entry step 5.0

Table 11: CD-500 action-level metrics, pooled over 5 vendors $\times$ 50 episodes per mode. *Mean rank* is the oracle-IG rank of the chosen action within the full 1024-action space (lower is better). *Repeat rate* is the fraction of turns repeating an earlier action.

Mode	turns	mean oracle IG (chosen)	mean rank / 1024	repeat rate	mean entropy drop
IG-constrain	2500	0.496	698.4	0.500	0.554
DR-IG	2481	**0.830**	**465.9**	**0.086**	**0.858**
Oracle IG	2490	0.898	57.0	0.000	1.001

In all 250 episodes the guidance belief is frozen after entry (maximum consecutive L_1 change exactly 0.0; $\leq 3.6 \times 10^{-10}$ for the oracle posterior), the regenerated shortlist is identical at every subsequent turn, and every subsequent action stays inside the entry-step shortlist. Within the observed horizon the closed loop — frozen belief $\rightarrow$ identical ranking $\rightarrow$ identical candidate set $\rightarrow$ repeated action — does not leave the visited state set. This is an empirical instantiation of forward invariance under the evaluated policy; it remains a finite-horizon observation rather than a proof of absorbingness for all admissible actions.

D.5 Action-level Mechanism: Where the Corrupted Ranking Sends the Agent

Pooling all turns over all vendors and episodes on CD-500 ($|\mathcal{A}| = 1024$):

The per-turn breakdown localizes the failure precisely. Through turn 5 the IG-constrain agent selects actions with near-maximal oracle IG ($1.00, 1.00, 0.997, 0.989, 0.979$); from turn 6 onward the chosen action's oracle IG is exactly 0.000 and its rank collapses to 1021.9, then 1024.0 for every remaining turn — the least informative actions in the *full* action space, not merely the worst within the shortlist — with within-window repetition rate 1.00. Under the identical interface DR-IG keeps positive oracle IG at every turn ($0.652, 0.443, 0.891, 0.797, 0.527$ for turns 6–10) with repetition rate 0.086, and oracle IG has repetition rate 0.000. This is the separation that the ablation in §3.5 predicts: breaking repetition is not the same as recovering an informative direction.

D.6 Elicitation-width Sweep

The verbalized proxy is obtained from a structured elicitation of width M: the model is asked for its top-M candidate hypotheses with confidences, and any hypothesis it does not name receives zero mass. M therefore bounds the size of the *active* support at the current step, not the dimensionality of Θ.

Table 12: CD-500 DR-IG under varying elicitation width ($n = 30$ per cell; $M = 10$ is the main-table $n = 50$ cell). Cells: solve% / BTR_{std}%.

Model	$M = 3$	$M = 5$	$M = 10$ (paper)	$M = 20$
`gpt-4o-mini`	26.7/0	26.7/0	30.0/0	26.7/0
`gemini-2.5-flash-lite`	16.7/0	16.7/0	14.0/0	16.7/0

Solve rates are flat across nearly an order of magnitude in M (differences within binomial noise at these cell sizes) and trap suppression is invariant at 0% throughout. This is consistent with the support-width reading: past a handful of candidates the model adds only low-confidence tail mass, which does not change the induced action ranking. Accordingly, M can be fixed to the available inference budget rather than tuned to the (typically unknown) hypothesis-space size.

D.7 Preference-elicitation Tasks: LP-500 and PP-500

LatentPreference-500 (LP-500). 500 user profiles over 30 abstract items A01–A30 (8 cluster prototypes plus $\mathcal{N}(0, 0.25^2)$ perturbation). Action: pairwise preference query 'A_i vs A_j' (435 unordered pairs); binary feedback. The primary metric is the exact-profile-ID solve.

Table 13: LP-500 exact-profile-ID solve ($n = 50$, 2 LLMs). Cells: solve% / $\mathrm{BTR}_{\mathrm{std}}$%.

