INTENT-AWARE OFF-POLICY EVALUATION FOR SHARED-REALIZER DIALOGUE CONTROLLERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Asking for clarification can rescue an ambiguous request, but each extra question costs the user a turn. We study how to compare dialogue controllers that make this choice using logged conversations, without treating a user simulator as ground truth. The controllers share an utterance realizer, and their decisions can use a belief over user intent. Our estimator, IM-DR, weights a compact state built from that belief and conversation progress, then corrects a user model's continuation value with logged rewards. We show how errors in the user model, density ratio, intent posterior and state abstraction enter the resulting bias. For the unnormalised estimator, we also establish a population double robustness identity. An oracle variance comparison holds with fixed independently fitted nuisance functions and bounded or clipped weights. On a synthetic dialogue testbed with Monte Carlo value references from 40,000 rollouts and 50 independent seeds, IM-DR has $0.23\times$ the MSE of per decision importance sampling and $0.50\times$ that of sequential DR using the same user model. Paired comparisons clustered by seed favour it over all four baselines. Its MSE advantage persists under a separate synthetic policy shift, but coverage of nominal 95% intervals falls to 0.760. Whether these results carry over to human dialogue requires future research.

1 INTRODUCTION

A dialogue assistant sometimes needs to ask before it acts. One question can reveal what the user meant; too many can make a simple task feel like an interview. Work on collaborative assistants has begun to train this choice with simulated users, intent hierarchies and rewards for reducing uncertainty (Wu et al., 2025; Kim et al., 2026; Deng et al., 2026; Sun et al., 2026; Bai et al., 2026). What matters to the user, though, is the outcome of the conversation, including the time spent clarifying.

Simulated users make it easy to try new controllers, but a good simulated outcome need not predict a good human one. A policy trained with one user simulator, for example, failed to improve human outcomes in a recent study, while a simulator tuned on dialogue logs transferred better (Suh et al., 2026). Prompted simulators have their own systematic biases (Zhang et al., 2026b). Human studies remain important, but running one for every change to a controller is rarely practical.

This motivates evaluation from logged conversations. ENIGMA framed dialogue evaluation as off policy evaluation (OPE) of human ratings without deploying the target policy (Jiang et al., 2021). Here we focus on controllers that choose a high level action, such as whether to ask, and then use the same utterance realizer as the logging controllers. The shared realizer lets us compare controller decisions while holding language generation fixed. We assume access to a belief over latent user intent and an imperfect user model that can simulate how the conversation might continue.

The statistical question is what to retain from a conversation history. Marginalized and doubly robust OPE offer ways to combine density ratios with value estimates (Xie et al., 2019; Kallus & Uehara, 2020; Shi et al., 2022). A full history is hard to use for ratio estimation; a short state can miss information that affects later rewards. We use the intent belief and progress counters as that state, and correct model based continuation values with logged rewards. Our main technical result decomposes bias into errors from the user model, the ratio estimate, the intent posterior and information omitted from the state. The ratio and value errors can interact because the same user model supplies the target rollouts used to fit the ratio. Figure 1 sketches this workflow. We also give a population double

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.

robustness identity and, for independently fitted nuisance functions with bounded or clipped weights, an oracle variance comparison (Section 5).

We test this construction in a controlled synthetic dialogue MDP with Monte Carlo value references. Across 50 independent seeds, IM-DR has $0.23\times$ the MSE of per decision importance sampling and $0.50\times$ that of sequential DR using the same synthetic user model. Its within task Kendall τ is 0.884, and paired comparisons clustered by seed favour it over four baselines (Table 1). A second experiment with different controller rules preserves the point estimate advantage, but coverage of nominal 95% intervals falls to 0.760 (Section 7.2). Exploratory ablations and stress tests are reported in the appendix, including tests that did not show the sensitivity we expected. The next question is whether the same abstraction and uncertainty estimates hold up on human dialogue logs.

2 Background and Notation

Conversations with a latent intent. A conversation starts from the user's first message x_0 and a latent intent $z \in \mathcal{Z}$ drawn from a task prior $p(z \mid x_0)$: what the user would have specified had they been perfectly explicit. At turn $t \le T$, a high-level controller chooses a dialogue act $a_t \sim \pi(\cdot \mid h_t)$ (question, attempt, revision or close). A fixed realizer shared by every policy draws an utterance $y_t \sim g(\cdot \mid h_t, a_t)$, and the user replies $u_t \sim \pi_U(\cdot \mid h_t, y_t, z)$; $h_t = (x_0, a_1, y_1, u_1, \ldots, u_{t-1}) \in \mathcal{H}$. A conversation ends when an attempt is accepted or at T. The evaluated policies differ only in $\pi(a_t \mid h_t)$: the realizer g and user kernel are policy-invariant. This action-sufficiency condition makes act-level reweighting valid. Arbitrary end-to-end LLM policies with different within-act utterance distributions are outside the identification result.

Conversation-level reward. Following Wu et al. (2025), $R = w_1\,\text{quality} + w_2\,\text{satisfaction} - w_3\,|\tau| - w_4\,\text{unnecessary}$ with $w = (1, 1, 0.03, 0.1)$ fixed in the testbed (task success, simulated satisfaction, turns, and questions whose answer h_t already determined). We write $R = \sum_{t \le T} r_t$ with $|r_t| \le R_{\max}$, turn costs paid per turn and quality and satisfaction at the final turn; $V(\pi) = \mathbb{E}_{z,\tau\sim\pi,\pi_U}[R]$.

Logs and the problem. Logs $\mathcal{D} = \{(\tau_j, z_j, b_j)\}_{j=1}^n$ hold conversations under behaviour policies b_j, their intents z_j and the probabilities $\pi_b(a_t \mid h_t)$; for a new π' we want $\hat{V}(\pi') \approx V(\pi')$ with an interval and, across candidates, the correct value ranking.

Importance sampling and DR. With $\rho_t = \pi'(a_t \mid h_t)/\pi_b(a_t \mid h_t)$ and $\rho_{1:t} = \prod_{k \le t} \rho_k$, PDIS estimates $V(\pi')$ by $\frac{1}{n}\sum_j \sum_t \rho_{1:t}^{(j)} r_t^{(j)}$ (Precup et al., 2000), unbiased under overlap and the shared-realizer condition but with $\mathrm{Var}(\rho_{1:t})$ growing as $\rho_{\max}^{2t}$. Sequential DR (Jiang & Li, 2016; Thomas & Brunskill, 2016) adds a model $\hat{Q}_t(h,a) \approx Q_t^{\pi'}(h,a)$, $\hat{V}_t(h) = \mathbb{E}_{a\sim\pi'}\hat{Q}_t(h,a)$, as a control variate:

$$\hat{V}_{\mathrm{DR}} = \tfrac{1}{n}\textstyle\sum_j \big[\hat{V}_1(h_1^{(j)}) + \sum_t \rho_{1:t}^{(j)}\big(r_t + \hat{V}_{t+1}(h_{t+1}) - \hat{Q}_t(h_t,a_t)\big)^{(j)}\big], \tag{1}$$

unbiased if either the ratios or $\hat{Q}$ are correct. Marginalised IS (Liu et al., 2018; Xie et al., 2019; Kallus & Uehara, 2020) replaces $\rho_{1:t}$ by the occupancy ratio $w_t(s,a) = d_t^{\pi'}(s,a)/d_t^{\pi_b}(s,a)$ on a suitable state s. The full history is a valid information state but can be too high-dimensional for reliable ratio estimation; a compact summary is useful only to the extent that its omitted history does not confound the residual correction.

Calibration. For a posterior $q(z \mid h)$ with confidence $c = \max_z q(z \mid h)$ and correctness y of its arg-max against the elicited intent, $\mathrm{ECE}(q) = \sum_{b=1}^B P(c \in I_b)\,|\mathbb{E}[y \mid c \in I_b] - \mathbb{E}[c \mid c \in I_b]|$ over B equal-width bins (Naeini et al., 2015; Guo et al., 2017); we also report Brier score (Brier, 1950) and NLL.

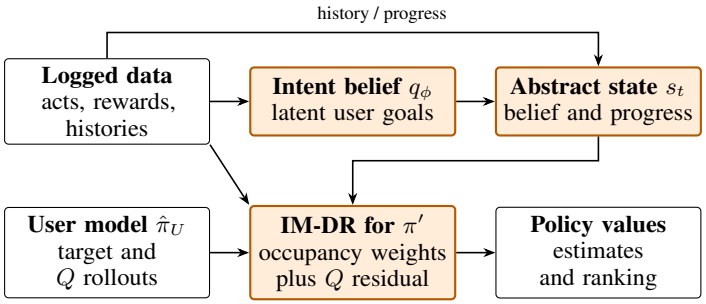

Figure 1: One evaluation pass for a fixed candidate controller π'. Logged histories feed both the intent belief and the abstract state's progress features; logged rewards and histories also enter the DR residual directly. The *same* user model supplies target-state occupancy rollouts for the density-ratio classifier and Q-continuation rollouts for the residual correction, coupling the fitted nuisances. Policy values are outputs, not inputs to the state. The current evaluation instantiates every block in a controlled synthetic MDP.

3 Related Work

Collaborative assistants and user simulation. Collaborative assistants increasingly learn when to clarify or act proactively by interacting with simulated users (Wu et al., 2025; Kim et al., 2026; Deng et al., 2026; Sun et al., 2026; Bai et al., 2026). Those systems motivate our setting, but we do not evaluate their checkpoints or logs. Simulator quality is a separate concern: role-play-trained policies may fail to transfer to people, while a dialogue-log-tuned simulator transferred better in one study (Suh et al., 2026; Zhang et al., 2026b; Zhao et al., 2024). We treat the user model as a fallible control variate in a controlled MDP, leaving correction on natural-language human logs open.

Dialogue and partially observed OPE. ENIGMA already estimates dialogue quality from logs using model-free, behaviour-agnostic distribution correction and language-model features (Jiang et al., 2021). We neither introduce dialogue OPE nor claim to outperform ENIGMA; we have not run that comparison. Instead, we use logged controller actions and an intent-belief state, with a fallible user model supplying the control variate. Marginalized ratios and DR are established tools (Xie et al., 2019; Kallus & Uehara, 2020), as are POMDP belief states and DR with bridge functions (Williams & Young, 2007; Shi et al., 2022). Bridge-function work addresses latent confounding; here we assume known behaviour-controller propensities and a shared realizer. With identical state and nuisance estimates, generic marginalized DR and IM-DR give identical estimates. What this paper adds is error accounting for the chosen dialogue abstraction under these assumptions; Section A discusses the distinction in more detail.

