A THEORETICAL PERSPECTIVE ON WHY LLMS GET LOST IN MULTI-TURN CONVERSATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work finds that LLM performance degrades when the same task information is distributed across multiple conversational turns rather than presented at once (Laban et al., 2026). We study this phenomenon theoretically by modeling multi-turn conversation as closed-loop inference over a fixed user intent: the model repeatedly updates its interpretation, generates a response, and may reuse that response as evidence later. This feedback loop creates a fundamental asymmetry between the two sources of conversational evidence: user messages provide exogenous information about the true intent, whereas model responses are generated from the model's current interpretation and therefore tend to reinforce it. We show that user evidence alone matches the corresponding single-turn error probability and becomes exponentially reliable as additional information arrives. Although model-generated evidence is Bayes-correct in a reference setting where user and model messages are generated from the same latent intent, in the closed loop it instead reinforces the current estimate, whether correct or incorrect. This competition creates distinct recovery and lock-in regimes: sufficiently strong feedback can make errors persist indefinitely, producing a non-vanishing multi-turn performance gap and coexistence of persistently correct and incorrect conversations. We further characterize how pruning or delaying model-history reuse and increasing the relative influence of user evidence mitigate these effects.

1 INTRODUCTION

Large language models (LLMs) are often used in multi-turn interactions, where users reveal task requirements, constraints, and clarifications gradually rather than in a single prompt. Recent evidence suggests that such interaction can degrade performance even when the underlying task information is unchanged. Specifically, Laban et al. (2026) split complete instructions into smaller pieces, or *shards*, and compare the original instruction presented at once (FULL), the same shards concatenated into one prompt (CONCAT), and shards revealed sequentially across turns with intermediate model responses (SHARDED). They find that SHARDED exhibits an average relative performance drop of 39% from FULL, whereas CONCAT retains 95.1% of FULL performance on average.

To further understand what causes such degradation and how to mitigate it, Liu et al. (2026) study failures to infer the intended task and reconstruct clearer instructions, while Huang et al. (2026) investigate error propagation through model-generated history. Other approaches improve multi-turn behavior through history-cleaned training references (Zheng et al., 2026), distillation across single-turn and multi-turn views (Chen et al., 2026), or rewards for withholding answers when information is insufficient (Li et al., 2026). We provide a broader discussion of related work in Appx. A. These advances, however, remain largely empirical: they document where multi-turn interaction fails and propose interventions that work in practice. We still lack a theoretical account of *why* LLMs get lost in multi-turn conversations: when accumulating user information overcomes the influence of model-generated history, when early errors instead persist, and how these dynamics produce lasting performance gaps and variability across conversations.

To address this gap, we start from the observation that a multi-turn conversation carries two qualitatively different sources of information: *user messages* provide evidence about a fixed underlying intent, whereas *model responses* are generated from the model's current interpretation of it. Reusing those responses as context creates a **closed loop**, in which an early interpretation shapes subsequent

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.

responses and, through them, later inference. Building on statistical views of in-context learning as inference from context (Xie et al., 2022; Zhang et al., 2024), we model this loop as closed-loop inference over a fixed user intent. As shown in Fig. 1, we compare matched single-turn and multi-turn protocols that receive the same user shards and share a common prior, with only the multi-turn protocol additionally incorporating model evidence from its own previous responses.

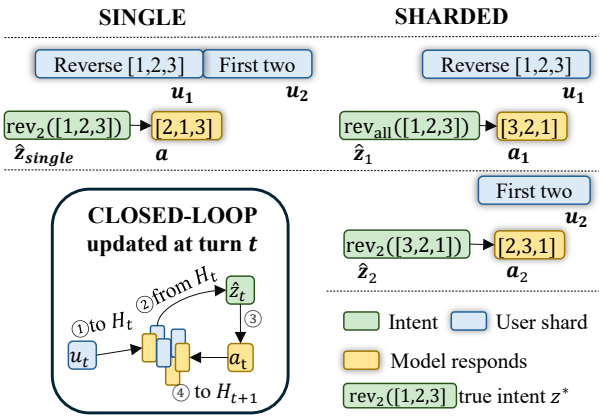


Figure 1: Matched single-turn and closed-loop multi-turn inference. Both protocols observe the same user shards $u_{1:T}$. Multi-turn interaction additionally produces intermediate estimates $\hat{z}_t$ and responses a_t that feed back into later history H. $\text{rev}_k(x)$ reverses the first k elements of x.

Given this framework, we study the roles of the two evidence sources, the dynamics they induce, and when multi-turn inference remains reliable. We first show that inference from user evidence alone matches the corresponding single-turn error probability and becomes exponentially reliable as information accumulates. Model evidence, by contrast, is not inherently invalid: in a self-consistent reference setting, where user messages and model responses are generated from the same latent intent, assigning it full likelihood weight is Bayes-correct. What breaks this guarantee is the closed loop, in which responses are generated from the model's current estimate rather than the true intent, so in expectation they reinforce that estimate whether or not it is correct. The two sources therefore work against each other once the current interpretation is wrong, separating distinct regimes: incorrect interpretations recover when user evidence dominates, yet sufficiently strong feedback can sustain them indefinitely with positive probability. Aggregated over conversations, these regimes produce a non-vanishing gap from matched single-turn inference and lasting variability across conversations. Our main contributions are:

- **Closed-loop framework (Sec. 2).** We model multi-turn conversation as closed-loop inference over a fixed user intent, with a matched single-turn protocol that isolates model-response feedback.
- **Role of model-generated history (Sec. 3).** We identify the conditions under which reusing model-generated response history is statistically valid, and show why the closed loop violates them.
- **Recovery and persistence dynamics (Sec. 4).** We derive sufficient conditions separating recovery from permanent lock-in, and conditions under which a correct interpretation remains stable.
- **Population-level consequences (Sec. 5).** We connect these trajectory-level dynamics to aggregate behavior, accounting for both the persistent multi-turn gap and its variability across conversations.
- **Unified view of mitigations (Sec. 6).** We show how the theoretical results can guide intervention design and explain why methods such as pruning, deferral, recap, and snowballing work.

2 Multi-turn Conversation as Closed Loop Intent Inference

Task, user intent, and sharded disclosure. Consider a user interacting with a model over a multi-turn conversation. The user has a fixed intended task interpretation $z^\star \in \mathbb{Z} := \{z_1, \ldots, z_K\}$, which the model does not observe directly and must infer from the user's messages. Each $z \in \mathbb{Z}$ represents a possible interpretation of the complete task specification, and we refer to $z^\star$ as the **user intent**. We take recovery of $z^\star$ as the criterion for task success.[1]

The model places a prior π over $\mathbb{Z}$, with $\pi(z) > 0$ for all $z \in \mathbb{Z}$. The user reveals information about $z^\star$ gradually through a sequence of **shards** $u_{1:T} = (u_1, \ldots, u_T)$, where $u_t \sim p_{\text{U,t}}(\cdot \mid z^\star)$. A shard may specify an objective, add a constraint or edge case, or clarify an earlier part of the request. We assume *pre-planned* sharded disclosure: conditional on $z^\star$, the user shards are independent and do

[1] We abstract from how the final answer is phrased once the intended task has been recovered. For example, $z^\star$ may represent the intended query semantics in text-to-SQL, the complete specification in code generation, or the relevant constraint set in mathematical reasoning.

not adapt to the model's intermediate responses, i.e., $p_{\mathrm{U_{1:T}}}(u_{1:T} \mid z^\star) = \prod_{t=1}^{T} p_{\mathrm{U,t}}(u_t \mid z^\star)$. Thus, subsequent user messages continue to provide evidence about the same underlying intent even if the model forms an incorrect intermediate interpretation. This assumption isolates model-history feedback from user-side adaptation, allowing us to study the single-turn/multi-turn accuracy gap even under a faithful, non-reactive user; reactive user corrections are left for future work.

Matched single-turn and multi-turn protocols. We compare two protocols for delivering these shards, as illustrated in Fig. 1. In the SINGLE protocol, all shards $u_{1:T}$ are presented together and the model infers the intent once, producing $\hat{z}_{\mathrm{single}}$. In the SHARDED protocol, the same shards arrive sequentially, with a model response generated between consecutive user turns.

At turn t of SHARDED, the available history is $H_t := (U_1, A_1, U_2, A_2, \ldots, U_{t-1}, A_{t-1}, U_t)$. From H_t, the model forms a current estimate $\hat{z}_t$ of the user's intent and then generates response A_t. That response becomes part of H_{t+1} and can therefore influence the next estimate. Repeating this process creates the **closed loop**: an early interpretation shapes the responses generated under it, and those responses subsequently become part of the context from which later interpretations are formed. We regard the task as successfully completed when the terminal estimate satisfies $\hat{z}_T = z^\star$.

Intent inference. We now specify how the model infers the latent intent. Both protocols use the same prior over candidate intents and the same user-provided shards; the distinction is that, under SHARDED, intermediate model responses may also influence subsequent intent inference.

Under SINGLE, no model response is generated before the terminal estimate, so inference uses only the complete sequence of user shards:

$$\hat{z}_{\mathrm{single}} \in \arg\max_{z \in \mathbb{Z}} \left[\log \pi(z) + \textstyle\sum_{t=1}^{T} \log p_{\mathrm{U,t}}(U_t \mid z) \right] \tag{1}$$

Under SHARDED, the model forms a new intent estimate after each user shard arrives, and previous model responses may additionally contribute to inference. We call $\log p_{\mathrm{U,t}}(U_t \mid z)$ *user-evidence* and $\log p_{\mathrm{A}}(A_t \mid z)$ *model-evidence*, where $p_{\mathrm{A}}(\cdot \mid z)$ denotes the response distribution when the model acts under interpretation z. At turn t, the SHARDED estimate is the maximum a posteriori estimate under the power posterior $p_\lambda(z \mid H_t) \propto \pi(z) \prod_{s=1}^{t} p_{\mathrm{U,s}}(U_s \mid z) \prod_{s=1}^{t-1} p_{\mathrm{A}}(A_s \mid z)^\lambda$. Equivalently,

$$\hat{z}_t^\lambda \in \arg\max_{z \in \mathbb{Z}} S_\lambda^{(t)}(z) := \log \pi(z) + \sum_{s=1}^{t} \log p_{\mathrm{U,s}}(U_s \mid z) + \lambda \sum_{s=1}^{t-1} \log p_{\mathrm{A}}(A_s \mid z) \tag{2}$$

Here $\lambda \geq 0$ controls the weight of model-evidence: at $\lambda = 0$, previous model responses remain in the history but do not affect the intent score; at $\lambda = 1$, they enter with full likelihood weight.

Randomness and initial ambiguity. Randomness enters both protocols through the realized user shards and, when needed, tie-breaking. When multiple candidate intents attain the same maximum score, the model breaks the tie uniformly at random. Under SHARDED, let $\{\xi_t\}_{t \geq 1}$ denote i.i.d. tie-breaking variables, independent of the user- and model-sampling randomness, with ξ_t selecting uniformly among the maximizing intents at turn t. Under SINGLE, the terminal estimate uses an independent tie-breaking variable ξ_T^{single} with the same rule. These variables contain no task information and affect the estimates only when the maximizer is non-unique.

Under SHARDED, intermediate model responses introduce an *additional* source of randomness. After the turn-t estimate $\hat{z}_t^\lambda$ has been formed, let $\mathcal{F}_t := \sigma(H_t, \xi_1, \ldots, \xi_t)$ denote the information generated jointly by the conversation history and the tie-breaking variables up to turn t. For $t < T$, the model generates its response according to $A_t \mid \mathcal{F}_t \sim p_{\mathrm{A}}(\cdot \mid \hat{z}_t^\lambda)$. The realized response then enters the subsequent conversation history and, when $\lambda > 0$, contributes model-evidence to later intent scores. Consequently, under SHARDED, an early random choice of intent can alter the distribution of subsequent responses and thereby the later inference trajectory. This path dependence becomes consequential when the user information is ambiguous at an early turn. We capture this possibility through the following assumption.

Assumption 2.1 (Initial ambiguity)**.** There exists a set $\mathbb{Z}_1 \subseteq \mathbb{Z}$ such that $z^\star \in \mathbb{Z}_1$, $|\mathbb{Z}_1| = m \geq 2$, and the prior π is constant on $\mathbb{Z}_1$. For every $u \in \mathrm{supp}\big(p_{\mathrm{U,1}}(\cdot \mid z^\star)\big)$, the set of first-turn maximizers is exactly $\mathbb{Z}_1$, i.e., $\arg\max_{z \in \mathbb{Z}} \{\log \pi(z) + \log p_{\mathrm{U,1}}(u \mid z)\} = \mathbb{Z}_1$.

Assump. 2.1 implies that the first shard leaves m candidate intents, including $z^\star$, tied for the highest score. Hence, under uniform tie-breaking, $\mathbb{P}(\hat{z}_1 = z^\star) = \frac{1}{m}$ and $\mathbb{P}(\hat{z}_1 \neq z^\star) = 1 - \frac{1}{m}$. The resulting

tie-break acts as a random seed: under SHARDED, an incorrect initial choice changes the distribution of A_1, which then enters the conversation history and can influence subsequent estimates. The closed-loop dynamics determine whether this deviation is later corrected or reinforced.

Research objective. Laban et al. (2026) show that LLM performance can degrade substantially when task information is revealed across multiple turns. Using the framework developed above, we study theoretically how such degradation can persist even as the user continues to provide informative evidence about a fixed intent. Specifically, we investigate the following questions:

1. *What is the statistical role of model-generated history?* Is user-evidence alone ($\lambda = 0$) sufficient to recover $z^\star$? When is $\lambda = 1$ statistically justified, and how does its role change when A_t is generated from the current estimate $\hat{z}_t^\lambda$ rather than from $z^\star$? (Sec. 3).
2. *When are early errors corrected or reinforced?* Starting from an incorrect estimate at some turn t_0, when is the recovery time $t_{\mathrm{rec}} := \inf\{t > t_0 : \hat{z}_t^\lambda = z^\star\}$ finite, and when can $t_{\mathrm{rec}} = \infty$? If the model currently favors $z^\star$, how likely is it to later leave the correct state? (Sec. 4)
3. *When does trajectory-level feedback produce persistent multi-turn degradation?* Under what conditions does the terminal error gap $\mathbb{P}(\hat{z}_T^\lambda \neq z^\star) - \mathbb{P}(\hat{z}_{\mathrm{single}} \neq z^\star)$ remain non-vanishing as T grows? When can persistently correct and incorrect trajectories coexist, yielding non-vanishing terminal variability $\mathrm{Var}(\mathbf{1}\{\hat{z}_T^\lambda \neq z^\star\})$? (Sec. 5)
4. *Why do existing mitigation methods work?* How do methods such as pruning model evidence, deferring its reuse, recap, and snowballing alter the feedback dynamics, and under what conditions can they reduce or eliminate persistent degradation? (Sec. 6)

Technical assumptions. We now state the assumptions used throughout the analysis. For results involving $T \to \infty$, we define the stochastic process $\{U_t, A_t, \xi_t\}_{t\geq 1}$ on a common probability space, with each finite-horizon conversation given by its restriction to the first T turns. The effects of relaxing these assumptions on the theoretical results are discussed in Appx. B.

Assumption 2.2 (User informativeness after the first turn)**.** From the second turn onward, the user-shard distribution under the true intent remains distinguishable from those under competing intents. We characterize this informativeness by its weakest and strongest levels:

$$\underline{\kappa} := \inf_{t\geq 2} \min_{z\neq z^\star} \mathrm{KL}\left(p_{\mathrm{U,t}}(\cdot \mid z^\star) \,\|\, p_{\mathrm{U,t}}(\cdot \mid z)\right) > 0, \quad \overline{\kappa} := \sup_{t\geq 2} \max_{z\neq z^\star} \mathrm{KL}\left(p_{\mathrm{U,t}}(\cdot \mid z^\star) \,\|\, p_{\mathrm{U,t}}(\cdot \mid z)\right) < \infty.$$

Assumption 2.3 (Model-response separation)**.** Model response distributions are intent-separated:

$$c := \min_{z\in\mathbb{Z}} \min_{z'\neq z} \mathrm{KL}\left(p_{\mathrm{A}}(\cdot \mid z) \,\|\, p_{\mathrm{A}}(\cdot \mid z')\right) > 0.$$

Assumption 2.4 (Bounded log-likelihood ratios)**.** For every turn t, the family $\{p_{\mathrm{U,t}}(\cdot \mid z) : z \in \mathbb{Z}\}$ admits densities with respect to a common dominating measure, and likewise, $\{p_{\mathrm{A}}(\cdot \mid z) : z \in \mathbb{Z}\}$ admits densities with respect to a common dominating measure. There exists $M < \infty$ such that, for every $z, z' \in \mathbb{Z}$ and $t \geq 1$, $\forall u, a$,

$$\left|\log \frac{p_{\mathrm{U,t}}(u \mid z)}{p_{\mathrm{U,t}}(u \mid z')}\right| \leq M, \quad \left|\log \frac{p_{\mathrm{A}}(a \mid z)}{p_{\mathrm{A}}(a \mid z')}\right| \leq M.$$

Assump. 2.4 limits the amount by which any single user shard or model response can change a pairwise intent score. In particular, $\overline{\kappa} \leq M$.

3 THE STATISTICAL ROLE OF MODEL-GENERATED HISTORY

Next, we examine the role of model-generated history. Both protocols receive the same user shards, but the SHARDED estimator additionally reuses previous model responses as evidence. Comparing Eq. (1) and Eq. (2) at $t = T$, the additional term is $\lambda \sum_{s=1}^{T-1} \log p_{\mathrm{A}}(A_s \mid z)$. This raises the central question: *when is model-evidence informative about the true intent, and when can reusing it distort inference?* We study this in three steps: (i) we show that user-evidence alone is sufficient to recover the true intent; (ii) we identify a self-consistent reference setting in which user shards and model responses are generated from the same fixed latent intent, and show that full model-evidence weight, $\lambda = 1$, is Bayes-correct; and (iii) we show what changes in the closed loop, where model responses are generated from the model's current estimate rather than the true intent.

User-evidence alone is sufficient for recovery. We first consider $\lambda = 0$, so model responses do not contribute to the intent score. At the terminal turn, the SHARDED and SINGLE estimators then have the same score for every realization of the user shards. Under the uniform tie-breaking rule defined in Sec. 2, they therefore induce the same terminal distribution over intent estimates.

Theorem 3.1 (Exponential consistency of user-only inference)**.** *Under Assumps. 2.1, 2.2, and 2.4, and the user-disclosure process defined in Sec. 2, when $\lambda = 0$, for every $T \geq 2$,*

$$\mathbb{P}(\hat{z}_T^0 \neq z^\star) = \mathbb{P}(\hat{z}_{\text{single}} \neq z^\star) \leq (K-1)\exp\left(-\frac{(T-1)\underline{\kappa}^2}{2M^2}\right). \tag{3}$$

The proof is in Appx. C.1. Thm. 3.1 shows that user-evidence alone is sufficient for recovery: when model-evidence is excluded, the terminal error probability decreases exponentially with T.

When is model-evidence statistically valid? We now ask when model-evidence is correctly specified for inference about the latent intent. To establish a reference case, we introduce a *self-consistent reference model* in which user shards and model responses are generated directly from the latent intent. Let $Z \sim \pi$, and conditional on $Z = z$, the likelihood of $H_t = h_t$ under this reference model is $p_{\text{sc}}(h_t \mid z) = \prod_{s=1}^{t} p_{\text{U},s}(u_s \mid z) \prod_{s=1}^{t-1} p_{\text{A}}(a_s \mid z)$. Let $\mathbb{P}_{\text{sc}}$ denote the resulting joint law of (Z, H_t). The following proposition shows that under this reference model, assigning full weight to model-evidence, $\lambda = 1$, exactly recovers the Bayesian posterior over intent.

Proposition 3.1 (Posterior identity under the self-consistent reference)**.** *Under $\mathbb{P}_{\text{sc}}$, for every $z \in \mathbb{Z}$, $\mathbb{P}_{\text{sc}}(Z = z \mid H_t) = p_1(z \mid H_t)$. Hence, the Bayesian maximum a posterior (MAP) argmax set coincides with that of Eq. (2) at $\lambda = 1$.*

The proof is in Appx. C.2. The posterior identity establishes that $\lambda = 1$ is Bayes-correct under the reference model. We next ask whether it is also optimal within the family $\{p_\lambda : \lambda \geq 0\}$. For $t \geq 2$, define $R_t(\lambda) := \mathbb{E}_{\text{sc}}\left[\text{KL}\left(p_1(\cdot \mid H_t) \,\|\, p_\lambda(\cdot \mid H_t)\right)\right]$, where $\mathbb{E}_{\text{sc}}$ denotes expectation under $\mathbb{P}_{\text{sc}}$.

Proposition 3.2 (Strict posterior optimality of $\lambda = 1$)**.** *Under Assumps. 2.3-2.4, for any $t \geq 2$, $R_t(\lambda)$ is uniquely minimized at $\lambda = 1$, with $R_t(1) = 0$. Moreover, in a neighborhood of $\lambda = 1$,*

$$R_t(\lambda) = \frac{(\lambda-1)^2}{2}\mathcal{I}_t + o\big((\lambda-1)^2\big), \text{ where } \mathcal{I}_t := \mathbb{E}_{\text{sc}}\left[\text{Var}_{z\sim p_1(\cdot\mid H_t)}\Big(\sum_{s=1}^{t-1}\log p_{\text{A}}(A_s \mid z)\Big)\right] > 0.$$

The proof is given in Appx. C.3. Props. 3.1 and 3.2 therefore establish the benchmark: when model responses are generated from the latent intent, $\lambda = 1$ recovers the Bayesian posterior and uniquely minimizes posterior discrepancy.

Closed-loop model-evidence reinforces the current estimate. We now return to the actual multi-turn process. Under the reference model, $A_t \mid Z \sim p_{\text{A}}(\cdot \mid Z)$, whereas in the closed loop, $A_t \mid \mathcal{F}_t \sim p_{\text{A}}(\cdot \mid \hat{z}_t^\lambda)$. Thus, the response is generated from the model's current estimate, while subsequent intent scores still evaluate that response through $\log p_{\text{A}}(A_t \mid z)$.

Lem. C.3 in Appx. C.4 shows that if the current estimate $\hat{z}_t^\lambda \neq z^\star$, then $\mathbb{E}\left[\log \frac{p_{\text{A}}(A_t \mid z^\star)}{p_{\text{A}}(A_t \mid \hat{z}_t^\lambda)} \,\middle|\, \mathcal{F}_t\right] = -\,\text{KL}\left(p_{\text{A}}(\cdot \mid \hat{z}_t^\lambda) \,\middle\|\, p_{\text{A}}(\cdot \mid z^\star)\right) \leq -c$. Hence, a response generated from an incorrect estimate favors that estimate over the truth in expectation. When $\hat{z}_t^\lambda = z^\star$, the direction reverses: model-evidence favors the truth over every competing intent in expectation. Fresh user evidence therefore favors the fixed true intent, whereas model-evidence favors the current estimate. Sec. 4 characterizes when this competition leads to recovery and when an incorrect estimate persists.

