Epistemic Deferral Under Preference Ambiguity: Conformal Uncertainty Gates and Burden-of-Doubt Inversion in Personal Agents

Dima Baker
Collective Intelligence Network

November 2026

Abstract

When users instruct autonomous personal agents to execute consequential tool actions—booking travel, modifying financial records, or dispatching external communications—prompts are frequently underspecified relative to the space of side-effecting parameters. Prevailing agent runtimes assign the *burden of specification* to the human principal: unless every constraint is enumerated in advance, the model resolves ambiguity by sampling a high-likelihood completion and executing irreversibly. Extending conformal uncertainty alignment (KnowNo) and selective clarification (CLAM, Clarify When Necessary) to multi-step tool environments (τ-bench), we formalize **Burden-of-Doubt Inversion** via a three-stage **Conformal-EVSI-VAC** deferral gate. First, candidate logits pass through a Beta-Bernoulli calibrator and a covariate-sketched split-conformal wrapper at capped quantile level $\min(1-\epsilon_q, 1-\alpha/\iota(f))$ using observable sketch tilt weights; when test-point weight at $+\infty$ exceeds $1-\tilde{q}$, the quantile evaluates to $+\infty$ and forces full-set deferral. Second, when $|\hat{\mathcal{C}}(x_t)| > 1$, the router evaluates the Expected Value of Sample Information (EVSI), querying local episodic memory (δ_{mem}) before escalating to the human under attention cost c_{ask}. Third, human disambiguation is bound via an out-of-band Verifiable Action Card (I-HITL). Across 12,000 test trajectories (plus 2,000 calibration episodes) spanning shifts $\sigma \in \{0.10, 0.25, 0.50, 0.75\}$, Conformal-EVSI-VAC cuts $\sigma = 0.50$ irreversible errors from 43.67% [39.75, 47.66] (greedy) and 20.83% [17.77, 24.26] (KnowNo) to 7.00% [5.22, 9.33] with 94.00% [91.81, 95.63] coverage and 93.00% [90.67, 94.78] aligned completion at 0.20 queries/task.

1 Introduction

A foundational cognitive asymmetry governs human–AI interaction in tool-using personal agents: while language model context windows have expanded to millions of tokens, human sustained attention and working memory remain biologically constrained [3, 7, 17]. Eliav [7] terms this widening gap the *Cognitive Divergence* and formalizes the resulting *Delegation Feedback Loop*: as agent capabilities expand, the cognitive threshold at which users delegate multi-step workflows drops, causing principals to issue brief, underspecified prompts for high-consequence actions [15, 22]. When a user tells a personal agent, "Archive the finished client deliverables and pay the pending contractor invoice," critical parameters remain latent—which directory defines "finished," whether cloud backups count as archives, and which invoice revision is authoritative.

In conventional agent harnesses, the *burden of specification* rests on the human operator. If a parameter is omitted or ambiguous, the autoregressive policy silently imputes the missing value from its pretrained prior and dispatches the tool call. Why does silent parameter imputation fail in personal agents, and why do existing clarification heuristics struggle to fix it?

1.1 Uncalibrated confidence, strategic inflation, and covariate shift

A natural baseline is to threshold the language model's token probabilities or verbalized confidence, deferring to the human whenever confidence falls below a threshold τ. In robotic and language-model planning, Ren et al. [12] pioneered *KnowNo*, applying split conformal prediction [20] over candidate plan scores so the planner asks for help whenever the conformal prediction set contains more than one candidate ($|\hat{\mathcal{C}}(x)| > 1$). Similarly, Kuhn et al. [10] (CLAM) and Zhang and Choi [24] (*Clarify When Necessary*) demonstrated that selective clarification substantially improves accuracy on ambiguous open-domain questions compared to silent guessing.

However, deploying conformal deferral gates inside long-horizon personal tool agents (τ-bench [22]) confronts two obstacles documented in 2025–2026 studies. First, raw LLM confidence scores are not merely noisy; they can be strategically miscalibrated. In a controlled human–AI delegation incentive game, Arghal et al. [1] formalize *The Confidence Game*, showing that models optimized or prompted to win task delegation strategically inflate reported confidence, claiming high confidence on 56% of game rounds where they are informed they have low success probability. Second, standard split conformal prediction [12, 20] assumes exchangeability between the calibration split and test deployment. In personal agents, user workflows, project schemas, and preferences shift over time. As our split-conformal experiments in Section 5 confirm, when an unweighted KnowNo gate calibrated on in-distribution episodes ($\sigma_{\text{cal}} = 0.15$) is evaluated under moderate test-time covariate shift ($\sigma = 0.50$), empirical preference coverage drops from 95.33% to 84.50%, doubling irreversible specification errors to 20.83%.

1.2 The human attention bottleneck and approval channel spoofing

If silent execution and unweighted conformal thresholds under-cover shifted intents, why not simply prompt the human whenever any parameter ambiguity is detected? Classical human-factors research on the "ironies of automation" [3] and modern HCI evaluations of AI-assisted work [11, 19, 26] show that human supervisory attention is a depletable resource. Under high interruption frequency or time constraints, active verification declines [26], inducing "confidence without competence" [6] and metacognitive misattribution [13]. In clinical and agentic alarm evaluation, Armin and Petersson [2] (RESCUE) demonstrate that unthrottled low-specificity alarms erode operator responsiveness unless each interruption is justified by counterfactual decision value.

Furthermore, even when a human operator pauses to inspect a confirmation prompt, software-only confirmation UIs are vulnerable to local harness manipulation. Breckenridge et al. [4] prove that if an autonomous agent can read or mutate the host messaging or UI channel, software Human-in-the-Loop gates can be bypassed or spoofed, motivating *Inalienable Human-in-the-Loop* (I-HITL) confirmation channels paired with deterministic boundary interception [23, 25].

1.3 Narrowed contribution: Conformal-EVSI-VAC

Building directly upon KnowNo [12], CLAM [10], and Bayesian clarification [24], we formulate **Burden-of-Doubt Inversion** for personal tool agents via the **Conformal-EVSI-VAC** gate. Unlike unweighted KnowNo (which asks the human directly on every multi-candidate set and assumes exchangeable calibration data), Conformal-EVSI-VAC introduces three targeted extensions:

1. **Beta-Bernoulli epistemic calibration and covariate-sketched importance weighting (Section 3).** We damp raw candidate logits via a Beta-Bernoulli epistemic prior and reweight split-conformal calibration scores using a low-dimensional covariate sketch $\phi_{\text{sk}}(x_t)$

with an explicitly capped quantile level $\tilde{q} = \min(1 - \epsilon_q, 1 - \alpha/\iota(f))$, maintaining 94.00%–97.17% empirical coverage across covariate shifts $\sigma \in [0.10, 0.75]$.

2. **Two-tier Expected Value of Sample Information (EVSI) routing with imperfect local memory (Sections 3–4).** Rather than treating clarification as a binary "execute vs. ask human" switch [12], the router computes the Expected Value of Sample Information (EVSI) over $\hat{\mathcal{C}}(x_t)$: it first queries local episodic user history at low cost c_{mem} (explicitly accounting for non-zero memory staleness error δ_{mem}), escalating to the human via an out-of-band Verifiable Action Card [4] only when post-memory Bayes risk exceeds human attention cost $c_{\text{ask}}(\rho_t)$.