4 Method: Intent-Marginalised Doubly-Robust Evaluation

Problem and key insight. The estimator receives logs $\mathcal{D}$, a candidate π', a user model $\hat{\pi}_U(u \mid h, a, z)$ and a posterior $q_\phi(z \mid h)$, and returns $\hat{V}(\pi')$ with a descriptive fixed-fit bootstrap interval; the objective is MSE against $V(\pi')$ and agreement with the ranking a study would produce. Under the shared-realizer and policy-invariant-user assumptions, a user model supplies continuation values for a candidate from a logged prefix, while only the assistant's high-level act probabilities need reweighting. We test whether a compact intent-belief/progress summary can support lower-variance occupancy weights than cumulative act ratios. Sufficiency of this summary is an assumption to examine, not a property guaranteed by calibration. IM-DR combines these weights with the user-model control variate. Either exact continuations or weights that balance the relevant residual recover the value at the population level (Theorem 5.4); their fitted errors are coupled in the implementation.

Intent posterior. In general, a learned head can map h_t to logits over intent slots or candidate intent descriptions, with $q_\phi^\eta(z \mid h_t) = \mathrm{softmax}(\ell_\phi(h_t)/\eta)_z$ and a temperature η fitted on held-out intent-labelled data. In the reported testbed this component is exact Bayesian filtering under a

deliberately misspecified user model (Section D); no LoRA posterior or human calibration data are evaluated. The calibration-transfer result in Theorem C.3 is therefore a theoretical bound, not an empirical human-calibration claim.

Abstract state and marginalised weights. The implemented abstract state $s_t = \varphi(h_t)$ comprises normalised posterior entropy (8 bins), MAP joint-intent probability (4 bins), turn index (one-hot) and questions asked so far (capped at 3 and one-hot). The classifier adds the act indicator and its interactions with each state block; it does not use the full posterior, resolved-slot fraction or previous-act type. The action in every weight and ratio is the high-level dialogue act a_t, not the realised utterance y_t. The shared realizer $g(y \mid h, a)$ makes the likelihood ratio between policies equal to the ratio of their controller probabilities. The testbed exposes these probabilities exactly; deployment requires logged behaviour-controller probabilities and an auditable target controller. The guarantee does not cover two end-to-end LLMs whose within-act utterance distributions differ. The weight we need is the occupancy ratio $w_t(s,a) = d_t^{\pi'}(s,a)/d_t^{\pi_b}(s,a)$ of (s_t, a_t) under the two controllers with the true user (ended conversations are absorbing). We estimate it by classifier-based density-ratio estimation (Sugiyama et al., 2012): roll out π' on $\hat{\pi}_U$, label rollout pairs $y = 1$ and logged pairs $y = 0$, fit a logistic classifier $c(s,a) \approx P(y = 1 \mid s, a)$ on one-hot bins of the state with action interactions, and set

$$\hat{w}_t(s,a) = \min\Big\{W_{\max},\ \frac{n_0}{n_1} \cdot \frac{\bar{L}'}{\bar{L}_b} \cdot \frac{c(s,a)}{1 - c(s,a)}\Big\}, \tag{2}$$

with n_0, n_1 the numbers of logged and rolled-out steps, $\bar{L}', \bar{L}_b$ the mean lengths and $W_{\max} = 20$. Class-prior odds first estimate the ratio of pooled-step distributions; multiplying by $\bar{L}'/\bar{L}_b$ converts it to an unnormalised occupancy ratio. Because turn index is a feature, this can represent per-turn occupancy ratios when the classifier is correctly specified. Neither classifier specification nor abstract-state sufficiency is guaranteed here. Because rollouts use $\hat{\pi}_U$, $\hat{w}_t$ inherits a user-model error (part of ε_w in Theorem 5.5), which multiplies, rather than adds to, the error of the control variate. The theoretically analysed estimator uses the unnormalised clipped weights $\hat{w}_t$. We also evaluate the common finite-sample heuristic $\tilde{w}_t^{(j)} = \hat{w}_t^{(j)}/(n^{-1}\sum_i \hat{w}_t^{(i)})$ with terminated conversations carried into an absorbing state (Swaminathan & Joachims, 2015); its random denominator is not covered by our concentration theorem and is reported as an ablation rather than inheriting that guarantee.

User model and control variate. The model-based term can be any conditional model $\hat{\pi}_U(u \mid h, a, z)$. The reported experiment uses controlled role-play and grounded testbed models obtained by interpolation toward the known true user dynamics (Section D); this uses privileged simulator knowledge and does not train or evaluate UserLM-8b or WildChat. From every logged step we continue the conversation $m = 8$ times with π' and $\hat{\pi}_U$ under the logged intent and average the returns:

$$\begin{aligned} \hat{Q}_t(h_t, a) &= \tfrac{1}{m}\textstyle\sum_{i=1}^{m} G^{(i)}(h_t, a; \pi', \hat{\pi}_U, z), \\ \hat{V}_t(h_t) &= \mathbb{E}_{a\sim\pi'(\cdot|h_t)}\hat{Q}_t(h_t, a), \end{aligned} \tag{3}$$

The implementation estimates both $\hat{Q}_t(h_t, \mathrm{ASK})$ and $\hat{Q}_t(h_t, \mathrm{ACT})$ using m independent continuations per act, then computes $\hat{V}_t$ as their exact two-act probability-weighted sum; $\hat{V}_{T_j+1} \equiv 0$ for a conversation that ended at T_j. The primary estimator conditions on logged true z, possible in this synthetic testbed but requiring verified intent labels in deployment. An exploratory variant instead samples z from the factorised synthetic posterior for each continuation (Table 11); no learned natural-language posterior or accompanying bias theorem is evaluated.

The estimator. IM-DR is the marginalised doubly-robust estimator

$$\begin{aligned} \hat{V}_{IM-DR}(\pi') &= \frac{1}{n}\sum_{j=1}^{n}\Big[\hat{V}_1(h_1^{(j)}) + \sum_{t=1}^{T_j}\hat{w}_t\big(s_t^{(j)}, a_t^{(j)}\big)\,\delta_t^{(j)}\Big], \\ \delta_t &= r_t + \hat{V}_{t+1}(h_{t+1}) - \hat{Q}_t(h_t, a_t). \end{aligned} \tag{4}$$

What is and is not estimator-specific. For a fixed representation φ, fitted $\hat{w}$ and $\hat{Q}$, (4) is exactly the generic marginalized-DR residual estimator. A comparator supplied with the *same* state, ratio, Q,

Algorithm 1 IM-DR

1: **Input:** logs $\mathcal{D}$, candidate π', user model $\hat{\pi}_U$, posterior q_ϕ, rollouts m, clip $W_{\max}$
2: Roll out π' on $\hat{\pi}_U$; fit classifier c between rollout and logged (s_t, a_t) pairs, $s_t = \varphi(h_t)$; unnormalised clipped weights by (2)
3: **for** each logged conversation j and turn $t \le T_j$ **do**
4: $\quad \hat{Q}_t, \hat{V}_t$ by m rollouts of π' on $\hat{\pi}_U$ given z_j (3); $\delta_t \leftarrow r_t + \hat{V}_{t+1} - \hat{Q}_t(h_t, a_t)$
5: **end for**
6: $\psi_j \leftarrow \hat{V}_1(h_1^{(j)}) + \sum_t \hat{w}_t(s_t^{(j)}, a_t^{(j)})\, \delta_t^{(j)}$
7: **Return:** $\frac{1}{n}\sum_j \psi_j$ and a fixed-nuisance bootstrap interval of $\{\psi_j\}$

logged sample and clipping therefore produces identical per-conversation ψ_j by algebra, not merely statistically indistinguishable estimates. "Intent" in IM-DR describes the chosen state construction and its error accounting, not a new DR update. The confirmatory comparison isolates the benefit of abstract weighting relative to cumulative-ratio methods in this simulator; it does *not* isolate an advantage over generic marginalized DR on the same state. Exploratory feature ablations (Tables 5 and 6) compare the intent/progress representation to reduced states, but do not establish a general representation advantage. Replacing $\hat{w}_t$ by $\rho_{1:t}$ recovers sequential DR (1); $\hat{w} \equiv 0$ gives the simulator-only estimate $\frac{1}{n}\sum_j \hat{V}_1(h_1^{(j)})$; $\hat{Q} \equiv 0$ gives marginalised IS on the abstract state (IM-IS); all three are ablations. Intervals are 1,000-resample percentile bootstraps over per-conversation contributions (Efron & Tibshirani, 1993). Leave-one-policy-out excludes the target policy's logs, but the density-ratio classifier is fitted on the same pooled evaluation logs used for the estimate; the user model is a fixed synthetic interpolation, not fitted from these logs. Thus these intervals hold fitted nuisances and the Monte-Carlo truth fixed, omit nuisance-fitting and reference-value uncertainty, and are not certified 95% coverage intervals. Pooled logs use each conversation's own logging probabilities (Algorithm 1).

Candidate policies. The estimator is agnostic to where candidates come from. The primary comparison uses four fixed stochastic testbed policies whose question probability depends on posterior entropy; the policy-family-shift experiment adds four unseen stochastic controllers based on time or prior rejection (Section 7.2). None is a trained LLM or DPO policy, and no deployment improvement is claimed.

5 THEORY

The results adapt known arguments for doubly-robust and marginalised estimators (Jiang & Li, 2016; Kallus & Uehara, 2020) to an estimated abstract state; full proofs are in Section C. The ancillary calibration-transfer diagnostic is stated in the appendix (Theorem C.3). $\mathbb{E}_b$ and $\mathbb{E}'$ denote expectations under the logging process and under π' with the true user, $\Delta_t(h, a) = \mathbb{E}[\delta_t \mid h_t = h, a_t = a]$ the conditional mean of the correction term (zero when $\hat{Q}$ is exact), $\|\cdot\|_2$ the $L_2(\mathbb{E}_b)$ norm, $\omega_t(h, a) = w_t^\star(\varphi^\star(h), a)$ the ideal weight on the abstract state built from the *true* posterior, and $\tilde{\omega}_t$ the same ratio function evaluated at q_ϕ.

Assumption 5.1 (Overlap on the abstract state)**.** $\omega_t(h, a) \le W_{\max}$ for all t, h, a with π_b-positive probability.

Assumption 5.2 (Action-level identification)**.** The logging and target controllers share the utterance realizer $g(y \mid h, a)$ and the user-response kernel is policy-invariant conditional on (h, y, z); logged behaviour-controller probabilities are known. Thus the likelihood ratio contributed by the assistant at a turn is $\pi'(a \mid h)/\pi_b(a \mid h)$.

Assumption 5.3 (Continuity)**.** $w_t^\star(\varphi(p), a)$ is L-Lipschitz in the posterior p under $\|\cdot\|_1$.

Let $\varepsilon_{\mathrm{abs},t} = \sup_{|f|\le 1} |\mathbb{E}'[f(h_t, a_t)] - \mathbb{E}_b[\omega_t(h_t, a_t) f(h_t, a_t)]|$ be the abstraction error. It is zero when the abstract ratio balances every bounded history–act function, a stronger property than balancing functions of the abstract state alone; it need not vanish merely because the full history is a valid information state. Let $\varepsilon_{w,t} = \|\hat{w}_t - \tilde{\omega}_t\|_2$ be density-ratio error, including rollout-model error and any clipping bias; $\varepsilon_{\mathrm{cal}} = \max_t \big\| \|q_\phi(\cdot \mid h_t) - p(\cdot \mid h_t)\|_1 \big\|_2$ be root-mean-square posterior

error, unobservable on humans and not bounded by ECE alone; and $\varepsilon_U = \sup_{h,a,z} \mathrm{TV}(\hat{\pi}_U(\cdot \mid h, a, z), \pi_U(\cdot \mid h, a, z))$ be user-model error.