4 CLOSED-LOOP STATE–DRIFT DYNAMICS

Sec. 3 establishes two opposing effects in the closed loop: fresh user evidence systematically favors the true intent, whereas model-evidence reinforces the intent that currently generates the model response. We now study how these effects shape the intent scores along a realized conversation trajectory, and determine when an incorrect estimate is corrected or persists. Throughout this section, the assumptions in Sec. 2 hold over an infinite horizon.

Pairwise margins. For every wrong intent $z \neq z^\star$, define the *score margin* of the truth against z by $\Delta_t(z) := S_\lambda^{(t)}(z^\star) - S_\lambda^{(t)}(z)$, and define the *leader gap* $G_t := \min_{z \neq z^\star} \Delta_t(z)$. Thus, $G_t > 0$

means that $z^\star$ strictly outperforms every competitor, whereas $G_t < 0$ means that at least one wrong intent has a larger score than the truth. For each $z \neq z^\star$, the pairwise margin evolves according to

$$\Delta_{t+1}(z) = \Delta_t(z) + \log \frac{p_{\mathrm{U,t+1}}(U_{t+1} \mid z^\star)}{p_{\mathrm{U,t+1}}(U_{t+1} \mid z)} + \lambda \log \frac{p_{\mathrm{A}}(A_t \mid z^\star)}{p_{\mathrm{A}}(A_t \mid z)}. \tag{4}$$

The two terms added in Eq. (4) correspond to fresh user evidence and model-evidence. By Assump. 2.2, the user term favors $z^\star$ in expectation. By Lem. C.3, the expected model term depends on the current estimate: it favors the truth when the current estimate is correct, but can oppose the truth when the current estimate is wrong. The results below use this margin evolution to study three cases: whether an incorrect trajectory can remain incorrect, when all pairwise margins are instead driven toward the truth, and whether a correct trajectory remains correct.

Wrong-state persistence. We first consider trajectories on which a wrong intent is ahead of the truth. To obtain a uniform bound on wrong-state reinforcement, define $\alpha_{\mathrm{err}} := \min_{z \neq z^\star} \mathrm{KL}\left(p_{\mathrm{A}}(\cdot \mid z) \,\|\, p_{\mathrm{A}}(\cdot \mid z^\star)\right)$. Whenever the current estimate is incorrect, the expected model contribution to the margin against that estimate is therefore at most $-\lambda\alpha_{\mathrm{err}}$. By Assump. 2.3, $\alpha_{\mathrm{err}} \geq c$. The next result asks whether a trajectory with $G_{t_0} < 0$ must eventually return to the truth, or whether model-evidence can sustain the initial score deficit.

Theorem 4.1 (Wrong-state lock-in). *Fix any turn $t_0 \geq 2$ and condition on a realization of $\mathcal{F}_{t_0}$ such that $G_{t_0} = -\delta < 0$. Let $t_{\mathrm{rec}} := \inf\{t > t_0 : \hat{z}_t^\lambda = z^\star\}$ be the first subsequent recovery time. If $\lambda\alpha_{\mathrm{err}} > \overline{\kappa}$, then*

$$\mathbb{P}\left(t_{\mathrm{rec}} = \infty \,|\, \mathcal{F}_{t_0}\right) \geq 1 - \exp\left(-\frac{2(\lambda\alpha_{\mathrm{err}} - \overline{\kappa})\delta}{M^2(1+\lambda^2)}\right).$$

Moreover, with probability one on $\{t_{\mathrm{rec}} = \infty\}$, $G_t \longrightarrow -\infty$.

The proof is in Appx. D.1. The condition $\lambda\alpha_{\mathrm{err}} > \overline{\kappa}$ compares the minimum expected reinforcement of an incorrect current estimate with the maximum expected correction supplied by user evidence. When the former dominates, a negative leader gap has positive probability of never returning to the correct side of the boundary. Moreover, on trajectories that never recover, the truth falls progressively farther behind. The initial deficit δ determines how far the trajectory already lies on the wrong side of this boundary, with larger δ yielding a stronger lower bound on permanent non-recovery.

Thm. 4.1 controls the margin between the truth and the wrong intent currently generating the response. When $K > 2$, however, becoming correct requires the truth to overcome every competing intent. We therefore next control the model contribution to all pairwise margins.

Uniformly corrective drift. Define

$$\beta := \max_{w, z \neq z^\star} \left[\mathrm{KL}\left(p_{\mathrm{A}}(\cdot \mid w) \,\|\, p_{\mathrm{A}}(\cdot \mid z^\star)\right) - \mathrm{KL}\left(p_{\mathrm{A}}(\cdot \mid w) \,\|\, p_{\mathrm{A}}(\cdot \mid z)\right)\right],$$

where $w = z$ is allowed. If the current estimate is $w \neq z^\star$, then for every competitor $z \neq z^\star$, $\mathbb{E}_{A \sim p_{\mathrm{A}}(\cdot|w)}\left[\log \frac{p_{\mathrm{A}}(A|z^\star)}{p_{\mathrm{A}}(A|z)}\right] \geq -\beta$. Thus, β upper-bounds the largest expected disadvantage that model-evidence can create for the truth across all incorrect current estimates and all competitors.

Fresh user evidence contributes at least $\underline{\kappa}$ in expectation to every pairwise margin. Hence, if $\lambda\beta < \underline{\kappa}$, the total expected change in every pairwise margin is positive even when the current estimate is wrong. When the current estimate is correct, the model term also favors the truth, so the same conclusion continues to hold. Thus, under this condition, every pairwise margin has positive expected drift toward the truth regardless of the current estimate.

Theorem 4.2 (Recovery under uniformly corrective drift). *Suppose $\lambda\beta < \underline{\kappa}$. Then, for every $t \geq 2$,*

$$\mathbb{P}(\hat{z}_t^\lambda \neq z^\star) \leq (K-1)\exp\left(-\frac{(\underline{\kappa} - \lambda\beta)^2(t-1)}{8(1+\lambda)^2M^2}\right).$$

And the same upper bound holds for $\mathbb{P}(t_{\mathrm{rec}} > t)$ where $t_{\mathrm{rec}} := \inf\{t \geq 2 : \hat{z}_t^\lambda = z^\star\}$. Consequently, $\mathbb{P}(t_{\mathrm{rec}} < \infty) = 1$ and $\mathbb{E}[t_{\mathrm{rec}}] < \infty$.

See proof in Appx. D.2. Under $\lambda\beta < \underline{\kappa}$, fresh user evidence dominates the largest expected disadvantage that model-evidence can create for the truth. The probability that some competitor continues to outrank $z^\star$ thus decreases, and an incorrect trajectory reaches $z^\star$ almost surely in finite time.

Correct-state persistence. When the current estimate is $z^\star$, model-evidence and fresh user evidence both favor the truth in expectation. Define $\alpha_{\text{true}} := \min_{z \neq z^\star} \mathrm{KL}\left(p_{\mathrm{A}}(\cdot \mid z^\star) \,\|\, p_{\mathrm{A}}(\cdot \mid z)\right)$. By Assump. 2.3, $\alpha_{\text{true}} \geq c$. While the current estimate remains correct, every pairwise margin therefore has conditional expected drift at least $\underline{\kappa} + \lambda\alpha_{\text{true}} > 0$. Starting from $G_{t_0} > 0$, becoming incorrect requires at least one positive pairwise margin to fall to zero or below.

Theorem 4.3 (Correct-state persistence). *Fix any turn t_0 such that $\hat{z}_{t_0}^\lambda = z^\star$ and $G_{t_0} > 0$, define $t_{\text{exit}} := \inf\{t > t_0 : \hat{z}_t^\lambda \neq z^\star\}$ as the first subsequent exit from the correct state. Then*

$$\mathbb{P}\left(t_{\text{exit}} < \infty \,|\, \mathcal{F}_{t_0}\right) \leq (K-1)\exp\left(-\frac{2(\underline{\kappa} + \lambda\alpha_{\text{true}})G_{t_0}}{M^2(1+\lambda^2)}\right). \tag{5}$$

See proof in Appx. D.3. Unlike the wrong-state case, the two evidence sources point in the same direction while the estimate remains correct, so no additional comparison between their strengths is required. Whenever the right-hand side of Eq. (5) is strictly below one, $\mathbb{P}\left(t_{\text{exit}} = \infty | \mathcal{F}_{t_0}\right) > 0$, so the process has positive probability of remaining correct at every subsequent turn.

Binary case. When $\mathbb{Z} = \{z^\star, z'\}$, there is only one wrong intent, so $\beta = \alpha_{\text{err}}$. If, in addition, the user divergence is constant across turns $t \geq 2$, so that $\underline{\kappa} = \overline{\kappa} = \kappa$, the recovery and lock-in conditions become $\lambda\alpha_{\text{err}} < \kappa$ and $\lambda\alpha_{\text{err}} > \kappa$, respectively. Thus, in the binary case, the two regimes meet at the drift threshold $\lambda\alpha_{\text{err}} = \kappa$. Below this threshold, user evidence dominates wrong-state model reinforcement; above it, model reinforcement can sustain an incorrect trajectory. The equality case is not characterized by the present results; see Appx. D.4.

5 Persistent Multi-Turn Degradation and Unreliability

Sec. 4 characterizes how individual conversation trajectories evolve, including conditions for persistent error, recovery, and persistence of a correct estimate. We now aggregate these trajectory-level results over the randomness of the closed-loop interaction to study performance across conversations. We first show that a positive mass of permanently incorrect trajectories yields a non-vanishing terminal error gap between Sharded and Single. We then show that persistently incorrect and persistently correct trajectories can coexist with positive probability, causing terminal correctness to remain variable across conversations.

Persistent multi-turn degradation. We first quantify the population mass of trajectories that remain persistently incorrect. Suppose $\lambda\alpha_{\text{err}} > \overline{\kappa}$. Define

$$p_{\text{lock}}(\lambda) = \mathbb{E}\left[\left(1 - \exp\left\{-\frac{2(\lambda\alpha_{\text{err}} - \overline{\kappa})(-G_2)}{M^2(1+\lambda^2)}\right\}\right)\mathbf{1}\{G_2 < 0\} \,\middle|\, \hat{z}_1 \neq z^\star\right], \tag{6}$$

which averages the conditional non-recovery bound from Thm. 4.1 over incorrect first-turn estimates. Under $\lambda\alpha_{\text{err}} > \overline{\kappa}$, $\mathbb{P}\big(G_2 < 0 \,\big|\, \hat{z}_1^\lambda \neq z^\star\big) > 0$ and hence $p_{\text{lock}}(\lambda) > 0$; see Appx. E.1.

Theorem 5.1 (Persistent degradation floor). *If $\lambda\alpha_{\text{err}} > \overline{\kappa}$, then $p_{\text{lock}}(\lambda) > 0$, and for every $T \geq 2$,*

$$\mathbb{P}\big(\hat{z}_T^\lambda \neq z^\star\big) - \mathbb{P}\big(\hat{z}^{\text{single}} \neq z^\star\big) \geq \left(1 - \frac{1}{m}\right)p_{\text{lock}}(\lambda) - (K-1)\exp\left(-\frac{(T-1)\underline{\kappa}^2}{2M^2}\right). \tag{7}$$

Consequently, $\liminf_{T\to\infty}\left(\mathbb{P}\big(\hat{z}_T^\lambda \neq z^\star\big) - \mathbb{P}\big(\hat{z}^{\text{single}} \neq z^\star\big)\right) \geq \left(1 - \frac{1}{m}\right)p_{\text{lock}}(\lambda) > 0.$

See proof in Appx. E.1. The factor $1 - \frac{1}{m}$ is the probability of an incorrect first-turn estimate, while $p_{\text{lock}}(\lambda)$ lower-bounds the conditional probability that such a trajectory never recovers. Their product therefore gives a non-vanishing contribution to the Sharded terminal error probability. By contrast, Thm. 3.1 shows that the error probability of the matched Single estimator vanishes exponentially with T. Hence, the terminal error gap remains positive at long horizons.

Persistent unreliability. Thm. 5.1 establishes a positive mass of trajectories that remain incorrect. We next ask whether a positive mass can simultaneously remain correct. If persistently correct and incorrect trajectories coexist, the terminal error probability remains bounded away from both zero and one, yielding non-vanishing variability in terminal correctness.

Theorem 5.2 (Persistent unreliability). *Suppose* $\lambda\alpha_{\text{err}} > \overline{\kappa}$ *and there exists* $\gamma > 0$ *such that*

$$r_\gamma := \mathbb{P}\left(G_2 \geq \gamma \,\middle|\, \hat{z}_1^\lambda = z^\star\right) > 0, \quad \text{and } (K-1)\exp\left(-\frac{2(\underline{\kappa}+\lambda\alpha_{\text{true}})\gamma}{M^2(1+\lambda^2)}\right) < 1. \tag{8}$$

Define $q_L := \left(1-\frac{1}{m}\right)p_{\text{lock}}(\lambda)$, $q_U := 1-\frac{r_\gamma}{m}\left[1-(K-1)\exp\left(-\frac{2(\underline{\kappa}+\lambda\alpha_{\text{true}})\gamma}{M^2(1+\lambda^2)}\right)\right]$. *Then,* $\forall T \geq 2$, *we have* $q_L \leq \mathbb{P}\left(\hat{z}_T^\lambda \neq z^\star\right) \leq q_U$. *Moreover,* $0 < q_L \leq q_U < 1$, *and therefore*

$$\inf_{T\geq 2} \operatorname{Var}\left(\mathbf{1}\{\hat{z}_T^\lambda \neq z^\star\}\right) \geq \min\{q_L(1-q_L), q_U(1-q_U)\} > 0.$$

See proof in Appx. E.2. Condition $r_\gamma > 0$ gives a positive mass of trajectories that, after a correct first-turn estimate, reach turn 2 with the truth ahead of every competitor by at least γ. The second condition, together with Thm. 4.3, gives these trajectories a positive probability of remaining correct. Thm. 5.1 simultaneously provides a positive mass that remains incorrect. Hence the terminal error probability stays uniformly bounded away from both zero and one, yielding non-vanishing variance.

6 Implications for Mitigating Multi-Turn Degradation

Our theory provides a common lens for understanding why existing interventions can mitigate multi-turn degradation. We analyze four representative strategies: pruning ($\lambda = 0$), deferred commitment, recap, and snowballing. Each preserves the generation of new user shards while changing how observed content enters the intent score, with model responses generated from the resulting estimate.

Pruning model-evidence. By Thm. 3.1, pruning model-generated history by setting $\lambda = 0$ reduces multi-turn inference to user evidence alone and exactly recovers the matched single-turn error probability: $\mathbb{P}(\hat{z}_T^{\text{prune}} \neq z^\star) = \mathbb{P}(\hat{z}_{\text{single}} \neq z^\star)$. This provides a theoretical explanation for history-removal strategies used in prior work. In the user-turn-only configuration of Huang et al. (2026), previous model responses are replaced with placeholders while user messages are retained. Similarly, MAIGO (Zheng et al., 2026) excludes earlier model replies from middle-turn reference contexts while preserving the student's multi-turn history. Although these methods differ in where history is removed, either from the inference context or from the training reference, their common effect is to prevent model-generated responses from serving as feedback evidence.

Deferring commitment. Rather than removing model-generated history throughout the conversation, deferral delays when it can influence subsequent inference. For an integer $r \geq 1$, the first r model responses are excluded from subsequent intent scores, while later responses contribute with weight λ. The estimate at turn $r+1$ therefore relies only on user evidence, and A_{r+1} first affects inference at turn $r+2$. By Thm. F.1 (Appx. F.1), for every fixed $\lambda \geq 0$ and $T > r+1$, $\mathbb{P}(\hat{z}_T^{\text{def}} \neq z^\star) \leq 2(K-1)e^{-c_{\text{def}} r}$, where $c_{\text{def}} > 0$ is independent of r and T. Thus, terminal error decreases exponentially with the deferral length: user evidence has time to establish a favorable state before model feedback becomes active, without removing feedback altogether.

This provides a mechanism-level interpretation of the association between later first answer attempts and higher performance reported by Laban et al. (2026). Related timing principles appear in the deferral and clarification training of Chen et al. (2026) and in the accuracy and abstention rewards of Li et al. (2026). Our analysis isolates the common feedback mechanism: delaying commitment prevents early model responses from reinforcing an under-informed interpretation.

A single recap. Unlike pruning and deferral, a recap retains model-generated history while changing the current balance of evidence. At turn $t_0 \geq 2$, the observed user shards are repeated once, which shifts every pairwise margin by an accumulated user-evidence term and produces a post-recap leader gap $G_{t_0}^+ := \min_{z\neq z^\star}[\Delta_{t_0}(z) + \sum_{s=1}^{t_0} \log \frac{p_{\text{U},s}(U_s|z^\star)}{p_{\text{U},s}(U_s|z)}]$, leaving the subsequent update rule unchanged. By Cor. F.1 (Appx. F.2), the correct-state exit bound applies when $G_{t_0}^+ > 0$ and the non-recovery bound when $G_{t_0}^+ < 0$ and $\lambda\alpha_{\text{err}} > \overline{\kappa}$: a recap changes the leader gap, and may change its sign, but not the drift regime that governs what follows. This provides a theoretical interpretation of the Recap strategy in Laban et al. (2026), where repeating prior user messages improves Sharded performance but does not fully restore single-turn performance.

Snowballing. Snowballing instead changes the long-run balance between user and model evidence: rather than applying a recap once, it reintroduces the accumulated recap at every turn. Thus, shard

U_s appears in turn $s, \ldots, T$ and receives terminal weight $T - s + 1$, so $U_2, \ldots, U_T$ collectively receive weight $T(T-1)/2$, compared with only $T - 1$ model-response terms. Under our assumptions, the expected user contribution toward the truth grows quadratically, eventually dominating fixed-weight model feedback. By Thm. F.2 (Appx. F.3), for every fixed $\lambda \geq 0$, the terminal error decreases exponentially in T for sufficiently large T. Repetition thus changes the weights assigned to existing observations, not the number of independent samples. This provides a natural interpretation of the SNOWBALL protocol of Laban et al. (2026): unlike a one-time recap, repeated user information changes the relative accumulation rates of user and model evidence, eventually overwhelming fixed-weight model feedback and ruling out permanent lock-in for every fixed λ. This guarantee is asymptotic rather than a finite-horizon ranking. Indeed, Laban et al. (2026) report stronger performance for RECAP over the conversation lengths they evaluate. This does not conflict with our result, which guarantees eventual dominance of user evidence under snowballing but does not imply that this dominance is already realized, or that snowballing must outperform recap, at finite horizons.

7 EXPERIMENTS

While prior work has already established the empirical phenomenon of multi-turn degradation, we conduct complementary experiments to test the mechanism predicted by our theory. **Synthetic experiments in Appx. G** test our theoretical predictions within the formal model.

We evaluate two models, Qwen2.5-7B-Instruct (Yang et al., 2024) and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), to test two predictions: model responses favor the interpretation used to generate them, whereas user clarifications favor the true intent. We use 54 binary-intent instances from 9 task families. Let h contain the initial ambiguous request, and let $g(h)$ be the difference in mean token log-likelihood between the true and competing intent descriptions, measured using a fixed scoring prompt. Appending a model response a or user clarification u gives $D_A = g(h + a) - g(h)$ or $D_U = g(h + u) - g(h)$, respectively, where $+$ denotes message appending. Thus, $D > 0$ indicates a shift toward the true intent, whereas $D < 0$ indicates a shift toward the competing intent. For each instance, we generate ten responses under each of two instructions: follow the true intent (Correct) or follow the competing intent (Wrong). These generation-only instructions are excluded from scoring. User clarifications express the true requirement at three preset levels of linguistic explicitness: Weak, Medium, and Strong.

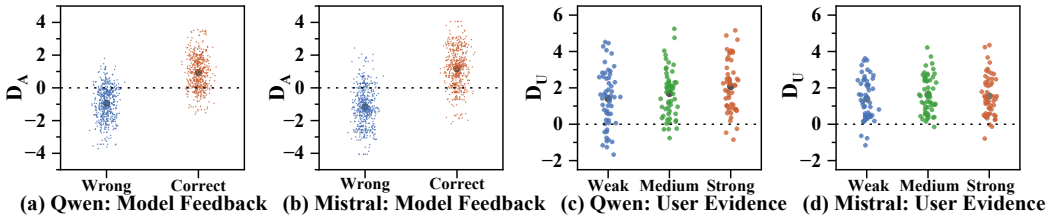


Figure 2: Preference shifts toward the true intent from model history (a–b) and user clarification (c–d). Colored points show sampled-response effects in (a–b) and instance-level clarification effects in (c–d); gray markers show means. $D > 0$ indicates increased preference for the true intent.

Fig. 2 shows, for Qwen/Mistral, mean D_A is $1.166/0.933$ under true-intent generation and $-1.154/-0.929$ under competing-intent generation; mean $D_U > 0$ in every clarification condition for both models. These results support the predicted local evidence directions. **Full methods and family-cluster confidence intervals are in Appx. H**.

8 CONCLUSION

We model multi-turn interaction as a closed-loop intent-inference process and show how model-generated responses can sustain errors even when user information alone supports reliable recovery. By abstracting from particular architectures and training procedures, the theory isolates a behavioral source of multi-turn degradation and suggests that improving reliability requires controlling how models reuse and revise conversational history, not only improving their single-turn task capability. Further discussion is provided in Appx. B

AI USE STATEMENT

In this work, we used generative AI tools for limited assistance. The theoretical framework, mathematical claims, and proofs were developed and completed by the authors. Generative AI was used only to review the proofs and help identify potential issues. For the experiments, generative AI was used to assist with writing and editing code, including code used to construct or generate the datasets employed in our experiments; the datasets themselves were produced by executing the resulting code. Generative AI was also used to improve the clarity and readability of the manuscript. We did not use generative AI for the other required-disclosure tasks, or the remaining required-disclosure tasks were not applicable to this work. We have reviewed all AI-assisted work before incorporating them into the final work. We take responsibility for the final content of this work, including all text, claims and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the complete theoretical framework, assumptions, and formal statements in the main text, with full proofs of all theoretical results in Appx. C-Appx. F. The role, limitations, and possible extensions of the assumptions are discussed in Appx. B. Appx. G and Appx. H provide detailed methods and results for the synthetic and real-LLM mechanism experiments. We additionally submit the core implementation code as supplementary material to facilitate reproduction of the reported experiments.

REFERENCES

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. Self-consuming generative models go MAD. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ShjMHfmPs0.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7421–7454, Bangkok, Thailand, August 2024. doi: 10.18653/v1/2024.acl-long.401. URL https://aclanthology.org/2024.acl-long.401/.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=liMSqUuVg9.