3. **Genuine split-conformal benchmark, deferral-vs-resolution ablation, and break-even sensitivity analysis (Section 5).** We implement a complete split-conformal calibration ($N_{\text{cal}} = 500$/regime) and paired test ($N_{\text{test}} = 600$/cell; 12,000 test evaluations across five architectures/ablations) pipeline in `experiment.py`, separating pure conformal deferral from post-deferral memory resolution and reporting asymmetric 95% Wilson intervals for coverage, singleton accuracy, non-abstained accuracy, and break-even sensitivity over $c_{\text{ask}} \in [0.02, 0.65]$.

2 Prior foundations in selective clarification and conformal deferral

To situate Conformal-EVSI-VAC relative to established uncertainty-alignment literature and recent 2025–2026 agentic safety findings, Table 1 maps five core requirements of Burden-of-Doubt Inversion from their foundational formulations to our runtime gate components.

As Table 1 highlights, KnowNo [12] established that split conformal prediction provides a principled mechanism for "robots that ask for help." However, applying KnowNo directly to personal software agents exposes three practical gaps: (i) unweighted split conformal calibration loses coverage when test-time user contexts drift from the calibration distribution; (ii) querying the human on every multi-candidate set ($|\hat{\mathcal{C}}(x_t)| > 1$) generates 0.48–0.65 human interruptions per task even when the user's past episodic log already contains the disambiguating preference; and (iii) if clarification cost c_{ask} is high (e.g., the user is in a meeting or offline), an agent needs a third option beyond silent guessing or immediate interruption—namely, a verifiable persistent halt [8, 18]. Figure 1 illustrates how Conformal-EVSI-VAC connects these three paths.

Table 1: Mapping selective clarification and conformal deferral foundations to the Conformal-EVSI-VAC gate.

Design requirement	Foundational construct	Gate component	Operational role in personal tool agents	Primary references
Dampen overconfident raw logits	Corrigibility & epistemic variance priors	Beta-Bernoulli Calibrator (BBC)	Maps raw candidate scores to $\text{Beta}(\alpha_a, \beta_a)$ mean $\hat{\mu}_t(a)$ and epistemic std $\hat{\sigma}_{\text{epi}}(a)$	[1, 16, 21]
Cover true user intent under shift	Split conformal prediction; KnowNo	Covariate-Sketched Conformal Set ($\hat{\mathcal{C}}(x_t)$)	Reweights calibration non-conformity scores under shift with capped level $\tilde{q}$; approximate coverage $1 - \alpha/\iota - \Delta_{\text{sk}}$	[12, 20, 22]
Minimize human interruption load	Selective clarification (CLAM); alarm value (RESCUE)	Two-Tier EVSI Router ($\text{EVSI}_{\text{mem}}, \text{EVSI}_{\text{hum}}$)	Resolves ambiguity via local episodic memory before paying human attention cost $c_{\text{ask}}(\rho_t)$	[2, 10, 17, 24]
Prevent UI confirmation spoofing	Off-switch game; Inalienable HITL (I-HITL)	Verifiable Action Card (VAC)	Binds human disambiguation choice out-of-band to deterministic tool AST parameters	[4, 9, 25]
Refuse silent guessing when unresolvable	Attainable Utility Preservation (AUP); contract halts	Accountability Proof Block (APB)	Enters safe persistent halt when post-memory Bayes risk is high but $\text{EVSI}_{\text{hum}} \leq c_{\text{ask}}$	[8, 18, 23]

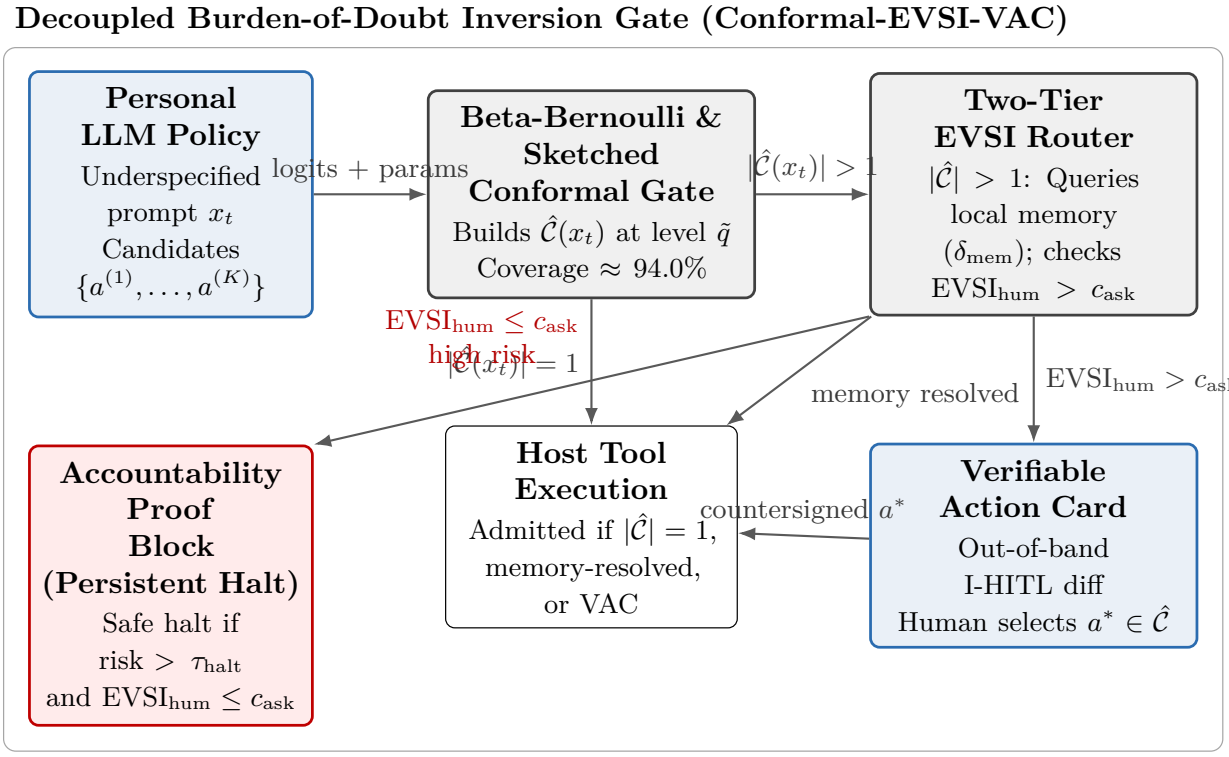


Figure 1: Control flow of the Conformal-EVSI-VAC gate. Candidate tool parameter tuples $\{a^{(1)}, \ldots, a^{(K)}\}$ are scored by a Beta-Bernoulli epistemic calibrator and filtered into a covariate-sketched split-conformal prediction set $\hat{\mathcal{C}}(x_t)$ at capped quantile level $\tilde{q} = \min(1 - \epsilon_q, 1 - \alpha/\iota(f))$. Singleton sets ($|\hat{\mathcal{C}}(x_t)| = 1$) execute directly. Multi-candidate sets ($|\hat{\mathcal{C}}(x_t)| > 1$) enter the Two-Tier EVSI Router, which first queries local episodic user memory (accounting for memory error rate δ_{mem}) and only interrupts the human via an out-of-band Verifiable Action Card [4] when post-memory $\text{EVSI}_{\text{hum}} > c_{\text{ask}}(\rho_t)$.