Theorem 5.4 (Double robustness)**.** *Under Assumptions 5.1–5.3, the population version of the unnormalised estimator in* (4) *equals $V(\pi')$ if either (i) $\varepsilon_U = 0$ or (ii) $\varepsilon_{w,t} = \varepsilon_{\mathrm{abs},t} = 0$ for all t and $\varepsilon_{\mathrm{cal}} = 0$ (double robustness in the sense that either nuisance suffices; the two conditions are not independent as in Kallus & Uehara (2020)). With finite m, the population identity still holds if the Q-continuation draws are conditionally independent of logged replies and of the draws used to fit weights, given the logged histories and fixed model parameters. Sharing the user model does not mean sharing Monte-Carlo draws. The result does not apply to the self-normalised heuristic, whose random denominator generally changes the estimand at finite n.*

Theorem 5.5 (Bias under misspecification)**.** *Under Assumptions 5.1–5.3, with* $\mathrm{Bias}(\hat{V}) = |\mathbb{E}_b[\hat{V}] - V(\pi')|$,

$$\begin{aligned} \mathrm{Bias}(\hat{V}_{IM-DR}) &\le \textstyle\sum_t \|\Delta_t\|_2(\varepsilon_{w,t} + L\varepsilon_{\mathrm{cal}}) + \|\Delta_t\|_\infty \varepsilon_{\mathrm{abs},t} \\ &\le T(T+1)\, R_{\max}\, \varepsilon_U \big(\bar{\varepsilon}_w + L\,\varepsilon_{\mathrm{cal}} + \bar{\varepsilon}_{\mathrm{abs}}\big), \end{aligned} \tag{5}$$

with $\bar{\varepsilon}_w = \max_t \varepsilon_{w,t}$, $\bar{\varepsilon}_{\mathrm{abs}} = \max_t \varepsilon_{\mathrm{abs},t}$; the second step uses $\|\Delta_t\|_\infty \le 2(T-t+1)R_{\max}\varepsilon_U$.

Proof sketch. The bias equals $\sum_t \mathbb{E}_b[(\hat{w}_t - \omega_t)\Delta_t] + (\mathbb{E}_b[\omega_t \Delta_t] - \mathbb{E}'[\Delta_t])$ by the telescoping identity; Cauchy–Schwarz, Lipschitz continuity and the definition of $\varepsilon_{\mathrm{abs}}$ give the first bound, and a simulation-lemma argument bounds Δ_t by the user model's TV error times the remaining horizon. □

The bound is a *product* of residual size and weight/abstraction errors: improving the user model may reduce the residual contribution, but does not by itself fix the posterior, classifier or abstraction. Better posterior L_1 accuracy can reduce $\varepsilon_{\mathrm{cal}}$; lower ECE alone need not. The weight fit can reduce $\bar{\varepsilon}_w$, although it depends on the user-model rollouts. The following comparison concerns oracle variance under fixed nuisances, not the fitted estimator's full sampling variance.

Theorem 5.6 (Variance and concentration)**.** *Let $|r_t| \le R_{\max}$, $G_{\max} = TR_{\max}$ and $\hat{w}_t \le W_{\max}$. Assume the nuisance functions and user model are fixed before, or fitted independently of, the n evaluation conversations, with independent rollout randomness across evaluation units. Conditional on these nuisances, $\mathrm{Var}(\hat{V}_{IM-DR}) \le \frac{2}{n}\big(\mathrm{Var}_b(\hat{V}_1) + T\,W_{\max}^2 \sum_t \mathbb{E}_b[\delta_t^2]\big)$, and with probability at least $1-\delta$ the estimate is within $\sqrt{2\mathrm{Var}_b(\psi)\log(2/\delta)/n} + O(TW_{\max}G_{\max}\log(2/\delta)/n)$ of its conditional mean (Bernstein; explicit constants in Section C). For sequential DR with cumulative ratios the same bounds hold with $W_{\max}$ replaced by $\rho_{\max}^T$, $\rho_{\max} = \max_{h,a} \pi'(a \mid h)/\pi_b(a \mid h)$.*

For untruncated cumulative ratios this comparison replaces the potentially exponential $\rho_{\max}^T$ factor by a bounded abstract weight; clipping a sequential estimator at the same threshold also bounds its variance but may introduce bias. The result does not quantify the variance of same-log nuisance fitting used in our experiments, and $W_{\max} = 20$ is imposed by clipping rather than established from an overlap test. The remaining bound grows polynomially through the sum over turns and $G_{\max} = TR_{\max}$; it does not claim horizon-independent variance. Figure 3 shows the realised effective-sample-size gap in the testbed.

Corollary 5.7 (Where to spend human data)**.** *If grounding improves the user model in TV, it can reduce the residual contribution to bias; if posterior learning improves L_1 accuracy, it can reduce $\varepsilon_{\mathrm{cal}}$; and bounded abstract-state weights limit the conditional variance term, potentially at the cost of clipping and abstraction bias. These are design implications of Section 4, not guarantees that an ECE or user-model training metric improves OPE; the components are explored in Table 5.*

6 Experimental Setup

Synthetic testbed with known truth. Because the human-rated value of a policy is never observed exactly, we evaluate estimators in a simulator with a finite Monte-Carlo value reference. For the 40,000-rollout references, the largest estimated Monte-Carlo standard error across primary task–policy pairs is 0.00371 reward units (the IM-DR RMSE is 0.0278); this uncertainty is not propagated into MSE or coverage intervals (Section G). The testbed (Section D) is a conversation MDP with a latent intent of M slots with V values: the assistant asks about its most uncertain slot (answered misleadingly with a per-user probability η_i) or commits to its MAP intent (accepted with probability

Table 1: Fresh-seed confirmatory comparison of unnormalised IM-DR on 600 synthetic testbed cells (50 seeds, three tasks, four target policies). MSE is shown in $\times 10^{-2}$. Baseline MSEs are reconstructed from the reported mean paired difference; intervals resample seeds and retain all task–policy cells. "Wins" is the fraction of cells in which IM-DR has lower squared error.

Estimator	MSE ↓	IM-DR minus baseline	Wins
IM-DR (unnormalised)	**0.0773 [0.0658, 0.0901]**	—	—
PDIS	0.3346	-0.2573 [-0.3034, -0.2139]	71.3%
Sequential DR, synthetic 0.85	0.1541	-0.0768 [-0.0974, -0.0590]	67.0%
Sequential DR, role-play	0.1633	-0.0860 [-0.1052, -0.0683]	65.3%
MAGIC, synthetic 0.85	0.3691	-0.2918 [-0.3164, -0.2670]	85.0%

0.85 if correct, 0.15 otherwise), so the turn type is the action; conversations end on acceptance or at $T = 12$, with rewards as in Section 2 and rating noise of SD 0.3. Three task families vary M and the user population (A: 3×3; B: 4×3; C: terse users); four behaviour policies P_1–P_4 ask with probability $\sigma((H_t - \theta)/\beta)$ of the posterior entropy, with values that are close and not monotone in question rate. The role-play simulator is an over-cooperative user; the primary control variate is a synthetic model interpolated 85% of the way from that simulator to the known true dynamics. This interpolation is a controlled misspecification level, not a model trained on WildChat or any other dialogue corpus. Each behaviour policy logs 200 conversations per task; leave-one-policy-out evaluates each policy from the 600 pooled conversations of the other three against a 40,000-rollout Monte-Carlo reference; 50 seeds. The finite reference is reused across seeds, and its Monte-Carlo error is not propagated into the reported intervals.

Evaluation scope. All reported experiments use the synthetic testbed above. The three families are controlled regimes rather than claims about MATH, coding, document editing or deployed users. Ground-truth values come from 40,000 Monte-Carlo rollouts, while each estimator receives only the pooled logged trajectories. No participant data, production conversations, trained LLM policy checkpoints or human calibration labels are used in the reported results.

Policies and baselines. The four stochastic testbed policies differ in their entropy threshold and question propensity. OPE baselines receive the same logs, user model, clipping rule and bootstrap budget: trajectory and per-decision IS (Precup et al., 2000), self-normalised PDIS (Swaminathan & Joachims, 2015), linear fitted-Q evaluation (Le et al., 2019), simulator-only evaluation, sequential DR (Jiang & Li, 2016) with the role-play and grounded testbed models, MAGIC (Thomas & Brunskill, 2016), and a fine-grid histogram-ratio MIS baseline inspired by marginalised OPE (Nachum et al., 2019). The last is not a DualDICE implementation. A fresh on-policy sample of 60 conversations is a finite-sample reference, not an estimator trained on the logged pool.

Metrics and protocol. OPE: MSE over (task, policy, seed) cells ($\times 10^{-2}$, 95% seed-clustered bootstrap CI), within-task Kendall τ (Kendall, 1938) of the four-policy ranking averaged over tasks and seeds, and coverage of the truth by the per-estimate 95% bootstrap interval. We also report seed-clustered paired MSE differences and an independent fresh-seed ablation follow-up. For transparency, the appendix retains exploratory effective-sample-size, horizon/data and component-ablation analyses, plus two 30-fresh-seed stress tests that did not produce the anticipated monotone degradation. Calibration quantities are synthetic diagnostics and are not evidence about human calibration. Three additional 50-seed checks use disjoint seed offsets: independent ratio fitting versus same-data fitting, four target controllers outside the entropy-threshold logging family, and posterior-sampled versus true logged intent in model continuations. The policy-family-shift MSE/ranking result is in Section 7.2; all three protocols, including negative findings and interval limitations, are detailed in Section E.3. These checks do not add language or human data.

Table 2: Synthetic policy-family shift: 50 fresh seeds, three tasks and four target controllers absent from the entropy-threshold logging family. MSE is in units of 10^{-2}; its 95% interval resamples seeds. Coverage is empirical coverage of nominal 95% per-cell intervals with fitted nuisances held fixed.

Estimator	MSE [seed CI] ↓	τ ↑	Coverage ↑
IM-DR	**0.1345 [0.1157, 0.1542]**	**0.996**	0.760
PDIS	15.9399 [15.0743, 16.8130]	0.222	0.183
Sequential DR	0.5705 [0.4947, 0.6521]	0.816	**0.865**

7 Results

7.1 Conversation-Level OPE

In the confirmatory 50-seed testbed (Table 1), unnormalised IM-DR reaches MSE 0.0773×10^{-2} [0.0658, 0.0901], within-task Kendall $\tau = 0.884$, and empirical coverage 0.950 for its fixed-fit intervals. Its MSE is $0.23\times$ PDIS and $0.50\times$ sequential DR when both use the same synthetic user model. Each of the four seed-clustered paired intervals excludes zero. IM-DR-minus-baseline MSE differences range from -0.0768 against sequential DR with the synthetic model to -0.2918 against MAGIC. This is an aggregate advantage, not a win in every cell: IM-DR has lower error in 65.3–71.3% of cells against the two DR variants and PDIS, and in 85.0% against MAGIC.

