Trust, Invert, or Verify? Decision-Making with an Uncertain Answer Source

Anonymous submission

Abstract

Information-seeking agents must decide not only what to ask, but whether the answers are worth using. We study this decision when question opportunities are limited, the task state is hidden, and the reliability of a single answer source is initially unknown. For binary questions, we model a stationary source as independently inverting each truthful answer with probability δ. At $\delta = 1/2$, the answers contain no information about the task state. Away from this boundary, independently verifiable probes identify whether to trust or invert the source with sign-error probability α using $O(\log(1/\alpha)/(1-2\delta)^2)$ probes. Under a finite question budget, the agent must balance questions that diagnose the source against those that distinguish plausible task states, with some questions advancing both goals. We instantiate a joint state–source belief in Collaborative Battleship, where a Captain with a limited question budget asks a fully informed Spotter yes/no questions about a hidden board while deciding where to fire. A deterministic inverting source reduces a source-blind QMD baseline from 0.770 to 0.510 final-board F1, while source-aware variants recover at least 0.735 F1. Across the measured deception rates, the value of source-aware asking reflects both the information carried by the channel and the questions spent identifying it. Gains over an ask-nothing baseline are small near $\delta = 1/2$, where answers carry little usable signal and distinguishing trust from inversion requires increasingly many probes. Farther from the boundary, the channel direction becomes easier to identify and each answer carries more usable information. Query-adaptive sources can recreate this bottleneck by passing public checks while corrupting harder-to-verify answers. Beyond the binary setting, we test the same decision principle in a synthetic transfer-verification POMDP with categorical observations and asymmetric losses. Joint inference improves simulated reward when calibrated, while source-model mismatch can misdirect belief updates and question selection. Source-aware planning therefore helps when the channel carries identifiable, task-relevant information, the source model matches the behavior encountered, and recovering that information is worth the cost of verification.

1 Introduction

Consider two answer sources. The first reports the truth with probability 0.9. The second reports the opposite of the truth with probability 0.9. Once an agent knows which source it faces, the two are equally informative: it should trust the first and invert the second. Before it knows the source, however, the same answer can support opposite conclusions. Determining the direction of the channel may require questions whose answers can be independently checked, and those questions consume resources that could otherwise be used to solve the task.

This creates a decision problem that is absent when source reliability is known. An information-seeking agent must decide how to divide limited question opportunities between distinguishing plausible task states and diagnosing the answer source, while acting under the task's own constraints. Individual questions may primarily serve one inferential goal or advance both. More verification can improve how later answers are interpreted, but it can also consume questions that would otherwise help solve the task. A source that adapts to the agent's public checks introduces a further difficulty because evidence collected to estimate reliability may no longer represent the answers that matter for the task. We ask:

> *Under a finite information budget, when should an agent spend resources identifying an answer source's reliability rather than using the source directly or acting without it, and what additional verification is needed when the source adapts to the agent's queries?*

For pedagogical background, Kaelbling, Littman, and Cassandra (1998) surveys POMDP planning, Ross, Chaib-draa, and Pineau (2007) introduces Bayes-adaptive POMDPs, and Pelc (2002) surveys searching with unreliable answers. Related work studies label noise, margin-based risk, and crowd models (Angluin and Laird 1988; Massart and Nédélec 2006; Dawid and Skene 1979), while searching-with-lies often assumes a known lie budget or an error rate below one half (Rényi 1961; Pelc 2002). Our joint task–source posterior follows the Bayes-adaptive view. We focus on what an agent must spend to learn whether one source should be trusted or inverted, with questions for source diagnosis competing directly with questions for solving the task. Unlike indirect prompt injection (Greshake et al. 2023; Debenedetti et al. 2024), the source changes the content of its answers rather