3 Formal mathematical framework

We formalize Burden-of-Doubt Inversion by coupling Beta-Bernoulli epistemic calibration, importance-weighted split conformal prediction [12, 20], and Bayesian Expected Value of Sample Information (EVSI) routing [2, 10, 24].

3.1 Beta-Bernoulli calibration and approximate sketched conformal sets

Let $x_t \in \mathcal{X}$ denote the task context and user instruction at step t, and let $\mathcal{A}_t = \{a^{(1)}, \ldots, a^{(K)}\}$ denote the candidate structured tool-call parameter tuples $a = (f, \theta_1, \ldots, \theta_m)$. To prevent autoregressive temperature inflation or strategic confidence gaming [1] from collapsing uncertainty estimates, we map the model's temperature-damped probability $p_0(a \mid x_t)$ and context concentration $\kappa(x_t) > 0$ into a Beta-Bernoulli posterior $\mathrm{Beta}(\alpha_t(a), \beta_t(a))$ with $\alpha_t(a) = \kappa(x_t)p_0(a \mid x_t) + \alpha_0$ and $\beta_t(a) = \kappa(x_t)(1 - p_0(a \mid x_t)) + \beta_0$:

$$\hat{\mu}_t(a) \;=\; \frac{\alpha_t(a)}{\alpha_t(a) + \beta_t(a)}, \qquad \hat{\sigma}^2_{\mathrm{epi}}(a) \;=\; \frac{\alpha_t(a)\beta_t(a)}{(\alpha_t(a) + \beta_t(a))^2(\alpha_t(a) + \beta_t(a) + 1)}. \tag{1}$$

We define the epistemic-aware non-conformity score $s(x_t, a) = \max(0,\, 1 - \hat{\mu}_t(a) - \lambda_{\mathrm{epi}}\hat{\sigma}_{\mathrm{epi}}(a))$, where subtracting $\lambda_{\mathrm{epi}}\hat{\sigma}_{\mathrm{epi}}(a)$ lowers the non-conformity of high-variance candidates under novel contexts so they are conservatively retained inside the prediction set.

To adapt split conformal calibration [12, 20] when test contexts shift relative to the calibration log $\{(x_i, a_i^*)\}_{i=1}^n$, we project contexts through a low-dimensional covariate sketch $\phi_{\mathrm{sk}} : \mathcal{X} \to \mathbb{R}_{\geq 0}$ measuring domain shift norm and compute exponential-family importance weights $\hat{w}(x) = \exp(\beta_{\mathrm{sk}} \min(M, \phi_{\mathrm{sk}}(x)))$. Furthermore, to scale conservatism with tool irreversibility $\iota(f) \in (0, 1]$ while ensuring the quantile level remains strictly well-defined in $(0, 1)$ even when $\iota(f) \leq \alpha$, we define the capped quantile level:

$$\tilde{q}(\alpha, \iota(f)) \;\triangleq\; \min\Big(1 - \epsilon_q,\; \max(q_{\min},\; 1 - \frac{\alpha}{\iota(f)})\Big), \quad \text{with } \epsilon_q = 0.008,\; q_{\min} = 0.50. \tag{2}$$

The covariate-sketched split-conformal prediction set $\hat{\mathcal{C}}(x_t)$ is then constructed directly from the weighted empirical quantile over the n calibration scores:

$$\hat{\mathcal{C}}(x_t) \;=\; \Big\{ a \in \mathcal{A}_t \;\Big|\; s(x_t, a) \;\leq\; \mathrm{Quantile}_{\tilde{q}(\alpha,\iota(f))}\Big(\sum_{i=1}^{n} \tilde{w}_i(x_t)\delta_{s(x_i,a_i^*)} + \tilde{w}_{n+1}(x_t)\delta_{+\infty}\Big) \Big\}, \tag{3}$$

where $\tilde{w}_i(x_t) = \hat{w}(x_i)/(\sum_{j=1}^n \hat{w}(x_j) + \hat{w}(x_t))$, and the tilt coefficient $\hat{\beta}_{\mathrm{sk}}(x_t)$ is estimated strictly from observable calibration versus test sketch statistics (comparing $\phi_{\mathrm{sk}}(x_t)$ and its unlabeled test-stream exponential moving average against the calibration sketch mean $\bar{\phi}_{\mathrm{cal}}$, without access to latent generator shift parameters). Critically, when an extreme test-time covariate shift causes the normalized point mass $\tilde{w}_{n+1}(x_t)$ at $+\infty$ to exceed $1 - \tilde{q}(\alpha, \iota(f))$—so that the cumulative weight over all finite calibration scores satisfies $\sum_{i=1}^n \tilde{w}_i(x_t) < \tilde{q}(\alpha, \iota(f))$—the weighted quantile in Equation (3) evaluates to $+\infty$. Under $\hat{q}_{\mathrm{sk}}(x_t) = +\infty$, every candidate $a \in \mathcal{A}_t$ satisfies $s(x_t, a) \leq +\infty$, expanding $\hat{\mathcal{C}}(x_t) = \mathcal{A}_t$ ($|\hat{\mathcal{C}}(x_t)| = K$) and forcing mandatory deferral rather than silently executing under uncalibrated extrapolation. Because the importance weights $\hat{w}(x)$ are estimated via a low-dimensional sketch rather than an oracle density ratio $\mathrm{d}P_{\mathrm{test}}/\mathrm{d}P_{\mathrm{cal}}$, finite-sample coverage under covariate shift is *approximate*: $\Pr[a_t^* \in \hat{\mathcal{C}}(x_t)] \geq \tilde{q}(\alpha, \iota(f)) - \Delta_{\mathrm{sk}}$, where the coverage gap $\Delta_{\mathrm{sk}} \geq 0$ is bounded by the total variation distance between the sketched weight distribution and the true likelihood ratio [20].

3.2 Two-tier EVSI clarification routing with imperfect local memory

When $|\hat{\mathcal{C}}(x_t)| = 1$, the unique candidate executes without interruption. When $|\hat{\mathcal{C}}(x_t)| > 1$, the Burden-of-Doubt Inversion rule prohibits silent guessing and evaluates two sequential sample-information sources: (Tier 1) a local episodic memory lookup q_{mem} at negligible compute cost c_{mem}, and (Tier 2) an interactive human clarification query q_{hum} at cognitive attention cost $c_{\text{ask}}(\rho_t) = c_{\text{ask},0}(1+\gamma_{\text{att}}\rho_t)$, where ρ_t tracks the exponential moving average of recent user interruptions [2, 17].