Quentin Bertrand, Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. On the stability of iterative retraining of generative models on their own data. In *The Twelfth International Conference on Learning Representations*, 2024.

Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 6045–6061, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/brown22a.html.

Tianlang Chen, Shirley Wu, and Jure Leskovec. Found in conversation: Llms teach themselves to close the multi-turn gap, 2026. URL https://arxiv.org/abs/2605.24432.

Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, Lifeng Jin, Ed-Yeremai Hernandez-Cardona, Dean Lee, Jeremy Kritz, Willow E. Primack, Summer Yue, and Chen Xing. MultiChallenge: A realistic multi-turn conversation evaluation benchmark challenging

to frontier LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18632–18702, Vienna, Austria, July 2025. doi: 10.18653/v1/2025.findings-acl.958. URL https://aclanthology.org/2025.findings-acl.958/.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Tomasz Korbak, Henry Sleight, Rajashree Agrawal, John Hughes, Dhruv Bhandarkar Pai, Andrey Gromov, Dan Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=5B2K4LRgmz.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, Shruti Bhosale, Chenguang Zhu, Karthik Abinav Sankararaman, Eryk Helenowski, Melanie Kambadur, Aditya Tayade, Hao Ma, Han Fang, and Sinong Wang. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following, 2024. URL https://arxiv.org/abs/2410.15553.

Jenny Y. Huang, Leshem Choshen, Wei Sun, Omar Khattab, Ramón Fernandez Astudillo, Mehul Damani, Tamara Broderick, and Jacob Andreas. Do LLMs benefit from their own words? In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=l6T6MrLgWp.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=IkmD3fKBPQ.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VKGTGGcwl6.

Ming Li, Pei Chen, Zhenhao Zhang, Tao Yang, Xinyang Zhang, Han Li, Tianyu Cao, Ming Zeng, Zhuofeng Wu, Meng Jiang, Huasheng Li, Lihong Li, and Bing Yin. Mitigating lost in multi-turn conversation via curriculum RL with verifiable accuracy and abstention rewards. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33364–33381, San Diego, California, United States, July 2026. doi: 10.18653/v1/2026.acl-long.1540. URL https://aclanthology.org/2026.acl-long.1540/.

Geng Liu, Fei Zhu, Rong Feng, Changyi Ma, Shiqi Wang, and Gaofeng Meng. Intent mismatch causes llms to get lost in multi-turn conversation, 2026. URL https://arxiv.org/abs/2602.07338.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7599–7609, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/perdomo20a.html.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, 11–13 Apr 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

Jyotika Singh, Fang Tu, Miguel Ballesteros, Weiyi Sun, Sandip Ghoshal, Michelle Yuan, Yassine Benajiba, Sujith Ravi, and Dan Roth. Mt-osc: Path for llms that get lost in multi-turn conversation, 2026. URL https://arxiv.org/abs/2604.08782.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. MINT: Evaluating LLMs in multi-turn interaction with tools and language feedback. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jp3gWrMuIZ.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization, 2024. URL https://openreview.net/forum?id=J9wzKfgZVK.

Haoyu Zheng, Yun Zhu, Shu Yuan, Shangming Chen, Qing Wang, Wenqiao Zhang, Jun Xiao, and Yueting Zhuang. Maigo: Mitigating lost-in-conversation with history-cleaned on-policy self-distillation, 2026. URL https://arxiv.org/abs/2605.27186.

APPENDIX CONTENTS

A Related Work **14**

B Discussion **14**

C Proofs for Section 3 **15**

C.1 Proof of Theorem 3.1. 15

C.2 Proof of Proposition 3.1. 17

C.3 Proof of Proposition 3.2. 19

C.4 Lemma C.3 and its Proof. 23

D Proofs for Section 4 **24**

D.1 Proof of Theorem 4.1. 25

D.2 Proof of Theorem 4.2. 28

D.3 Proof of Theorem 4.3. 33

D.4 Proof of Binary Case 36

E Proofs for Section 5 **37**

E.1 Proof of Theorem 5.1 37

E.2 Proof of Theorem 5.2 40

F Proofs for Section 6 **43**

F.1 Proof of Theorem F.1. 43

F.2 Proof of Corollary F.1. 46

F.3 Proof of Theorem F.2. 46

G Synthetic Validation of the Theoretical Model **51**

G.1 Statistical Role of Model-Generated History 52

G.2 Closed-Loop State–Drift Dynamics 53

G.3 Persistent Multi-Turn Degradation and Unreliability 55

G.4 Effects of Interaction Modifications 56

H Real-LLM Mechanism Validation **58**

H.1 State-Dependent Model Feedback 59

H.2 User-Evidence Effects 59

A Related Work

Multi-turn task completion. *Multi-turn* evaluation studies instruction following, contextual reasoning, and interaction with tools and language feedback (Bai et al., 2024; He et al., 2024; Deshpande et al., 2025; Wang et al., 2024). For tasks whose requirements are disclosed gradually, Laban et al. (2026) identify a substantial performance gap between single-turn and sharded multi-turn interaction, together with premature answers and persistent reliance on early assumptions. Liu et al. (2026) analyze intent mismatch and introduce a mediator that reconstructs explicit task instructions. Training-based approaches further address the multi-turn gap through paired-view self-distillation and rewards for accurate answers and appropriate abstention (Chen et al., 2026; Li et al., 2026).

Studies of model-generated history examine how earlier responses influence subsequent interaction. Huang et al. (2026) compare full-context prompting with reduced history configurations and identify the propagation of errors through model-generated context. MAIGO uses history-cleaned references to reduce self-contamination in multi-turn task completion (Zheng et al., 2026), while MT-OSC selectively condenses conversation history to retain relevant information with fewer tokens(Singh et al., 2026). Related analyses examine failures of iterative self-correction (Huang et al., 2024) and sensitivity to the placement of relevant information in long contexts (Liu et al., 2024).

We develop a quantitative theory of multi-turn task degradation by modeling how responses generated from the current intent estimate are reused as model-evidence in later inference. The resulting closed-loop analysis derives sufficient conditions for recovery and permanent non-recovery, connects these dynamics to persistent population-level performance gaps, and characterizes how protocol interventions alter the feedback process.

In-context inference of latent tasks. Statistical accounts of in-context learning interpret contextual adaptation through inference over latent concepts and tasks. Xie et al. (2022) analyze implicit Bayesian inference over a shared latent concept, while Zhang et al. (2024) characterize in-context learning through Bayesian model averaging. Related algorithmic analyses connect in-context learning to gradient descent, regression estimators, and algorithm selection (Akyürek et al., 2023; Von Oswald et al., 2023; Bai et al., 2023). These perspectives motivate our treatment of multi-turn conversation as inference over a fixed user intent. Our theory couples task inference with response generation: the current estimate generates model responses, which are then reused as model-evidence in subsequent intent scores. We characterize how this feedback interacts with accumulating user information to produce recovery or persistent errors.

Training-time feedback. Feedback in model training arises when a model influences the data used in subsequent training rounds. Performative prediction studies how deployed predictions alter data distributions and analyzes convergence under retraining (Perdomo et al., 2020). Stateful extensions allow these changes to depend on both the deployed model and the existing population state (Brown et al., 2022). Related feedback appears in imitation learning, where the current policy shapes the states visited during data collection and the resulting examples are used to update the policy (Ross et al., 2011). In self-consuming models, generated samples become training data for subsequent model generations. Studies of recursive training identify conditions under which this process degrades distributional fidelity or diversity (Bertrand et al., 2024; Alemohammad et al., 2024; Shumailov et al., 2024). Retaining real data and accumulating synthetic data can prevent collapse in the settings studied by Gerstgrasser et al. (2024).

Our theory studies feedback within multi-turn task inference: the true intent and likelihood families remain fixed, while responses generated from the current estimate enter later intent scores. We characterize how this model-evidence competes with fresh user information, deriving recovery and permanent non-recovery guarantees and connecting these dynamics to persistent multi-turn degradation without retraining.

B Discussion

Beyond fixed and non-reactive intent. Our work assumes that the user's intent $z^\star$ is fixed throughout the conversation and that future user shards U_{t+1} do not react to intermediate model responses. These assumptions **deliberately isolate** feedback generated by the model itself. An important extension is to allow U_{t+1} to depend on conversation history, Defining $\mathcal{G}_t := \sigma(\mathcal{F}_t, A_t)$, one may replace

$U_{t+1} \sim p_{\mathrm{U,t+1}}(\cdot \mid z^\star)$ with $U_{t+1} \mid \mathcal{G}_t \sim p_{\mathrm{U,t+1}}(\cdot \mid z^\star, \mathcal{G}_t)$. The user-evidence drift in Assump. 2.2 then becomes history dependent: $\underline{\kappa}_{t+1}(\mathcal{G}_t) := \min_{z \neq z^\star} D_{\mathrm{KL}}(p_{\mathrm{U,t+1}}(\cdot \mid z^\star, \mathcal{G}_t) \,\|\, p_{\mathrm{U,t+1}}(\cdot \mid z, \mathcal{G}_t))$. If this conditional informativeness remains uniformly positive, the recovery analysis may extend using conditional drift and martingale concentration rather than independent increments. Reactive corrections could increase this drift after an error, whereas accommodation to an incorrect model interpretation could decrease it.

A more substantial extension replaces the fixed target $z^\star$ with a time-varying intent $z_t^\star$. The objective then becomes tracking $z_t^\star$ rather than recovering a single latent intent, with natural quantities including $\mathbb{P}(\hat{z}_t \neq z_t^\star)$ and the adaptation delay following an intent change.

Relaxing technical assumptions. Assumptions in our paper can be relaxed in ways that preserve the basic feedback mechanism while changing the form of the guarantees.

First, exact initial ambiguity in Assump. 2.1 could be replaced by a distribution over the initial leader gap G_1. The factor $1 - \frac{1}{m}$ in our population bounds in Sec. 5 would then become the probability mass assigned to incorrect or near-ambiguous initial states.

The uniform user-informativeness in Assump. 2.2 condition $\kappa_t \geq \kappa > 0$ could instead be replaced by a cumulative condition such as K_T. User-only consistency would then depend on accumulated information rather than directly on T, while the fixed recovery comparison $\kappa > \lambda\beta$ would become a comparison between cumulative user correction and cumulative model feedback.

The uniform separation of model-response distributions in Assump. 2.3 could likewise be replaced by history-dependent quantities $c_t(H_t; z, z')$. The constants α_{err} and α_{true} would then become state dependent: weak separation would reduce self-reinforcement, whereas stronger commitment would amplify it.

The common bounded-evidence constant in Assump. 2.4 uses the same M. More generally, user- and model-evidence can satisfy separate bounds M_U and M_A on their log-likelihood ratios. The drift conditions are unchanged, while concentration terms such as $M^2(1 + \lambda^2)$ are replaced by source-specific scales of the form $M_U^2 + \lambda^2 M_A^2$. User-only bounds depend only on M_U, whereas closed-loop bounds depend on both sources.

Finally, the fixed source weight λ could be replaced by predictable history-dependent weights $\lambda_t = \lambda(H_t)$. The wrong-state drift would then be governed locally by quantities of the form $\overline{\kappa}_{t+1} - \lambda_t \alpha_{\mathrm{err},t}$, rather than a single global threshold. This extension would formalize models that discount self-generated history when their current interpretation is uncertain.

These extensions do not alter the core feedback mechanism, but allow a broader range of state-dependent and nonstationary conversational dynamics. Our work therefore provides a starting point for a broader theory of multi-turn interaction beyond the simplifying assumptions studied here.

Empirical scope and implications. Our theory is motivated by empirical studies showing that LLMs can degrade in multi-turn interaction despite receiving the same task information, and that model responses and interaction modifications systematically affect performance. The experiments in this work serve two complementary roles. First, synthetic experiments in Appx. G instantiate the formal model directly and test whether the predicted recovery, lock-in, and intervention dynamics emerge under controlled conditions. Second, the real-LLM experiments in Appx. H test the key behavioral prediction of the model: model-generated history tends to reinforce the interpretation under which it was generated, while fresh user clarification shifts preference toward the true intent. The synthetic results therefore validate the implications of the formal model, whereas the real-model results provide evidence that its central feedback mechanism is behaviorally plausible in instruction-tuned LLMs. They do not establish the full long-horizon guarantees for deployed models, but connect the theoretical abstraction to the empirical phenomena it is designed to explain.

C Proofs for Section 3

C.1 Proof of Theorem 3.1.

Proposition C.1 (Protocol equivalence). *Under Assumps. 2.1, 2.2, and 2.4, and the user-disclosure process defined in Sec. 2, when $\lambda = 0$, for every $T \geq 2$, $\mathbb{P}\big(\hat{z}_T^0 \neq z^\star\big) = \mathbb{P}\big(\hat{z}^{\mathrm{single}} \neq z^\star\big)$.*

Proof. At $\lambda = 0$, the score of the multi-turn estimator at $t = T$ is, for every $z \in \mathbb{Z}$ and every realization

$$S_0^{(T)}(z) = \log \pi(z) + \sum_{t=1}^{T} \log p_{\mathrm{U,t}}(U_t \mid z)$$

which is exactly the objective maximized in Eq. (1).

$S_0^{(T)}$ is a function of $U_{1:T}$ alone and the responses $A_{1:T-1}$, do not enter it.

Therefore, the multi-turn estimator maximizes

$$\log \pi(z) + \sum_{t=1}^{T} \log p_{\mathrm{U,t}}(U_t \mid z)$$

Thus, for every realization of the user shards $U_{1:T}$, the two protocols have the same argmax set.

Since both estimators break ties uniformly over this same set, using their respective tie-breaking variables ξ_T and ξ_T^{single}, they have the same conditional error probability:

$$\mathbb{P}(\hat{z}_T^0 \neq z^\star \mid U_{1:T}) = \mathbb{P}(\hat{z}_{\mathrm{single}} \neq z^\star \mid U_{1:T}).$$

Taking expectations over $U_{1:T}$ gives

$$\mathbb{P}(\hat{z}_T^0 \neq z^\star) = \mathbb{P}(\hat{z}_{\mathrm{single}} \neq z^\star).$$

□

Theorem 3.1 (Exponential consistency of user-only inference)**.** *Under Assumps. 2.1, 2.2, and 2.4, and the user-disclosure process defined in Sec. 2, when $\lambda = 0$, for every $T \geq 2$,*

$$\mathbb{P}(\hat{z}_T^0 \neq z^\star) = \mathbb{P}(\hat{z}_{\mathrm{single}} \neq z^\star) \leq (K-1) \exp\left(-\frac{(T-1)\underline{\kappa}^2}{2M^2}\right). \tag{3}$$

Proof. Fix an incorrect intent $z \neq z^\star$. Define its first-turn margin relative to the truth by

$$\Delta_1(z) := \log \pi(z^\star) + \log p_{\mathrm{U,1}}(U_1 \mid z^\star) - \log \pi(z) - \log p_{\mathrm{U,1}}(U_1 \mid z)$$

By Assump. 2.1, the first-turn argmax set is exactly $\mathbb{Z}_1$ and contains $z^\star$.

Hence for $z \in \mathbb{Z}_1$, $\Delta_1(z) = 0$. For $z \notin \mathbb{Z}_1$, $\Delta_1(z) > 0$.

Therefore, for every $z \neq z^\star$, $\Delta_1(z) \geq 0$

For every $s \geq 2$, define the user-side log-likelihood ratio

$$X_s(z) := \log \frac{p_{\mathrm{U,s}}(U_s \mid z^\star)}{p_{\mathrm{U,s}}(U_s \mid z)}$$

At $\lambda = 0$, the final score margin between the true intent and z is

$$\begin{aligned} S_0^{(T)}(z^\star) - S_0^{(T)}(z) &= \log \pi(z^\star) + \sum_{t=1}^{T} \log p_{\mathrm{U,t}}(U_t \mid z^\star) - \log \pi(z) - \sum_{t=1}^{T} \log p_{\mathrm{U,t}}(U_t \mid z) \\ &= \log \pi(z^\star) + \log p_{\mathrm{U,1}}(U_1 \mid z^\star) - \log \pi(z) - \log p_{\mathrm{U,1}}(U_1 \mid z) \\ &+ \sum_{s=2}^{T} \left(\log p_{\mathrm{U,s}}(U_s \mid z^\star) - \log p_{\mathrm{U,s}}(U_s \mid z)\right) \\ &= \Delta_1(z) + \sum_{s=2}^{T} X_s(z) \end{aligned}$$

If the final estimator selects an incorrect intent, then $\exists$ some $z \neq z^\star$, $S_0^{(T)}(z^\star) - S_0^{(T)}(z) \leq 0$.

Hence

$$\Delta_1(z) + \sum_{s=2}^{T} X_s(z) \le 0$$

By $\Delta_1(z) \ge 0$, this implies

$$\sum_{s=2}^{T} X_s(z) \le 0.$$

Conditional on the fixed true intent $z^\star$, $U_s \sim p_{\mathrm{U,s}}(\cdot \mid z^\star)$ and the random variables $\{X_s(z)\}_{s=2}^{T}$ are independent.

Their expectations are

$$\begin{aligned}\mathbb{E}[X_s(z)] &= \mathbb{E}_{U_s \sim p_{\mathrm{U,s}}(\cdot \mid z^\star)}\left[\log \frac{p_{\mathrm{U,s}}(U_s \mid z^\star)}{p_{\mathrm{U,s}}(U_s \mid z)}\right] \\ &= \mathrm{KL}\left(p_{\mathrm{U,s}}(\cdot \mid z^\star) \,\|\, p_{\mathrm{U,s}}(\cdot \mid z)\right) \ge \underline{\kappa}\end{aligned}$$

where the final inequality follows from Assump. 2.2.

Thus,

$$\sum_{s=2}^{T} \mathbb{E}[X_s(z)] \ge (T-1)\underline{\kappa}$$

So

$$\begin{aligned}\sum_{s=2}^{T} X_s(z) \le 0 &\Rightarrow \sum_{s=2}^{T}(X_s(z) - \mathbb{E}[X_s(z)]) \le -\sum_{s=2}^{T}\mathbb{E}[X_s(z)] \\ &\Rightarrow \sum_{s=2}^{T}(X_s(z) - \mathbb{E}[X_s(z)]) \le -(T-1)\underline{\kappa}\end{aligned}$$

By Assump. 2.4,

$$|X_s(z)| \le M$$

Thus each $X_s(z)$ lies in an interval of length at most $2M$.

Hoeffding's inequality gives

$$\begin{aligned}\mathbb{P}\left(\sum_{s=2}^{T} X_s(z) \le 0\right) &\le \mathbb{P}\left(\sum_{s=2}^{T}\big(X_s(z) - \mathbb{E}[X_s(z)]\big) \le -(T-1)\underline{\kappa}\right) \\ &\le \exp\left(-\frac{2((T-1)\underline{\kappa})^2}{(T-1)(2M)^2}\right) \\ &= \exp\left(-\frac{(T-1)\underline{\kappa}^2}{2M^2}\right)\end{aligned}$$

Finally, taking a union bound over the $K-1$ incorrect intents,

$$\mathbb{P}(\hat{z}_T^0 \ne z^\star) \le (K-1)\exp\left(-\frac{(T-1)\underline{\kappa}^2}{2M^2}\right)$$

□

C.2 Proof of Proposition 3.1.

Proposition C.2 (Posterior identity under the self-consistent reference)**.** *Under $\mathbb{P}_{\mathrm{sc}}$, for every $z \in \mathbb{Z}$, $\mathbb{P}_{\mathrm{sc}}(Z = z \mid H_t) = p_1(z \mid H_t)$. Consequently, the Bayes predictive distribution of the model response is $\sum_{z\in\mathbb{Z}} p_1(z \mid H_t)\, p_A(\cdot \mid z)$, and the maximum a posterior (MAP) argmax set coincides with that of Eq. (2) at $\lambda = 1$.*

Proof. Fix a history $h_t = (u_{1:t}, a_{1:t-1})$ generated under $\mathbb{P}_{\mathrm{sc}}$. Under the conditional-independence structure

$$p_{\mathrm{sc}}(h_t \mid z) = \prod_{s=1}^{t} p_{\mathrm{U,s}}(u_s \mid z) \prod_{s=1}^{t-1} p_{\mathrm{A}}(a_s \mid z)$$

Multiplying by the prior probability $\pi(z)$ gives the joint density

$$\begin{aligned} p_{\mathrm{sc}}(z, h_t) &= \pi(z) p_{\mathrm{sc}}(h_t \mid z) \\ &= \pi(z) \prod_{s=1}^{t} p_{\mathrm{U,s}}(u_s \mid z) \prod_{s=1}^{t-1} p_{\mathrm{A}}(a_s \mid z) \end{aligned}$$

We now compute the posterior distribution of the latent intent. By Bayes' rule,

$$p_{\mathrm{sc}}(z \mid h_t) = \frac{p_{\mathrm{sc}}(z, h_t)}{p_{\mathrm{sc}}(h_t)}$$

Since Z takes values in $\mathbb{Z}$, the marginal history density is

$$p_{\mathrm{sc}}(h_t) = \sum_{\tilde{z} \in \mathbb{Z}} p_{\mathrm{sc}}(\tilde{z}, h_t)$$

Substituting the joint density above yields

$$p_{\mathrm{sc}}(z \mid h_t) = \frac{\pi(z) \prod_{s=1}^{t} p_{\mathrm{U,s}}(u_s \mid z) \prod_{s=1}^{t-1} p_{\mathrm{A}}(a_s \mid z)}{\sum_{\tilde{z} \in \mathbb{Z}} \pi(\tilde{z}) \prod_{s=1}^{t} p_{\mathrm{U,s}}(u_s \mid \tilde{z}) \prod_{s=1}^{t-1} p_{\mathrm{A}}(a_s \mid \tilde{z})}$$

Recall that the source-weighted posterior is

$$p_{\lambda}(z \mid h_t) \propto \pi(z) \prod_{s=1}^{t} p_{\mathrm{U,s}}(u_s \mid z) \prod_{s=1}^{t-1} p_{\mathrm{A}}(a_s \mid z)^{\lambda}$$

Setting $\lambda = 1$ gives exactly

$$p_{\mathrm{sc}}(z \mid h_t) = p_1(z \mid h_t)$$

Under $\mathbb{P}_{\mathrm{sc}}$, conditional on $Z = z$, the next response is generated from $p_{\mathrm{A}}(\cdot \mid z)$ and is independent of the previous history H_t.

Thus, by conditioning on the latent intent,

$$\begin{aligned} \mathbb{P}_{\mathrm{sc}}(A_t \in \mathcal{B} \mid H_t = h_t) &= \sum_{z \in \mathbb{Z}} \mathbb{P}_{\mathrm{sc}}(Z = z \mid H_t = h_t) \times \mathbb{P}_{\mathrm{sc}}(A_t \in \mathcal{B} \mid Z = z, H_t = h_t) \\ &= \sum_{z \in \mathbb{Z}} p_1(z \mid h_t) p_{\mathrm{A}}(\mathcal{B} \mid z) \end{aligned}$$

where the last equality uses the posterior identity proved above and the conditional independence of A_t and H_t given Z.