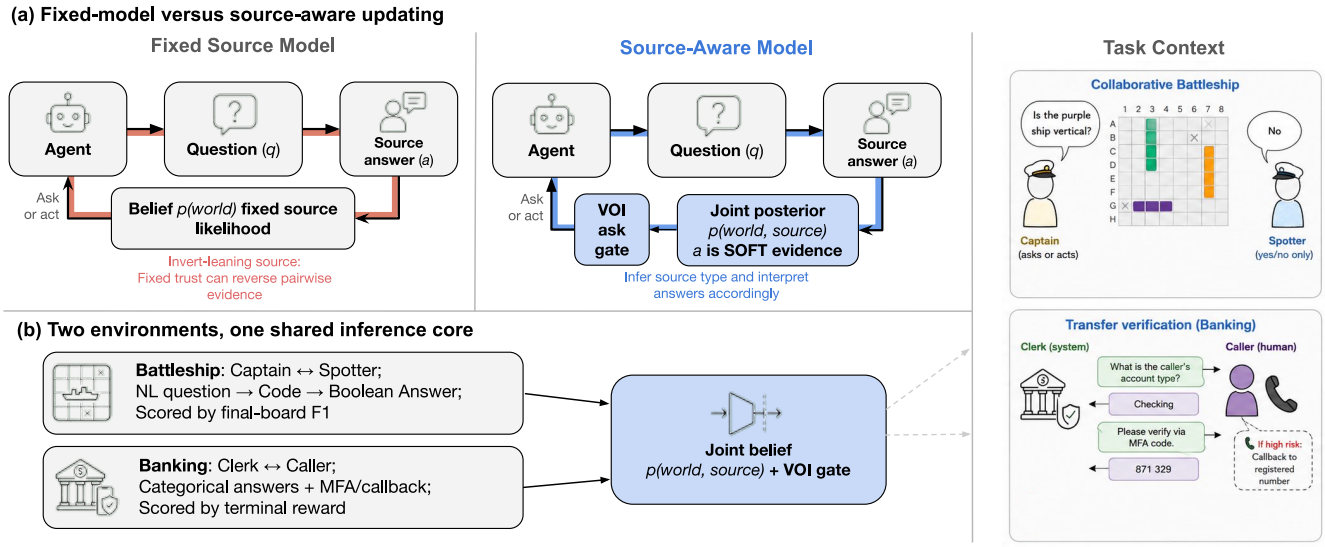

Figure 1: Fixed-likelihood and source-aware updating. The source-aware model jointly infers task and source, gates questions by value of information, and is evaluated in Collaborative Battleship and transfer verification.

than inserting instructions, so the defense operates through belief updating and question selection.

Our primary controlled environment is Collaborative Battleship (Grand et al. 2026). At each turn, a *Captain* chooses whether to fire at the hidden fleet or ask a fully informed *Spotter* a natural-language yes/no question. The benchmark was introduced to study cooperative information seeking under a fixed trust-leaning Spotter model. We preserve the task and its Bayesian QMD planner but change the observation model. The Spotter may be honest, noisy, systematically inverting, or query-adaptive. Because the hidden board supplies ground truth for every answer, Collaborative Battleship lets us vary the Spotter's behavior while holding the task and planner fixed. This isolates how source uncertainty and verification affect the Captain's questions and decisions.

We augment the task belief with a source variable (Figure 1). The resulting joint posterior over world state and source type treats each answer as evidence whose interpretation depends on the inferred source. A value-of-information rule can then credit a question for distinguishing task states, diagnosing the source, or doing both. In Battleship, a deterministic inverting source reduces the source-blind QMD Captain from 0.770 to 0.510 final-board F1, while source-aware variants recover at least 0.735 F1. The recovery is not free: diagnosing the source can require additional questions, and the value of doing so depends on how much usable information remains in the channel.

In sum, our work provides theoretical, methodological, and empirical contributions. Theoretically, we characterize when an uncertain binary source is identifiable, how verification cost grows near $\delta = 1/2$, and why source identification alone does not guarantee task-state recovery. Methodologically, we extend QMD with a joint state–source belief that values questions for task information, source diagnosis, or both. Empirically, we evaluate this approach against stationary and query-adaptive sources in Collaborative Battleship and test the same decision principle in a categorical transfer-verification POMDP, showing both the gains from calibrated source inference and its sensitivity to source-model mismatch.

Together, these results show that source diagnosis helps when the source is identifiable, the available questions distinguish relevant task states, and the resulting information repays the cost of verification. Adaptive or misspecified sources can violate these conditions.

2 Background and Problem Setup

2.1 Collaborative Battleship and QMD

In Collaborative Battleship (Grand et al. 2026), a Captain faces an 8×8 grid concealing a small fleet. At each stage, the Captain chooses whether to fire at a cell or ask a fully informed Spotter a free-form yes/no question. A shot reveals ship or water. Each question is translated into a program that is executed against the hidden board, producing the truthful boolean answer available to the Spotter. An episode permits at most $M_{\max}{=}40$ firing moves and $Q_{\max}{=}15$ questions. These limits are separate, so a full episode may contain up to 55 stages.

Performance is measured by *final-board F1*. If the Captain records h hits and m misses on a board with s ship cells, then $\mathrm{F1} = 2h/(s + h + m)$. Hits increase the score and misses decrease it. Questions do not enter raw F1 directly, but source-diagnostic questions use slots that could otherwise be spent asking task-directed questions. We therefore also report question counts and a cost-adjusted score where noted.

Grand et al.'s +Bayes-QMD Captain combines Bayesian strategies for question selection, move selection, and the decision of whether to ask. The upstream Planner uses a fixed

trust-leaning observation model with error rate $\hat{\delta} = 0.1$: sampled boards agreeing with an answer receive weight 0.9 and boards disagreeing receive weight 0.1. It asks when the discounted expected value of the best question exceeds that of firing. This source-blind model is appropriate for a mostly reliable Spotter but cannot infer that a source is systematically inverting.