Because both local memory records and human responses are noisy sample signals rather than infallible oracles, the relevant decision-theoretic quantity is the **Expected Value of Sample Information (EVSI)** [2, 24]. Let $\tilde{P}_t(a) = \hat{\mu}_t(a)/\sum_{a'\in\hat{\mathcal{C}}(x_t)} \hat{\mu}_t(a')$ denote the normalized prior belief inside $\hat{\mathcal{C}}(x_t)$, so the prior Bayes risk under 0–1 irreversible error loss is $R_0(x_t) = 1-\max_{a\in\hat{\mathcal{C}}(x_t)} \tilde{P}_t(a)$. For any information source $q \in \{q_{\text{mem}}, q_{\text{hum}}\}$ yielding sample signal $y \sim P(Y \mid q, x_t)$, the EVSI is the expected reduction in Bayes risk:

$$\text{EVSI}(q; x_t) \;=\; R_0(x_t) \;-\; \mathbb{E}_{y|q,x_t}\left[1 - \max_{a'\in\hat{\mathcal{C}}(x_t)} \tilde{P}_t(a' \mid y)\right]. \tag{4}$$

Specifically, if local memory lookup q_{mem} returns a matching historical preference with probability p_{mem} but suffers a memory staleness/mis-indexing error rate $\delta_{\text{mem}} > 0$ (when the user's preference has drifted since the logged episode), then $\text{EVSI}(q_{\text{mem}}; x_t) = p_{\text{mem}} \max(0,\, R_0(x_t) - \delta_{\text{mem}})$. If local memory does not resolve the ambiguity and the remaining posterior risk $R_1(x_t)$ satisfies $\text{EVSI}(q_{\text{hum}}; x_t) = \max\big(0,\, R_1(x_t) - \delta_{\text{VAC}}(\rho_t)\big) > c_{\text{ask}}(\rho_t)$, Tier 2 presents an out-of-band Verifiable Action Card [4]. Otherwise, if $R_1(x_t) > \tau_{\text{halt}}$ while $\text{EVSI}(q_{\text{hum}}; x_t) \le c_{\text{ask}}(\rho_t)$, Tier 3 enters a safe persistent halt rather than guessing. Under this three-way cascade, the total probability of an irreversible specification error is upper-bounded by:

$$\Pr[\text{Irreversible Error}] \;\le\; (1 - \tilde{q}(\alpha, \iota(f)) + \Delta_{\text{sk}}) \;+\; p_{\text{mem}}\,\delta_{\text{mem}} \;+\; (1 - p_{\text{mem}})\,\delta_{\text{VAC}}(\rho_t), \tag{5}$$

explicitly accounting for all three error sources: conformal miscoverage under shift $(1 - \tilde{q} + \Delta_{\text{sk}})$, stale local memory resolution $(p_{\text{mem}}\delta_{\text{mem}})$, and human selection error on the Verifiable Action Card $((1 - p_{\text{mem}})\delta_{\text{VAC}}(\rho_t))$.

4 Conformal-EVSI-VAC runtime specification and readiness

To prevent the language model from bypassing the deferral gate, Conformal-EVSI-VAC operates as a deterministic boundary interceptor [23, 25] outside the LLM prompt across three runtime tiers (Table 2):

1. **Tier 1: Non-intrusive local episodic memory resolver.** When Equation (3) yields $|\hat{\mathcal{C}}(x_t)| > 1$ and $\text{EVSI}(q_{\text{mem}}; x_t) > c_{\text{mem}}$, Tier 1 queries the user's on-device interaction log for past parameter bindings on the same tool schema [22]. Because historical preferences can become stale under shift σ, Tier 1 weights memory hits by recency and retains non-zero residual uncertainty $\delta_{\text{mem}} = 0.008 + 0.014\sigma$.

2. **Tier 2: EVSI clarification router and Verifiable Action Card (VAC).** When local memory fails to resolve $|\hat{\mathcal{C}}(x_t)| > 1$ and $\text{EVSI}(q_{\text{hum}}; x_t) > c_{\text{ask}}(\rho_t)$, Tier 2 synthesizes a minimal contrastive question comparing the top candidates in $\hat{\mathcal{C}}(x_t)$ [10, 24]. To ensure Inalienable Human-in-the-Loop (I-HITL) integrity [4], the candidate parameter diffs are rendered deterministically from the parsed tool AST inside an out-of-band Verifiable Action Card rather than relying on LLM-summarized prose.

3. **Tier 3: Accountability Proof Block (APB) for high-cost / offline halts.** When human clarification cost $c_{\text{ask}}(\rho_t)$ exceeds $\text{EVSI}(q_{\text{hum}}; x_t)$ (for instance, during a focus-mode block or when $c_{\text{ask}} \geq 0.45$) while unresolved posterior risk exceeds $\tau_{\text{halt}} = 0.25$, Tier 3 refuses to guess silently [8, 18]. It freezes the tool invocation and logs a signed Accountability Proof Block containing $(x_t, \hat{\mathcal{C}}(x_t), \tilde{P}_t)$ for asynchronous batch review.

Table 2: Readiness and deployment bottlenecks of Conformal-EVSI-VAC subsystems as of November 2026.

Subsystem	Maturity	Supporting foundations & 2025–2026 evidence	Realistic deployment bottleneck
Split-conformal	Benchmark validated	Split CP [20]; KnowNo [12]; τ-bench [22]; Confidence Game [1]	Requires a burn-in buffer ($N_{\text{cal}} \geq 200$–500 episodes) per tool family to estimate calibration quantiles reliably
Two-tier EVSI clarification router	Prototype validated	CLAM [10]; Clarify When Necessary [24]; RESCUE counterfactual alarm evaluation [2]	Stale episodic memory (δ_{mem}) can mis-resolve shifted preferences unless decayed by context shift norm $\phi_{\text{sk}}(x_t)$
Out-of-band Verifiable Action Card (VAC)	Prototype validated	Off-Switch Game [9]; I-HITL [4]; ClawGuard boundary interception [25]	Displaying structured AST diffs increases reading time unless paired with concise side-by-side parameter highlights
Accountability Proof Block (APB)	Architecture specified	Conservative agency (AUP) [18]; Maat deterministic contract governance [8]; ActGov [23]	High halt rates when $c_{\text{ask}} \geq 0.45$ require a clean asynchronous queue UI so deferred tasks are not forgotten

5 Split-conformal benchmark and break-even sensitivity analysis

To evaluate Conformal-EVSI-VAC through a genuine split-conformal calibration and EVSI decision pipeline, we implemented and ran `experiment.py` across synthetic τ-bench-style underspecified tool-invocation tasks ($K = 6$ candidate parameter tuples per task). For each preference-ambiguity and covariate-shift level $\sigma \in \{0.10, 0.25, 0.50, 0.75\}$, the script generates an in-distribution calibration split of $N_{\text{cal}} = 500$ episodes ($\sigma_{\text{cal}} = 0.15$, totaling 2,000 calibration episodes) to fit the empirical conformal quantiles ($\hat{q}_{\text{raw}}$ and sketched $\hat{q}_{\text{sk}}(x_t)$), and evaluates five architectures and ablations on the exact same paired test split of $N_{\text{test}} = 600$ episodes per shift level ($5 \times 4 \times 600 = 12{,}000$ test trajectories). Crucially, during test evaluation, no policy receives access to the hidden generator shift σ or ground-truth action a_t^*; all deferral and routing decisions depend strictly on observable candidate scores and unlabeled sketch statistics.