Therefore,

$$p_{\mathrm{sc}}(A_t \mid H_t) = \sum_{z \in \mathbb{Z}} p_1(z \mid H_t)\, p_{\mathrm{A}}(A_t \mid z)$$

Finally, the MAP estimate of the latent intent under $\mathbb{P}_{\mathrm{sc}}$ is

$$\arg\max_{z \in \mathbb{Z}} p_{sc}(z \mid H_t) = \arg\max_{z \in \mathbb{Z}} p_1(z \mid H_t)$$

The normalizing denominator of $p_1(z \mid H_t)$ does not depend on z. Taking logarithms therefore shows that this is equivalent to maximizing

$$\log \pi(z) + \sum_{s=1}^{t} \log p_{\mathrm{U,s}}(U_s \mid z) + \sum_{s=1}^{t-1} \log p_{\mathrm{A}}(A_s \mid z)$$

This is exactly the score in Eq. (2) with $\lambda = 1$. □

C.3 Proof of Proposition 3.2.

Proposition 3.2 (Strict posterior optimality of $\lambda = 1$)**.** *Under Assumps. 2.3-2.4, for any $t \geq 2$, $R_t(\lambda)$ is uniquely minimized at $\lambda = 1$, with $R_t(1) = 0$. Moreover, in a neighborhood of $\lambda = 1$,*

$$R_t(\lambda) = \frac{(\lambda-1)^2}{2}\mathcal{I}_t + o\big((\lambda-1)^2\big), \text{ where } \mathcal{I}_t := \mathbb{E}_{\mathrm{sc}}\left[\mathrm{Var}_{z\sim p_1(\cdot|H_t)}\Big(\sum_{s=1}^{t-1}\log p_{\mathrm{A}}(A_s \mid z)\Big)\right] > 0.$$

Proof. Fix a history $H_t = h_t$. Define

$$B_t(z; h_t) := \log \pi(z) + \sum_{s=1}^{t} \log p_{\mathrm{U,s}}(u_s \mid z)$$

and

$$L(z) := \sum_{s=1}^{t-1} \log p_{\mathrm{A}}(a_s \mid z)$$

For $\lambda \geq 0$, define the normalizing constant

$$Z_{h_t}(\lambda) := \sum_{z\in\mathbb{Z}} \exp\big(B_t(z; h_t) + \lambda L(z)\big)$$

The source-weighted posterior can then be written as

$$p_\lambda(z \mid h_t) = \frac{\exp\big(B_t(z; h_t) + \lambda L(z)\big)}{Z_{h_t}(\lambda)}$$

We first establish that $\lambda = 1$ is the unique minimizer of the pretraining risk, and then derive its local expansion.

Step 1: Nonnegativity of the risk. Define the history-wise risk

$$r_{h_t}(\lambda) := \mathrm{KL}\left(p_1(\cdot \mid h_t) \,\|\, p_\lambda(\cdot \mid h_t)\right)$$

Since KL divergence is nonnegative,$r_{h_t}(\lambda) \geq 0$.

Therefore,

$$R_t(\lambda) = \mathbb{E}_{\mathbb{P}_{\mathrm{sc}}}\left[r_{H_t}(\lambda)\right] \geq 0.$$

At $\lambda = 1$, the two distributions inside the KL divergence are identical:

$$p_\lambda(\cdot \mid h_t) = p_1(\cdot \mid h_t)$$

Hence for every history,

$$r_{h_t}(1) = 0$$

consequently

$$R_t(1) = 0$$

Thus $\lambda = 1$ is a global minimizer. It remains to prove that it is the unique minimizer.

Step 2: Characterizing when the history-wise risk is zero. By the positive-prior and common-support assumptions, all posterior probabilities $p_\lambda(z \mid h_t)$ are strictly positive.

Thus $r_{h_t}(\lambda) = 0$ if and only if for every $z \in \mathbb{Z}$, $p_\lambda(z \mid h_t) = p_1(z \mid h_t)$

For any two intents $z, z' \in \mathbb{Z}$, their posterior odds under p_λ are

$$\begin{aligned}\frac{p_\lambda(z \mid h_t)}{p_\lambda(z' \mid h_t)} &= \frac{\exp(B_t(z; h_t) + \lambda L(z))}{\exp(B_t(z'; h_t) + \lambda L(z'))} \\ &= \exp\big(B_t(z; h_t) - B_t(z'; h_t) + \lambda\big(L(z) - L(z')\big)\big)\end{aligned}$$

Similarly,

$$\frac{p_1(z \mid h_t)}{p_1(z' \mid h_t)} = \exp(B_t(z; h_t) - B_t(z'; h_t) + L(z) - L(z')) .$$

If $p_\lambda(\cdot \mid h_t) = p_1(\cdot \mid h_t)$, these two odds must be equal.

Cancelling the common term $B_t(z; h_t) - B_t(z'; h_t)$ gives

$$\exp\big(\lambda\big(L(z) - L(z')\big)\big) = \exp(L(z) - L(z'))$$

Equivalently,

$$\exp\big((\lambda - 1)\big(L(z) - L(z')\big)\big) = 1.$$

Since the exponent is real,

$$(\lambda - 1)\big(L(z) - L(z')\big) = 0.$$

Thus, if $\lambda \neq 1$, equality of the two posteriors requires for every $z, z' \in \mathbb{Z}$

$$L(z) = L(z')$$

In other words, for $\lambda \neq 1$, the history-wise risk can vanish only when the accumulated likelihood $L(z)$ is constant across all intents.

Step 3: Such equality fails with positive probability. We now show that under $\mathbb{P}_{\text{sc}}$, there is positive probability that $L(z)$ is not constant across intents.

Choose any two distinct intents $z_0, z_1 \in \mathbb{Z}$, and define

$$g(a) := \log \frac{p_{\text{A}}(a \mid z_0)}{p_{\text{A}}(a \mid z_1)}.$$

If $A \sim p_{\text{A}}(\cdot \mid z_0)$, then

$$\begin{aligned}\mathbb{E}[g(A)] &= \mathbb{E}_{A \sim p_{\text{A}}(\cdot \mid z_0)}\left[\log \frac{p_{\text{A}}((A \mid z_0)}{p_{\text{A}}((A \mid z_1)}\right] \\ &= \text{KL}\left(p_{\text{A}}((\cdot \mid z_0) \,\|\, p_{\text{A}}((\cdot \mid z_1)\right)\end{aligned}$$

By Assumption 2.3, this KL divergence is strictly positive. Hence

$$\mathbb{E}[g(A)] > 0.$$

This implies

$$\mathbb{P}_{A \sim p_{\text{A}}((\cdot \mid z_0)}\big(g(A) > 0\big) > 0$$

Conditional on $Z = z_0$, the model responses $A_1, \ldots, A_{t-1}$ are independent under $\mathbb{P}_{\text{sc}}$. Therefore,

$$\mathbb{P}\left(g(A_s) > 0 \text{ for all } s = 1, \ldots, t-1 \mid Z = z_0\right) = \left[\mathbb{P}_{A \sim p_{\text{A}}((\cdot \mid z_0)}\big(g(A) > 0\big)\right]^{t-1} > 0$$

Because $\pi(z_0) > 0$, the event

$$\mathcal{E} := \{Z = z_0,\ g(A_s) > 0 \text{ for all } s = 1, \ldots, t-1\}$$

has probability

$$\mathbb{P}_{\mathbb{P}_{\text{sc}}}(\mathcal{E}) = \pi(z_0)\left[\mathbb{P}_{A \sim p_{\text{A}}((\cdot \mid z_0)}\big(g(A) > 0\big)\right]^{t-1} > 0$$

On $\mathcal{E}$,

$$\begin{aligned}L(z_0) - L(z_1) &= \sum_{s=1}^{t-1} \left[\log p_{\text{A}}(A_s \mid z_0) - \log p_{\text{A}}(A_s \mid z_1)\right] \\ &= \sum_{s=1}^{t-1} \log \frac{p_{\text{A}}(A_s \mid z_0)}{p_{\text{A}}(A_s \mid z_1)} \\ &= \sum_{s=1}^{t-1} g(A_s) > 0\end{aligned}$$

Thus $L(z_0) \neq L(z_1)$ on an event of strictly positive probability.

For any $\lambda \neq 1$, Step 2 then implies

$$p_\lambda(\cdot \mid H_t) \neq p_1(\cdot \mid H_t)$$

on $\mathcal{E}$, and therefore on $\mathcal{E}$

$$r_{H_t}(\lambda) > 0$$

Since $r_{H_t}(\lambda) \geq 0$ everywhere and is strictly positive on an event of positive probability, for every $\lambda \neq 1$

$$R_t(\lambda) = \mathbb{E}[r_{H_t}(\lambda)] > 0$$

Together with $R_t(1) = 0$, this proves that $\lambda = 1$ is the unique minimizer.

Step 4: Rewriting the history-wise KL risk. We next derive the local expansion around $\lambda = 1$. By definition,

$$r_{h_t}(\lambda) = \sum_{z \in \mathbb{Z}} p_1(z \mid h_t) \log \frac{p_1(z \mid h_t)}{p_\lambda(z \mid h_t)}$$

Using

$$p_1(z \mid h_t) = \frac{\exp(B_t(z; h_t) + L(z))}{Z_{h_t}(1)}$$

and

$$p_\lambda(z \mid h_t) = \frac{\exp(B_t(z; h_t) + \lambda L(z))}{Z_{h_t}(\lambda)}$$

their log ratio is

$$\begin{aligned} \log \frac{p_1(z \mid h_t)}{p_\lambda(z \mid h_t)} &= [B_t(z; h_t) + L(z) - \log Z_{h_t}(1)] - [B_t(z; h_t) + \lambda L(z) - \log Z_{h_t}(\lambda)] \\ &= \log Z_{h_t}(\lambda) - \log Z_{h_t}(1) - (\lambda - 1)L(z) \end{aligned}$$

Substituting this expression into the KL divergence gives

$$r_{h_t}(\lambda) = \sum_{z \in \mathbb{Z}} p_1(z \mid h_t) \Big[\log Z_{h_t}(\lambda) - \log Z_{h_t}(1) - (\lambda - 1)L(z) \Big]$$

The first two terms do not depend on z, and

$$\sum_{z \in \mathbb{Z}} p_1(z \mid h_t) = 1.$$

Therefore,

$$\begin{aligned} r_{h_t}(\lambda) &= \log Z_{h_t}(\lambda) - \log Z_{h_t}(1) - (\lambda - 1) \sum_{z \in \mathbb{Z}} p_1(z \mid h_t) L(z) \\ &= \log Z_{h_t}(\lambda) - \log Z_{h_t}(1) - (\lambda - 1)\mathbb{E}_{z \sim p_1(\cdot \mid h_t)}[L(z)] \end{aligned}$$

Step 5: Derivatives of the log-partition function. We now compute the derivatives of $Z_{h_t}(\lambda)$. From its definition,

$$Z_{h_t}(\lambda) = \sum_{z \in \mathbb{Z}} \exp\left(B_t(z; h_t) + \lambda L(z)\right)$$

Differentiating with respect to λ gives

$$Z'_{h_t}(\lambda) = \sum_{z \in \mathbb{Z}} L(z) \exp\left(B_t(z; h_t) + \lambda L(z)\right)$$

Hence

$$\begin{aligned} \frac{d}{d\lambda} \log Z_{h_t}(\lambda) &= \frac{Z'_{h_t}(\lambda)}{Z_{h_t}(\lambda)} \\ &= \sum_{z \in \mathbb{Z}} \frac{\exp(B_t(z; h_t) + \lambda L(z))}{Z_{h_t}(\lambda)} L(z) \\ &= \sum_{z \in \mathbb{Z}} p_\lambda(z \mid h_t) L(z) \\ &= \mathbb{E}_{z \sim p_\lambda(\cdot \mid h_t)}[L(z)] \end{aligned}$$

Differentiating once more,

$$Z''_{h_t}(\lambda) = \sum_{z\in\mathbb{Z}} L(z)^2 \exp\left(B_t(z;h_t) + \lambda L(z)\right)$$

Therefore,

$$\begin{aligned}\frac{d^2}{d\lambda^2}\log Z_{h_t}(\lambda) &= \frac{Z''_{h_t}(\lambda)}{Z_{h_t}(\lambda)} - \left(\frac{Z'_{h_t}(\lambda)}{Z_{h_t}(\lambda)}\right)^2 \\ &= \mathbb{E}_{z\sim p_\lambda(\cdot|h_t)}[L(z)^2] - \left(\mathbb{E}_{z\sim p_\lambda(\cdot|h_t)}[L(z)]\right)^2 \\ &= \mathrm{Var}_{z\sim p_\lambda(\cdot|h_t)}\big(L(z)\big)\end{aligned}$$

Step 6: Taylor expansion around $\lambda = 1$. Differentiating the expression for $r_{h_t}(\lambda)$ gives

$$\begin{aligned}r'_{h_t}(\lambda) &= \frac{d}{d\lambda}\log Z_{h_t}(\lambda) - \mathbb{E}_{z\sim p_1(\cdot|h_t)}[L(z)] \\ &= \mathbb{E}_{z\sim p_\lambda(\cdot|h_t)}[L(z)] - \mathbb{E}_{z\sim p_1(\cdot|h_t)}[L(z)]\end{aligned}$$

Thus at $\lambda = 1$,

$$r'_{h_t}(1) = 0.$$

We already know

$$r_{h_t}(1) = 0.$$

The second derivative is

$$r''_{h_t}(\lambda) = \mathrm{Var}_{z\sim p_\lambda(\cdot|h_t)}\big(L(z)\big),$$

and hence

$$r''_{h_t}(1) = \mathrm{Var}_{z\sim p_1(\cdot|h_t)}\big(L(z)\big)$$

Taylor's theorem around $\lambda = 1$ therefore gives

$$\begin{aligned}r_{h_t}(\lambda) &= r_{h_t}(1) + (\lambda-1)r'_{h_t}(1) + \frac{(\lambda-1)^2}{2}r''_{h_t}(1) + o\big((\lambda-1)^2\big) \\ &= \frac{(\lambda-1)^2}{2}\,\mathrm{Var}_{z\sim p_1(\cdot|h_t)}\big(L(z)\big) + o\big((\lambda-1)^2\big)\,.\end{aligned}$$

Step 7: Taking expectation over histories. It remains to justify taking expectation of this expansion over H_t.

By Assumption 2.4, for every pair $z, z' \in \mathbb{Z}$ and every response a,

$$\left|\log\frac{p_{\mathrm{A}}(a \mid z)}{p_{\mathrm{A}}(a \mid z')}\right| \le M$$

Therefore, for every history,

$$\begin{aligned}|L(z) - L(z')| &= \left|\sum_{s=1}^{t-1}\log\frac{p_{\mathrm{A}}(a_s \mid z)}{p_{\mathrm{A}}(a_s \mid z')}\right| \\ &\le \sum_{s=1}^{t-1}\left|\log\frac{p_{\mathrm{A}}(a_s \mid z)}{p_{\mathrm{A}}(a_s \mid z')}\right| \\ &\le (t-1)M\end{aligned}$$

Hence the range of $L(z)$ over $z \in \mathbb{Z}$ is uniformly bounded by $(t-1)M$.

In particular,

$$\mathrm{Var}_{z\sim p_\lambda(\cdot|h_t)}(L(z)) \le \frac{(t-1)^2M^2}{4}$$

for every history and every λ.

Moreover, the third derivative of the log-partition function is the third centered moment of L under p_λ:

$$\frac{d^3}{d\lambda^3}\log Z_{h_t}(\lambda) = \mathbb{E}_{z\sim p_\lambda(\cdot|h_t)}\left[(L(z)-\mathbb{E}_{p_\lambda}[L])^3\right]$$

Since the range of L is at most $(t-1)M$, its absolute value is uniformly bounded:

$$\left|\frac{d^3}{d\lambda^3}\log Z_{h_t}(\lambda)\right| \le (t-1)^3M^3.$$

Thus the Taylor remainder above is uniformly $O(|\lambda-1|^3)$ over all histories.

We may therefore take expectations to obtain

$$R_t(\lambda) = \frac{(\lambda-1)^2}{2}\mathbb{E}_{\mathbb{P}_{\mathrm{sc}}}\left[\mathrm{Var}_{z\sim p_1(\cdot|H_t)}\big(L(z)\big)\right] + o\big((\lambda-1)^2\big)$$

By the definition of $\mathcal{I}_t$,

$$R_t(\lambda) = \frac{(\lambda-1)^2}{2}\mathcal{I}_t + o\big((\lambda-1)^2\big)$$

Step 8: Strict positivity of $\mathcal{I}_t$. Finally, on the positive-probability event $\mathcal{E}$ constructed in Step 3,

$$L(z_0) \neq L(z_1).$$

The posterior $p_1(\cdot \mid H_t)$ assigns strictly positive probability to every intent.

Hence, whenever two values of $L(z)$ differ,

$$\mathrm{Var}_{z\sim p_1(\cdot|H_t)}\big(L(z)\big) > 0$$

Therefore the random variable $\mathrm{Var}_{z\sim p_1(\cdot|H_t)}\big(L(z)\big)$ is nonnegative everywhere and strictly positive on an event of positive probability.

Its expectation is consequently strictly positive:

$$\mathcal{I}_t = \mathbb{E}_{\mathrm{sc}}\left[\mathrm{Var}_{z\sim p_1(\cdot|H_t)}\big(L(z)\big)\right] > 0.$$

□

C.4 Lemma C.3 and its Proof.

Lemma C.3 (Model-evidence reinforces the current estimate)**.** *If the estimate is an incorrect intent $\hat{z}_t^\lambda \neq z^\star$, then $\mathbb{E}\left[\log \frac{p_{\mathrm{A}}(A_t|z^\star)}{p_{\mathrm{A}}(A_t|\hat{z}_t^\lambda)} \mid \mathcal{F}_t\right] = -\mathrm{KL}\left(p_{\mathrm{A}}(\cdot \mid \hat{z}_t^\lambda)\,\|\,p_{\mathrm{A}}(\cdot \mid z^\star)\right) \le -c$. If instead $\hat{z}_t^\lambda = z^\star$, then for every $z \neq z^\star$, $\mathbb{E}\left[\log \frac{p_{\mathrm{A}}(A_t|z^\star)}{p_{\mathrm{A}}(A_t|z)} \mid \mathcal{F}_t\right] = \mathrm{KL}\left(p_{\mathrm{A}}(\cdot \mid z^\star)\,\|\,p_{\mathrm{A}}(\cdot \mid z)\right) \ge c$.*

Proof. Conditional on $\mathcal{F}_t$, the estimate $\hat{z}_t$ is fixed. And

$$A_t \mid \mathcal{F}_t \sim p_{\mathrm{A}}(\cdot \mid \hat{z}_t)$$

Suppose first that $\hat{z}_t = w \neq z^\star$. Then

$$\begin{aligned}\mathbb{E}\left[\log\frac{p_{\mathrm{A}}(A_t \mid z^\star)}{p_{\mathrm{A}}(A_t \mid w)}\,\middle|\,\mathcal{F}_t\right] &= \mathbb{E}_{A\sim p_{\mathrm{A}}(\cdot|w)}\left[\log\frac{p_{\mathrm{A}}(A \mid z^\star)}{p_{\mathrm{A}}(A \mid w)}\right]\\ &= -\mathbb{E}_{A\sim p_{\mathrm{A}}(\cdot|w)}\left[\log\frac{p_{\mathrm{A}}(A \mid w)}{p_{\mathrm{A}}(A \mid z^\star)}\right]\\ &= -\mathrm{KL}\left(p_{\mathrm{A}}(\cdot \mid w)\,\|\,p_{\mathrm{A}}(\cdot \mid z^\star)\right) \le -c\end{aligned}$$

where the last inequality follows from Assump. 2.3.

Suppose instead that $\hat{z}_t = z^\star$. Then

$$A_t \mid \mathcal{F}_t \sim p_{\mathrm{A}}(\cdot \mid z^\star)$$

Hence, for every $z \neq z^\star$,

$$\begin{aligned}\mathbb{E}\left[\log\frac{p_{\mathrm{A}}(A_t \mid z^\star)}{p_{\mathrm{A}}(A_t \mid z)}\,\middle|\,\mathcal{F}_t\right] &= \mathbb{E}_{A\sim p_{\mathrm{A}}(\cdot|z^\star)}\left[\log\frac{p_{\mathrm{A}}(A \mid z^\star)}{p_{\mathrm{A}}(A \mid z)}\right]\\ &= \mathrm{KL}\left(p_{\mathrm{A}}(\cdot \mid z^\star)\,\|\,p_{\mathrm{A}}(\cdot \mid z)\right) \ge c\end{aligned}$$

This proves both statements. □

D PROOFS FOR SECTION 4

We first record the conditional concentration bound used throughout the proofs.

Lemma D.1 (Conditional moment-generating-function bound)**.** *Let $\mathcal{G}$ be a sigma-algebra, and let X and Y be conditionally independent given $\mathcal{G}$. Suppose $|X| \le M$ and $|Y| \le M$. Then, for every $s \in \mathbb{R}$,*

$$\mathbb{E}\left[e^{s(X+\lambda Y)} \,\middle|\, \mathcal{G}\right] \le \exp\left(s\,\mathbb{E}[X+\lambda Y \mid \mathcal{G}] + \frac{s^2M^2(1+\lambda^2)}{2}\right) \tag{9}$$

Proof. Since

$$e^{s(X+\lambda Y)} = e^{sX}e^{s\lambda Y}$$

and X and Y are conditionally independent given $\mathcal{G}$, we have

$$\begin{aligned}\mathbb{E}\left[e^{s(X+\lambda Y)} \,\middle|\, \mathcal{G}\right] &= \mathbb{E}\left[e^{sX}e^{s\lambda Y} \,\middle|\, \mathcal{G}\right] \\ &= \mathbb{E}\left[e^{sX} \,\middle|\, \mathcal{G}\right]\mathbb{E}\left[e^{s\lambda Y} \,\middle|\, \mathcal{G}\right]\end{aligned}$$

We bound the two factors separately.

Since $|X| \le M$, conditional on $\mathcal{G}$, the random variable X takes values in the interval $[-M, M]$, whose length is $2M$.