2.2 Source-Uncertain Decision Model

Let $B \in \mathcal{B}$ denote the hidden task state, such as the ship configuration in Battleship. Let $S \in \mathcal{S}$ denote the unknown source type, such as an honest, noisy, or systematically inverting source. The agent begins with a joint prior $P_0(B, S)$ over both.

When the agent asks question q_t, its correct answer under state B is $T_{q_t}(B) \in \{0, 1\}$, while $a_t \in \{0, 1\}$ is the answer returned by the source. Given the preceding history h_{t-1}, we write

$$\ell_S(a_t \mid B, q_t, h_{t-1}) := P(a_t \mid T_{q_t}(B), S, q_t, h_{t-1})$$

for the probability that source type S would return answer a_t. The history dependence allows the source to adapt to earlier interactions. For a stationary source, it can be omitted.

Our theoretical analysis considers a stationary binary source that independently flips each correct answer with probability δ:

$$a_t = T_{q_t}(B) \oplus z_t, \qquad z_t \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\delta).$$

Thus answers are more often correct when $\delta < 1/2$ and more often inverted when $\delta > 1/2$. At $\delta = 1/2$, the returned answer is independent of the correct answer and supplies no information about the task state. The next section studies what can and cannot be identified from this channel.

3 The Identifiability Frontier

For the stationary binary channel, $|\delta - \frac{1}{2}|$ governs the answer-side difficulty of learning whether answers should be trusted or inverted. For N asked questions, define $m_N(B) = \#\{t \leq N : a_t \neq T_{q_t}(B)\}$. The likelihood under world B and rate δ is $L_N(B, \delta) = \delta^{m_N(B)}(1-\delta)^{N-m_N(B)}$. When both channel orientations remain admissible, a world that complements the asked truths can be paired with the reflected rate $1 - \delta$. This reflection ambiguity is separate from the question of whether the selected queries distinguish the task states. The propositions below make these distinctions precise. Formal statements and proofs appear in Appendix A.

Proposition 1 (zero answer information at the half-point, informal). *At $\delta = \frac{1}{2}$, each source answer is a fair coin conditional on the question and prior history. Observing the answer therefore leaves the task posterior unchanged from its pre-answer value.* The result concerns the incremental information supplied by the answer stream. Task-grounded observations, such as revealed Battleship cells, can still update the task posterior.

Proposition 2a (source-orientation identification, informal). *For fixed $\delta \neq \frac{1}{2}$, a majority test on independently verifiable probes identifies whether the source is trust-leaning or invert-leaning with error at most $\exp[-N(1-2\delta)^2/2]$. Thus $O(\log(1/\alpha)/(1-2\delta)^2)$ probes suffice for error probability α, and the $(1-2\delta)^{-2}$ dependence is necessary at fixed confidence near the half-point.*

Proposition 2b (reflection and task-state identification, informal). *If both channel orientations are profiled over, or integrated under an exactly reflection-symmetric source prior, a task state that complements every asked truth has the same evidence when paired with rate $1-\delta$. Fixing the correct orientation removes this exact tie. In either case, task-state recovery additionally requires the policy to accumulate distinguishing evidence against every positive-prior rival.* For known δ, a sufficient finite-state condition is that every rival be separated by infinitely many selected queries. The literal cellwise complement of a legal Battleship board is illegal, but recovery still depends on the actual query sequence rather than that structural fact alone.

Proposition 3 (mis-signed soft evidence, informal). *Suppose a source-blind planner uses a fixed trust-leaning likelihood $0 < \hat{\delta} < \frac{1}{2}$ while the true source is invert-leaning, $\delta > \frac{1}{2}$. On any query that separates the true task state from a rival, the expected log-posterior-odds increment favoring the truth is $(1-2\delta)\log[(1-\hat{\delta})/\hat{\delta}] < 0$.* The upstream Planner is this soft case with $\hat{\delta} = 0.1$. The proposition establishes reversed pairwise evidence on separating queries, not a universal floor on task utility. Its effect on final-board F1 also depends on the query policy, task geometry, and information obtained from firing.

Implications for the experiments. Honest ($\delta = 0$) and noisy ($\delta = 0.25$) sources are trust-leaning, while a deterministic inverter ($\delta = 1$) is invert-leaning. Rates δ and $1 - \delta$ have the same binary-channel capacity, but finite-task performance can differ because orientation must be learned from limited probes and the selected questions must distinguish task states. Source diagnosis becomes more expensive as δ approaches $\frac{1}{2}$, motivating the full-axis sweep in Section 5.3. Query-adaptive sources fall outside this stationary analysis and are evaluated separately.