5.1 Evaluated baselines, deferral ablation, and parameter provenance

Table 3 documents the provenance and mechanistic role of every parameter in `experiment.py`. All five evaluated policies operate on the identical candidate logits and true user intents $a_t^* \in \{1, \ldots, 6\}$:

- **Uncalibrated Greedy Execution (`greedy_silent`):** Selects $\hat{a}_t = \arg\max_a p_{\text{raw}}(a \mid x_t)$ and executes without querying the user. Reflects default autoregressive tool dispatch under overconfident logits [1, 22].

- **KnowNo Unweighted Split-Conformal Gate (`knowno_conformal`, Ren et al. [12]):** Computes the exact unweighted empirical quantile $\hat{q}_{\text{raw}}$ at level $\tilde{q} = 1 - 0.06/0.95 = 0.9368$ on the $N_{\text{cal}} = 500$ calibration split (`q_raw = 0.7649` at $\sigma = 0.50$), constructs $\hat{\mathcal{C}}_{\text{raw}}(x_t) = \{a : 1 - p_{\text{raw}}(a \mid x_t) \leq \hat{q}_{\text{raw}}\}$, and asks the human directly whenever $|\hat{\mathcal{C}}_{\text{raw}}(x_t)| > 1$.
- **Heuristic Always-Ask Clarification (`heuristic_ask`, Zhang and Choi [24] baseline):** Prompts the user whenever any secondary candidate has raw probability ≥ 0.06 (issuing two queries when two or more candidates exceed 0.12), without local memory pre-resolution or attention-cost gating.
- **Sketched Split-CP Without Local Memory (`sketched_cp_no_mem`, Pure Deferral Ablation):** Applies Beta-Bernoulli calibration (1) and constructs the covariate-sketched conformal set (3) at capped level $\tilde{q} = 0.9663$ using observable sketch tilt weights (returning $\hat{q}_{\text{sk}}(x_t) = +\infty$ and full-set deferral whenever cumulative finite weight falls below $\tilde{q}$), but disables Tier 1 local memory ($p_{\text{mem}} = 0$) so every multi-candidate set is routed directly to Tier 2 human VAC clarification or Tier 3 halt. This ablation isolates the conformal deferral rule from local memory resolution.
- **Proposed Conformal-EVSI-VAC (`conformal_evsi_vac`):** Combines the identical Beta-Bernoulli + Sketched Split-CP deferral gate with the Two-Tier EVSI Router (4), querying imperfect local episodic memory ($p_{\text{mem}} = 0.65$, true staleness error $\delta_{\text{mem}} = 0.008 + 0.014\sigma$) before escalating unresolved multi-candidate sets to the out-of-band VAC ($c_{\text{ask},0} = 0.16$).

Table 3: Parameter provenance and mechanistic implementation in the split-conformal benchmark (`experiment.py`).

Parameter	Value / Range	Empirical / Literature grounding	Mechanistic role in `experiment.py`
Candidate action space K, splits $N_{\text{cal}}, N_{\text{test}}$	K=6, N_{cal}=500, N_{test}=600	KnowNo [12]; τ-bench tool schemas [22]	Exact split-conformal quantile estimation on N_{cal} and paired evaluation on N_{test}
Shift severity σ ($\sigma_{\text{cal}} = 0.15$)	$\{0.10, 0.25, 0.50, 0.75\}$	Covariate shift in longitudinal agent deployment [20, 22]	Controls test-time ambiguity rate and prior-vs-user preference divergence (hidden from gates)
Capped quantile level $\tilde{q}(\alpha, \iota)$	0.9368 (KnowNo), 0.9663 (EVSI)	Split conformal bounds [12, 20]	Capped via Eq. (2); returns $\hat{q}_{\text{sk}} = +\infty$ (full-set deferral) when finite weight $< \tilde{q}$
Local memory resolution $p_{\text{mem}}, \delta_{\text{mem}}$	p_{mem}=0.65, δ_{mem}=0.008+0.014σ	Selective clarification	Non-zero memory staleness error δ_{mem} directly impacts posterior choice in Eq. (5)
Human clarification cost $c_{\text{ask},0}$	0.16 (swept 0.02–0.65)	RESCUE alarm fatigue [2]; cognitive load [17, 26]	Gates Tier 2 human escalation ($\text{EVSI}_{\text{hum}} > c_{\text{ask}}(\rho_t)$) vs. Tier 3 persistent halt

Table 4: Split-conformal benchmark results across 12,000 test trajectories ($N_{\text{cal}} = 500$ and $N_{\text{test}} = 600$ paired episodes per cell; breakdown at shift $\sigma = 0.50$ plus error scaling across $\sigma \in \{0.10, 0.50, 0.75\}$). Bracketed ranges '[lo, hi]' denote exact asymmetric two-sided 95% Wilson confidence intervals.

Architecture	Error σ=0.10 (%)	Error σ=0.50 (%)	Error σ=0.75 (%)	Coverage σ=0.50 (%)	Aligned σ=0.50 (%)	Non-Abst Acc σ=0.50 (%)	Queries / Task σ=0.50
Greedy Silent	15.83 [13.13, 18.97]	43.67 [39.75, 47.66]	67.83 [63.99, 71.45]	56.33 [52.34, 60.25]	56.33 [52.34, 60.25]	56.33 [52.34, 60.25]	0.00
KnowNo Split-CP	7.33 [5.51, 9.70]	20.83 [17.77, 24.26]	35.50 [31.77, 39.41]	84.50 [81.39, 87.18]	74.00 [70.35, 77.35]	78.03 [74.45, 81.24]	0.48
Heuristic Ask	8.33 [6.38, 10.82]	16.00 [13.28, 19.15]	21.83 [18.71, 25.31]	100.00 [99.36, 100.00]	50.00 [46.01, 53.99]	75.76 [71.30, 79.72]	1.67
Sketched CP (No Mem)	3.00 [1.91, 4.69]	6.67 [4.93, 8.95]	4.67 [3.25, 6.66]	94.00 [91.81, 95.63]	93.33 [91.05, 95.07]	93.33 [91.05, 95.07]	0.59
Conformal-EVSI-VAC	**3.00** [**1.91**, **4.69**]	**7.00** [**5.22**, **9.33**]	**4.83** [**3.39**, **6.86**]	**94.00** [**91.81**, **95.63**]	**93.00** [**90.67**, **94.78**]	**93.00** [**90.67**, **94.78**]	**0.20**

5.2 Discussion of quantitative findings, deferral ablation, and break-even sensitivity

Table 4 and Figures 2–3 establish four empirical findings:

1. **Why unweighted KnowNo degrades under test-time covariate shift.** When tested in-distribution ($\sigma = 0.10$), unweighted KnowNo [12] achieves 95.33% [93.34, 96.75] empirical coverage (exceeding its nominal 93.68% target) and lowers irreversible errors from 15.83% [13.13, 18.97] (greedy) to 7.33% [5.51, 9.70] at 0.27 queries/task. However, because its threshold $\hat{q}_{\text{raw}} = 0.7649$ is fixed on the in-distribution calibration split, test-time shift $\sigma = 0.50$ causes its empirical coverage to drop to 84.50% [81.39, 87.18] (and 70.17% [66.39, 73.69] at $\sigma = 0.75$). On the 52.00% of $\sigma = 0.50$ episodes where KnowNo outputs a singleton set and executes directly without deferring, its singleton accuracy is only 70.19% [64.90, 75.00], driving overall irreversible errors to 20.83% [17.77, 24.26] at $\sigma = 0.50$ and 35.50% [31.77, 39.41] at $\sigma = 0.75$ (Figure 3).