Estimator choice. Before the confirmatory run, an independent 50-seed paired follow-up showed that removing self-normalisation improved MSE from 0.1031 to 0.0896 ($\Delta = +0.0135$ [$+0.0057$, $+0.0217$], self-normalised minus unnormalised). We therefore used the unnormalised estimator for confirmation, matching the object analysed in Section 5. Exploratory ablations, horizon/data sweeps and two non-confirmatory stress tests remain in Section E.

7.2 Synthetic Policy-Family Shift

We next test four high-level target controllers outside the entropy-threshold logging family: constant, front-loaded, back-loaded and rejection-reactive asking. Over 50 fresh seeds and three tasks (Table 2), IM-DR has MSE 0.1345×10^{-2}, compared with 0.5705×10^{-2} for sequential DR; the paired difference is -0.4360 [-0.5077, -0.3687] in the same units. Within-task ranking τ reaches 0.996. Yet empirical coverage of the nominal-95% fixed-fit intervals falls to 0.760, and is worst for constant asking (Table 12). These intervals do not provide reliable uncertainty estimates under this shift. Front-loaded asking is truly top-ranked in all three tasks, making top-1 accuracy a limited ranking test. The lower error applies to these *synthetic* controller rules; it says nothing yet about transfer to language policies or people. Additional diagnostics and two weaker follow-ups appear in Section E.3.

Limitations. The evidence is limited to a controlled synthetic conversation MDP. The primary behaviour/target policies are aligned with the estimator's abstract state; a separate test with four different controller rules improves the policy-family-shift evidence but has only 0.760 nominal-95%-interval coverage. None of these experiments establishes accuracy on natural-language histories, learned LLM posteriors, trained user models or human users; see Section B.

8 Conclusion

Logged conversations could let us compare clarification strategies before asking people to try every new controller. We studied that possibility in a setting where controllers share an utterance realizer. An intent belief and progress counters make occupancy weighting tractable, while logged rewards correct continuation values from a user model. The analysis shows where errors in those ingredients, and in the state abstraction itself, enter the estimate.

In the synthetic testbed, the unnormalised estimator has lower MSE than PDIS, both sequential DR variants and MAGIC across 50 independent seeds. Its point estimate advantage also persists for a different family of controller rules. The uncertainty estimates are less dependable: coverage of nominal 95% intervals drops to 0.760 under that shift. The appendix reports weaker follow up results

and stress tests that did not reveal the expected sensitivity. Testing the method on human dialogue logs, with a learned intent posterior and a fitted user model, is the next step.

AI USE STATEMENT

Generative AI tools assisted with drafting and editing text, literature discovery, and implementation and debugging of the synthetic artifact. The authors reviewed all AI-assisted text and claims, checked cited sources, and tested generated code. The authors take responsibility for the final content, including text, claims, and artifacts produced with AI assistance.

ETHICS STATEMENT

This work aims to make evaluation of collaborative assistants cheaper and more reliable by ranking candidates from logged conversations. The reported experiments use only synthetic trajectories and involve no participants or personal data. Any future human validation requires institutional ethics approval before recruitment, informed consent, fair compensation, and release only of scrubbed conversations with separate sharing consent. Intent inference could be repurposed to manipulate users; the method therefore should not be used to optimise engagement or infer sensitive attributes.

REPRODUCIBILITY STATEMENT

The anonymous artifact is available at `https://anonymous.4open.science/r/IM-DR/`. It contains source code for the estimator, deterministic synthetic testbed, tests and a Bash reproduction guide with exact seed settings for the primary comparison, ablation follow-up and three additional 50-seed checks. No results are bundled; the scripts generate the outputs used for comparison with the reported tables. Algorithmic details and proofs appear in Sections C, D and G. The artifact contains no human logs, trained language model checkpoints, human study outputs or projected human results.

REFERENCES

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019.

Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. STaR-GATE: Teaching language models to ask clarifying questions. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2403.19154.

Cedar Site Bai, Duanshun Li, Zhenyu Liao, Sheikh Sarwar, Huiyuan Chen, Yuan Chen, Changhe Yuan, Haiyang Zhang, and Qilin Qi. Ask to be sure: Informative interactions for confident multi-turn LLM recommendation. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM)*, 2026. doi: 10.1145/3799682.3839868. arXiv:2608.15949.

Krisztian Balog and ChengXiang Zhai. User simulation in the era of generative AI: User modeling, synthetic data generation, and system evaluation. *arXiv preprint arXiv:2501.04410*, 2025. arXiv:2501.04410.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Maximillian Chen, Ruoxi Sun, Tomas Pfister, and Sercan Ö. Arık. Learning to clarify: Multi-turn conversations with action-based contrastive self-training. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2406.00222.

Mengyi Deng, Zhiwei Li, Xin Li, Tingyu Zhu, Ying Zhao, Zhijiang Guo, and Wei Wang. Uncertainty-aware clarification in LLM agents with information gain. *arXiv preprint arXiv:2606.03135*, 2026. arXiv:2606.03135.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 1097–1104, 2011.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

Justin Fu, Mohammad Norouzi, Ofir Nachum, George Tucker, Ziyu Wang, Alexander Novikov, Mengjiao Yang, Michael R. Zhang, Yutian Chen, Aviral Kumar, Cosmin Paduraru, Sergey Levine, and Tom Le Paine. Benchmarks for deep off-policy evaluation. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2103.16596.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*, 2019.

Natasha Jaques, Judy Hanwen Shen, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Shane Gu, and Rosalind Picard. Human-centric dialog training via offline reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Haoming Jiang, Bo Dai, Mengjiao Yang, Tuo Zhao, and Wei Wei. Towards automatic evaluation of dialog systems: A model-free off-policy evaluation approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7419–7451, 2021. doi: 10.18653/v1/2021.emnlp-main.589. URL https://aclanthology.org/2021.emnlp-main.589/.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 652–661, 2016.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020.

Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

Tae Soo Kim, Yoonjoo Lee, Jaesang Yu, John Joon Young Chung, and Juho Kim. DiscoverLLM: From executing intents to discovering them. In *Proceedings of the International Conference on Machine Learning*, 2026. arXiv:2602.03429.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. CLAM: Selective clarification for ambiguous questions with generative language models. *arXiv preprint arXiv:2212.07769*, 2023.

Hoang M. Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. Eliciting human preferences with language models. *arXiv preprint arXiv:2310.11589*, 2023.

Fangzhou Lin, Peiran Li, Shuo Xing, Siyuan Yang, Qianwen Ge, Kazunori Yamada, Ziming Zhang, Haichong Zhang, and Zhengzhong Tu. AdaptFuse: Training-free sequential preference learning via externalized Bayesian inference. *arXiv preprint arXiv:2604.03925*, 2026. arXiv:2604.03925.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022. arXiv:2205.14334.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2015.

Doina Precup, Richard S. Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*, pp. 759–766, 2000.

Manik Rana, Calissa Man, Anotida Expected Msiiwa, Jeffrey Paine, Kevin Zhu, Sunishchal Dev, Vasu Sharma, and Ahan M R. AgentChangeBench: A multi-dimensional evaluation framework for goal-shift robustness in conversational AI. *arXiv preprint arXiv:2510.18170*, 2025. arXiv:2510.18170.

Sudha Rao and Hal Daumé III. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

Ankur Samanta, Akshayaa Magesh, Tal Lancewicki, Ayush Jain, Youliang Yu, Paul Sajda, Kaveh Hassani, Aditya Modi, Daniel R. Jiang, and Yonathan Efroni. BayesBench: Evaluating LLM belief trajectories under multi-turn evidence accumulation. *arXiv preprint arXiv:2606.30850*, 2026. arXiv:2606.30850.

Jost Schatzmann, Karl Weilhammer, Matt Stuttle, and Steve Young. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The Knowledge Engineering Review*, 21(2):97–126, 2006.

Chengchun Shi, Masatoshi Uehara, Jiawei Huang, and Nan Jiang. A minimax learning approach to off-policy evaluation in confounded partially observable markov decision processes. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 2022. URL `https://proceedings.mlr.press/v162/shi22f.html`.

Seongho Son, William Bankes, Sayak Ray Chowdhury, Brooks Paige, and Ilija Bogunovic. Right now, wrong then: Non-stationary direct preference optimization under preference drift. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 56173–56203. PMLR, 2025. URL `https://proceedings.mlr.press/v267/son25a.html`.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

Joseph Suh, Ayush Raj, Minwoo Kang, and Serina Chang. Quantifying the utility of user simulators for building collaborative LLM assistants. *arXiv preprint arXiv:2605.09808*, 2026. arXiv:2605.09808.

Weiwei Sun, Xuhui Zhou, Weihua Du, Xingyao Wang, Sean Welleck, Graham Neubig, Maarten Sap, and Yiming Yang. Training proactive and personalized LLM agents. In *Conference on Language Modeling (COLM)*, 2026. arXiv:2511.02208.

Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems 28 (NeurIPS)*, 2015.

Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 2139–2148, 2016.

Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2015.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. arXiv:2305.14975.

Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*, 2022.

Cameron Voloshin, Hoang M. Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. In *NeurIPS Datasets and Benchmarks Track*, 2021. arXiv:1911.06854.

Jason D. Williams and Steve Young. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422, 2007.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. CollabLLM: From passive responders to active collaborators. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. arXiv:2502.00640.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

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. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2306.13063.

Ziyi Xuan, Yiwen Wu, Zhaoyang Yan, Vinod Namboodiri, and Yu Yang. After talking with 1,000 personas: Learning preference-aligned proactive assistants from large-scale persona interactions. *arXiv preprint arXiv:2602.04000*, 2026. arXiv:2602.04000.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D. Williams. POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179, 2013.

Caiqi Zhang, Ruihan Yang, Xiaochen Zhu, Chengzu Li, Tiancheng Hu, Yijiang River Dong, Deqing Yang, and Nigel Collier. Confidence estimation for LLMs in multi-turn interactions. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25661–25676, 2026a. doi: 10.18653/v1/2026.findings-acl.1280.

Feng Zhang, Shijia Li, Chunmao Zhang, Zhanyu Ma, Jun Xu, Jiuchong Gao, Jinghua Hao, Renqing He, Jingwen Xu, and Han Liu. UserLM-R1: Modeling human reasoning in user language models with multi-reward reinforcement learning. *arXiv preprint arXiv:2601.09215*, 2026b. arXiv:2601.09215.

Michael J. Q. Zhang, W. Bradley Knox, and Eunsol Choi. Modeling future conversation turns to teach LLMs to ask clarifying questions. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.13788.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1m ChatGPT interaction logs in the wild. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2405.01470.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS), Datasets and Benchmarks Track*, 2023.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. ArCHer: Training language model agents via hierarchical multi-turn RL. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.19446.