4 Method: Source-Aware QMD

Joint belief and answer likelihood. Using the notation of Section 2.2, the source-aware method maintains the joint posterior $p_t(B, S) = P(B, S \mid h_t)$, initialized by $p_0(B, S) = P_0(B, S)$. For a stationary component $S \in \{\text{honest}, \text{noisy}_\epsilon, \text{lying}\}$ and a translated boolean question q_t, the answer likelihood is

$$\ell_S(a_t \mid B, q_t) = \begin{cases} 1 & S = \text{honest},\ a_t = T_{q_t}(B), \\ 1-\epsilon & S = \text{noisy}_\epsilon,\ a_t = T_{q_t}(B), \\ \epsilon & S = \text{noisy}_\epsilon,\ a_t \neq T_{q_t}(B), \\ 1 & S = \text{lying},\ a_t = 1 - T_{q_t}(B), \\ 0 & \text{otherwise.} \end{cases}$$

Adaptive components may additionally condition this likelihood on the history h_{t-1} and on question properties such as public checkability. After an answer, the posterior becomes

$$p_{t+1}(B, S) \propto \ell_S(a_t \mid B, q_t, h_{t-1}) p_t(B, S).$$

Task-grounded reveals use their own observation likelihood. For firing, the source variable is marginalized. If $X_c(B)$ indicates whether state B contains a ship at cell c, then

$$p_t(X_c = 1) = \sum_{B,S} p_t(B, S) X_c(B).$$

Question scoring and ask gate. Each candidate question receives the net score

$$\mathrm{VOI}(q) = \mathrm{AV}(q) + w_{\mathrm{eff}} \mathrm{IG}_S(q) - \kappa,$$

$$w_{\mathrm{eff}} = W \frac{H[p_t(S)]}{H_{\max}}.$$

Here, $\mathrm{AV}(q)$ is the expected improvement in the best next-shot hit probability after observing the answer. $\mathrm{IG}_S(q)$ is the expected reduction in source entropy, W controls the maximum value assigned to source information, and $H_{\max}$ is the maximum entropy over the configured source components. Entropy scaling reduces the source bonus as the source posterior becomes resolved.

The flagship gate asks only when the best candidate has positive net value, $\mathrm{VOI}(q^\star) > 0$. Older comparison variants retain the upstream Planner's gamma gate with $\gamma = 0.95$. Because source diagnosis can improve raw F1 by consuming more questions, we also report question counts and cost-adjusted F1, $\overline{\mathrm{F1}} - \kappa \bar{n}_q$, with $\kappa = 0.05$.

Policy taxonomy. PlannerCaptain is the upstream source-blind baseline. It uses a fixed $0.9/0.1$ answer likelihood and does not infer the source. **SymmetricMixture**, our flagship, spans both channel orientations and combines revealed-cell probes with the cost-aware ask gate. **NoisyMixture** is an asymmetric ablation that represents trust-leaning noise and a deterministic liar, but not intermediate invert-leaning rates. **OracleJoint** receives the true source component. Exact mixture grids and older gate variants appear in Appendix C. Other defenses are introduced where they are evaluated.

Live-channel implementation. Live runs use a conservative translation-noise rate of $\rho = 0.02$, based on estimates near 0.01. A failed question execution on a candidate board receives an uninformative likelihood rather than eliminating that board. Source-diagnostic probes ask only about already revealed cells, so they provide no new board information. Further details appear in Appendix B.

5 Controlled Study 1: Battleship – Attacks and Defenses

Unless noted otherwise, experiments use the upstream 18-board setting, the budgets of Section 2.1, and deepseek/deepseek-v4-flash through OpenRouter with the provider pinned to DeepSeek. We run each Captain at three seeds. LLM completions are unseeded, so each seed produces fresh completions rather than a deterministic rerun. We evaluate the policies in Section 4 against six source behaviors:

honest (δ=0) — answers truthfully;
noisy$_\epsilon$ (δ=ϵ) — independently flips each answer with probability ϵ (0.25 unless noted);
lying (δ=1) — always inverts the truth;
sleeper — answers honestly five times, then always lies;
probe-dodger — answers honestly only when the answer is publicly checkable, and lies otherwise;
best-response — answers honestly when publicly checkable and returns a fair coin otherwise.
The first three are stationary points on the δ axis. The adaptive sources vary their behavior with the interaction history or question type.

5.1 Static Corners: Honest, Noisy, and Lying

Source awareness recovers performance under systematic inversion. Table 1 first compares the three stationary conditions. Under honest and noisy sources, the policies remain close. Under a deterministic liar, the source-blind Planner falls to 0.510 F1 and a 65% win rate, while every source-aware variant remains at 0.735 F1 or above and wins every game. NoisyMixture reaches 0.781, and OracleJoint reaches 0.776. Their 0.22–0.27 F1 advantages over the Planner exceed 12 reported standard errors. The source-aware captains also select *lying* as the MAP source in every game. The adaptive columns are discussed in Section 5.2.