The conditional form of Hoeffding's lemma therefore gives

$$\mathbb{E}\left[e^{s(X-\mathbb{E}[X|\mathcal{G}])} \,\middle|\, \mathcal{G}\right] \le \exp\left(\frac{s^2(2M)^2}{8}\right) = \exp\left(\frac{s^2M^2}{2}\right)$$

Since

$$sX = s\mathbb{E}[X \mid \mathcal{G}] + s(X - \mathbb{E}[X \mid \mathcal{G}])$$

then

$$e^{sX} = e^{s\mathbb{E}[X|\mathcal{G}]}e^{s(X-\mathbb{E}[X|\mathcal{G}])}$$

and $\mathbb{E}[X \mid \mathcal{G}]$ is $\mathcal{G}$-measurable,

$$\begin{aligned}\mathbb{E}[e^{sX} \mid \mathcal{G}] &= \mathbb{E}\left[e^{s\mathbb{E}[X|\mathcal{G}]}e^{s(X-\mathbb{E}[X|\mathcal{G}])} \,\middle|\, \mathcal{G}\right] \\ &= e^{s\mathbb{E}[X|\mathcal{G}]}\mathbb{E}\left[e^{s(X-\mathbb{E}[X|\mathcal{G}])} \,\middle|\, \mathcal{G}\right] \\ &\le e^{s\mathbb{E}[X|\mathcal{G}]}\exp\left(\frac{s^2M^2}{2}\right) \\ &= \exp\left(s\mathbb{E}[X \mid \mathcal{G}] + \frac{s^2M^2}{2}\right)\end{aligned}$$

Similarly, $|Y| \le M$ implies

$$|\lambda Y| \le |\lambda| M$$

Thus λY lies in an interval of length $2|\lambda|M$, and applying conditional Hoeffding's lemma again gives

$$\begin{aligned}\mathbb{E}[e^{s\lambda Y} \mid \mathcal{G}] &\le \exp\left(s\lambda\mathbb{E}[Y \mid \mathcal{G}] + \frac{s^2(2|\lambda|M)^2}{8}\right) \\ &= \exp\left(s\lambda\mathbb{E}[Y \mid \mathcal{G}] + \frac{s^2\lambda^2M^2}{2}\right)\end{aligned}$$

Multiplying the two bounds yields

$$\mathbb{E}\left[e^{s(X+\lambda Y)} \,\middle|\, \mathcal{G}\right] \quad \le \exp\left(s\mathbb{E}[X \mid \mathcal{G}] + s\lambda\mathbb{E}[Y \mid \mathcal{G}] + \frac{s^2M^2(1+\lambda^2)}{2}\right)$$

Finally, by linearity of conditional expectation,

$$\mathbb{E}[X \mid \mathcal{G}] + \lambda\mathbb{E}[Y \mid \mathcal{G}] = \mathbb{E}[X + \lambda Y \mid \mathcal{G}]$$

Therefore,

$$\mathbb{E}\left[e^{s(X+\lambda Y)} \,\middle|\, \mathcal{G}\right] \le \exp\left(s\mathbb{E}[X+\lambda Y \mid \mathcal{G}] + \frac{s^2M^2(1+\lambda^2)}{2}\right)$$

which proves Eq. (9). □

D.1 Proof of Theorem 4.1.

Theorem 4.1 (Wrong-state lock-in). *Fix any turn $t_0 \geq 2$ and condition on a realization of $\mathcal{F}_{t_0}$ such that $G_{t_0} = -\delta < 0$. Let $t_{\text{rec}} := \inf\{t > t_0 : \hat{z}_t^\lambda = z^\star\}$ be the first subsequent recovery time. If $\lambda\alpha_{\text{err}} > \overline{\kappa}$, then*

$$\mathbb{P}\left(t_{\text{rec}} = \infty \,|\, \mathcal{F}_{t_0}\right) \geq 1 - \exp\left(-\frac{2(\lambda\alpha_{\text{err}} - \overline{\kappa})\delta}{M^2(1+\lambda^2)}\right).$$

Moreover, with probability one on $\{t_{\text{rec}} = \infty\}$, $G_t \longrightarrow -\infty$.

Proof. Condition on $\mathcal{F}_{t_0}$ and we use

$$t_\sigma = t_{\text{rec}} = \inf\{t > t_0 : \hat{z}_t^\lambda = z^\star\}$$

for this proof. For notational convenience, write

$$h := \lambda\alpha_{\text{err}} - \overline{\kappa} > 0, \quad \nu := \frac{2h}{M^2(1+\lambda^2)}$$

Fix a turn t with $t_0 \leq t < t_\sigma$ and let $z_t := \hat{z}_t^\lambda$.

Since the model is wrong before t_σ, $z_t \neq z^\star$. Moreover, because z_t is the selected wrong maximizer,

$$\Delta_t(z_t) = G_t$$

Define

$$W_t := \log \frac{p_{\text{U,t+1}}(U_{t+1} \mid z^\star)}{p_{\text{U,t+1}}(U_{t+1} \mid z_t)} + \lambda \log \frac{p_{\text{A}}(A_t \mid z^\star)}{p_{\text{A}}(A_t \mid z_t)}$$

By the margin recursion,

$$\Delta_{t+1}(z_t) = G_t + W_t$$

Since G_{t+1} is the minimum margin over all wrong intents, for $t < t_\sigma$

$$G_{t+1} \leq G_t + W_t \tag{10}$$

Conditional on $\mathcal{F}_t$, the user term has expectation

$$\mathbb{E}\left[\log \frac{p_{\text{U,t+1}}(U_{t+1} \mid z^\star)}{p_{\text{U,t+1}}(U_{t+1} \mid z_t)} \,\middle|\, \mathcal{F}_t\right] = \text{KL}\left(p_{\text{U,t+1}}(\cdot \mid z^\star) \,\|\, p_{\text{U,t+1}}(\cdot \mid z_t)\right) \leq \overline{\kappa}$$

Because $A_t \mid \mathcal{F}_t \sim p_{\text{A}}(\cdot \mid z_t)$,

$$\mathbb{E}\left[\log \frac{p_{\text{A}}(A_t \mid z^\star)}{p_{\text{A}}(A_t \mid z_t)} \,\middle|\, \mathcal{F}_t\right] = -\,\text{KL}\left(p_{\text{A}}(\cdot \mid z_t) \,\|\, p_{\text{A}}(\cdot \mid z^\star)\right) \leq -\alpha_{\text{err}}$$

Therefore,

$$\mathbb{E}[W_t \mid \mathcal{F}_t] \leq \overline{\kappa} - \lambda\alpha_{\text{err}} = -h$$

Conditional on $\mathcal{F}_t$, the user and terms in W_t are independent and each has absolute value at most M. Therefore, Lem. D.1 with $s = \nu$ gives

$$\begin{aligned}
\mathbb{E}[e^{\nu W_t} \mid \mathcal{F}_t] &\leq \exp\left(-\nu h + \frac{\nu^2 M^2(1+\lambda^2)}{2}\right) \\
&= \exp\left(\frac{-\nu^2 M^2(1+\lambda^2)}{2} + \frac{\nu^2 M^2(1+\lambda^2)}{2}\right) \\
&= 1
\end{aligned}$$

Now define the stopped process, for $t \geq t_0$

$$L_t := e^{\nu G_{\min\{t, t_\sigma\}}}$$

On $\{t < t_\sigma\}$, by Eq. (10) implies

$$\begin{aligned}\mathbb{E}[L_{t+1} \mid \mathcal{F}_t] &= \mathbb{E}[e^{\nu G_{t+1}} \mid \mathcal{F}_t] \\ &\le \mathbb{E}[e^{\nu G_t + \nu W_t} \mid \mathcal{F}_t] \\ &\le e^{\nu G_t}\mathbb{E}[e^{\nu W_t} \mid \mathcal{F}_t] \le L_t\end{aligned}$$

On $\{t \ge t_\sigma\}$, the stopped process remains constant $e^{\nu G_{t_\sigma}}$.

For every integer $n > t_0$,

$$\mathbb{E}[L_n \mid \mathcal{F}_{t_0}] \le L_{t_0} = e^{\nu G_{t_0}} = e^{-\nu\delta}$$

Now consider the event $\{t_\sigma \le n\}$. At time t_σ, the true intent is selected, so for every $z \ne z^\star$.

$$\Delta_{t_\sigma}(z) \ge 0$$

Hence

$$G_{t_\sigma} \ge 0.$$

Because the process is stopped at t_σ, on $\{t_\sigma \le n\}$

$$L_n = e^{\nu G_{t_\sigma}} \ge 1$$

Therefore,

$$\mathbf{1}\{t_\sigma \le n\} \le L_n$$

Taking conditional expectations gives

$$\begin{aligned}\mathbb{P}(t_\sigma \le n \mid \mathcal{F}_{t_0}) &= \mathbb{E}[\mathbf{1}\{t_\sigma \le n\} \mid \mathcal{F}_{t_0}] \\ &\le \mathbb{E}[L_n \mid \mathcal{F}_{t_0}] \\ &\le e^{-\nu\delta}\end{aligned}$$

Finally, since

$$\{t_\sigma < \infty\} = \bigcup_{n > t_0} \{t_\sigma \le n\}$$

letting $n \to \infty$ yields

$$\mathbb{P}(t_\sigma < \infty \mid \mathcal{F}_{t_0}) \le e^{-\nu\delta}$$

Equivalently,

$$\mathbb{P}(t_\sigma = \infty \mid \mathcal{F}_{t_0}) \ge 1 - e^{-\nu\delta}$$

For the divergence of the leader gap, define

$$D_t := \mathbf{1}\{t < t_\sigma\}\,(W_t - \mathbb{E}[W_t \mid \mathcal{F}_t])$$

and

$$N_n := \sum_{t=t_0}^{n-1} D_t.$$

Since t_σ is a stopping time, $\{t < t_\sigma\} \in \mathcal{F}_t$, and hence $\mathbf{1}\{t < t_\sigma\}$ is $\mathcal{F}_t$-measurable. Therefore,

$$\begin{aligned}\mathbb{E}[D_t \mid \mathcal{F}_t] &= \mathbf{1}\{t < t_\sigma\}\mathbb{E}\left[W_t - \mathbb{E}[W_t \mid \mathcal{F}_t] \mid \mathcal{F}_t\right] \\ &= \mathbf{1}\{t < t_\sigma\}\left(\mathbb{E}[W_t \mid \mathcal{F}_t] - \mathbb{E}\left[\mathbb{E}[W_t \mid \mathcal{F}_t] \mid \mathcal{F}_t\right]\right) \\ &= \mathbf{1}\{t < t_\sigma\}\left(\mathbb{E}[W_t \mid \mathcal{F}_t] - \mathbb{E}[W_t \mid \mathcal{F}_t]\right) \\ &= 0\end{aligned}$$

Hence the increments D_t have zero conditional mean, and $\{N_n\}_{n \ge t_0}$ is a martingale.

Moreover, since

$$|W_t| \le (1+\lambda)M,$$

we also have

$$|\mathbb{E}[W_t \mid \mathcal{F}_t]| \le (1+\lambda)M$$

Therefore,

$$|D_t| \le 2(1+\lambda)M$$

Azuma–Hoeffding's inequality states that a martingale with increments bounded by c_t satisfies

$$\mathbb{P}(N_n - N_{t_0} \ge x) \le \exp\left(-\frac{x^2}{2\sum_{t=t_0}^{n-1} c_t^2}\right)$$

Here $N_{t_0} = 0$, $c_t = 2(1+\lambda)M$, and we choose

$$x = \frac{h}{2}(n - t_0)$$

Thus

$$\begin{aligned}\mathbb{P}\left(N_n \ge \frac{h}{2}(n-t_0)\right) &\le \exp\left(-\frac{\left[\frac{h}{2}(n-t_0)\right]^2}{2(n-t_0)[2(1+\lambda)M]^2}\right) \\ &= \exp\left(-\frac{h^2(n-t_0)}{32(1+\lambda)^2M^2}\right)\end{aligned}$$

Define the events

$$E_n := \left\{N_n \ge \frac{h}{2}(n-t_0)\right\}$$

By Azuma–Hoeffding's inequality,

$$\mathbb{P}(E_n) \le \exp\left(-\frac{h^2(n-t_0)}{32(1+\lambda)^2M^2}\right)$$

Let

$$C := \frac{h^2}{32(1+\lambda)^2M^2} > 0.$$

Then

$$\begin{aligned}\sum_{n=t_0+1}^{\infty} \mathbb{P}(E_n) &\le \sum_{n=t_0+1}^{\infty} e^{-C(n-t_0)} \\ &= \sum_{k=1}^{\infty} e^{-Ck} < \infty\end{aligned}$$

By the first Borel–Cantelli Lemma,

$$\mathbb{P}(E_n \text{ infinitely often}) = 0$$

We obtain

$$\mathbb{P}\left(\exists\, n_0 < \infty \text{ such that } \forall n \ge n_0,\ N_n < \frac{h}{2}(n-t_0)\right) = 1$$

Equivalently,

$$N_n < \frac{h}{2}(n-t_0)$$

for all sufficiently large n.

Now consider the lock-in event $\{t_\sigma = \infty\}$.

On this event the model remains wrong at every turn, so Eq. (10) holds for all $t \ge t_0$.

Iterating it gives

$$G_n \le G_{t_0} + \sum_{t=t_0}^{n-1} W_t$$

Since $G_{t_0} = -\delta$ and the stopping indicator in D_t equals one on $\{t_\sigma = \infty\}$,

$$G_n \le -\delta + \sum_{t=t_0}^{n-1} (W_t - \mathbb{E}[W_t \mid \mathcal{F}_t]) + \sum_{t=t_0}^{n-1} \mathbb{E}[W_t \mid \mathcal{F}_t]$$
$$= -\delta + N_n + \sum_{t=t_0}^{n-1} \mathbb{E}[W_t \mid \mathcal{F}_t]$$

Using

$$\mathbb{E}[W_t \mid \mathcal{F}_t] \le -h$$

at every turn,

$$G_n \le -\delta + N_n - h(n - t_0)$$

For all sufficiently large n,

$$N_n < \frac{h}{2}(n - t_0)$$

and hence

$$G_n \le -\delta - \frac{h}{2}(n - t_0)$$

So when $n \to \infty$, the right side tends to $-\infty$, so when $\{t_\sigma = \infty\}$

$$G_n \to -\infty$$

□

D.2 Proof of Theorem 4.2.

Theorem 4.2 (Recovery under uniformly corrective drift)**.** *Suppose $\lambda\beta < \underline{\kappa}$. Then, for every $t \ge 2$,*

$$\mathbb{P}(\hat{z}_t^\lambda \ne z^\star) \le (K-1) \exp\left(-\frac{(\underline{\kappa} - \lambda\beta)^2 (t-1)}{8(1+\lambda)^2 M^2}\right).$$

And the same upper bound holds for $\mathbb{P}(t_{\mathrm{rec}} > t)$ where $t_{\mathrm{rec}} := \inf\{t \ge 2 : \hat{z}_t^\lambda = z^\star\}$. Consequently, $\mathbb{P}(t_{\mathrm{rec}} < \infty) = 1$ and $\mathbb{E}[t_{\mathrm{rec}}] < \infty$.

Proof. Fix a wrong intent $z \ne z^\star$ and write

$$h := \underline{\kappa} - \lambda\beta > 0$$

For $s \ge 1$, define the one-step increment of its pairwise margin by

$$I_s(z) := \log \frac{p_{\mathrm{U},s+1}(U_{s+1} \mid z^\star)}{p_{\mathrm{U},s+1}(U_{s+1} \mid z)} + \lambda \log \frac{p_{\mathrm{A}}(A_s \mid z^\star)}{p_{\mathrm{A}}(A_s \mid z)}$$

Then

$$\Delta_t(z) = \Delta_1(z) + \sum_{s=1}^{t-1} I_s(z) \tag{11}$$

Step 1: User-side contribution. Conditional on $\mathcal{F}_s$, the future user shard U_{s+1} is generated from $p_{U,s+1}(\cdot \mid z^\star)$.

Therefore,

$$\mathbb{E}\left[\log \frac{p_{\mathrm{U},s+1}(U_{s+1} \mid z^\star)}{p_{\mathrm{U},s+1}(U_{s+1} \mid z)} \,\middle|\, \mathcal{F}_s\right] = \mathbb{E}_{U \sim p_{\mathrm{U},s+1}(\cdot \mid z^\star)}\left[\log \frac{p_{\mathrm{U},s+1}(U \mid z^\star)}{p_{\mathrm{U},s+1}(U \mid z)}\right]$$
$$= \mathrm{KL}\left(p_{\mathrm{U},s+1}(\cdot \mid z^\star) \,\|\, p_{\mathrm{U},s+1}(\cdot \mid z)\right)$$
$$\ge \underline{\kappa}$$

Step 2: Model-side contribution. Suppose first that $\hat{z}_s = z^\star$. Conditional on $\mathcal{F}_s$,

$$A_s \sim p_{\mathrm{A}}(\cdot \mid z^\star).$$

Therefore,

$$\mathbb{E}\left[\log \frac{p_{\mathrm{A}}(A_s \mid z^\star)}{p_{\mathrm{A}}(A_s \mid z)} \,\middle|\, \mathcal{F}_s\right] = \mathbb{E}_{A\sim p_{\mathrm{A}}(\cdot\mid z^\star)}\left[\log \frac{p_{\mathrm{A}}(A \mid z^\star)}{p_{\mathrm{A}}(A \mid z)}\right] = \mathrm{KL}\left(p_{\mathrm{A}}(\cdot \mid z^\star) \,\|\, p_{\mathrm{A}}(\cdot \mid z)\right) \geq 0$$

By the definition of β,

$$\beta := \max_{w, z' \neq z^\star} \left[\mathrm{KL}\left(p_{\mathrm{A}}(\cdot \mid w) \,\|\, p_{\mathrm{A}}(\cdot \mid z^\star)\right) - \mathrm{KL}\left(p_{\mathrm{A}}(\cdot \mid w) \,\|\, p_{\mathrm{A}}(\cdot \mid z')\right)\right]$$

Taking $z' = w$ in the maximization gives

$$\beta \geq \mathrm{KL}\left(p_{\mathrm{A}}(\cdot \mid w) \,\|\, p_{\mathrm{A}}(\cdot \mid z^\star)\right) \geq 0$$

Thus

$$0 \geq -\beta,$$

and therefore, when $\hat{z}_s = z^\star$,

$$\mathbb{E}\left[\log \frac{p_{\mathrm{A}}(A_s \mid z^\star)}{p_{\mathrm{A}}(A_s \mid z)} \,\middle|\, \mathcal{F}_s\right] \geq -\beta \tag{12}$$

Suppose instead that $\hat{z}_s = w \neq z^\star$. Conditional on $\mathcal{F}_s$,

$$A_s \sim p_{\mathrm{A}}(\cdot \mid w)$$

Hence

$$\mathbb{E}\left[\log \frac{p_{\mathrm{A}}(A_s \mid z^\star)}{p_{\mathrm{A}}(A_s \mid z)} \,\middle|\, \mathcal{F}_s\right] = \mathbb{E}_{A\sim p_{\mathrm{A}}(\cdot\mid w)}\left[\log \frac{p_{\mathrm{A}}(A \mid z^\star)}{p_{\mathrm{A}}(A \mid z)}\right]$$

Insert $p_{\mathrm{A}}(A \mid w)$ into the likelihood ratio:

$$\log \frac{p_{\mathrm{A}}(A \mid z^\star)}{p_{\mathrm{A}}(A \mid z)} = \log \frac{p_{\mathrm{A}}(A \mid z^\star)}{p_{\mathrm{A}}(A \mid w)} + \log \frac{p_{\mathrm{A}}(A \mid w)}{p_{\mathrm{A}}(A \mid z)}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\log \frac{p_{\mathrm{A}}(A_s \mid z^\star)}{p_{\mathrm{A}}(A_s \mid z)} \,\middle|\, \mathcal{F}_s\right] &= \mathbb{E}_{A\sim p_{\mathrm{A}}(\cdot\mid w)}\left[\log \frac{p_{\mathrm{A}}(A \mid z^\star)}{p_{\mathrm{A}}(A \mid w)}\right] + \mathbb{E}_{A\sim p_{\mathrm{A}}(\cdot\mid w)}\left[\log \frac{p_{\mathrm{A}}(A \mid w)}{p_{\mathrm{A}}(A \mid z)}\right] \\
&= -\mathrm{KL}\left(p_{\mathrm{A}}(\cdot \mid w) \,\|\, p_{\mathrm{A}}(\cdot \mid z^\star)\right) + \mathrm{KL}\left(p_{\mathrm{A}}(\cdot \mid w) \,\|\, p_{\mathrm{A}}(\cdot \mid z)\right) \\
&= -\left[\mathrm{KL}\left(p_{\mathrm{A}}(\cdot \mid w) \,\|\, p_{\mathrm{A}}(\cdot \mid z^\star)\right) - \mathrm{KL}\left(p_{\mathrm{A}}(\cdot \mid w) \,\|\, p_{\mathrm{A}}(\cdot \mid z)\right)\right] \\
&\geq -\beta
\end{aligned}$$

Hence

$$\mathbb{E}\left[\log \frac{p_{\mathrm{A}}(A_s \mid z^\star)}{p_{\mathrm{A}}(A_s \mid z)} \,\middle|\, \mathcal{F}_s\right] \geq -\beta. \tag{13}$$

Combining Eq. (12) and Eq. (13), we obtain, regardless of the current selected intent,

$$\mathbb{E}\left[\log \frac{p_{\mathrm{A}}(A_s \mid z^\star)}{p_{\mathrm{A}}(A_s \mid z)} \,\middle|\, \mathcal{F}_s\right] \geq -\beta \tag{14}$$

Step 3: Positive drift of each pairwise margin.

By linearity of conditional expectation and the definition of $I_s(z)$,

$$\mathbb{E}[I_s(z) \mid \mathcal{F}_s] = \mathbb{E}\left[\log \frac{p_{\mathrm{U,s+1}}(U_{s+1} \mid z^\star)}{p_{\mathrm{U,s+1}}(U_{s+1} \mid z)} \,\middle|\, \mathcal{F}_s\right] + \lambda \mathbb{E}\left[\log \frac{p_{\mathrm{A}}(A_s \mid z^\star)}{p_{\mathrm{A}}(A_s \mid z)} \,\middle|\, \mathcal{F}_s\right]$$

Recalling that $\lambda \geq 0$,

$$\begin{aligned}
\mathbb{E}[I_s(z) \mid \mathcal{F}_s] &\geq \underline{\kappa} + \lambda(-\beta) \\
&= \underline{\kappa} - \lambda\beta \\
&= h
\end{aligned}$$

Thus,

$$\mathbb{E}[I_s(z) \mid \mathcal{F}_s] \geq h > 0. \tag{15}$$

Step 4: Initial margin and martingale fluctuations.

By Assump. 2.1, the true intent $z^\star$ belongs to the first-turn argmax set.

Hence its first-turn score is at least that of every wrong intent z, so

$$\Delta_1(z) \geq 0$$

For competitors in the initial ambiguity set the margin is exactly zero, whereas competitors outside that set have a strictly positive margin.

Now define the centered increment

$$D_s(z) := I_s(z) - \mathbb{E}[I_s(z) \mid \mathcal{F}_s],$$

and its cumulative sum

$$N_t(z) := \sum_{s=1}^{t-1} D_s(z).$$

We verify that $\{N_t(z)\}_{t\geq 1}$ is a martingale.