A Extended Related Work

Training collaborative and clarifying assistants. CollabLLM (Wu et al., 2025) introduced a multi-turn-aware reward computed by forward simulation with an LLM user and validated the resulting policy with a 201-person study; it motivates the target setting, but we use neither its checkpoint nor its task data. Its successors decompose the problem: DiscoverLLM (Kim et al., 2026) tracks an intent hierarchy and rewards concretisation; Deng et al. (2026) maintain a Bayesian belief over the user's goal and reward questions by expected information gain; Sun et al. (2026) optimise proactivity, personalisation and perceptiveness jointly against a simulated user population; Bai et al. (2026) ask when the policy's own entropy is high; Xuan et al. (2026) steer proactivity by persona. Earlier work learned when and what to ask from single-turn or short-horizon signals: expected value of perfect information (Rao & Daumé III, 2018), clarification retrieval (Aliannejadi et al., 2019), selective clarification (Kuhn et al., 2023), preference elicitation with language models (Li et al., 2023), self-taught questioning (Andukuri et al., 2024), future-turn simulation (Zhang et al., 2025) and contrastive self-training on dialogue actions (Chen et al., 2025). Multi-turn RL for language agents (Zhou et al., 2024) supplies the optimisers. All of these papers contribute policies; none focuses on the latent-intent, conversation-level OPE formulation studied here. We evaluate the estimator with controlled testbed policies rather than claiming a new trained language policy.

User simulation and its validity. Statistical user simulation for dialogue management predates LLMs (Schatzmann et al., 2006), and its central question, whether a policy that is good against the simulator is good with people, is as old as the technique. For LLM-based simulators the answer is increasingly negative: Suh et al. (2026) find that RL against a role-play simulator does not transfer to 283 real users while a WildChat-fine-tuned simulator does; Zhang et al. (2026b) train user language models and document the biases of prompted role play; Balog & Zhai (2025) survey the evaluation use of simulators. The result of Suh et al. (2026) motivates treating the user model as a fallible control variate rather than as ground truth. Our testbed varies synthetic model misspecification explicitly; validating a learned language user model and IM-DR against human outcomes remains future work. Unlike simulator-only evaluation, IM-DR uses logged-reward residuals as a correction. The user model still affects both the control variate and the target-state rollouts used to fit occupancy ratios, so an inaccurate model can bias both nuisance components; correction is guaranteed only under the conditions of Theorem 5.4.

Off-policy evaluation. Trajectory and per-decision importance sampling are unbiased under overlap but can suffer from the curse of horizon; self-normalisation and clipping trade some bias for variance control (Precup et al., 2000; Swaminathan & Joachims, 2015; Thomas et al., 2015). Doubly-robust estimators (Dudík et al., 2011; Jiang & Li, 2016) and MAGIC (Thomas & Brunskill, 2016) use a model to reduce error; marginalised importance sampling (Liu et al., 2018; Xie et al., 2019), DualDICE (Nachum et al., 2019) and double reinforcement learning (Kallus & Uehara, 2020) replace trajectory ratios by occupancy ratios on a Markov state. Le et al. (2019) study fitted-Q evaluation, and Voloshin et al. (2021); Fu et al. (2021); Uehara et al. (2022) benchmark and review the field. Kallus & Uehara (2020) already combine cross-fitted marginalized density ratios and value functions in DR OPE. Our testbed implementation uses a coarser, hand-specified summary of an exact Bayesian intent belief and a user model for continuation values; it does not establish a new generic DR primitive or an efficiency theorem. Under matched state, Q and ratio estimates, generic marginalized DR has exactly the same sample-level formula as IM-DR; the paper's empirical comparison cannot attribute its MSE gain to novel DR algebra. More directly, ENIGMA (Jiang et al., 2021) estimates human dialogue scores from logged interactions without target-policy conversations, using DICE-style distribution correction, pseudo-state padding for variable horizons and pretrained language representations. It is model-free and does not require behaviour-policy probabilities. Our estimator trades those properties for an explicit high-level controller model: logged act propensities and a shared realizer, a user-model continuation term, and a bias decomposition for posterior and abstraction misspecification. Neither paper subsumes the other under the evidence here, and we have not run ENIGMA as a baseline. Dialogue POMDPs already use belief states (Williams & Young, 2007; Young et al., 2013); Shi et al. (2022) derive marginalized and DR estimators with bridge functions for partially observed environments with latent confounding. Our action-level logging assumption rules out that confounding problem, while our analysis quantifies error in a specified intent-belief compression of an observable history. Earlier dialogue OPE also includes Jaques et al. (2019; 2020). Relative to this

literature, the residual-based error accounting in Theorem 5.5 is the principal theoretical distinction; whether it helps on real language trajectories is untested.

Calibrated beliefs about the user. Belief tracking over user goals appears in Deng et al. (2026), BayesBench (Samanta et al., 2026) and AdaptFuse (Lin et al., 2026), and drift over time is benchmarked by AgentChangeBench (Rana et al., 2025) and modelled at the population level by NS-DPO (Son et al., 2025); these works evaluate task success, not whether the stated belief is calibrated against the user's actual intent. LLM confidence elicitation (Lin et al., 2022; Tian et al., 2023; Xiong et al., 2024) measures calibration of a model about *its own answers*. Multi-turn confidence work additionally studies how that confidence evolves as evidence accumulates (Zhang et al., 2026a). Our theoretical object is different: a posterior about another agent's hidden intent. The present stress test reports standard ECE, Brier and NLL only for synthetic intents (Naeini et al., 2015; Guo et al., 2017; Brier, 1950); no prompted LLM calibration baseline is evaluated. LLM-as-a-judge evaluation (Zheng et al., 2023) is relevant to future reward construction, but no judge output or human rating enters the current synthetic reward.

B Extended Limitations

Only the synthetic testbed is evaluated. The paper contains no human-log, human-calibration or deployment result, so its empirical conclusions are restricted to the controlled MDP. **Six hundred conversations is small:** on the testbed the advantage over sequential DR vanishes at $N = 1200$ and MAGIC wins at $N = 75$ (Figure 2b); the crossovers on real text are unknown. **The testbed favours abstract-state weights**, whose base policies are functions of that state by construction. A policy-family-shift test adds four target controller rules outside that family and finds lower MSE for IM-DR, but the controllers still share a synthetic ASK/ACT interface and fixed realizer, and coverage falls to 0.760. A separate stress test adds mild omitted rejection-history dependence without monotone degradation, so neither check validates the abstraction-error term empirically. Real policies may react much more strongly to phrasing beyond the state. Likewise, severe temperature distortion changes synthetic ECE without resolving any change in OPE MSE; measured ECE neither bounds $\varepsilon_{\mathrm{cal}}$ nor establishes practical sensitivity here. **The posterior and user models are stylised rather than learned language models**; their misspecification is controlled using known simulator dynamics and cannot reproduce the full errors of open-ended dialogue. The primary continuations condition on the true logged intent, which is privileged information and not generally available in human logs. An exploratory posterior-sampled continuation avoids reading true z for its Q term without a resolved MSE change here, but has no separate bias theorem or natural-language validation. **Reward weights are fixed** and alternative preferences may change policy rankings. The unnormalised estimator is the object covered by the population theory; the earlier self-normalised run is retained only as a labelled finite-sample ablation. **The conditional variance theorem is not an in-sample guarantee:** the primary ratio classifier uses the evaluation logs; an exploratory cross-fit has similar MSE but neither it nor the primary estimator uses a refitted-nuisance bootstrap, and the finite Monte-Carlo reference is treated as fixed. Consequently reported per-cell intervals and their empirical coverage cannot be interpreted as calibrated deployment uncertainty. Candidates outside the support of the logs still violate overlap; clipping bounds variance by construction but can introduce bias, even though the clipping variants coincide on this testbed. Deployment additionally requires logged behaviour-controller probabilities, a fixed or separately modelled utterance realizer, an auditable target controller, and either verified intents or a validated intent-marginal continuation model; these requirements are not satisfied by arbitrary black-box chat APIs.

C Proofs

Throughout, all model-based quantities are conditioned on the elicited intent z recorded in the logs; we suppress z from the notation. The initial distribution of (x_0, z), hence of h_1, does not depend on the assistant policy. Under Assumption 5.2, integrating the shared realizer $g(y_t \mid h_t, a_t)$ into the user kernel makes the reply law $u_t \mid (h_t, a_t, z)$ policy-invariant as well. For a conversation ending at turn T_j we set $\hat{V}_{T_j+1} \equiv 0$ and $\delta_t = 0$ for $t > T_j$, so all sums may be written to T.

C.1 A TELESCOPING IDENTITY

Lemma C.1. *For any bounded $\hat{Q}_t$ and $\hat{V}_t(h) = \mathbb{E}_{a\sim\pi'(\cdot|h)}\hat{Q}_t(h,a)$, $V(\pi') = \mathbb{E}'\big[\hat{V}_1(h_1)\big] + \sum_{t=1}^{T}\mathbb{E}'\big[\delta_t\big]$ with $\delta_t = r_t + \hat{V}_{t+1}(h_{t+1}) - \hat{Q}_t(h_t,a_t)$.*

Proof. Under π', $a_t \sim \pi'(\cdot \mid h_t)$, so $\mathbb{E}'[\hat{Q}_t(h_t,a_t) \mid h_t] = \hat{V}_t(h_t)$ and $\mathbb{E}'[\hat{Q}_t(h_t,a_t)] = \mathbb{E}'[\hat{V}_t(h_t)]$. Hence $\sum_t \mathbb{E}'[\hat{V}_{t+1}(h_{t+1}) - \hat{Q}_t(h_t,a_t)] = \sum_t \mathbb{E}'[\hat{V}_{t+1}(h_{t+1})] - \mathbb{E}'[\hat{V}_t(h_t)] = \mathbb{E}'[\hat{V}_{T+1}] - \mathbb{E}'[\hat{V}_1] = -\mathbb{E}'[\hat{V}_1(h_1)]$, and the claim follows from $V(\pi') = \mathbb{E}'[\sum_t r_t]$. □

Lemma C.2. *Let $\Delta_t(h,a) = \mathbb{E}[\delta_t \mid h_t = h, a_t = a]$, where the conditional expectation is over the true user's reply and the Monte-Carlo noise of the rollouts. Then Δ_t is the same function under π_b and under π', $\mathbb{E}_b[g(h_t,a_t)\delta_t] = \mathbb{E}_b[g(h_t,a_t)\Delta_t(h_t,a_t)]$ for any bounded g, and $\mathbb{E}'[\delta_t] = \mathbb{E}'[\Delta_t(h_t,a_t)]$.*

Proof. Given (h_t,a_t) the reply u_t, hence r_t and h_{t+1}, is drawn from $\pi_U(\cdot \mid h_t,a_t,z)$ regardless of the assistant policy, and the rollout noise is independent of it; the tower property gives both identities. □

C.2 PROOF OF THEOREM 5.4

Write the population version of (4) as $\bar{V} = \mathbb{E}_b[\hat{V}_1(h_1)] + \sum_t \mathbb{E}_b[\hat{w}_t\delta_t]$. Since the law of h_1 is policy-independent, $\mathbb{E}_b[\hat{V}_1(h_1)] = \mathbb{E}'[\hat{V}_1(h_1)]$.