A soft likelihood is not a substitute for source inference. To isolate the value of source inference, we retain the flagship's soft answer update and VOI gate but remove its source hypotheses. This fixed-channel ablation assumes a single trust-leaning error rate, $\hat{\epsilon} = 0.25$. Against the deterministic liar, it scores 0.545 ± 0.019 F1 ($n = 54$), comparable to the Planner and below the measured 0.625 ask-nothing baseline of Section 5.3, despite asking 7.6 questions per game. This behavior agrees with Proposition 3: a trust-leaning likelihood interprets systematically inverted answers in the wrong direction. Adding a three-hypothesis source prior raises the same captain's score to 0.666 ($n = 18$), even without probes or a mixture grid.

The flagship learns the channel orientation. From a prior spanning both orientations, SymmetricMixture scores 0.783 F1 under lying, not significantly different from its 0.734 honest result. This is consistent with the equal channel capacity of correctly oriented honest and inverting sources, although the theory does not guarantee equal finite-task F1. Under an honest source, it also remains close to the Planner (0.734 versus 0.770, n.s.).

Recovery can require more questions. Across the source-aware policies, average question use ranges from roughly 6 to 15, so the raw-F1 leaders are not always the cost-adjusted leaders. Table 3 reports question counts and adjusted scores. In the budget sweep of Appendix E, the source-aware advantage under lying grows from $+0.08$ at $Q_{\max} = 2$ to $+0.25$ at $Q_{\max} = 15$, showing that recovery improves when the agent has enough opportunities to diagnose and invert the source.

Backbone generalization. The static-corner pattern persists with grok-4.1-fast, qwen3-235b, and a partial gemini-3.1-pro run. Under lying, the Planner remains near 0.5 F1 while the source-aware policies recover to 0.64–0.79 (Appendix D).

Captain	honest	noisy	lying	sleeper	probe-dodger
Planner	0.770±0.015	0.663±0.018	0.510±0.018	0.547±0.018	0.563±0.017
Warned Planner†	0.747±0.031	—	0.532±0.026	0.583±0.034	0.538±0.033
Heuristic consistency†	0.691±0.034	—	0.683±0.025	0.601±0.039	0.561±0.038
NoisyMixture QMD	0.779±0.017	0.654±0.019	0.781±0.014	0.562±0.018	0.570±0.019
SymmetricMixture QMD	0.734±0.019	**0.682**‡	**0.783**±0.017	**0.605**±0.017	**0.591**±0.016
OracleJoint	**0.802**±0.013	0.630±0.018	0.776±0.014	—	—

Table 1: Battleship final-board F1 (mean ± SEM) over 18 boards with $Q_{\max} = 15$. Bold and underline mark the best and second-best result in each column. Unmarked results pool seeds 42–44 ($n = 54$). Rows marked † and cells marked ‡ use seed 42 ($n = 18$). "—" denotes an unrun condition. Table 3 reports question counts and cost-adjusted F1.

5.2 The Defense Ladder: Adaptive Sources and Public Defenses

Simple public defenses fail under adaptation. The stationary model assumes that source behavior does not change with the interaction. We therefore test the sleeper and probe-dodger against two simple countermeasures. The *warned* Planner is told in its prompt that the source may deceive, but receives no source-inference machinery. *Heuristic consistency* asks probes whose answers are already determined by revealed cells and begins inverting answers once their contradiction rate exceeds 0.6. All runs use $Q_{\max} = 15$ and the same 18-board set. The Warned Planner and Heuristic consistency baselines use seed 42 ($n = 18$). The Planner, NoisyMixture, and SymmetricMixture results against the sleeper and probe-dodger pool seeds 42–44 ($n = 54$). Figure 2 summarizes each defense's gain against its target source and the gain that remains after a tested adaptive response. Paired statistics appear in Table 8.

Relative to its honest performance, the source-blind Planner loses 0.22–0.29 F1 under each adversarial source. Heuristic consistency performs well against the static liar, improving by 0.191 over the Planner and identifying the source in every game, but it incurs a 0.101 F1 cost under an honest source. It also fails under adaptation. Its advantage falls to a non-significant 0.061 against the sleeper and 0.012 against the probe-dodger. Warning-only prompting is likewise ineffective ($|\Delta\text{F1}| \leq 0.04$, all $|t| < 1.5$).

Adaptive sources create question-dependent channels. Pooled across three seeds, SymmetricMixture retains a 0.058 advantage over the Planner against the sleeper and a 0.029 advantage against the probe-dodger, much smaller than its 0.258 gain against the static liar. The sleeper advantage remains significant after Holm correction over the four adaptive-arm comparisons. The probe-dodger advantage is positive in every seed but does not remain significant after the same correction.

Replay analysis explains this compression. On the questions actually asked, the adaptive sources produce aggregate deception rates of 0.51–0.69, closer to the half-point than the static liar. These aggregates do not define a stationary channel because behavior varies with time for the sleeper and with public checkability for the probe-dodger. SymmetricMixture can at least represent intermediate invert-leaning rates, while NoisyMixture supports only trust-leaning rates and a separate deterministic-liar component. NoisyMixture therefore retains no significant advantage against either adaptive source, and neither mixture captures the underlying adaptive structure. Absolute scores remain at or slightly below the measured 0.625 ask-nothing baseline.