The increment $I_s(z)$ depends on U_{s+1} and A_s, and is therefore measurable with respect to $\mathcal{F}_{s+1}$. Moreover,

$$\begin{aligned}\mathbb{E}[D_s(z) \mid \mathcal{F}_s] &= \mathbb{E}\left[I_s(z) - \mathbb{E}[I_s(z) \mid \mathcal{F}_s] \mid \mathcal{F}_s\right] \\ &= \mathbb{E}[I_s(z) \mid \mathcal{F}_s] - \mathbb{E}\left[\mathbb{E}[I_s(z) \mid \mathcal{F}_s] \mid \mathcal{F}_s\right]\end{aligned}$$

Since $\mathbb{E}[I_s(z) \mid \mathcal{F}_s]$ is $\mathcal{F}_s$-measurable,

$$\mathbb{E}\left[\mathbb{E}[I_s(z) \mid \mathcal{F}_s] \mid \mathcal{F}_s\right] = \mathbb{E}[I_s(z) \mid \mathcal{F}_s]$$

Therefore,

$$\mathbb{E}[D_s(z) \mid \mathcal{F}_s] = 0$$

Hence

$$\begin{aligned}\mathbb{E}[N_{t+1}(z) \mid \mathcal{F}_t] &= \mathbb{E}[N_t(z) + D_t(z) \mid \mathcal{F}_t] \\ &= N_t(z) + \mathbb{E}[D_t(z) \mid \mathcal{F}_t] \\ &= N_t(z)\end{aligned}$$

so $\{N_t(z)\}$ is a martingale.

We next bound its increments. By Assump. 2.4,

$$\left|\log \frac{p_{\mathrm{U,s+1}}(U_{s+1} \mid z^\star)}{p_{\mathrm{U,s+1}}(U_{s+1} \mid z)}\right| \leq M$$

and

$$\left|\log \frac{p_{\mathrm{A}}(A_s \mid z^\star)}{p_{\mathrm{A}}(A_s \mid z)}\right| \leq M$$

Since $\lambda \geq 0$,

$$|I_s(z)| \leq M + \lambda M = (1+\lambda)M$$

Therefore,

$$\begin{aligned}|D_s(z)| &= |I_s(z) - \mathbb{E}[I_s(z) \mid \mathcal{F}_s]| \\ &\leq |I_s(z)| + |\mathbb{E}[I_s(z) \mid \mathcal{F}_s]| \\ &\leq |I_s(z)| + \mathbb{E}[|I_s(z)| \mid \mathcal{F}_s] \\ &\leq 2(1+\lambda)M.\end{aligned}$$

Step 5: Concentration of the pairwise margin.

Using the definition of $N_t(z)$,

$$\sum_{s=1}^{t-1} I_s(z) = \sum_{s=1}^{t-1} \mathbb{E}[I_s(z) \mid \mathcal{F}_s] + N_t(z)$$

Substituting this into Eq. (11) gives

$$\Delta_t(z) = \Delta_1(z) + \sum_{s=1}^{t-1} \mathbb{E}[I_s(z) \mid \mathcal{F}_s] + N_t(z)$$

By $\Delta_1(z) \geq 0$, and by Eq. (15), for every s

$$\mathbb{E}[I_s(z) \mid \mathcal{F}_s] \geq h$$

Therefore,

$$\Delta_t(z) \geq 0 + \sum_{s=1}^{t-1} h + N_t(z) = h(t-1) + N_t(z)$$

Hence, if

$$\Delta_t(z) \leq 0$$

then necessarily

$$h(t-1) + N_t(z) \leq 0$$

or equivalently,

$$N_t(z) \leq -h(t-1)$$

Thus

$$\{\Delta_t(z) \leq 0\} \subseteq \{N_t(z) \leq -h(t-1)\} \tag{16}$$

We now apply the lower-tail form of the Azuma–Hoeffding inequality.

If a martingale has increments bounded in absolute value by c_s, then

$$\mathbb{P}\left(N_t - N_1 \leq -x\right) \leq \exp\left(-\frac{x^2}{2\sum_{s=1}^{t-1} c_s^2}\right)$$

Here,

$$N_1(z) = 0 \qquad c_s = 2(1+\lambda)M$$

and we choose

$$x = h(t-1)$$

Therefore,

$$\begin{aligned}
\mathbb{P}\left(N_t(z) \leq -h(t-1)\right) &\leq \exp\left(-\frac{[h(t-1)]^2}{2\sum_{s=1}^{t-1}[2(1+\lambda)M]^2}\right) \\
&= \exp\left(-\frac{h^2(t-1)^2}{2(t-1)4(1+\lambda)^2M^2}\right) \\
&= \exp\left(-\frac{h^2(t-1)}{8(1+\lambda)^2M^2}\right)
\end{aligned}$$

Combining this with Eq. (16),

$$\mathbb{P}(\Delta_t(z) \leq 0) \leq \exp\left(-\frac{h^2(t-1)}{8(1+\lambda)^2M^2}\right) \tag{17}$$

Step 6: From pairwise margins to terminal error.

Suppose $\hat{z}_t^\lambda \neq z^\star$. Let $z_t := \hat{z}_t^\lambda$ be the selected wrong intent. Since the estimator selects a maximizer of the score,

$$S_\lambda^{(t)}(z_t) \geq S_\lambda^{(t)}(z^\star)$$

By the definition of the pairwise margin,

$$\Delta_t(z_t) = S_\lambda^{(t)}(z^\star) - S_\lambda^{(t)}(z_t) \leq 0$$

Thus, whenever the final estimate is wrong, at least one wrong competitor has a nonpositive pairwise margin:

$$\{\hat{z}_t^\lambda \neq z^\star\} \subseteq \bigcup_{z \neq z^\star} \{\Delta_t(z) \leq 0\}$$

Applying the union bound,

$$\begin{aligned}\mathbb{P}(\hat{z}_t^\lambda \neq z^\star) &\leq \sum_{z \neq z^\star} \mathbb{P}(\Delta_t(z) \leq 0) \\ &\leq (K-1)\exp\left(-\frac{h^2(t-1)}{8(1+\lambda)^2 M^2}\right)\end{aligned}$$

Substituting $h = \underline{\kappa} - \lambda\beta$ gives

$$\mathbb{P}(\hat{z}_t^\lambda \neq z^\star) \leq (K-1)\exp\left(-\frac{(\underline{\kappa} - \lambda\beta)^2(t-1)}{8(1+\lambda)^2 M^2}\right)$$

Step 7: Almost-sure recovery and finite expected recovery time.

By the definition of t_{rec}, the event $\{t_{\mathrm{rec}} > t\}$ means that the true intent has not been selected by time t.

In particular, the estimate at time t must still be wrong. Therefore,

$$\{t_{\mathrm{rec}} > t\} \subseteq \{\hat{z}_t^\lambda \neq z^\star\}.$$

Hence

$$\begin{aligned}\mathbb{P}(t_{\mathrm{rec}} > t) &\leq \mathbb{P}(\hat{z}_t^\lambda \neq z^\star) \\ &\leq (K-1)\exp\left(-\frac{h^2(t-1)}{8(1+\lambda)^2 M^2}\right)\end{aligned}$$

The event $\{t_{\mathrm{rec}} = \infty\}$ is contained in $\{t_{\mathrm{rec}} > t\}$ for every finite t. Therefore, for every t.

$$\mathbb{P}(t_{\mathrm{rec}} = \infty) \leq \mathbb{P}(t_{\mathrm{rec}} > t)$$

Since $h > 0$, the upper bound above tends to zero as $t \to \infty$. Consequently,

$$\begin{aligned}\mathbb{P}(t_{\mathrm{rec}} = \infty) &\leq \lim_{t\to\infty} (K-1)\exp\left(-\frac{h^2(t-1)}{8(1+\lambda)^2 M^2}\right) \\ &= 0\end{aligned}$$

Thus

$$\mathbb{P}(t_{\mathrm{rec}} < \infty) = 1.$$

Finally, since t_{rec} is a nonnegative integer-valued random variable, the tail-sum formula gives

$$\mathbb{E}[t_{\mathrm{rec}}] = \sum_{t=0}^{\infty} \mathbb{P}(t_{\mathrm{rec}} > t)$$

For $t \geq 1$, the preceding bound gives

$$\mathbb{P}(t_{\mathrm{rec}} > t) \leq (K-1)e^{-C(t-1)}$$

where

$$C := \frac{h^2}{8(1+\lambda)^2 M^2} > 0$$

Therefore,

$$\begin{aligned}\mathbb{E}[t_{\mathrm{rec}}] &= \mathbb{P}(t_{\mathrm{rec}} > 0) + \sum_{t=1}^{\infty} \mathbb{P}(t_{\mathrm{rec}} > t) \\ &\leq 1 + (K-1)\sum_{t=1}^{\infty} e^{-C(t-1)} \\ &= 1 + (K-1)\sum_{j=0}^{\infty} e^{-Cj}\end{aligned}$$

Since $C > 0$, we have $0 < e^{-C} < 1$, so the geometric series converges:

$$\sum_{j=0}^{\infty} e^{-Cj} = \frac{1}{1 - e^{-C}} < \infty$$

Hence

$$\mathbb{E}[t_{\text{rec}}] < \infty$$

□

D.3 Proof of Theorem 4.3.

Theorem 4.3 (Correct-state persistence). *Fix any turn t_0 such that $\hat{z}_{t_0}^{\lambda} = z^{\star}$ and $G_{t_0} > 0$, define $t_{\text{exit}} := \inf\{t > t_0 : \hat{z}_t^{\lambda} \neq z^{\star}\}$ as the first subsequent exit from the correct state. Then*

$$\mathbb{P}\left(t_{\text{exit}} < \infty \,|\, \mathcal{F}_{t_0}\right) \leq (K-1) \exp\left(-\frac{2(\underline{\kappa} + \lambda\alpha_{\text{true}})G_{t_0}}{M^2(1+\lambda^2)}\right). \tag{5}$$

Proof. Condition on $\mathcal{F}_{t_0}$ and define $t_{\text{exit}} := \inf\{t > t_0 : \hat{z}_t^{\lambda} \neq z^{\star}\}$

Thus t_{exit} is the first time after t_0 at which the model leaves the true intent.

For every $t_0 \leq t < t_{\text{exit}}$, $\hat{z}_t = z^{\star}$.

And, we have

$$A_t \mid \mathcal{F}_t \sim p_{\text{A}}(\cdot \mid z^{\star})$$

Fix a wrong intent $z \neq z^{\star}$. Define its one-step pairwise margin increment by

$$I_t(z) := \log \frac{p_{\text{U,t+1}}(U_{t+1} \mid z^{\star})}{p_{\text{U,t+1}}(U_{t+1} \mid z)} + \lambda \log \frac{p_{\text{A}}(A_t \mid z^{\star})}{p_{\text{A}}(A_t \mid z)}$$

By the pairwise-margin recursion,

$$\Delta_{t+1}(z) = \Delta_t(z) + I_t(z) \tag{18}$$

Step 1: Positive drift while the model follows the truth.

We first compute the conditional expectation of the user contribution.

Conditional on $\mathcal{F}_t$, the next user shard U_{t+1} is generated according to $p_{\text{U,t+1}}(\cdot \mid z^{\star})$. Therefore,

$$\begin{aligned}
\mathbb{E}\left[\log \frac{p_{\text{U,t+1}}(U_{t+1} \mid z^{\star})}{p_{\text{U,t+1}}(U_{t+1} \mid z)} \,\middle|\, \mathcal{F}_t\right] &= \mathbb{E}_{U \sim p_{\text{U,t+1}}(\cdot \mid z^{\star})}\left[\log \frac{p_{\text{U,t+1}}(U \mid z^{\star})}{p_{\text{U,t+1}}(U \mid z)}\right] \\
&= \mathrm{KL}\left(p_{\text{U,t+1}}(\cdot \mid z^{\star}) \,\|\, p_{\text{U,t+1}}(\cdot \mid z)\right) \\
&\geq \underline{\kappa}
\end{aligned}$$

For the contribution, since $t < t_{\text{exit}}$, $\hat{z}_t = z^{\star}$, and hence

$$A_t \mid \mathcal{F}_t \sim p_{\text{A}}(\cdot \mid z^{\star})$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\log \frac{p_{\text{A}}(A_t \mid z^{\star})}{p_{\text{A}}(A_t \mid z)} \,\middle|\, \mathcal{F}_t\right] &= \mathbb{E}_{A \sim p_{\text{A}}(\cdot \mid z^{\star})}\left[\log \frac{p_{\text{A}}(A \mid z^{\star})}{p_{\text{A}}(A \mid z)}\right] \\
&= \mathrm{KL}\left(p_{\text{A}}(\cdot \mid z^{\star}) \,\|\, p_{\text{A}}(\cdot \mid z)\right) \\
&\geq \alpha_{\text{true}}
\end{aligned}$$

By linearity of conditional expectation,

$$\begin{aligned}
\mathbb{E}[I_t(z) \mid \mathcal{F}_t] &= \mathbb{E}\left[\log \frac{p_{\text{U,t+1}}(U_{t+1} \mid z^{\star})}{p_{\text{U,t+1}}(U_{t+1} \mid z)} \,\middle|\, \mathcal{F}_t\right] + \lambda \mathbb{E}\left[\log \frac{p_{\text{A}}(A_t \mid z^{\star})}{p_{\text{A}}(A_t \mid z)} \,\middle|\, \mathcal{F}_t\right] \\
&\geq \underline{\kappa} + \lambda\alpha_{\text{true}}.
\end{aligned}$$

For notational convenience, write

$$h := \underline{\kappa} + \lambda \alpha_{\text{true}} > 0$$

and

$$\nu := \frac{2h}{M^2(1+\lambda^2)}$$

Thus, for every $t < t_{\text{exit}}$,

$$\mathbb{E}[I_t(z) \mid \mathcal{F}_t] \geq h \tag{19}$$

Step 2: Conditional MGF bound.

By Assump. 2.4, conditional on $\mathcal{F}_t$, the user and terms in $I_t(z)$ are independent and each has absolute value at most M.

Therefore, Lem. D.1 with $s = -\nu$ gives

$$\begin{aligned}
\mathbb{E}\left[e^{-\nu I_t(z)} \,\middle|\, \mathcal{F}_t\right] &\leq \exp\left(-\nu \mathbb{E}[I_t(z) \mid \mathcal{F}_t] + \frac{\nu^2 M^2(1+\lambda^2)}{2}\right) \\
&\leq \exp\left(-\nu h + \frac{\nu^2 M^2(1+\lambda^2)}{2}\right) \\
&= 1
\end{aligned}$$

Step 3: Probability that a fixed competitor catches the truth.

Define

$$\tau_z := \inf\{t > t_0 : \Delta_t(z) \leq 0\}.$$

Thus τ_z is the first time after t_0 at which the fixed competitor z has score at least as large as the truth.

Since $\Delta_t(z)$ is determined by the history available at time t, τ_z is a stopping time. The same is true of t_{exit}.

Now define the stopped process for $t \geq t_0$

$$L_t^{(z)} := \exp\left(-\nu \Delta_{t \wedge \tau_z \wedge t_{\text{exit}}}(z)\right)$$

We verify that $\{L_t^{(z)}\}_{t \geq t_0}$ is a nonnegative supermartingale.

Consider first the event

$$\{t < \tau_z \wedge t_{\text{exit}}\}.$$

On this event neither stopping time has yet occurred. Hence

$$L_t^{(z)} = e^{-\nu \Delta_t(z)}$$

and

$$L_{t+1}^{(z)} = e^{-\nu \Delta_{t+1}(z)}$$

Using Eq. (18),

$$\begin{aligned}
\mathbb{E}\left[L_{t+1}^{(z)} \,\middle|\, \mathcal{F}_t\right] &= \mathbb{E}\left[e^{-\nu(\Delta_t(z) + I_t(z))} \,\middle|\, \mathcal{F}_t\right] \\
&= e^{-\nu \Delta_t(z)} \mathbb{E}\left[e^{-\nu I_t(z)} \,\middle|\, \mathcal{F}_t\right] \\
&\leq e^{-\nu \Delta_t(z)} \\
&= L_t^{(z)}
\end{aligned}$$

On the complementary event $\{t \geq \tau_z \wedge t_{\text{exit}}\}$, the process has already been stopped, so

$$L_{t+1}^{(z)} = L_t^{(z)}$$

Therefore, in all cases

$$\mathbb{E}\left[L_{t+1}^{(z)} \,\middle|\, \mathcal{F}_t\right] \leq L_t^{(z)}$$

Hence $\{L_t^{(z)}\}_{t\geq t_0}$ is a nonnegative supermartingale.

For every integer $n > t_0$, the supermartingale property implies

$$\begin{aligned}\mathbb{E}\left[L_n^{(z)} \,\middle|\, \mathcal{F}_{t_0}\right] &\leq L_{t_0}^{(z)} \\ &= e^{-\nu\Delta_{t_0}(z)}\end{aligned}$$

Now consider the event

$$E_{z,n} := \{\tau_z \leq t_{\text{exit}},\ \tau_z \leq n\}$$

On $E_{z,n}$, the process reaches τ_z before leaving the true-intent regime, and by the definition of τ_z,

$$\Delta_{\tau_z}(z) \leq 0$$

Because the process is stopped at τ_z, on $E_{z,n}$

$$L_n^{(z)} = e^{-\nu\Delta_{\tau_z}(z)} \geq 1$$

Therefore,

$$\mathbf{1}\{E_{z,n}\} \leq L_n^{(z)}$$

Taking conditional expectations,

$$\begin{aligned}\mathbb{P}\left(E_{z,n} \,|\, \mathcal{F}_{t_0}\right) &= \mathbb{E}\left[\mathbf{1}\{E_{z,n}\} \,|\, \mathcal{F}_{t_0}\right] \\ &\leq \mathbb{E}\left[L_n^{(z)} \,\middle|\, \mathcal{F}_{t_0}\right] \\ &\leq e^{-\nu\Delta_{t_0}(z)}\end{aligned}$$

As n increases, the events $E_{z,n}$ are increasing, and

$$\bigcup_{n>t_0} E_{z,n} = \{\tau_z \leq t_{\text{exit}},\ \tau_z < \infty\}.$$

Therefore, by continuity from below,

$$\mathbb{P}\left(\tau_z \leq t_{\text{exit}},\ \tau_z < \infty \,|\, \mathcal{F}_{t_0}\right) \leq e^{-\nu\Delta_{t_0}(z)}$$

Step 4: From a fixed competitor to leaving the true state.

Suppose $t_{\text{exit}} < \infty$. By the definition of t_{exit}, $\hat{z}_{t_{\text{exit}}} \neq z^\star$

Let

$$z_{t_{\text{exit}}} := \hat{z}_{t_{\text{exit}}}$$

Since $z_{t_{\text{exit}}}$ is selected as a score maximizer,

$$S_\lambda^{(t_{\text{exit}})}(z_{t_{\text{exit}}}) \geq S_\lambda^{(t_{\text{exit}})}(z^\star).$$

Therefore,

$$\Delta_{t_{\text{exit}}}(z_{t_{\text{exit}}}) = S_\lambda^{(t_{\text{exit}})}(z^\star) - S_\lambda^{(t_{\text{exit}})}(z_{t_{\text{exit}}}) \leq 0$$

By the definition of $\tau_{z_{t_{\text{exit}}}}$,

$$\tau_{z_{t_{\text{exit}}}} \leq t_{\text{exit}} < \infty$$

Therefore,

$$\{t_{\text{exit}} < \infty\} \subseteq \bigcup_{z\neq z^\star} \{\tau_z \leq t_{\text{exit}},\ \tau_z < \infty\}$$

Applying the union bound,

$$\begin{aligned}\mathbb{P}\left(t_{\text{exit}} < \infty \,|\, \mathcal{F}_{t_0}\right) &\leq \sum_{z\neq z^\star} \mathbb{P}\left(\tau_z \leq t_{\text{exit}},\ \tau_z < \infty \,|\, \mathcal{F}_{t_0}\right) \\ &\leq \sum_{z\neq z^\star} e^{-\nu\Delta_{t_0}(z)}\end{aligned}$$

where

$$\nu = \frac{2h}{M^2(1+\lambda^2)} = \frac{2(\kappa + \lambda\alpha_{\text{true}})}{M^2(1+\lambda^2)}$$

Finally, by the definition of the leader gap,

$$G_{t_0} = \min_{z \neq z^\star} \Delta_{t_0}(z)$$

Hence, for every wrong intent,

$$\Delta_{t_0}(z) \geq G_{t_0}$$

Since $\nu > 0$, the function $x \mapsto e^{-\nu x}$ is decreasing, and therefore

$$e^{-\nu\Delta_{t_0}(z)} \leq e^{-\nu G_{t_0}}$$

Summing over the $K-1$ wrong intents gives

$$\sum_{z \neq z^\star} e^{-\nu\Delta_{t_0}(z)} \leq (K-1)e^{-\nu G_{t_0}}$$

□

D.4 Proof of Binary Case

Corollary D.2 (Binary drift threshold)**.** *Suppose* $\mathbb{Z} = \{z^\star, z'\}$ *and for* $t \geq 2$, $\mathrm{KL}\left(p_{\mathrm{U,t}}(\cdot \mid z^\star) \,\|\, p_{\mathrm{U,t}}(\cdot \mid z')\right) = \kappa$. *Consequently, the recovery and lock-in conditions reduce to the same drift threshold:*

- *If* $\lambda\alpha_{\text{err}} < \kappa$, *then*

$$\mathbb{P}(\hat{z}_t^\lambda \neq z^\star) \leq \exp\left(-\frac{(\kappa - \lambda\alpha_{\text{err}})^2(t-1)}{8(1+\lambda)^2M^2}\right)$$

  and t_{rec} *is finite with finite expectation.*

- *If* $\lambda\alpha_{\text{err}} > \kappa$, *then, conditional on any time* t_0 *at which* $G_{t_0} = -\delta < 0$,

$$\mathbb{P}\left(t_{\text{rec}} = \infty \mid \mathcal{F}_{t_0}\right) \geq 1 - \exp\left(-\frac{2(\lambda\alpha_{\text{err}} - \kappa)\delta}{M^2(1+\lambda^2)}\right)$$

  and $G_t \to -\infty$ *on the permanent lock-in event.*

Proof. Since there is only one wrong intent z', the Assump. 2.2 is

$$\underline{\kappa} = \overline{\kappa}$$

Moreover, in the definition of β, the only possible choices of the wrong intents are $w = z = z'$. Hence

$$\begin{aligned}\beta &= \mathrm{KL}\left(p_{\mathrm{A}}(\cdot \mid z') \,\|\, p_{\mathrm{A}}(\cdot \mid z^\star)\right) - \mathrm{KL}\left(p_{\mathrm{A}}(\cdot \mid z') \,\|\, p_{\mathrm{A}}(\cdot \mid z')\right) \\ &= \mathrm{KL}\left(p_{\mathrm{A}}(\cdot \mid z') \,\|\, p_{\mathrm{A}}(\cdot \mid z^\star)\right) \\ &= \alpha_{\text{err}}\end{aligned}$$

where the second equality uses $\mathrm{KL}(P\|P) = 0$, and the last equality follows because z' is the unique wrong intent.