Case (i), $\varepsilon_U = 0$. With exact rollouts, $\hat{Q}_t(h,a) = \mathbb{E}_{\pi_U}[r_t + \hat{V}_{t+1}(h_{t+1}) \mid h_t = h, a_t = a]$ by definition of the rollout value (the continuation follows π' and the true user), so $\Delta_t \equiv 0$. By Theorem C.2, $\mathbb{E}_b[\hat{w}_t\delta_t] = \mathbb{E}_b[\hat{w}_t\Delta_t] = 0$ for every t and every bounded weight function, and $\bar{V} = \mathbb{E}'[\hat{V}_1(h_1)] = \mathbb{E}'[V^{\pi'}(h_1)] = V(\pi')$, because with an exact model $\hat{V}_1 = V_1^{\pi'}$.

Case (ii), $\varepsilon_{w,t} = \varepsilon_{\mathrm{abs},t} = 0$ and $\varepsilon_{\mathrm{cal}} = 0$. Then $\hat{w}_t = \tilde{\omega}_t = \omega_t$. By Theorem C.2 and $\varepsilon_{\mathrm{abs},t} = 0$ applied to $f = \Delta_t$, $\mathbb{E}_b[\omega_t\delta_t] = \mathbb{E}_b[\omega_t\Delta_t] = \mathbb{E}'[\Delta_t] = \mathbb{E}'[\delta_t]$. Therefore $\bar{V} = \mathbb{E}'[\hat{V}_1(h_1)] + \sum_t \mathbb{E}'[\delta_t] = V(\pi')$ by Theorem C.1.

Finite m. Condition on logged histories, fixed user-model parameters and any ratio-fitting draws. If the Q-continuation draws are fresh and conditionally independent of both the logged replies and ratio-fitting draws, their Monte-Carlo errors in $\hat{Q}_t$ and $\hat{V}_t$ have conditional mean zero. Hence the expectations above are unchanged. The two nuisance estimates may still be coupled through the *same fitted user model*; this statement separates that shared-parameter dependence from reused Monte-Carlo draws. The implementation consumes fresh, non-reused random draws for the two rollout stages, although its ratio classifier is fitted on the evaluation logs, so the fixed-nuisance variance theorem does not apply to that run. A per-turn self-normalised variant has a random, data-dependent denominator and is not implied by this argument; it is treated only as an empirical ablation. □

C.3 PROOF OF THEOREM 5.5

By Theorems C.1 and C.2, $V(\pi') = \mathbb{E}'[\hat{V}_1] + \sum_t \mathbb{E}'[\Delta_t]$ and $\bar{V} = \mathbb{E}'[\hat{V}_1] + \sum_t \mathbb{E}_b[\hat{w}_t\Delta_t]$, so

$$\bar{V} - V(\pi') = \sum_t \mathbb{E}_b\big[(\hat{w}_t - \omega_t)\Delta_t\big] + \sum_t \big(\mathbb{E}_b[\omega_t\Delta_t] - \mathbb{E}'[\Delta_t]\big).$$

The second sum is bounded by $\sum_t \varepsilon_{\mathrm{abs},t}\|\Delta_t\|_\infty$ by the definition of $\varepsilon_{\mathrm{abs},t}$ applied to $f = \Delta_t/\|\Delta_t\|_\infty$. For the first, Cauchy–Schwarz gives $|\mathbb{E}_b[(\hat{w}_t - \omega_t)\Delta_t]| \le \|\hat{w}_t - \omega_t\|_2\|\Delta_t\|_2$. By the triangle inequality and Assumption 5.3, $\|\hat{w}_t - \omega_t\|_2 \le \|\hat{w}_t - \tilde{\omega}_t\|_2 + \|\tilde{\omega}_t - \omega_t\|_2 \le \varepsilon_{w,t} + L\,\big\|\,\|q_\phi(\cdot \mid h_t) - p(\cdot \mid h_t)\|_1\big\|_2 \le \varepsilon_{w,t} + L\varepsilon_{\mathrm{cal}}$, which is the first inequality of (5).

For the second, let $F_t(h,a,u) = r_t + \hat{V}_{t+1}(h,a,u)$, which is bounded by $(T-t+1)R_{\max}$ in absolute value because $|r_t| \le R_{\max}$ and $|\hat{V}_{t+1}| \le (T-t)R_{\max}$. With exact rollouts $\hat{Q}_t(h,a) = \mathbb{E}_{u\sim\hat{\pi}_U}F_t(h,a,u)$ and $\Delta_t(h,a) = \mathbb{E}_{u\sim\pi_U}F_t(h,a,u) - \mathbb{E}_{u\sim\hat{\pi}_U}F_t(h,a,u)$, so $|\Delta_t(h,a)| \le 2\sup|F_t| \cdot \mathrm{TV}(\pi_U(\cdot \mid h,a,z), \hat{\pi}_U(\cdot \mid h,a,z)) \le 2(T-t+1)R_{\max}\varepsilon_U$. Hence $\|\Delta_t\|_2 \le \|\Delta_t\|_\infty \le 2(T-t+1)R_{\max}\varepsilon_U$ and $\sum_{t=1}^{T} 2(T-t+1)R_{\max}\varepsilon_U = T(T+1)R_{\max}\varepsilon_U$. □

C.4 Proof of Theorem 5.6

$|\delta_t| \le |r_t| + |\hat{V}_{t+1}| + |\hat{Q}_t| \le R_{\max} + 2G_{\max} \le 3G_{\max}$, so $|\psi_j| \le |\hat{V}_1| + \sum_t \hat{w}_t|\delta_t| \le G_{\max} + 3TW_{\max}G_{\max}$. Condition on nuisance functions fitted independently of the evaluation conversations. The evaluation contributions are then i.i.d. if rollout randomness is independent across conversations. The inequalities $\mathrm{Var}(X+Y) \le 2\mathrm{Var}X + 2\mathrm{Var}Y$ and $\mathrm{Var}(\sum_{t=1}^{T} X_t) \le T\sum_t \mathbb{E}X_t^2$ (Cauchy–Schwarz) give $\mathrm{Var}_b(\psi \mid \mathcal{N}) \le 2\mathrm{Var}_b(\hat{V}_1 \mid \mathcal{N}) + 2T\sum_t \mathbb{E}_b[\hat{w}_t^2\delta_t^2 \mid \mathcal{N}] \le 2\mathrm{Var}_b(\hat{V}_1 \mid \mathcal{N}) + 2TW_{\max}^2\sum_t \mathbb{E}_b[\delta_t^2 \mid \mathcal{N}]$, and $\mathrm{Var}(\hat{V}_{IM-DR} \mid \mathcal{N}) = \mathrm{Var}_b(\psi \mid \mathcal{N})/n$ for the fixed nuisance collection $\mathcal{N}$. Bernstein's inequality for these conditionally i.i.d. variables with $|\psi - \mathbb{E}[\psi \mid \mathcal{N}]| \le M = 2G_{\max}(1+3TW_{\max})$ states that with conditional probability at least $1-\delta$,

$$\left|\tfrac{1}{n}\sum_j \psi_j - \mathbb{E}[\psi \mid \mathcal{N}]\right| \le \sqrt{\tfrac{2\mathrm{Var}_b(\psi|\mathcal{N})\log(2/\delta)}{n}} + \tfrac{4G_{\max}(1+3TW_{\max})\log(2/\delta)}{3n}, \tag{6}$$

which is the concentration statement of Theorem 5.6. For untruncated sequential DR the weight at turn t is $\rho_{1:t} = \prod_{k\le t}\rho_k \le \rho_{\max}^t \le \rho_{\max}^T$, and the same three steps apply with $W_{\max}$ replaced by $\rho_{\max}^T$. Clipping either estimator can force a similar finite bound while adding bias. Because our classifier is fitted on the evaluation logs, conditioning on it does not make the observed contributions independent; this proof does not apply directly to the reported in-sample bootstrap intervals. □

C.5 Diagnostic calibration transfer

Theorem C.3 (Held-out calibration transfer)**.** *Fix q_ϕ and any temperature before examining the calibration sample. Let P_S, P_H be the joint laws of (confidence, correctness) on simulated and human intents, B the number of bins, and $\widehat{\mathrm{ECE}}_S$ empirical ECE on m_S i.i.d. simulated intents independent of posterior and temperature selection. With probability at least $1-\delta$,*

$$\mathrm{ECE}_{P_H}(q_\phi) \le \widehat{\mathrm{ECE}}_S(q_\phi) + 2\,\mathrm{TV}(P_S, P_H) + \sqrt{\tfrac{2(B\ln 2 + \ln(2/\delta))}{m_S}}. \tag{7}$$

The unknown $\mathrm{TV}(P_S, P_H)$ prevents a human-calibration certificate without additional information. Selecting the posterior or temperature on these same m_S examples would require a separate held-out split or uniform bound. This result is a diagnostic inequality, not an OPE estimator guarantee.

Proof. Let $I_1, \ldots, I_B$ be the bins and, for a sign pattern $\sigma \in \{-1,+1\}^B$, $g_\sigma(c,y) = \sum_b \sigma_b(y-c)\mathbb{1}\{c \in I_b\} \in [-1,1]$. Then $\mathrm{ECE}_P(q) = \sum_b |\mathbb{E}_P[(y-c)\mathbb{1}\{c\in I_b\}]| = \max_\sigma \mathbb{E}_P[g_\sigma]$. For any σ and any two laws P, Q on (c,y), $|\mathbb{E}_P g_\sigma - \mathbb{E}_Q g_\sigma| \le (\sup g_\sigma - \inf g_\sigma)\,\mathrm{TV}(P,Q) \le 2\,\mathrm{TV}(P,Q)$. Hence $\mathrm{ECE}_{P_H} = \max_\sigma \mathbb{E}_{P_H} g_\sigma \le \max_\sigma \mathbb{E}_{P_S} g_\sigma + 2\mathrm{TV}(P_S, P_H) = \mathrm{ECE}_{P_S} + 2\mathrm{TV}(P_S, P_H)$. For the estimation term, Hoeffding's inequality for m_S i.i.d. samples of a variable with range 2 gives $P(|\hat{\mathbb{E}} g_\sigma - \mathbb{E}_{P_S} g_\sigma| \ge \epsilon) \le 2\exp(-m_S\epsilon^2/2)$ for each fixed σ; a union bound over the 2^B patterns yields $\max_\sigma |\hat{\mathbb{E}} g_\sigma - \mathbb{E}_{P_S} g_\sigma| \le \epsilon$ with probability at least $1 - 2^{B+1}\exp(-m_S\epsilon^2/2)$, and solving for ϵ at level δ gives $\epsilon = \sqrt{2(B\ln 2 + \ln(2/\delta))/m_S}$. Since $\mathrm{ECE}_{P_S} = \max_\sigma \mathbb{E}_{P_S} g_\sigma \le \max_\sigma \hat{\mathbb{E}} g_\sigma + \epsilon = \widehat{\mathrm{ECE}}_S + \epsilon$, the bound follows. □

C.6 Corollaries

Corollary C.4 (What the sim-to-human gap measures)**.** *With probability at least $1-\delta$,* $\mathrm{ECE}_{P_H} - \widehat{\mathrm{ECE}}_S - \sqrt{2(B\ln 2 + \ln(2/\delta))/m_S} \le 2\,\mathrm{TV}(P_S, P_H)$.