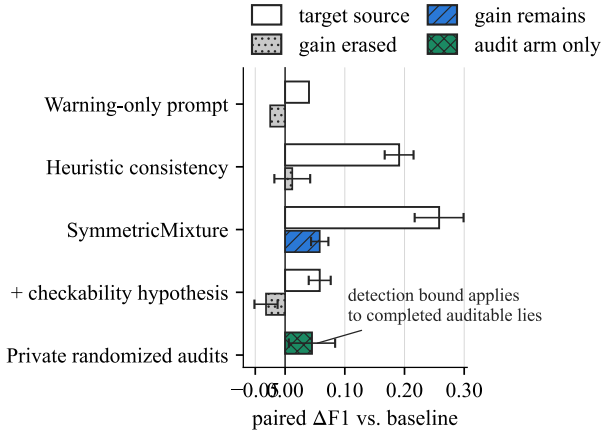

Figure 2: Defense ladder measured as paired ΔF1 from each arm's matched baseline. White bars show gains against the target source. Filled bars show the gain remaining after a tested adaptive response. The audit row contains only the measured audit arm. Whiskers are ±1 SEM where available. Statistics appear in Table 8.

Public checkability helps, but invites a best response. The probe-dodger lies on 97–100% of answers that public information cannot check, while its lie rate on checkable answers is about 2%. This mirrors gold-question attacks on crowdsourcing quality control (Checco, Bates, and Demartini 2020). In both settings, the verification rule is public, allowing the adversary to identify answers likely to escape the public checks. Because checkability is also visible to the Captain, the source model can condition on it. Adding a pair of checkability-conditioned hypotheses to the flagship improves F1 against the probe-dodger by 0.058. The gain is positive in all three seeds and remains significant after Holm correction ($t(17) = 3.15$, corrected $p = .017$), with no measurable cost under an honest source (-0.004, n.s.). This shows that an omitted source behavior explains part of

the earlier loss.

The same signal is also visible to the source. A best-response source answers honestly when checkable and returns a fair coin otherwise, making the uncheckable stream information-free. Against this source, the informed Captain ties both its uninformed counterpart and the Planner. Public conditioning can therefore be countered, motivating private verification.

Private audits can detect lies, but do not guarantee task recovery. Public probes can be evaded because the source knows which answers are checkable. A private audit hides that decision. After each answer, the Captain draws a private audit coin with nominal rate $r = 0.5$. If selected, the Captain searches for a *decisive* cell whose revealed content would settle the claim and schedules that cell as its next shot. The shot appears to the source as an ordinary task action. Claims that no single cell can resolve are skipped, and only about 35% of the questions asked in these runs are auditable by one shot. This is a form of randomized inspection (Dresher 1962; Avenhaus, von Stengel, and Zamir 2002).

Proposition 4 (detecting auditable lies, informal). *Suppose that, conditional on each auditable lie and the preceding history, a completed private audit resolves the claim with probability at least $p > 0$. After k such lies, the probability that none is detected is at most $(1-p)^k$.* The proof appears in Appendix A. Here, p is the effective probability of scheduling, completing, and resolving an audit, not the nominal audit rate r. The proposition bounds detection only. It does not cover claims that no available task action can settle, and detecting a lie need not correct earlier belief updates or improve final-board F1.

The audit policy also includes two changepoint source hypotheses. Each models a source whose error rate changes from 0.05 to 0.90 after answer $\tau \in \{3, 6\}$. We compare the full policy with an ablation that retains these hypotheses but sets the audit rate to zero (Table 9). Against the sleeper, changepoint hypotheses alone improve F1 by 0.084 and identify the switch in every game because ordinary shots already provide delayed ground truth. Audits provide no further improvement in this condition. Against the probe-dodger, changepoint hypotheses alone do not help, while private audits add 0.045 F1. In a pure-liar diagnostic, every completed audit of a lie detects it, with no false detections across 36 honest games. The probe-dodger F1 improvement is positive at $n = 18$ but is not statistically significant, so evidence for task recovery remains preliminary.