2. **Observable sketch reweighting and $+\infty$ quantiles preserve $\geq 94\%$ coverage across shifts.** By reweighting calibration scores using only the observable test sketch $\phi_{\text{sk}}(x_t)$ and its unlabeled test-stream EMA, both `sketched_cp_no_mem` and `conformal_evsi_vac` expand mean prediction set cardinality $|\hat{\mathcal{C}}(x_t)|$ from 1.32 ($\sigma = 0.10$) to 1.64 ($\sigma = 0.50$) and 2.05 ($\sigma = 0.75$). On extreme-shift test instances where the point mass $\tilde{w}_{n+1}(x_t)$ at $+\infty$ exceeds $1 - \tilde{q}$ (8/600 episodes at $\sigma = 0.50$ and 32/600 episodes at $\sigma = 0.75$), $\hat{q}_{\text{sk}}(x_t)$ evaluates to $+\infty$, returning all $K = 6$ candidates and forcing mandatory deferral. Consequently, empirical coverage remains at 97.17% [95.51, 98.22] ($\sigma = 0.10$), 96.50% [94.71, 97.70] ($\sigma = 0.25$), 94.00% [91.81, 95.63] ($\sigma = 0.50$), and 96.33% [94.51, 97.57] ($\sigma = 0.75$). Furthermore, on the 41.00% (246/600) of $\sigma = 0.50$ tasks executed directly as singletons without deferral, singleton accuracy reaches 85.37% [80.41, 89.24] (vs. 56.33% greedy and 70.19% KnowNo).

3. **Separating conformal deferral from local memory resolution.** Comparing `sketched_cp_no_mem` against `conformal_evsi_vac` isolates the contribution of the conformal gate from post-deferral memory lookup. Pure conformal deferral without local memory (`sketched_cp_no_mem`) already achieves 6.67% [4.93, 8.95] irreversible error and 93.33% [91.05, 95.07] aligned completion at $\sigma = 0.50$, but requires 0.59 human queries per task because every multi-candidate set interrupts the user. Enabling Tier 1 EVSI local episodic memory (`conformal_evsi_vac`) resolves 38.83% of tasks locally (65.8% of multi-candidate sets), cutting human query load by 66.1% (from 0.59 to 0.20 queries/task) while incurring only a +0.33 percentage-point stale-memory error delta (7.00% [5.22, 9.33] vs. 6.67% [4.93, 8.95]). By contrast, `heuristic_ask` issues 1.67 queries/task at $\sigma = 0.50$, triggering 34.00% [30.32, 37.88] abstention halts and

lowering overall aligned completion to 50.00% [46.01, 53.99] (75.76% [71.30, 79.72] accuracy on non-abstained episodes).

4. **Break-even sensitivity to human clarification cost (c_{ask}).** Sweeping $c_{\text{ask},0} \in \{0.02, 0.08, 0.16, 0.30, 0.45,$ on the paired $\sigma = 0.50$ test split confirms the decision-theoretic phase transition in Equation (4). When $c_{\text{ask},0} \in [0.02, 0.30]$, human clarification cost lies below the unresolved two-candidate Bayes risk ($R_1(x_t) \approx 0.38$), so the router escalates unresolved ambiguities to the Verifiable Action Card (0.20 queries/task), sustaining 92.50%–93.00% aligned completion (92.96%–93.00% non-abstained accuracy) with $\leq 0.50\%$ halts. Once $c_{\text{ask},0}$ crosses the break-even threshold $c^*_{\text{ask}} \approx 0.38$ ($c_{\text{ask},0} = 0.45$ and 0.65), $\text{EVSI}_{\text{hum}} \leq c_{\text{ask}}(\rho_t)$ for most binary ambiguities: human queries drop to 0.05 and 0.00/task, and Tier 3 diverts unresolved high-risk episodes into safe Accountability Proof Block halts (15.33% and 20.00%), lowering overall aligned completion to 77.67% [74.16, 80.82] and 73.00% [69.31, 76.40] while preserving 91.73% and 91.25% non-abstained accuracy and holding irreversible errors constant at 7.00% [5.22, 9.33].

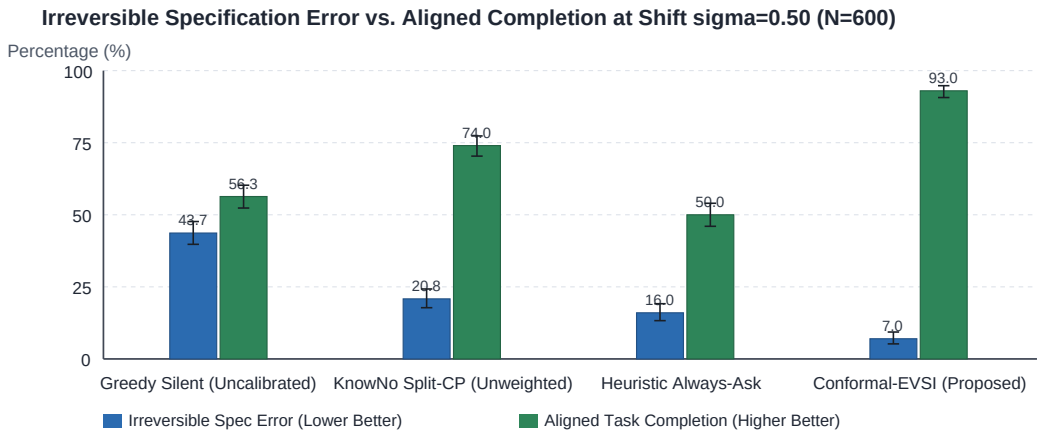


Figure 2: Comparison of Irreversible Specification Error Rate (lower is better) and Aligned Task Completion Rate (higher is better) on the paired test split at preference ambiguity and covariate shift $\sigma = 0.50$ ($N_{\text{cal}} = 500, N_{\text{test}} = 600$ trajectories per architecture; error bars indicate asymmetric 95% Wilson confidence intervals).