Therefore, the condition of Thm. 4.2,

$$\lambda\beta < \underline{\kappa}$$

reduces to

$$\lambda\alpha_{\text{err}} < \kappa$$

Substituting $\beta = \alpha_{\text{err}}$ and $\underline{\kappa} = \kappa$ into the bound of Thm. 4.2 gives

$$\mathbb{P}(\hat{z}_t^\lambda \neq z^\star) \leq \exp\left(-\frac{(\kappa - \lambda\alpha_{\text{err}})^2(t-1)}{8(1+\lambda)^2M^2}\right)$$

and the same theorem gives almost-sure recovery with finite expected recovery time.

Similarly, the condition of Thm. 4.1,

$$\lambda\alpha_{\mathrm{err}} > \overline{\kappa}$$

reduces, since $\overline{\kappa} = \kappa$, to

$$\lambda\alpha_{\mathrm{err}} > \kappa$$

Substituting $\overline{\kappa} = \kappa$ into Thm. 4.1 gives the stated lock-in probability and the divergence of G_t on the permanent lock-in event.

Thus the strict recovery and lock-in regimes are separated by the common boundary

$$\lambda\alpha_{\mathrm{err}} = \kappa$$

Neither preceding theorem covers the equality case. □

E Proofs for Section 5

E.1 Proof of Theorem 5.1

Theorem 5.1 (Persistent degradation floor)**.** *If $\lambda\alpha_{\mathrm{err}} > \overline{\kappa}$, then $p_{\mathrm{lock}}(\lambda) > 0$, and for every $T \geq 2$,*

$$\mathbb{P}\big(\hat{z}_T^{\lambda} \neq z^\star\big) - \mathbb{P}\big(\hat{z}^{\mathrm{single}} \neq z^\star\big) \geq \left(1 - \frac{1}{m}\right) p_{\mathrm{lock}}(\lambda) - (K-1)\exp\left(-\frac{(T-1)\underline{\kappa}^2}{2M^2}\right). \tag{7}$$

Consequently, $\liminf_{T\to\infty}\left(\mathbb{P}\big(\hat{z}_T^{\lambda} \neq z^\star\big) - \mathbb{P}\big(\hat{z}^{\mathrm{single}} \neq z^\star\big)\right) \geq \left(1 - \frac{1}{m}\right) p_{\mathrm{lock}}(\lambda) > 0.$

Proof. We first show that $p_{\mathrm{lock}}(\lambda) > 0$.

Step 1: A wrong first-turn seed enters a negative-gap state with positive probability.

Condition on a wrong first-turn estimate

$$\hat{z}_1^{\lambda} = w \neq z^\star.$$

By Assump. 2.1, both w and $z^\star$ belong to the first-turn argmax set. Hence they have equal first-turn scores, so

$$\Delta_1(w) = 0$$

Using the pairwise-margin recursion from turn one to turn two,

$$\Delta_2(w) = \Delta_1(w) + \log\frac{p_{\mathrm{U},2}(U_2 \mid z^\star)}{p_{\mathrm{U},2}(U_2 \mid w)} + \lambda\log\frac{p_{\mathrm{A}}(A_1 \mid z^\star)}{p_{\mathrm{A}}(A_1 \mid w)}$$

So

$$\Delta_2(w) = \log\frac{p_{\mathrm{U},2}(U_2 \mid z^\star)}{p_{\mathrm{U},2}(U_2 \mid w)} + \lambda\log\frac{p_{\mathrm{A}}(A_1 \mid z^\star)}{p_{\mathrm{A}}(A_1 \mid w)}$$

We now compute its conditional expectation.

The event $\{\hat{z}_1^{\lambda} = w\}$ is determined by the first-turn history and tie-breaking, while the next user shard is generated from the true intent.

Hence

$$U_2 \,|\, \{\hat{z}_1^{\lambda} = w\} \sim p_{\mathrm{U},2}(\cdot \mid z^\star)$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\log\frac{p_{\mathrm{U},2}(U_2 \mid z^\star)}{p_{\mathrm{U},2}(U_2 \mid w)} \,\middle|\, \hat{z}_1^{\lambda} = w\right] &= \mathbb{E}_{U\sim p_{\mathrm{U},2}(\cdot|z^\star)}\left[\log\frac{p_{\mathrm{U},2}(U \mid z^\star)}{p_{\mathrm{U},2}(U \mid w)}\right] \\ &= \mathrm{KL}\left(p_{\mathrm{U},2}(\cdot \mid z^\star) \,\|\, p_{\mathrm{U},2}(\cdot \mid w)\right) \\ &\leq \overline{\kappa}\end{aligned}$$

Since the first-turn estimate is w, the assistant response is generated according to

$$A_1 \,|\, \{\hat{z}_1^{\lambda} = w\} \sim p_{\mathrm{A}}(\cdot \mid w)$$

Thus,

$$\begin{aligned}\mathbb{E}\left[\log \frac{p_{\mathrm{A}}(A_1 \mid z^\star)}{p_{\mathrm{A}}(A_1 \mid w)} \,\middle|\, \hat{z}_1^\lambda = w\right] &= \mathbb{E}_{A\sim p_{\mathrm{A}}(\cdot\mid w)}\left[\log \frac{p_{\mathrm{A}}(A \mid z^\star)}{p_{\mathrm{A}}(A \mid w)}\right] \\ &= -\,\mathrm{KL}\left(p_{\mathrm{A}}(\cdot \mid w) \,\|\, p_{\mathrm{A}}(\cdot \mid z^\star)\right) \\ &\le -\alpha_{\mathrm{err}}.\end{aligned}$$

By linearity of conditional expectation,

$$\begin{aligned}\mathbb{E}\left[\Delta_2(w) \,\middle|\, \hat{z}_1^\lambda = w\right] &= \mathrm{KL}\left(p_{\mathrm{U},2}(\cdot \mid z^\star) \,\|\, p_{\mathrm{U},2}(\cdot \mid w)\right) - \lambda\,\mathrm{KL}\left(p_{\mathrm{A}}(\cdot \mid w) \,\|\, p_{\mathrm{A}}(\cdot \mid z^\star)\right) \\ &\le \overline{\kappa} - \lambda\alpha_{\mathrm{err}} \\ &< 0\end{aligned}$$

where the final inequality follows from the supercritical condition $\lambda\alpha_{\mathrm{err}} > \overline{\kappa}$

We next show that the strict negativity implies that $\Delta_2(w) < 0$ with positive probability. Suppose, to the contrary, that

$$\mathbb{P}\left(\Delta_2(w) < 0 \,\middle|\, \hat{z}_1^\lambda = w\right) = 0$$

Then

$$\Delta_2(w) \ge 0$$

Then $\Delta_2(w) \ge 0$ conditional on $\hat{z}_1^\lambda = w$.

Since Assump. 2.4 guarantees integrability, this would imply

$$\mathbb{E}\left[\Delta_2(w) \,\middle|\, \hat{z}_1^\lambda = w\right] \ge 0$$

contradicting $\mathbb{E}\left[\Delta_2(w) \,\middle|\, \hat{z}_1^\lambda = w\right] < 0$. Therefore,

$$\mathbb{P}\left(\Delta_2(w) < 0 \,\middle|\, \hat{z}_1^\lambda = w\right) > 0$$

By definition,

$$G_2 = \min_{z \ne z^\star} \Delta_2(z)$$

In particular,

$$G_2 \le \Delta_2(w)$$

Hence

$$\{\Delta_2(w) < 0\} \subseteq \{G_2 < 0\},$$

and consequently

$$\mathbb{P}\left(G_2 < 0 \,\middle|\, \hat{z}_1^\lambda = w\right) > 0$$

This argument applies to every wrong intent that can be selected from the initial ambiguity set.

Averaging over the possible wrong first-turn estimates therefore gives

$$\mathbb{P}\left(G_2 < 0 \,\middle|\, \hat{z}_1^\lambda \ne z^\star\right) > 0 \tag{20}$$

Step 2: The conditional lock-in mass is strictly positive.

Under the supercritical condition, define

$$c_{\mathrm{lock}} := \frac{2(\lambda\alpha_{\mathrm{err}} - \overline{\kappa})}{M^2(1+\lambda^2)} > 0.$$

Then the integrand in Eq. (6) can be written as

$$\left(1 - e^{-c_{\mathrm{lock}}(-G_2)}\right)\mathbf{1}\{G_2 < 0\}.$$

Whenever $G_2 < 0$, we have $-G_2 > 0$, and therefore

$$0 < e^{-c_{\mathrm{lock}}(-G_2)} < 1$$

Hence

$$1 - e^{-c_{\text{lock}}(-G_2)} > 0$$

on the event $\{G_2 < 0\}$.

By Eq. (20), this event has strictly positive probability conditional on a wrong first-turn estimate.

Therefore the nonnegative integrand in Eq. (6) is strictly positive on an event of positive conditional probability, which implies

$$p_{\text{lock}}(\lambda) > 0$$

Step 3: Permanent lock-in creates a terminal-error floor.

By Assump. 2.1 and uniform tie-breaking,

$$\Pr\left(\hat{z}_1^\lambda \neq z^\star\right) = 1 - \frac{1}{m}$$

Now condition on a wrong first-turn estimate.

On a realization with $G_2 < 0$, the true intent cannot be a score maximizer at turn two, because

$$G_2 < 0$$

means that at least one wrong intent has strictly larger score than $z^\star$.

Hence

$$\hat{z}_2^\lambda \neq z^\star$$

Applying Thm. 4.1 at $t_0 = 2$ gives, conditional on $\mathcal{F}_2$, on $\{G_2 < 0\}$

$$\mathbb{P}\left(\hat{z}_t^\lambda \neq z^\star \text{ for all } t \geq 2 \,\middle|\, \mathcal{F}_2\right) \geq 1 - \exp\left(-\frac{2(\lambda\alpha_{\text{err}} - \overline{\kappa})(-G_2)}{M^2(1+\lambda^2)}\right)$$

Outside $\{G_2 < 0\}$ we use the trivial lower bound zero.

Hence, for every realization,

$$\mathbb{P}\left(\hat{z}_t^\lambda \neq z^\star \text{ for all } t \geq 2 \,\middle|\, \mathcal{F}_2\right) \geq \left[1 - \exp\left(-\frac{2(\lambda\alpha_{\text{err}} - \overline{\kappa})(-G_2)}{M^2(1+\lambda^2)}\right)\right] \mathbf{1}\{G_2 < 0\}$$

Taking conditional expectation with respect to the wrong first-turn seed and using the tower property,

$$\begin{aligned}
\mathbb{P}\left(\hat{z}_t^\lambda \neq z^\star \text{ for all } t \geq 2 \,\middle|\, \hat{z}_1^\lambda \neq z^\star\right) &= \mathbb{E}\left[\mathbb{P}\left(\hat{z}_t^\lambda \neq z^\star \text{ for all } t \geq 2 \,\middle|\, \mathcal{F}_2\right) \,\middle|\, \hat{z}_1^\lambda \neq z^\star\right] \\
&\geq \mathbb{E}\left[\left(1 - \exp\left\{-\frac{2(\lambda\alpha_{\text{err}} - \overline{\kappa})(-G_2)}{M^2(1+\lambda^2)}\right\}\right) \mathbf{1}\{G_2 < 0\} \,\middle|\, \hat{z}_1^\lambda \neq z^\star\right] \\
&= p_{\text{lock}}(\lambda).
\end{aligned}$$

Multiplying by the probability of a wrong first-turn estimate gives

$$\begin{aligned}
&\mathbb{P}\left(\hat{z}_1^\lambda \neq z^\star,\ \hat{z}_t^\lambda \neq z^\star \text{ for all } t \geq 2\right) \\
&= \mathbb{P}\left(\hat{z}_1^\lambda \neq z^\star\right) \mathbb{P}\left(\hat{z}_t^\lambda \neq z^\star \text{ for all } t \geq 2 \,\middle|\, \hat{z}_1^\lambda \neq z^\star\right) \\
&\geq \left(1 - \frac{1}{m}\right) p_{\text{lock}}(\lambda)
\end{aligned}$$

Every trajectory in this event is wrong at every terminal horizon $T \geq 2$. Therefore,

$$\begin{aligned}
\mathbb{P}\left(\hat{z}_T^\lambda \neq z^\star\right) &\geq \mathbb{P}\left(\hat{z}_1^\lambda \neq z^\star,\ \hat{z}_t^\lambda \neq z^\star \text{ for all } t \geq 2\right) \\
&\geq \left(1 - \frac{1}{m}\right) p_{\text{lock}}(\lambda)
\end{aligned}$$

Step 4: From the terminal-error floor to degradation. Thm. 3.1 gives

$$\mathbb{P}\left(\hat{z}_{\text{single}} \neq z^\star\right) \leq (K-1)\exp\left(-\frac{(T-1)\underline{\kappa}^2}{2M^2}\right)$$

Combining this,

$$\mathbb{P}\left(\hat{z}_T^\lambda \neq z^\star\right) - \mathbb{P}\left(\hat{z}_{\text{single}} \neq z^\star\right) \geq \left(1-\frac{1}{m}\right)p_{\text{lock}}(\lambda) - (K-1)\exp\left(-\frac{(T-1)\underline{\kappa}^2}{2M^2}\right)$$

which proves Eq. (7).

Finally, as $T \to \infty$

$$(K-1)\exp\left(-\frac{(T-1)\underline{\kappa}^2}{2M^2}\right) \longrightarrow 0$$

Hence

$$\liminf_{T\to\infty}\left(\mathbb{P}\left(\hat{z}_T^\lambda \neq z^\star\right) - \mathbb{P}\left(\hat{z}_{\text{single}} \neq z^\star\right)\right) \geq \left(1-\frac{1}{m}\right)p_{\text{lock}}(\lambda) > 0$$

□

E.2 Proof of Theorem 5.2

Theorem 5.2 (Persistent unreliability). *Suppose $\lambda\alpha_{\text{err}} > \overline{\kappa}$ and there exists $\gamma > 0$ such that*

$$r_\gamma := \mathbb{P}\left(G_2 \geq \gamma \,\middle|\, \hat{z}_1^\lambda = z^\star\right) > 0, \quad \textit{and } (K-1)\exp\left(-\frac{2(\underline{\kappa}+\lambda\alpha_{\text{true}})\gamma}{M^2(1+\lambda^2)}\right) < 1. \tag{8}$$

Define $q_L := \left(1-\frac{1}{m}\right)p_{\text{lock}}(\lambda)$, $q_U := 1 - \frac{r_\gamma}{m}\left[1-(K-1)\exp\left(-\frac{2(\underline{\kappa}+\lambda\alpha_{\text{true}})\gamma}{M^2(1+\lambda^2)}\right)\right]$. Then, $\forall T \geq 2$, we have $q_L \leq \mathbb{P}\left(\hat{z}_T^\lambda \neq z^\star\right) \leq q_U$. Moreover, $0 < q_L \leq q_U < 1$, and therefore

$$\inf_{T\geq 2}\operatorname{Var}\left(\mathbf{1}\{\hat{z}_T^\lambda \neq z^\star\}\right) \geq \min\{q_L(1-q_L), q_U(1-q_U)\} > 0.$$

Proof. **Step 1: A positive mass of trajectories remains permanently wrong.**

In Thm. 5.1, for every $T \geq 2$,

$$\mathbb{P}\left(\hat{z}_T^\lambda \neq z^\star\right) \geq \left(1-\frac{1}{m}\right)p_{\text{lock}}(\lambda)$$

By the definition of q_L,

$$q_L = \left(1-\frac{1}{m}\right)p_{\text{lock}}(\lambda)$$

and therefore

$$\mathbb{P}\left(\hat{z}_T^\lambda \neq z^\star\right) \geq q_L \tag{21}$$

Step 2: A positive mass of trajectories remains permanently correct.

By Assump. 2.1,

$$\mathbb{P}\left(\hat{z}_1^\lambda = z^\star\right) = \frac{1}{m}$$

By the definition of r_γ,

$$\mathbb{P}\left(G_2 \geq \gamma \,\middle|\, \hat{z}_1^\lambda = z^\star\right) = r_\gamma$$

Hence, by the multiplication rule for conditional probabilities,

$$\begin{aligned}\mathbb{P}\left(\hat{z}_1^\lambda = z^\star,\, G_2 \geq \gamma\right) &= \mathbb{P}\left(\hat{z}_1^\lambda = z^\star\right)\mathbb{P}\left(G_2 \geq \gamma \,\middle|\, \hat{z}_1^\lambda = z^\star\right) \\ &= \frac{r_\gamma}{m}\end{aligned}$$

Now consider a realization satisfying $G_2 \geq \gamma > 0$.

By definition,

$$G_2 = \min_{z \neq z^\star} \Delta_2(z).$$

Thus, for every wrong intent z,

$$\Delta_2(z) \geq G_2 \geq \gamma > 0$$

Since

$$\Delta_2(z) = S_\lambda^{(2)}(z^\star) - S_\lambda^{(2)}(z)$$

we obtain for every $z \neq z^\star$

$$S_\lambda^{(2)}(z^\star) > S_\lambda^{(2)}(z)$$

Hence the true intent is the unique maximizer at turn two: $\hat{z}_2^\lambda = z^\star$.

Therefore apply Thm. 4.3 at $t_0 = 2$. Conditional on $\mathcal{F}_2$,

$$\mathbb{P}\left(\hat{z}_t^\lambda = z^\star \text{ for all } t \geq 2 \,\middle|\, \mathcal{F}_2\right) \geq 1 - (K-1)\exp\left(-\frac{2(\underline{\kappa} + \lambda\alpha_{\text{true}})G_2}{M^2(1+\lambda^2)}\right)$$

On the event $\{G_2 \geq \gamma\}$, we have $G_2 \geq \gamma$.

Since $x \mapsto \exp\left(-\frac{2(\underline{\kappa}+\lambda\alpha_{\text{true}})x}{M^2(1+\lambda^2)}\right)$ is decreasing in x, it follows that

$$\mathbb{P}\left(\hat{z}_t^\lambda = z^\star \text{ for all } t \geq 2 \,\middle|\, \mathcal{F}_2\right) \geq 1 - (K-1)\exp\left(-\frac{2(\underline{\kappa} + \lambda\alpha_{\text{true}})\gamma}{M^2(1+\lambda^2)}\right)$$

whenever $\hat{z}_1^\lambda = z^\star$ and $G_2 \geq \gamma$.

Let $\mathcal{C} := \left\{\hat{z}_t^\lambda = z^\star \text{ for all } t \geq 2\right\}$.

Using the tower property,

$$\mathbb{P}(\mathcal{C}) = \mathbb{E}\left[\mathbb{P}\left(\mathcal{C} \,|\, \mathcal{F}_2\right)\right]$$

Since conditional probabilities are nonnegative, restricting the expectation to the event

$$\left\{\hat{z}_1^\lambda = z^\star,\, G_2 \geq \gamma\right\}$$

gives

$$\mathbb{P}(\mathcal{C}) \geq \mathbb{E}\Big[\mathbf{1}\{\hat{z}_1^\lambda = z^\star,\, G_2 \geq \gamma\}\mathbb{P}\left(\mathcal{C} \,|\, \mathcal{F}_2\right)\Big]$$

Using the uniform lower bound above,

$$\mathbb{P}(\mathcal{C}) \geq \mathbb{E}\left[\mathbf{1}\{\hat{z}_1^\lambda = z^\star,\, G_2 \geq \gamma\} \times \left[1 - (K-1)\exp\left(-\frac{2(\underline{\kappa} + \lambda\alpha_{\text{true}})\gamma}{M^2(1+\lambda^2)}\right)\right]\right]$$

The term in square brackets is deterministic, so it can be taken outside the expectation:

$$\mathbb{P}(\mathcal{C}) \geq \mathbb{P}\left(\hat{z}_1^\lambda = z^\star,\, G_2 \geq \gamma\right) \times \left[1 - (K-1)\exp\left(-\frac{2(\underline{\kappa} + \lambda\alpha_{\text{true}})\gamma}{M^2(1+\lambda^2)}\right)\right]$$

Substituting

$$\mathbb{P}\left(\hat{z}_1^\lambda = z^\star,\, G_2 \geq \gamma\right) = \frac{r_\gamma}{m}$$

we obtain

$$\mathbb{P}(\mathcal{C}) \geq \frac{r_\gamma}{m}\left[1 - (K-1)\exp\left(-\frac{2(\underline{\kappa} + \lambda\alpha_{\text{true}})\gamma}{M^2(1+\lambda^2)}\right)\right] \tag{22}$$

Step 3: The persistent correct mass gives an upper bound on terminal error.

Every trajectory in $\mathcal{C}$ is correct at every terminal horizon $T \geq 2$. Hence

$$\mathcal{C} \subseteq \{\hat{z}_T^\lambda = z^\star\}$$

Therefore,

$$\begin{aligned}\mathbb{P}\left(\hat{z}_T^\lambda = z^\star\right) &\geq \mathbb{P}(\mathcal{C}) \\ &\geq \frac{r_\gamma}{m}\left[1 - (K-1)\exp\left(-\frac{2(\underline{\kappa} + \lambda\alpha_{\text{true}})\gamma}{M^2(1+\lambda^2)}\right)\right]\end{aligned}$$

Taking complements,

$$\begin{aligned}\mathbb{P}\left(\hat{z}_T^\lambda \neq z^\star\right) &= 1 - \mathbb{P}\left(\hat{z}_T^\lambda = z^\star\right) \\ &\leq 1 - \frac{r_\gamma}{m}\left[1 - (K-1)\exp\left(-\frac{2(\underline{\kappa} + \lambda\alpha_{\text{true}})\gamma}{M^2(1+\lambda^2)}\right)\right] \\ &= q_U\end{aligned}$$

Together with Eq. (21), this proves that for every $T \geq 2$,

$$q_L \leq \mathbb{P}\left(\hat{z}_T^\lambda \neq z^\star\right) \leq q_U$$

Step 4: Both endpoints lie strictly inside $[0, 1]$.

By Thm. 5.1,

$$p_{\text{lock}}(\lambda) > 0$$

Since $m \geq 2$,

$$1 - \frac{1}{m} > 0,$$

and therefore

$$q_L = \left(1 - \frac{1}{m}\right) p_{\text{lock}}(\lambda) > 0$$

Moreover, the integrand defining $p_{\text{lock}}(\lambda)$ lies in $[0, 1]$, so

$$p_{\text{lock}}(\lambda) \leq 1.$$

Hence

$$q_L \leq 1 - \frac{1}{m} < 1$$

Next, Eq. (8) gives

$$1 - (K-1)\exp\left(-\frac{2(\underline{\kappa} + \lambda\alpha_{\text{true}})\gamma}{M^2(1+\lambda^2)}\right) > 0$$

Therefore, the quantity subtracted from 1 in the definition of q_U is strictly positive, and hence

$$q_U < 1$$

Finally, since for every $T \geq 2$,

$$q_L \leq \mathbb{P}\left(\hat{z}_T^\lambda \neq z^\star\right) \leq q_U$$

we necessarily have

$$q_L \leq q_U$$

Because $q_L > 0$, it follows that

$$q_U \geq q_L > 0$$

Thus,

$$0 < q_L \leq q_U < 1$$

Step 5: Bistability yields a positive variance floor.