5.3 The Full Deception Axis: Both Slopes of the Wall

Figure 3 traces final-board F1 across the entire deception axis $\delta \in [0, 1]$. Sweeping the benign side ($\epsilon \in \{0.05, 0.10, 0.25, 0.40\}$, seed 42, 18 boards per condition), the flagship degrades smoothly from 0.753 to 0.659, tracking the thinning channel rather than paying a symmetric-prior penalty: hardening against a liar does not tax benign operation. The adversarial side fixes the source at $\delta \in \{0.55, 0.6, 0.7, 0.8, 0.9\}$ and pools seeds 42–43 over the same 18-board set, giving 36 runs per Captain and rate with

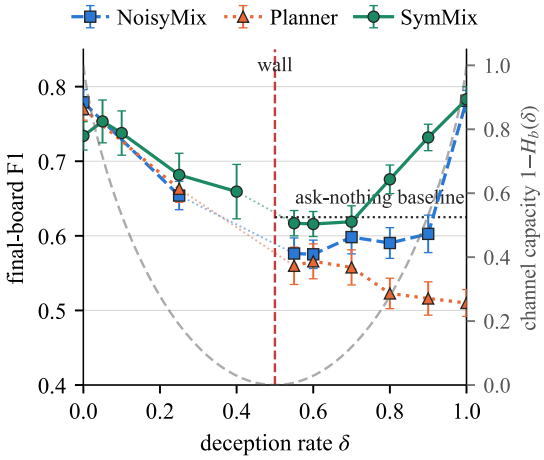

Figure 3: Final-board F1 across the stationary deception rate δ ($Q_{\max} = 15$, 18 boards per seed). Honest and lying endpoints pool seeds 42–44 ($n = 54$); adversarial interior points pool seeds 42–43 ($n = 36$); benign interior points use seed 42 ($n = 18$) and were measured only for SymmetricMixture. The grey curve shows binary-channel capacity for reference. Whiskers are ± 1 SEM over runs. The dotted bridge marks an unmeasured region, the vertical line marks $\delta = \frac{1}{2}$, and the horizontal line is the measured ask-nothing baseline.

$Q_{\max} = 15$. Mixture captains use the flagship configuration.

The three curves fan out from the wall (Figure 3). The Planner spends essentially the full budget and declines overall from 0.560 F1 at δ=0.55 to 0.516 at δ=0.9. NoisyMixture QMD remains in a narrow, nonmonotonic 0.575–0.603 band rather than tracking the strengthening inverting channel. Its trust-side prior survives by treating the channel as maximally noisy rather than learning its sign (Section 4). SymmetricMixture QMD shows the opposite empirical trend, rising overall from 0.617 to 0.732 as δ runs from 0.55 to 0.9. This is qualitatively consistent with orientation becoming easier to infer farther from the half-point, although the probe-complexity result does not by itself predict an F1 curve.

At δ=0.9, SymmetricMixture reaches 0.732 F1, compared with 0.516 for the Planner. Relative to the measured ask-nothing baseline of 0.625, its advantage is 0.107 F1. This descriptive comparison is consistent with source diagnosis becoming worthwhile once the channel orientation can be identified reliably within the question budget.

6 Controlled Study 2: Transfer Verification

We test this source-aware decision principle beyond Battleship in a banking-inspired POMDP where an agent must approve, block, hold, or escalate a transfer. Approximate, domain-informed probabilities and asymmetric rewards provide a controlled stress test.

6.1 Transfer Environment and Policies

Each episode's hidden state $x = (z, s)$ comprises case type z and caller type s. Cases are legitimate, legitimate-unusual,

Caller backbone	Policy	Reward	case acc	src acc	Q
deepseek-v4-flash	Planner	55.5	0.82	0.05	1.6
	value-ranked	60.4	0.88	0.62	1.9
	entropy-ranked	52.7	**0.92**	**0.65**	3.7
grok-4.20	Planner	58.4	0.82	0.05	1.7
	entropy-ranked	57.0	**0.93**	**0.63**	3.5
qwen3-235b	Planner	50.0	0.83	0.05	1.5
	entropy-ranked	50.6	**0.87**	**0.58**	3.3

Table 2: Live transfer verification across caller backbones (`appscam_v1`, $n = 60$ per row), each supplying the caller and classifier. Planner source accuracy indicates whether its fixed assumption is correct; Q counts questions.

unauthorized-transfer, or scam-or-coercion; callers are informed or mistaken owners, scam victims, unauthorized actors, or a noisy channel. Scam victims are genuine owners under social-engineering pressure; unauthorized actors answer strategically. The agent initially receives only non-diagnostic public context.

Costly information actions include caller questions, internal risk lookups, and identity-verification challenges. Caller-answer likelihoods depend on (z, s); non-caller evidence depends on the case. Free-text answers are parsed into soft observation distributions. Rewards are asymmetric: approving an unauthorized transfer costs -2500 versus -400 for a scam or coercion case (Table 10).

Policies. The *Planner* fixes caller noise at ϵ=0.25 and updates only case belief. Both source-aware policies maintain a joint (z, s) posterior and ask only when an information action justifies its cost. The *value-ranked* policy (`action_cond.`) selects by expected decision value; the *entropy-ranked* policy (`joint_info`) uses the same value gate but ranks eligible actions by expected joint-entropy reduction.

Their distinction is clearest for authorized-push-payment scams: the genuine owner is pressured, so MFA and callback succeed although the transfer is unsafe. Remaining conversational clues include unusual urgency or failure to recognize the recipient. Reassured by successful identity checks, the Planner approves; the joint posterior can infer a pressured scam victim and hold the transfer.