Theorem C.4 rearranges (7). Theorem 5.7 reads off the factors of (5) and the oracle comparison $W_{\max} \leftrightarrow \rho_{\max}^T$ in Theorem 5.6. The residual-size bound contributes ε_U without implying that grounding changes abstraction or posterior error; $\varepsilon_{\mathrm{cal}}$ enters through L, and the horizon enters the conditional variance bound through $G_{\max} = TR_{\max}$ and the factor T. □

D Testbed Details

Table 3 lists the testbed parameters; the implementation is `artifact/imdr/testbed.py`. The assistant's posterior is exact Bayesian filtering under an assumed user model; the rater-side posterior

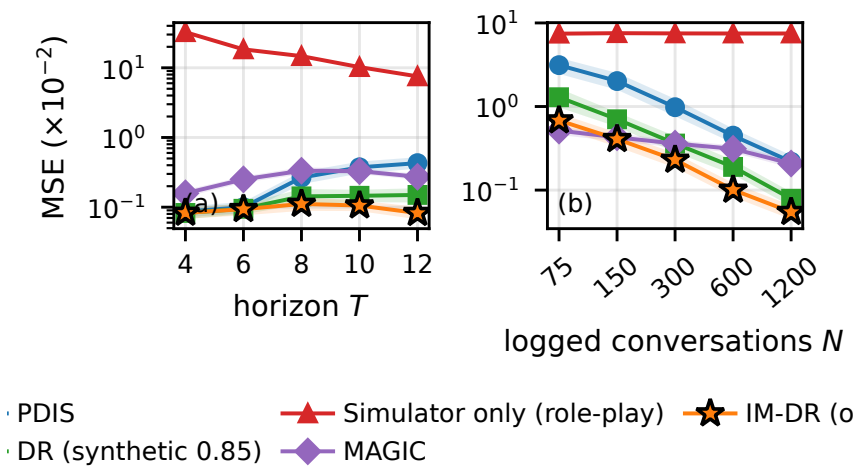

Figure 2: Computed synthetic-testbed MSE ($\times 10^{-2}$, 95% seed-clustered bootstrap CI) vs. (a) horizon T (task A, 600 pooled logs) and (b) pooled logged conversations N (task A, $T = 12$).

that judges whether a question was unnecessary uses the true user parameters. The abstract state binned by the density-ratio classifier is (normalised entropy in 8 bins, MAP probability in 4 bins, turn index, questions asked so far capped at 3) with action interactions, fitted by L_2-regularised logistic regression by Newton's method; the "fine-grained state" baseline uses a histogram ratio on the posterior rounded to 0.05.

Table 3: Parameters of the computed synthetic testbed.

Intent	M slots $\times$ $V = 3$ values; $M = 3$ (A, C), $M = 4$ (B)
User (true)	$\eta \sim U(\eta_0 \pm 0.15)$ with $\eta_0 = 0.25/0.20/0.35$ (A/B/C); accept 0.85 (correct) / 0.15 (wrong); hint noise 0.4/0.4/0.5
Role-play model	$\eta = 0.06$, no heterogeneity, accept 0.96/0.06, hint noise 0.10
Synthetic model levels	linear interpolation role-play $\rightarrow$ true: weak 0.50, moderate 0.85, strong 0.93
Policies	$P(\text{ask}) = \text{clip}(\sigma((H - \theta)/\beta), 0.05, 0.95)$, $(\theta, \beta) \in \{(0.15, 0.25), (0.25, 0.08), (0.33, 0.15), (0.42, 0.10)\}$; forced commit at T
Reward	$w = (1, 1, 0.03, 0.1)$; satisfaction $= \text{clip}(1 - 0.15\,\#\text{rej} - 0.05\,\#\text{ask}, 0, 1)$, -0.3 if unresolved; rating noise SD 0.3
Estimation	$m = 8$ rollouts, clip 20, unnormalised primary estimator, classifier on $4N$ rollouts, 1,000 within-cell bootstrap resamples; seed-clustered MSE intervals

E Additional Results

E.1 Exploratory comparisons and sweeps

The following complete estimator comparison, component ablations and sweeps were used during estimator development. They precede the fresh-seed confirmatory run in Table 1; in particular, the initial full comparison used the self-normalised implementation. We retain them to disclose model-selection evidence rather than treating them as confirmatory tests. Table 4 gives the exploratory full testbed estimator comparison; Table 5 is the full exploratory ablation; and Table 7 breaks its MSE down by testbed family and reports the pooled Kendall τ over all twelve (task, policy) cells. The 50-seed ablation uses common model-rollout draws across variants; an independent 50-seed follow-up with fresh seeds found that removing self-normalisation reduces MSE from 0.1031 to 0.0896 ($\Delta = +0.0135$ [$+0.0057$, $+0.0217$], self-normalised minus unnormalised). This prospective observation motivates the unnormalised revision run and aligns the primary implementation with the

Table 4: Exploratory full estimator comparison using the initial self-normalised IM-DR implementation. Conversation-level OPE on the computed synthetic testbed, leave-one-policy-out. MSE ($\times 10^{-2}$, 95% bootstrap CI), within-task Kendall τ and 95%-CI coverage. The fresh-seed unnormalised confirmatory result is in Table 1.

Estimator	MSE ↓	τ ↑	Cov.
Trajectory IS (conversation-level IPS)	0.27 [0.24, 0.31]	0.76	0.94
Per-decision IS (PDIS)	0.33 [0.30, 0.38]	0.74	0.93
Self-normalised PDIS	0.21 [0.18, 0.23]	0.79	0.98
Direct method (linear FQE)	0.46 [0.42, 0.50]	-0.54	0.00
Simulator only, role-play	12.46 [11.84, 13.04]	**1.00**	0.00
Simulator only, synthetic 0.85	0.94 [0.90, 0.99]	0.95	0.00
Sequential DR, role-play model	0.14 [0.13, 0.16]	0.82	0.94
Sequential DR, synthetic 0.85	0.15 [0.13, 0.17]	0.79	0.94
MAGIC (synthetic 0.85)	0.40 [0.38, 0.43]	0.87	0.26
Fine-grid histogram-ratio MIS	11.37 [11.15, 11.59]	0.73	0.00
IM-IS (ours, $\hat{Q} \equiv 0$)	0.79 [0.73, 0.86]	0.67	0.38
IM-DR (ours)	**0.10 [0.09, 0.12]**	0.86	0.94
Fresh on-policy sample (n=60), reference	0.97 [0.87, 1.09]	0.38	0.95

Table 5: Exploratory full ablation used during estimator development on computed testbed task A: 50 seeds, 4 folds, MSE ($\times 10^{-2} \pm$ 95% half-width), τ and coverage. Small differences among posterior-only, progress-only and the full state are not treated as confirmatory evidence.

Variant	MSE	τ	Cov.
IM-DR (full)	0.08±0.02	0.93	0.95
abstraction: text-space per-decision ratios	0.14±0.03	0.81	0.96
abstraction: posterior only	0.09±0.02	0.91	0.94
abstraction: progress features only	0.09±0.02	0.94	0.95
posterior calibration off (distorted)	0.08±0.01	0.93	0.95
user model: role-play	0.54±0.05	0.91	0.24
synthetic model: weak grounding (0.50)	0.08±0.02	0.89	0.94
synthetic model: strong grounding (0.93)	0.08±0.02	0.89	0.98
no model term ($\hat{Q} \equiv 0$)	0.95±0.12	0.74	0.31
clip 5	0.08±0.02	0.93	0.95
no clipping	0.08±0.02	0.93	0.95
unnormalised estimator (follow-up)	**0.07±0.01**	0.96	0.95
$n = 2$ rollouts	0.11±0.02	0.80	0.96
$n = 32$ rollouts	0.08±0.01	0.91	0.97

estimator analysed in Section 5. Clipping at 5, clipping at 20 and no clipping are identical because no fitted weight reaches the clipping threshold in this testbed.

E.2 Non-confirmatory stress tests

Omitted history. Table 8 introduces a genuine violation of the fitted abstraction: after a rejected commit, the controller's ask logit receives a shift γ, but rejection history is excluded from the density-ratio features. IM-DR remains more accurate than PDIS and sequential DR throughout the tested range, and its ranking metrics remain high. Crucially, its MSE does *not* increase monotonically with γ. The experiment therefore supports robustness only over this narrow perturbation range; it does not empirically validate the abstraction-error term in Theorem 5.5 or show that stronger misspecification is harmless.

Posterior temperature. In Table 9, temperature distortion raises ECE from 0.0281 to 0.1227–0.1568 and posterior L_1 error to 0.2247–0.2955, yet IM-DR MSE stays between 0.105 and 0.115 in units of 10^{-2} and all intervals overlap. Thus this testbed provides no empirical evidence that better calibration improves OPE accuracy. A plausible explanation is that temperature scaling preserves the MAP intent and the coarse binned density-ratio features absorb much of the change. The calibration-transfer result remains a theoretical diagnostic, not an empirical performance claim.

Table 6: Independent paired follow-up on synthetic testbed task A: 50 fresh seeds (offset 100), 4 policy folds, and common random numbers across variants. MSE and paired default-minus-variant MSE difference are scaled by 10^{-2}; intervals are 95% bootstrap CIs.

Variant	MSE ↓	Default − variant	Paired conclusion
IM-DR (default)	0.1031 [0.0842, 0.1229]	–	Reference
No self-normalisation	**0.0896 [0.0728, 0.1072]**	+0.0135 [+0.0057, +0.0217]	Variant better
Clip at 5	0.1031 [0.0842, 0.1229]	+0.0000 [+0.0000, +0.0000]	No observed effect
No clipping	0.1031 [0.0842, 0.1229]	+0.0000 [+0.0000, +0.0000]	No observed effect
Posterior only	0.1077 [0.0880, 0.1295]	-0.0046 [-0.0088, -0.0007]	Default better
Progress features only	0.1054 [0.0875, 0.1246]	-0.0023 [-0.0082, +0.0037]	Inconclusive

Table 7: Computed synthetic-testbed MSE ($\times 10^{-2}$) per task family (A: 3 slots; B: 4 slots; C: terse users), pooled Kendall τ and Spearman ρ over all 12 (task, policy) cells, and per-task τ (A/B/C).