LLM	Oracle	DR-IG	IG	Vanilla
GPT-4o-mini	100/0	2/32	0/100	0/44
Gemini 2.5 Flash Lite	100/0	2/34	0/100	2/ 4
Mean	**100/0**	**2/33**	**0/100**	**1/24**

The CD-500 pattern (Oracle 100%, IG 100% BTR, Vanilla $\sim 1\%$) reproduces on a semantically distinct preference-elicitation task. DR-IG is direction-correct ($2\% > 1\%$), but the absolute solve is small at this $|\Theta|$.

PreferencePrediction-500 (PP-500). Same environment as LP-500, but the primary metric is replaced by held-out preference prediction: at episode end we sample 10 held-out (i, j) pairs (fixed seed 1234, per-episode filtered), and the LLM predicts each pair's preference via a forced post-hoc call.

Table 14: PP-500 LLM held-out preference accuracy ($n = 10$ smoke, 2 LLMs). Cells: LLM-acc% / $\mathrm{BTR}_{\mathrm{pref\text{-}std}}$%.

LLM	Oracle	DR-IG	IG	Vanilla
GPT-4o-mini	55/0	47/70	45/80	54/60
Gemini 2.5 Flash Lite	49/0	39/70	41/80	46/70
Mean	**52/0**	**43/70**	**43/80**	**50/65**

All modes hover near the binary-chance baseline (50%) at the LLM read-out level. The Bayes-optimal posterior, in contrast, reflects DR-IG's superior posterior concentration (DR-IG $\sim 87\%$ vs IG $\sim 70\%$): LLM read-out noise erases the posterior gap. The IG $\mathrm{BTR}_{\mathrm{pref\text{-}std}}$ rises to 80%, confirming that the framework signal is robust across both ID and predictive metrics, while the primary user-facing accuracy is non-discriminating.

D.8 Verbalized Proxy on Free-Text Situation Puzzles

We evaluate the verbalized belief proxy from §2.3 on *SituationPuzzles (SP)* (Zhou et al., 2025). The agent is expected to unravel a paradoxical puzzle by posing yes/no questions to a judge model, where the hypothesis space Θ is genuinely open-ended. Unlike finite-Θ benchmarks, the hypothesis space here is unbounded and cannot be explicitly enumerated. Accordingly, we employ the agent-side information gain objective (Eq. (3)) with an exploratory placeholder potential, and report *keypoint coverage* alongside strict solve rate.

Despite the absence of an explicit hypothesis set, the verbalized proxy remains operational: at each step, $\hat{b}_t$ forms a normalized distribution over sampled hypotheses, and DR-IG planning runs to completion in every episode. However, overall performance remains limited. Solve rates are low (3–10%), and $\mathrm{BTR}_{\mathrm{std}}$ remains close to unity, indicating that belief-trap behaviour persists in the open-vocabulary regime. At the same time, keypoint coverage (Tab. 15) reveals a more nuanced

Table 15: SP evaluation with the verbalized belief proxy ($M = 10$, $n = 30$ episodes per agent). *Keypoint coverage* is the mean fraction of ground-truth solution keypoints recovered by the agent's final answer. BTR_{std} is computed using a placeholder IG potential, since no closed-form posterior exists on free-text Θ.

Model	n	Solve (%)	Keypoint coverage	BTR_{std}
GPT-4o-mini	30	3.3	0.47	1.00
Claude Haiku 4.5	30	10.0	0.34	0.93

picture. Even when the agent fails to produce a fully correct explanation, the proxy consistently guides partial information acquisition, recovering 34–47% of the ground-truth solution keypoints on average. This suggests that the proxy retains a meaningful guidance signal beyond binary success metrics.

Overall, the verbalized belief proxy extends naturally to free-text settings, but also exposes a limitation: without a posterior-consistent potential, the agent remains vulnerable to persistent traps. Designing such potentials for open-ended Θ is an important direction for future work.