Offline controlled results. Across 1024 episodes per case stratum, value ranking raises mean reward from 44.71 (Planner) to 52.73, with non-overlapping 95% intervals and a better lower tail (Table 12).

Live language channel. We use one LLM to roleplay the caller and another to classify free-text answers. Across three caller backbones, entropy ranking improves case accuracy and identifies caller type with 0.58–0.65 accuracy (Table 2). Rewards are not significantly different at n=60: no policy approves a fraudulent transfer in the random review queue. The runs expose question selection and source-model fit.

Question selection should follow decision value. In DeepSeek live runs, replacing entropy with value ranking cuts question use from 3.5 to 2.0. On an unauthorized-transfer slice (n=150 per policy), it also cuts false approvals from 8.7% to 3.3% and raises reward by 147 ± 64 per episode, despite reducing caller-type accuracy from about 0.65 to 0.55. Source accuracy and decision quality can diverge.

Mimicry exposes source-model mismatch. These results assume source likelihoods cover test-time behavior. An unauthorized actor can violate this by imitating an honest caller. Mimicking an honest account of the disputed transfer backfires by shifting the policy toward evidence the impostor cannot fake. Mimicking a legitimate owner lowers callback value and raises offline false approvals 1.8–3.3×. Live impostors amplify this pattern; recalibration to measured live behavior removes the inversion (Appendix H).

7 Discussion, Limitations, and Conclusion

Across both studies, source inference matters only when it changes a decision. In Battleship, channel orientation makes reliable inversion useful; in transfer verification, caller type determines answer interpretation. Source accuracy is not the objective: entropy ranking classifies callers better but asks more and makes costlier errors. The transferable principle is decision-aware source inference, not a specific mixture.

Results also separate stationary uncertainty from strategic adaptation. Diagnosis is expensive near the half-point, and sources can selectively answer recognizable public checks. Delayed task feedback or private audits can expose what public probes miss only when the task can verify the claim. Proposition 4 guarantees detection only for lies resolvable by a completed hidden audit.

Limitations and future work. Several Battleship sweeps and the audit study use one seed; evidence for task recovery is preliminary. The theory excludes strategic sources. The transfer environment has hand-specified likelihoods and is not a deployment study; live results require language-channel calibration. Future work should examine richer adaptive models, multiple sources, and verification robust to model mismatch.

Conclusion. A reliably wrong source becomes useful once its orientation is known. Trust, invert, discount, or verify its answers according to their decision value.

References

Angluin, D.; and Laird, P. D. 1988. Learning from Noisy Examples. *Machine Learning*, 2(4): 343–370.

Avenhaus, R.; von Stengel, B.; and Zamir, S. 2002. Inspection Games. In Aumann, R. J.; and Hart, S., eds., *Handbook of Game Theory with Economic Applications*, volume 3, chapter 51, 1947–1987. Amsterdam: Elsevier.

Checco, A.; Bates, J.; and Demartini, G. 2020. Adversarial Attacks on Crowdsourcing Quality Control. *Journal of Artificial Intelligence Research*, 67: 375–408.

Dawid, A. P.; and Skene, A. M. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1): 20–28.

Debenedetti, E.; Zhang, J.; Balunović, M.; Beurer-Kellner, L.; Fischer, M.; and Tramèr, F. 2024. AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. In *Advances in Neural Information Processing Systems*, volume 37, 82895–82920. Datasets and Benchmarks Track.

Dresher, M. 1962. A Sampling Inspection Problem in Arms Control Agreements: A Game-Theoretic Analysis. Technical Report RM-2972-ARPA, RAND Corporation, Santa Monica, CA.

Grand, G.; Pepe, V.; Tenenbaum, J. B.; and Andreas, J. 2026. Shoot First, Ask Questions Later? Building Rational Agents that Explore and Act Like People. In *The Fourteenth International Conference on Learning Representations*.

Greshake, K.; Abdelnabi, S.; Mishra, S.; Endres, C.; Holz, T.; and Fritz, M. 2023. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, 79–90. ACM.

Kaelbling, L. P.; Littman, M. L.; and Cassandra, A. R. 1998. Planning and Acting in Partially Observable Stochastic Domains. *Artificial Intelligence*, 101(1–2): 99–134.

Massart, P.; and Nédélec, É. 2006. Risk Bounds for Statistical Learning. *The Annals of Statistics*, 34(5): 2326–2366.

Pelc, A. 2002. Searching Games with Errors—Fifty Years of Coping with Liars. *Theoretical Computer Science*, 270(1–2): 71–109.

Rényi, A. 1961. On a Problem of Information Theory. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 6(B): 505–516.

Ross, S.; Chaib-draa, B.; and Pineau, J. 2007. Bayes-Adaptive POMDPs. In *Advances in Neural Information Processing Systems 20*, 1225–1232.