Estimator	Task A	Task B	Task C	pooled τ	pooled ρ	τ A/B/C
PDIS	0.35	0.40	0.25	0.85	0.93	0.81/0.71/0.69
SN-PDIS	0.22	0.22	0.19	0.89	0.96	0.87/0.74/0.77
DM	0.44	0.42	0.52	0.54	0.78	-0.53/-0.56/-0.53
Sim-RP	7.57	7.31	22.49	0.82	0.93	1.00/1.00/1.00
Sim-WC	0.52	0.71	1.60	0.94	0.99	0.99/0.94/0.91
DR-RP	0.14	0.15	0.14	0.91	0.97	0.86/0.79/0.82
DR-WC	0.14	0.16	0.16	0.90	0.96	0.81/0.76/0.79
MAGIC	0.32	0.42	0.47	0.93	0.98	0.94/0.86/0.80
MIS-hist	13.24	11.43	9.43	0.88	0.96	0.88/0.78/0.53
IM-IS	0.95	0.85	0.58	0.85	0.95	0.74/0.76/0.50
IM-DR	0.09	0.10	0.12	0.93	0.98	0.89/0.84/0.85

E.3 Additional fresh-seed checks

Ratio fitting on independent logs. Table 10 compares the original same-data estimator, a matched-size same-data half, an independent train/evaluation split, and two-fold cross-fitting. On the matched evaluation half, independent splitting has slightly *higher* MSE than same-data fitting: split minus same-data-half $+0.0083$ $[+0.0044, +0.0123]$ in 10^{-2} units. The full-sample cross-fit minus same-data difference is -0.0007 $[-0.0015, +0.0000]$; this does not resolve a material gain. Its MSE remains below sequential DR by -0.0812 $[-0.0976, -0.0652]$. The split arm gives independence between the ratio-training set and its evaluation set, but discards half of the data for evaluation. The pooled two-fold arm reuses both halves and is not covered verbatim by Theorem 5.6; neither experiment supplies a refitted-nuisance bootstrap or a general finite-sample coverage guarantee.

Removing privileged logged intent from continuation values. In Table 11, each model continuation draws a slot-wise intent from the stored factorised posterior $q_\phi(z \mid h_t)$ instead of conditioning on true logged z_j. The posterior-sampled IM-DR MSE is 0.0866 versus 0.0876 for the oracle-z version ($\times 10^{-2}$); the paired posterior-minus-oracle difference is -0.0010 $[-0.0093, +0.0072]$, so neither a gain nor equivalence is established. It remains more accurate than sequential DR with the same posterior-sampled continuation (-0.1039 $[-0.1282, -0.0813]$). This variant removes true z from the implemented Q term, but the posterior remains an exact filter in a known synthetic model; the current bias theorem has not been extended to posterior-sampled continuation values. The result is therefore an exploratory feasibility check, not validation on unlabelled natural-language logs.

Policy-shift failure mode behind the main-text MSE gain. Table 12 shows that empirical interval coverage is only 0.567 for constant asking and 0.740 for front-loaded asking. The mean normalised effective sample size is as low as 0.214, despite zero observed weight clipping. These deficits explain why the strong MSE/ranking result in Section 7.2 must not be read as reliable uncertainty quantification. The four targets still use the same synthetic ASK/ACT interface and fixed realizer; this is a policy-family shift, not an end-to-end LLM test.

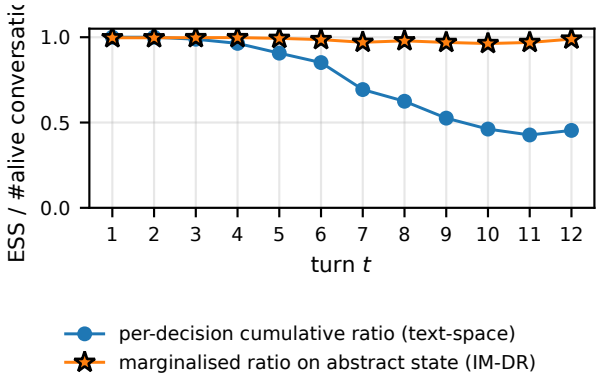

Figure 3: Effective sample size per turn, relative to the number of conversations still alive at that turn, of the cumulative per-decision ratio and of the marginalised abstract-state weight (computed testbed task A, target P_1, 600 pooled logs, mean over seeds).

Table 8: History-dependence stress test (30 fresh seeds). The controller reacts to whether an earlier commit was rejected with strength γ, while this history variable is omitted from the fitted density-ratio state. MSE is in units of 10^{-2}, with 95% seed-clustered bootstrap intervals. The anticipated monotone degradation does not occur.

γ	Estimator	MSE [seed CI]	τ	Top-1	Regret
0.00	IM-DR	0.112 [0.074, 0.155]	0.956	0.900	0.0029
	PDIS	0.453 [0.332, 0.584]	0.644	0.633	0.0173
	DR	0.180 [0.125, 0.247]	0.856	0.767	0.0068
0.05	IM-DR	0.091 [0.061, 0.125]	0.922	0.833	0.0047
	PDIS	0.320 [0.239, 0.412]	0.678	0.600	0.0207
	DR	0.175 [0.125, 0.232]	0.856	0.767	0.0066
0.10	IM-DR	0.068 [0.047, 0.092]	0.956	0.900	0.0034
	PDIS	0.293 [0.217, 0.395]	0.689	0.633	0.0183
	DR	0.148 [0.105, 0.197]	0.856	0.800	0.0084
0.20	IM-DR	0.069 [0.046, 0.099]	0.956	0.900	0.0033
	PDIS	0.324 [0.247, 0.408]	0.689	0.667	0.0162
	DR	0.125 [0.085, 0.175]	0.889	0.833	0.0054

F Scope of Current Evidence

The reported evidence is entirely generated by the synthetic conversation MDP in Section D. No external dialogue dataset, participant log, trained language-model checkpoint, proprietary API output or human annotation enters a reported number. MATH, BigCodeBench, CollabLLM, τ^2-bench, WildChat and UserLM motivate the intended application but are not evaluated here. A future human study would require a separate ethics-approved protocol and cannot be inferred from the testbed results.

G Hyper-parameters and Reproducibility

Posterior: exact Bayesian filtering under the assumed testbed user model; the rater-side posterior uses the true user parameters only to define unnecessary questions. **User models:** controlled interpolations from the role-play model toward the known testbed dynamics, as specified in Table 3; no language model is trained. **Density-ratio classifier:** logistic regression on the binned abstract state, L_2 weight 10^{-2}, fitted on $4N$ rollouts; clip $W_{\max} = 20$. **Rollouts:** $m = 8$ continuations per logged step and act, with an exact two-act probability sum for $\hat{V}_t$. **Interval scope:** the primary 1,000 per-cell bootstrap resamples the contributions after fitting the density-ratio classifier; it does not refit that classifier, cross-fit nuisances, or propagate uncertainty in the 40,000-rollout Monte-Carlo reference. The exploratory cross-fit in Section E.3 is separate and still uses a fixed-fit bootstrap. The primary reference uses fixed random seed 123 per task–policy pair and is reused across experimental seeds. The seed-clustered intervals for aggregate MSE likewise treat these references as fixed. **Reference Monte-Carlo audit:** an audit of the completed runs recomputed all 24 primary and policy-shift reference means

Table 9: Posterior-temperature stress test (30 fresh seeds). Temperature changes calibration substantially but has no resolved effect on IM-DR MSE ($\times 10^{-2}$); the 95% seed-clustered MSE intervals overlap.

Temperature	Posterior L_1	ECE	IM-DR MSE	Seed CI
0.5	0.2910	0.1420	0.113	[0.071, 0.163]
1.0	0.0000	0.0281	0.110	[0.070, 0.161]
2.0	0.2247	0.1227	0.115	[0.073, 0.167]
3.0	0.2768	0.1476	0.114	[0.073, 0.161]
5.0	0.2955	0.1568	0.105	[0.068, 0.150]

Table 10: Ratio-fitting check (50 fresh seeds, 600 task–policy–seed cells). MSE is in 10^{-2} units; 95% MSE intervals cluster-resample the 50 seeds, not the 600 cells independently. "Same-data half" and "sample split" share the evaluation half. Coverage uses the nominal 95% fixed-fit per-cell bootstrap.

Ratio fit	MSE [seed CI]	τ	Coverage
Same-data full	0.0929 [0.0790, 0.1077]	0.893	0.922
Same-data half	0.1701 [0.1528, 0.1883]	0.724	0.940
Independent sample split	0.1784 [0.1598, 0.1978]	0.713	0.948
Two-fold cross-fit	0.0921 [0.0783, 0.1070]	0.889	0.932
PDIS	0.3398 [0.2993, 0.3827]	0.704	0.943
Sequential DR	0.1733 [0.1517, 0.1957]	0.798	0.917

from their 40,000-rollout return vectors. Across the primary task–target cells, the largest reference standard error is 0.00371 reward units (normal-approximation 95% half-width 0.00727), versus primary IM-DR RMSE 0.0278. For the policy-shift cells, the corresponding maximum is 0.00347 (half-width 0.00680), versus IM-DR RMSE 0.0367. Thus these references are estimates, not exact values; their uncertainty is not propagated into the reported MSE or coverage. The supplementary `audit_reference_mc.py` script reruns this audit after generating the primary and policy-shift JSON outputs. **Additional checks:** The supplementary runner performs 50 seeds per task at offsets 400 (ratio fitting), 500 (policy shift) and 600 (posterior-sampled intent). Each uses 200 logs per behaviour policy, 40,000 on-policy reference rollouts per task–target pair, $m = 8$ model continuations per act, $4N$ ratio-classifier target rollouts and 1,000 per-cell fixed-fit bootstrap resamples. The tables in Tables 2 and 10 to 12 are transcribed from the completed run; the corresponding cell-level JSONs are generated by the supplement scripts and are not included in the source-only ZIP. These references use fixed seeds per task–target pair and are not pooled with the primary reference; all comparisons are within a run. **Compute:** the reported 50-seed testbed, sweeps and ablations run on CPU; no GPU experiment is reported. **Artifact.** The source-only supplementary ZIP contains the estimator library, synthetic testbed, tests and a Markdown guide with Bash commands to rerun the reported experiments. It contains no manuscript, cached results, PDF figures or projected human-result files.

Table 11: Posterior-sampled intent in continuation rollouts (50 fresh seeds, 600 cells). The posterior variant does not read the true logged intent for its Q term; all rewards and value references still come from the synthetic simulator. MSE is in 10^{-2} units, with seed-clustered 95% intervals.

Estimator	MSE [seed CI]	τ	Coverage
IM-DR, logged true z	0.0876 [0.0764, 0.0997]	0.876	0.947
IM-DR, posterior-sampled z	0.0866 [0.0754, 0.0981]	0.867	0.940
Sequential DR, posterior-sampled z	0.1905 [0.1665, 0.2172]	0.800	0.947
PDIS	0.3401 [0.2938, 0.3883]	0.727	0.933

Table 12: Policy-family-shift diagnostics pooled over three tasks and 50 seeds per target controller. MSE is in 10^{-2} units; ESS is normalised by the number of logged steps. No fitted IM-DR weight hit the clipping threshold of 20.

Target controller	IM-DR MSE	Coverage	Normalised ESS
Constant asking	0.1946	0.567	0.214
Front-loaded asking	0.1740	0.740	0.330
Back-loaded asking	0.0914	0.793	0.276
Rejection-reactive asking	0.0779	0.940	0.842