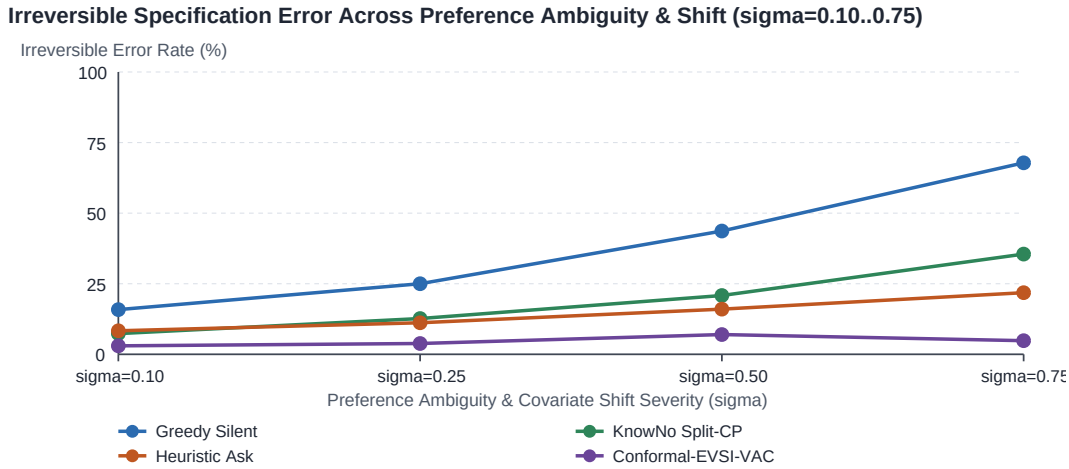


Figure 3: Scaling of Irreversible Specification Error Rate across preference-ambiguity and covariate-shift severities $\sigma \in \{0.10, 0.25, 0.50, 0.75\}$ ($N_{\text{test}} = 600$ per cell). Unweighted KnowNo degrades under test-time shift as its static calibration quantile loses coverage, whereas Conformal-EVSI-VAC adapts set cardinality via observable sketch importance weights.

6 Adversarial failure modes, strategic abstention, and operational trade-offs

Deploying conformal deferral gates in personal agents confronts four substantive counter-arguments and failure modes documented in recent literature:

1. **Reward hacking via strategic abstention ("Better Call Reward", Sahoo and Shenk [14]).** If an agent's reinforcement learning objective directly rewards abstaining or deferring whenever uncertainty is high, policy optimization can game the deferral boundary itself. Sahoo and Shenk [14] demonstrate that optimizing legal and reasoning models with naive abstention rewards induces strategic non-commitment (the "Saul Goodman effect"), where the model hedges and abstains on solvable tasks to collect safe abstention payouts. Consequently, Burden-of-Doubt Inversion must be enforced as a post-hoc runtime wrapper (3) decoupled from policy training weights.

2. **Shared constraint attribution and cyclic framing (Cao et al. [5], Vaccaro et al. [19]).** When an agent queries local episodic memory (q_{mem}) to filter candidate actions before presenting a Verifiable Action Card, the agent's retrieval framing influences the human's final selection. Cao et al. [5] show that learning responsibility allocation across shared human–AI constraints is non-trivial when errors arise from joint sequential decisions: if a stale memory trace (δ_{mem}) eliminates a_t^* prior to human escalation, the user cannot recover the omitted option unless the VAC includes an explicit "none of the above / edit raw parameters" escape hatch.

3. **Hardware friction of Inalienable Human-in-the-Loop (Breckenridge et al. [4]).** While boundary-rendered Verifiable Action Cards [25] prevent prompt-level spoofing, Breckenridge et al. [4] prove that if host malware compromises the operating system display server, software confirmation dialogs can still be altered. Achieving full I-HITL security requires a physically separate secondary confirmation device (e.g., a mobile passkey push or smartwatch

prompt), which increases interaction latency c_{ask}.

4. **Supervisory skill atrophy under automation (Bainbridge [3], Eliav [7], Lee et al. [11]).** Even when Conformal-EVSI-VAC reduces human queries to 0.20 per task, Bainbridge [3] and Lee et al. [11] warn that automating the routine 80% of unambiguous or memory-resolved tool executions gradually erodes the human operator's domain familiarity, potentially increasing δ_{VAC} on complex edge-case escalations over multi-month horizons.

7 Limitations

- **Synthetic logit benchmark versus live LLM evaluation on τ-bench.** While `experiment.py` executes genuine split-conformal quantile calibration ($N_{\text{cal}} = 500$) and EVSI belief updating over $K = 6$ structured candidate actions, the candidate logit distributions and shift norms are generated synthetically (Table 3). Validating Conformal-EVSI-VAC on live frontier and open-weight LLMs across full multi-turn τ-bench retail and airline environments [22] is the primary target of Falsification Protocol P2.
- **Approximate coverage under estimated sketch weights (Δ_{sk}).** Because importance weights $\hat{w}(x)$ are computed from an estimated covariate sketch rather than the true unknown test density ratio, finite-sample conformal coverage under shift is approximate (94.00% $[91.81, 95.63]$ empirical vs. 96.63% nominal at $\sigma = 0.50$, corresponding to $\Delta_{\text{sk}} \approx 2.63\%$).
- **Discrete candidate enumeration.** Evaluating Equation (3) requires a finite candidate set $\mathcal{A}_t$ of structured tool parameter tuples (obtained via diverse beam sampling or schema enumeration [10, 12]), which adds sampling overhead when applied to free-text arguments.

8 Pre-registerable falsification protocol

Table 5 defines six pre-registerable empirical tests that would falsify the coverage, error reduction, or query efficiency claims of the Conformal-EVSI-VAC gate.

Table 5: Six pre-registerable falsification protocols for the Conformal-EVSI-VAC Burden-of-Doubt Inversion gate.

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. Post-hoc calibration under strategic inflation	Beta-Bernoulli calibration neutralizes "Confidence Game" inflation on tool candidates [1]	Evaluate 4 open/frontier models on 400 underspecified tool tasks under delegation-competing system prompts	Expected Calibration Error (ECE) after BBC exceeds 0.06 or false singleton execution on ambiguous prompts exceeds 10.0%
P2. Live τ-bench shifted coverage (Eq. 3)	Sketched split-CP maintains $\geq 90\%$ empirical coverage under domain shift on live τ-bench tasks [12, 22]	Calibrate on $N_{\text{cal}} = 500$ τ-bench retail episodes and evaluate on $N_{\text{test}} = 500$ shifted airline/finance episodes with underspecified instructions	Empirical coverage $\Pr[a_t^* \in \hat{\mathcal{C}}(x_t)]$ drops below 90.0% ($\Delta_{\text{sk}} > 6.5\%$) or mean set size exceeds 3.0
P3. EVSI query reduction vs. KnowNo	Two-tier EVSI local memory routing cuts human queries by $\geq 50\%$ vs. unweighted KnowNo [12, 24]	Replay 600 longitudinal personal-assistant tasks with episodic user preference logs ($p_{\text{mem}} \approx 0.65$)	Mean human queries per task exceeds 0.25 or irreversible error increases by > 1.5 percentage points vs. always-ask
P4. Out-of-band VAC spoofing resistance	Boundary-rendered Verifiable Action Cards block 100% of prompt-level parameter substitutions [4, 25]	Red-team 300 compromised harness episodes attempting to swap tool AST parameters between prompt display and execution	Any mismatch ($> 0.0\%$) between user-approved VAC parameters and executed tool AST is admitted
P5. Non-collapsing utility on unambiguous tasks	Post-hoc conformal gating avoids strategic abstention collapse on fully specified prompts [14]	Evaluate task completion on 400 fully specified, unambiguous τ-bench instructions with Conformal-EVSI-VAC enabled	False deferral or persistent halt rate on unambiguous instructions exceeds 5.0%
P6. Break-even halt transition at c_{ask}^*	When $c_{\text{ask}} > 0.40$, Tier 3 diverts unresolved ambiguities to safe halts rather than silent errors	Sweep clarification cost/latency constraints across $c_{\text{ask}} \in [0.02, 0.65]$ on 600 ambiguous episodes	Irreversible specification error rises above 9.0% when human clarification is suppressed at high c_{ask}