E Task Descriptions and Interaction Examples

For each evaluated task we give a 2–3 sentence formal description of the hypothesis space, action space, observation channel, and termination criterion, followed by one concrete example trajectory step. All tasks share the finite-Θ active-reasoning template: the environment samples a hidden $\theta^\star$ uniformly from Θ, the agent issues an action a_t at each turn, observes $x_t \sim \mathbb{P}(\cdot \mid \theta^\star, a_t)$, and after $\max$_turns commits a final guess (or, for behavior-prediction tasks, a held-out output prediction). Solve is defined as a correct final commitment unless otherwise noted; BTR_{std} is the 5-step stagnation-length detector defined in §3.1.

CD-500 (CircuitDecoding). Hypothesis space: $|\Theta| = 500$ Boolean circuits, each represented by a truth table $\Theta_i : \{0,1\}^{10} \to \{0,1\}$ on 10-bit inputs. Action space: $\mathcal{A} = \{0,1\}^{10}$ (1024 binary queries). Observation: deterministic 1-bit output; $\max$_turns $= 10$, information-theoretic floor $\lceil \log_2 500 \rceil = 9$. *Concrete example.* The hidden circuit is $\theta^\star = \#147$. Turn 1: agent queries `1010101010`; environment returns `1`, eliminating the ~ 250 candidates that disagree on this input. Turn 2: agent queries `0011110011`; returns `0`; etc. until at most one consistent circuit remains.

CI-500 (ConceptInduction). Hypothesis space: $|\Theta| = 500$ subset-threshold rules $f_{S,k}(x) = \mathbb{1}[|x \cap S| \geq k]$, where $S \subseteq \{A, \ldots, L\}$ is a hidden non-empty subset of 12 abstract items and $k \in \{1, \ldots, |S|\}$ is a hidden threshold. Action space: $\mathcal{A} = 2^{\{A,\ldots,L\}} \setminus \{\emptyset\}$ (4,095 non-empty subsets); observation: binary yes/no; $\max$_turns $= 12$. *Concrete example.* Hidden rule has $S = \{B, F, J\}$ and $k = 2$ ("the rule fires iff the query contains at least 2 of B, F, J"). Turn 1: agent queries $\{A, B, F\}$; environment computes $|\{A, B, F\} \cap \{B, F, J\}| = 2 \geq 2$, returns YES. Turn 2: agent queries $\{C, J\}$; intersection size is $1 < 2$, returns NO. The agent must induce both S and k from such yes/no traces.

LP-500 (LatentPreference). Hypothesis space: 500 latent user profiles, each with a hidden utility vector $u_\theta \in \mathbb{R}^{30}$ over 30 items $A_{01}..A_{30}$; profiles are drawn from 8 cluster prototypes plus $\mathcal{N}(0, 0.25^2)$ perturbation, so within-cluster profiles are correlated. Action space: pairwise preference queries (i, j) with $i < j$ (435 pairs); observation: 1 if $u_\theta(i) > u_\theta(j)$ else 0; $\max$_turns $= 10$. Solve = exact-profile-ID match. *Concrete example.* Hidden profile prefers $A_{05} \succ A_{12}$ and $A_{20} \succ A_{05}$. Turn 1: agent queries "A_{05} vs A_{12}" $\to$ `1`. Turn 2: queries "A_{05} vs A_{20}" $\to$ `0`.

PP-500 (PreferencePrediction). Same environment and hidden Θ as LP-500, but the primary metric is *held-out preference prediction*: at episode end the environment samples 10 unseen (i, j) pairs (fixed seed, per-episode filtered), and the agent's final posterior must predict each pair's preference via a forced post-hoc call. *Concrete example.* After 10 turns of pairwise queries the agent's final belief assigns 0.7 mass to a small set of consistent profiles; the held-out pair "A_{08} vs A_{17}" is then scored against the LLM's predicted preference.