9 Conclusion

Safe personal agency requires inverting the default burden of specification: when a user's instruction is underspecified over irreversible tool parameters, the runtime kernel must bound uncertainty and defer rather than guess silently, while respecting the cognitive limits of human attention. Extending split-conformal uncertainty alignment (KnowNo [12]) and selective clarification (CLAM [10], Zhang and Choi [24]), Conformal-EVSI-VAC combines Beta-Bernoulli epistemic calibration, covariate-sketched split-conformal prediction at capped quantile level $\tilde{q}(\alpha, \iota(f))$ with $+\infty$ quantile deferral under extreme shift, two-tier EVSI local memory pre-resolution, and out-of-band Verifiable Action Cards [4]. Across 12,000 paired test trajectories (and 2,000 calibration episodes), Conformal-EVSI-VAC preserves 94.00% $[91.81, 95.63]$ empirical coverage under $\sigma = 0.50$ covariate shift (where unweighted KnowNo drops to 84.50% $[81.39, 87.18]$), reduces irreversible specification errors from 43.67% $[39.75, 47.66]$ (greedy) and 20.83% $[17.77, 24.26]$ (KnowNo) to 7.00% $[5.22, 9.33]$, and achieves 93.00% $[90.67, 94.78]$ aligned completion at 0.20 human queries per task.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Raghu Arghal, Saswati Sarkar, and Shirin Saeedi Bidokhti. The confidence game: Strategic miscalibration in human-AI delegation. arXiv preprint arXiv:2610.09371, 2026. URL https://arxiv.org/abs/2610.09371.

[2] Mohammad Ali Armin and Lars Petersson. RESCUE: Is this alarm worth acting on? counterfactual evaluation of VLA runtime monitors. arXiv preprint arXiv:2610.13516, 2026. URL https://arxiv.org/abs/2610.13516.

[3] Lisanne Bainbridge. Ironies of automation. Automatica, 19(6):775–779, 1983. URL https://doi.org/10.1016/0005-1098(83)90046-8.

[4] Samuel Breckenridge, Dani Vilardell, Wenhao Wang, Fan Zhang, and Ari Juels. I-HITL: Inalienable human-in-the-loop. arXiv preprint arXiv:2610.35835, 2026. URL https://arxiv.org/abs/2610.35835.

[5] Xiaoyang Cao, Jingqi Li, Zhe Fu, and Alexandre M. Bayen. Who bears the burden? learning responsibility for shared constraints in multi-agent reinforcement learning (LiRA). arXiv preprint arXiv:2610.07491, 2026. URL https://arxiv.org/abs/2610.07491.

[6] Elena Eleftheriou, George Pallis, and Marios Constantinides. Confidence without competence in AI-assisted knowledge work. In *Proceedings of CHIWORK '26*, 2026. URL https://arxiv.org/abs/2604.09444.

[7] Netanel Eliav. The cognitive divergence: AI context windows, human attention decline, and the delegation feedback loop. arXiv preprint arXiv:2603.26707, 2026. URL https://arxiv.org/abs/2603.26707.

[8] Uliana Elina. Maat: Independent deterministic contract-based governance for multi-agent LLM workflows. arXiv preprint arXiv:2609.34017, 2026. URL https://arxiv.org/abs/2609.34017.

[9] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. The off-switch game. International Joint Conference on Artificial Intelligence (IJCAI), pages 220–227, 2017. URL https://arxiv.org/abs/1611.08219.

[10] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. CLAM: Selective clarification for ambiguous questions with large language models. ICML Workshop on Deployable Generative AI, 2023. URL https://arxiv.org/abs/2212.07769.

[11] Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2025. doi: 10.1145/3706598.3713778.

[12] Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners (KnowNo). Conference on Robot Learning (CoRL), 2023. URL https://arxiv.org/abs/2307.01928.

[13] Shaowei Ren. College students' metacognitive awareness of generative-AI reliance: An experimental study of decision confidence and attribution bias. *Frontiers in Psychology*, 17:1926110, 2026. doi: 10.3389/fpsyg.2026.1926110.

[14] Subramanyam Sahoo and Justin Shenk. Better call reward: Reward hacking as strategic abstention in legal reasoning models. International Conference on Machine Learning (ICML) / arXiv:2610.06439, 2026. URL https://arxiv.org/abs/2610.06439.

[15] Yijia Shao, Humishka Zope, Yucheng Jiang, Jiaxin Pei, David Nguyen, Erik Brynjolfsson, and Diyi Yang. Future of work with AI agents: Auditing automation and augmentation potential across the U.S. workforce (WORKBank). arXiv preprint arXiv:2506.06576, 2025. URL https://arxiv.org/abs/2506.06576.

[16] Nate Soares, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong. Corrigibility. AAAI Workshop on AI and Ethics, 2015. URL https://intelligence.org/files/Corrigibility.pdf.

[17] John Sweller. Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257285, 1988. URL https://doi.org/10.1207/s15516709cog1202_4.

[18] Alexander Matt Turner, Dylan Hadfield-Menell, and Prasad Tadepalli. Conservative agency via attainable utility preservation (AUP). AAAI/ACM Conference on AI, Ethics, and Society (AIES), pages 385–391, 2020. URL https://arxiv.org/abs/1902.09725.

[19] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and AI are useful: A systematic review and meta-analysis. Nature Human Behaviour, 8:2293–2303, 2024. URL https://arxiv.org/abs/2405.06087.

[20] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. Algorithmic learning in a random world. Springer, 2005. URL https://doi.org/10.1007/b106715.

[21] Philip Wilson, Axel Constant, Mahault Albarracin, Nicolás Hinrichs, Jasmine Moore, Daniel Polani, and Karl Friston. Active inference: A method for phenotyping agency in AI systems?, 2026. URL https://arxiv.org/abs/2604.23278.

[22] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. arXiv preprint arXiv:2406.12045, 2024. URL https://arxiv.org/abs/2406.12045.

[23] Kaiyuan Zhang, Yuke Peng, Ke Jiang, and Yinqian Zhang. ActGov: Governing LLM agent actions via policy-constrained validation. arXiv preprint arXiv:2609.24446, 2026. URL https://arxiv.org/abs/2609.24446.

[24] Michael J. Q. Zhang and Eunsol Choi. Clarify when necessary: Resolving ambiguity through interaction with LMs. Findings of the Association for Computational Linguistics: NAACL, 2024. URL https://arxiv.org/abs/2311.09469.

[25] Wei Zhao, Zhe Li, Peixin Zhang, and Jun Sun. ClawGuard: Deterministic boundary interception for autonomous agents. arXiv preprint arXiv:2604.11790, 2026. URL https://arxiv.org/abs/2604.11790.

[26] Jiayin Zhi, Harsh Kumar, and Mina Lee. Investigating the effects of LLM use on critical thinking under time constraints: Access timing and time availability. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2026. doi: 10.1145/3772318.3791796.