Recall that

$$E_T^\lambda := \mathbf{1}\{\hat{z}_T^\lambda \neq z^\star\}$$

This is a Bernoulli random variable with success probability $\mathbb{P}\left(\hat{z}_T^\lambda \neq z^\star\right)$

Therefore,

$$\text{Var}(E_T^\lambda) = \mathbb{P}\left(\hat{z}_T^\lambda \neq z^\star\right)\left[1 - \Pr\left(\hat{z}_T^\lambda \neq z^\star\right)\right]$$

Define

$$f(x) := x(1-x)$$

Since

$$f''(x) = -2 < 0,$$

the function f is concave on $[0,1]$. Therefore, its minimum over the interval $[q_L, q_U]$ is attained at one of the two endpoints, for every $x \in [q_L, q_U]$

$$f(x) \geq \min\{f(q_L), f(q_U)\}$$

Since

$$q_L \leq \mathbb{P}\left(\hat{z}_T^{\lambda} \neq z^{\star}\right) \leq q_U$$

we obtain for every $T \geq 2$

$$V_T(\lambda) \geq \min\{q_L(1-q_L), q_U(1-q_U)\}$$

Because

$$0 < q_L \leq q_U < 1,$$

both $q_L(1-q_L) > 0$ and $q_U(1-q_U) > 0$.

Hence

$$\min\{q_L(1-q_L), q_U(1-q_U)\} > 0$$

Since the same lower bound holds for every $T \geq 2$, taking the infimum over T gives

$$\inf_{T \geq 2} \mathrm{Var}(E_T^{\lambda}) \geq \min\{q_L(1-q_L), q_U(1-q_U)\} > 0$$

□

F Proofs for Section 6

F.1 Proof of Theorem F.1.

Theorem F.1 (Deferred commitment suppresses bad seeds)**.** *Let G_{r+1}^{def} be the leader gap. For any $0 < \gamma < r\underline{\kappa}$, $\mathbb{P}\left(G_{r+1}^{\mathrm{def}} \leq \gamma\right) \leq (K-1)\exp\left(-\frac{(r\underline{\kappa}-\gamma)^2}{2rM^2}\right)$.Moreover, for every terminal horizon $T > r+1$,*

$$\mathbb{P}\left(\hat{z}_T^{\mathrm{def}} \neq z^{\star}\right) \leq (K-1)\exp\left(-\frac{(r\underline{\kappa}-\gamma)^2}{2rM^2}\right) + (K-1)\exp\left(-\frac{2(\underline{\kappa}+\lambda\alpha_{\mathrm{true}})\gamma}{M^2(1+\lambda^2)}\right) \tag{23}$$

In particular, taking $\gamma = \frac{r\underline{\kappa}}{2}$ and defining $c_{\mathrm{def}} := \min\left\{\frac{\underline{\kappa}^2}{8M^2}, \frac{\underline{\kappa}(\underline{\kappa}+\lambda\alpha_{\mathrm{true}})}{M^2(1+\lambda^2)}\right\}$, we obtain $\mathbb{P}\left(\hat{z}_T^{\mathrm{def}} \neq z^{\star}\right) \leq 2(K-1)e^{-c_{\mathrm{def}} r}$. Hence the terminal error probability decays exponentially in the deferred-commitment horizon r.

Proof. For this modification, the score is

$$S_{\mathrm{def}}^{(t)}(z) := \log \pi(z) + \sum_{s=1}^{t} \log p_{\mathrm{U,s}}(U_s \mid z) + \lambda \sum_{s=r+1}^{t-1} \log p_{\mathrm{A}}(A_s \mid z),$$

Fix a wrong intent $z \neq z^{\star}$.

Because the first r assistant responses are excluded from the intent score, the pairwise margin at turn $r+1$ contains only the first-turn margin and the r post-seed user contributions:

$$\Delta_{r+1}^{\mathrm{def}}(z) = \Delta_1(z) + \sum_{s=2}^{r+1} \log \frac{p_{\mathrm{U,s}}(U_s \mid z^{\star})}{p_{\mathrm{U,s}}(U_s \mid z)}$$

By Assump. 2.1,

$$\Delta_1(z) \geq 0 \tag{24}$$

For $s = 2, \ldots, r+1$, define

$$X_s(z) := \log \frac{p_{\mathrm{U,s}}(U_s \mid z^{\star})}{p_{\mathrm{U,s}}(U_s \mid z)}$$

Since the post-seed user shards are independent under the fixed true intent $z^\star$,

$$\begin{aligned}\mathbb{E}[X_s(z)] &= \mathbb{E}_{U\sim p_{\mathrm{U,s}}(\cdot\mid z^\star)}\left[\log\frac{p_{\mathrm{U,s}}(U\mid z^\star)}{p_{\mathrm{U,s}}(U\mid z)}\right]\\ &= \mathrm{KL}\left(p_{\mathrm{U,s}}(\cdot\mid z^\star)\,\|\,p_{\mathrm{U,s}}(\cdot\mid z)\right)\\ &\geq \underline{\kappa}\end{aligned}$$

where the final inequality follows from Assump. 2.2.

Hence

$$\sum_{s=2}^{r+1}\mathbb{E}[X_s(z)] \geq r\underline{\kappa} \tag{25}$$

Assump. 2.4 also gives $|X_s(z)| \leq M$.

Fix $0 < \gamma < r\underline{\kappa}$. Suppose $\Delta^{\mathrm{def}}_{r+1}(z) \leq \gamma$.

Using Eq. (24),

$$\sum_{s=2}^{r+1} X_s(z) = \Delta^{\mathrm{def}}_{r+1}(z) - \Delta_1(z) \leq \gamma$$

Therefore,

$$\begin{aligned}\sum_{s=2}^{r+1}\left(X_s(z)-\mathbb{E}[X_s(z)]\right) &= \sum_{s=2}^{r+1}X_s(z) - \sum_{s=2}^{r+1}\mathbb{E}[X_s(z)]\\ &\leq \gamma - r\underline{\kappa}\\ &= -(r\underline{\kappa}-\gamma).\end{aligned}$$

Thus,

$$\left\{\Delta^{\mathrm{def}}_{r+1}(z)\leq\gamma\right\} \subseteq \left\{\sum_{s=2}^{r+1}\left(X_s(z)-\mathbb{E}[X_s(z)]\right) \leq -(r\underline{\kappa}-\gamma)\right\} \tag{26}$$

Since the $X_s(z)$ are independent and each lies in $[-M, M]$, Hoeffding's inequality gives

$$\begin{aligned}\mathbb{P}\left(\sum_{s=2}^{r+1}\left(X_s(z)-\mathbb{E}[X_s(z)]\right)\leq -(r\underline{\kappa}-\gamma)\right) &\leq \exp\left(-\frac{2(r\underline{\kappa}-\gamma)^2}{r(2M)^2}\right)\\ &= \exp\left(-\frac{(r\underline{\kappa}-\gamma)^2}{2rM^2}\right)\end{aligned}$$

Combining this with Eq. (26),

$$\mathbb{P}\left(\Delta^{\mathrm{def}}_{r+1}(z)\leq\gamma\right) \leq \exp\left(-\frac{(r\underline{\kappa}-\gamma)^2}{2rM^2}\right)$$

Now recall that

$$G^{\mathrm{def}}_{r+1} = \min_{z\neq z^\star}\Delta^{\mathrm{def}}_{r+1}(z).$$

If $G^{\mathrm{def}}_{r+1} \leq \gamma$, then at least one wrong intent z must satisfy

$$\Delta^{\mathrm{def}}_{r+1}(z) \leq \gamma$$

Therefore,

$$\left\{G^{\mathrm{def}}_{r+1}\leq\gamma\right\} \subseteq \bigcup_{z\neq z^\star}\left\{\Delta^{\mathrm{def}}_{r+1}(z)\leq\gamma\right\}$$

Taking a union bound over the $K-1$ wrong intents gives

$$\mathbb{P}\left(G^{\mathrm{def}}_{r+1}\leq\gamma\right) \leq (K-1)\exp\left(-\frac{(r\underline{\kappa}-\gamma)^2}{2rM^2}\right) \tag{27}$$

Now define

$$\mathcal{E}_\gamma := \{G_{r+1}^{\mathrm{def}} > \gamma\}$$

On $\mathcal{E}_\gamma$, for every wrong intent z

$$\Delta_{r+1}^{\mathrm{def}}(z) > \gamma > 0$$

Hence for every $z \neq z^\star$

$$S_{\mathrm{def}}^{(r+1)}(z^\star) > S_{\mathrm{def}}^{(r+1)}(z)$$

so the true intent is the unique maximizer:

$$\hat{z}_{r+1}^{\mathrm{def}} = z^\star$$

From turn $r+1$ onward, future assistant responses again enter the score with weight λ.

Therefore, for every $t \geq r+1$ and every wrong intent z, the deferred pairwise margin satisfies

$$\Delta_{t+1}^{\mathrm{def}}(z) = \Delta_t^{\mathrm{def}}(z) + \log \frac{p_{U,t+1}(U_{t+1} \mid z^\star)}{p_{U,t+1}(U_{t+1} \mid z)} + \lambda \log \frac{p_{\mathrm{A}}(A_t \mid z^\star)}{p_{\mathrm{A}}(A_t \mid z)}$$

This is exactly the same one-step margin recursion used in Thm. 4.3. Thus that theorem can be applied from the state at turn $r+1$.

On $\mathcal{E}_\gamma$,

$$G_{r+1}^{\mathrm{def}} > \gamma$$

so Thm. 4.3 gives

$$\begin{aligned}
&\mathbb{P}\left(\text{the process ever leaves } z^\star \mid \mathcal{F}_{r+1}\right) \\
&\leq (K-1)\exp\left(-\frac{2(\underline{\kappa} + \lambda\alpha_{\mathrm{true}})G_{r+1}^{\mathrm{def}}}{M^2(1+\lambda^2)}\right) \\
&\leq (K-1)\exp\left(-\frac{2(\underline{\kappa} + \lambda\alpha_{\mathrm{true}})\gamma}{M^2(1+\lambda^2)}\right)
\end{aligned}$$

We next convert this conditional persistence guarantee into a terminal-error bound.

For any $T > r+1$, if $\hat{z}_T^{\mathrm{def}} \neq z^\star$, then either $\mathcal{E}_\gamma$ failed at turn $r+1$, or $\mathcal{E}_\gamma$ occurred and the process subsequently left the true state.

Therefore,

$$\left\{\hat{z}_T^{\mathrm{def}} \neq z^\star\right\} \subseteq \mathcal{E}_\gamma^c \cup \left(\mathcal{E}_\gamma \cap \{\text{the process ever leaves } z^\star\}\right)$$

Taking probabilities,

$$\mathbb{P}\left(\hat{z}_T^{\mathrm{def}} \neq z^\star\right) \leq \mathbb{P}(\mathcal{E}_\gamma^c) + \mathbb{P}\left(\mathcal{E}_\gamma \cap \{\text{the process ever leaves } z^\star\}\right)$$

Since $\mathcal{E}_\gamma \in \mathcal{F}_{r+1}$, the tower property gives

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}_\gamma \cap \{\text{the process ever leaves } z^\star\}\right) &= \mathbb{E}\left[\mathbf{1}\{\mathcal{E}_\gamma\}\mathbb{P}\left(\text{the process ever leaves } z^\star \mid \mathcal{F}_{r+1}\right)\right] \\
&\leq (K-1)\exp\left(-\frac{2(\underline{\kappa} + \lambda\alpha_{\mathrm{true}})\gamma}{M^2(1+\lambda^2)}\right)\mathbb{P}(\mathcal{E}_\gamma) \\
&\leq (K-1)\exp\left(-\frac{2(\underline{\kappa} + \lambda\alpha_{\mathrm{true}})\gamma}{M^2(1+\lambda^2)}\right)
\end{aligned}$$

Therefore,

$$\mathbb{P}\left(\hat{z}_T^{\mathrm{def}} \neq z^\star\right) \leq \mathbb{P}(\mathcal{E}_\gamma^c) + (K-1)\exp\left(-\frac{2(\underline{\kappa} + \lambda\alpha_{\mathrm{true}})\gamma}{M^2(1+\lambda^2)}\right)$$

Since $\mathcal{E}_\gamma^c = \left\{G_{r+1}^{\mathrm{def}} \leq \gamma\right\}$, substituting Eq. (27) yields

$$\mathbb{P}\left(\hat{z}_T^{\mathrm{def}} \neq z^\star\right) \leq (K-1)\exp\left(-\frac{(r\underline{\kappa} - \gamma)^2}{2rM^2}\right) + (K-1)\exp\left(-\frac{2(\underline{\kappa} + \lambda\alpha_{\mathrm{true}})\gamma}{M^2(1+\lambda^2)}\right)$$

Finally, choose $\gamma = \frac{r\underline{\kappa}}{2}$.

Then

$$\frac{(r\underline{\kappa}-\gamma)^2}{2rM^2} = \frac{(r\underline{\kappa}/2)^2}{2rM^2} = \frac{r\underline{\kappa}^2}{8M^2}$$

while the second exponent becomes

$$\frac{2(\underline{\kappa}+\lambda\alpha_{\text{true}})\gamma}{M^2(1+\lambda^2)} = \frac{r\underline{\kappa}(\underline{\kappa}+\lambda\alpha_{\text{true}})}{M^2(1+\lambda^2)}$$

Both exponents therefore grow linearly in r.

Defining

$$c_{\text{def}} := \min\left\{\frac{\underline{\kappa}^2}{8M^2}, \frac{\underline{\kappa}(\underline{\kappa}+\lambda\alpha_{\text{true}})}{M^2(1+\lambda^2)}\right\},$$

each of the two terms is at most $(K-1)e^{-c_{\text{def}}r}$. Therefore,

$$\mathbb{P}(\hat{z}_T^{\text{def}} \neq z^\star) \leq 2(K-1)e^{-c_{\text{def}}r}$$

Hence there exists a constant $c > 0$, independent of r, such that

$$\mathbb{P}\left(\hat{z}_T^{\text{def}} \neq z^\star\right) = O(e^{-cr})$$

Thus the terminal error probability decays exponentially with the deferred-commitment horizon r. □

F.2 Proof of Corollary F.1.

Corollary F.1 (A single recap shifts state but not drift)**.** *Suppose a single recap is applied at turn t_0, producing post-recap margins $\Delta_{t_0}^+(z)$ and leader gap $G_{t_0}^+$. The recap changes the current state but not the subsequent drift regime. Therefore, Theorems 4.1 and 4.3 hold from the post-recap state with G_{t_0} replaced by $G_{t_0}^+$.*

Proof. The recap modifies the margins only at turn t_0, replacing $\Delta_{t_0}(z)$ by $\Delta_{t_0}^+(z)$ and hence G_{t_0} by $G_{t_0}^+$.

After the recap, the ordinary closed-loop protocol resumes, so the future pairwise-margin recursion is identical to that analyzed in Sec. 4.

In particular, neither $\underline{\kappa}$, $\overline{\kappa}$, α_{true}, α_{err}, nor λ is changed.

Therefore, if $G_{t_0}^+ > 0$, Thm. 4.3 applies directly from the post-recap state; if $G_{t_0}^+ < 0$ and $\lambda\alpha_{\text{err}} > \overline{\kappa}$, Thm. 4.1 applies instead. □

F.3 Proof of Theorem F.2.

Theorem F.2 (Snowballing eliminates asymptotic lock-in)**.** *For every fixed $\lambda \geq 0$, there exist constants $C > 0$ and $T_0 < \infty$, depending only on $\underline{\kappa}, \beta, M$, and λ, such that for every $T \geq T_0$, $\mathbb{P}\left(\hat{z}_T^{\text{snow}} \neq z^\star\right) \leq (K-1)e^{-CT}$. Consequently, $\mathbb{P}\left(\hat{z}_T^{\text{snow}} \neq z^\star \text{ infinitely often}\right) = 0$. In particular, permanent lock-in has probability zero under snowballing.*

Proof. For this modification, the terminal score is

$$S_{\text{snow}}^{(T)}(z) := \log \pi(z) + \sum_{s=1}^{T}(T-s+1)\log p_{\text{U,s}}(U_s \mid z) + \lambda \sum_{s=1}^{T-1} \log p_{\text{A}}(A_s \mid z).$$

Fix a wrong intent $z \neq z^\star$ and define

$$X_s(z) := \log \frac{p_{\text{U,s}}(U_s \mid z^\star)}{p_{\text{U,s}}(U_s \mid z)}$$

For the assistant contribution, define

$$Y_s(z) := \log \frac{p_{\mathrm{A}}(A_s \mid z^\star)}{p_{\mathrm{A}}(A_s \mid z)}$$

Under snowballing, user shard U_s is re-presented at turns $s, s+1, \ldots, T$, and therefore has terminal weight $T - s + 1$.

Hence the terminal pairwise margin between the true intent and z is

$$\Delta_T^{\mathrm{snow}}(z) = \log \frac{\pi(z^\star)}{\pi(z)} + \sum_{s=1}^{T}(T - s + 1)X_s(z) + \lambda \sum_{s=1}^{T-1} Y_s(z) \tag{28}$$

Step 1: Lower-bound the accumulated user evidence.

The first-turn margin satisfies

$$\Delta_1(z) = \log \frac{\pi(z^\star)}{\pi(z)} + X_1(z)$$

Therefore, the prior term and the T copies of the first user shard in Eq. (28) can be rewritten as

$$\log \frac{\pi(z^\star)}{\pi(z)} + TX_1(z) = \Delta_1(z) + (T-1)X_1(z)$$

By Assump. 2.1, $\Delta_1(z) \geq 0$, and by Assump. 2.4, $X_1(z) \geq -M$.

Hence

$$\Delta_1(z) + (T-1)X_1(z) \geq -(T-1)M \tag{29}$$

For every post-seed shard $s \geq 2$, Assump. 2.2 gives

$$\mathbb{E}[X_s(z)] = \mathrm{KL}\left(p_{\mathrm{U,s}}(\cdot \mid z^\star) \,\|\, p_{\mathrm{U,s}}(\cdot \mid z)\right) \geq \underline{\kappa}$$

Since shard U_s has weight $T - s + 1$, the expected weighted post-seed user evidence is at least

$$\begin{aligned}\sum_{s=2}^{T}(T - s + 1)\mathbb{E}[X_s(z)] &\geq \underline{\kappa} \sum_{s=2}^{T}(T - s + 1) \\ &= \underline{\kappa} \sum_{j=1}^{T-1} j \\ &= \frac{\underline{\kappa} T(T-1)}{2}\end{aligned} \tag{30}$$

Thus snowballing causes the expected post-seed user evidence to grow quadratically with the horizon.

Step 2: Lower-bound the assistant contribution.

At any turn s, conditional on $\mathcal{F}_s$, the current snowball estimate $\hat{z}_s^{\mathrm{snow}}$ is fixed.

Suppose first that

$$\hat{z}_s^{\mathrm{snow}} = z^\star.$$

Then

$$A_s \mid \mathcal{F}_s \sim p_{\mathrm{A}}(\cdot \mid z^\star),$$

and therefore

$$\begin{aligned}\mathbb{E}[Y_s(z) \mid \mathcal{F}_s] &= \mathrm{KL}\left(p_{\mathrm{A}}(\cdot \mid z^\star) \,\|\, p_{\mathrm{A}}(\cdot \mid z)\right) \\ &\geq 0\end{aligned}$$

Suppose instead that

$$\hat{z}_s^{\mathrm{snow}} = w \neq z^\star$$

Then

$$A_s \mid \mathcal{F}_s \sim p_{\mathrm{A}}(\cdot \mid w)$$

so

$$\begin{aligned}\mathbb{E}[Y_s(z) \mid \mathcal{F}_s] &= \mathbb{E}_{A\sim p_{\mathrm{A}}(\cdot\mid w)}\left[\log \frac{p_{\mathrm{A}}(A \mid z^\star)}{p_{\mathrm{A}}(A \mid z)}\right] \\ &= -\operatorname{KL}\left(p_{\mathrm{A}}(\cdot \mid w) \,\|\, p_{\mathrm{A}}(\cdot \mid z^\star)\right) + \operatorname{KL}\left(p_{\mathrm{A}}(\cdot \mid w) \,\|\, p_{\mathrm{A}}(\cdot \mid z)\right) \\ &\geq -\beta\end{aligned}$$

where the final inequality follows from the definition of β.

Since $\beta \geq 0$, the same lower bound also holds when the current estimate is correct. Hence, in every state,

$$\mathbb{E}[Y_s(z) \mid \mathcal{F}_s] \geq -\beta \tag{31}$$

Summing over the $T-1$ assistant responses gives the predictable lower bound

$$\lambda \sum_{s=1}^{T-1} \mathbb{E}[Y_s(z) \mid \mathcal{F}_s] \geq -\lambda\beta(T-1)$$

Combining Eq. (29), Eq. (30), and Eq. (31), define

$$\mu_T := (T-1)\left(\frac{\kappa T}{2} - M - \lambda\beta\right) \tag{32}$$

Step 3: Construct the martingale fluctuations.

We next control deviations around this deterministic lower bound.

For each fixed terminal horizon T, consider the refined filtration that reveals the random variables chronologically: $\mathcal{F}_1,\ A_1,\ U_2,\ \mathcal{F}_2,\ A_2,\ U_3,\ \ldots,\ A_{T-1},\ U_T$ More precisely, for $s = 1, \ldots, T-1$, let

$$\mathcal{G}_{2s-2} := \mathcal{F}_s$$

and

$$\mathcal{G}_{2s-1} := \sigma(\mathcal{F}_s, A_s).$$

After U_{s+1} and the next tie-breaking randomness are revealed, the information becomes $\mathcal{F}_{s+1}$.

For the assistant response A_s, define the centered contribution

$$\lambda\left(Y_s(z) - \mathbb{E}[Y_s(z) \mid \mathcal{F}_s]\right)$$

Its conditional expectation given $\mathcal{F}_s$ is zero.

For the subsequent user shard U_{s+1}, its terminal snowball weight is $T-s$.

Since U_{s+1} remains generated from the true intent even after A_s has been observed,

$$\mathbb{E}\left[X_{s+1}(z) \,|\, \mathcal{G}_{2s-1}\right] = \operatorname{KL}\left(p_{U,s+1}(\cdot \mid z^\star) \,\|\, p_{U,s+1}(\cdot \mid z)\right) \geq \underline{\kappa}$$

Thus its centered weighted contribution is

$$(T-s)\left(X_{s+1}(z) - \mathbb{E}[X_{s+1}(z) \mid \mathcal{G}_{2s-1}]\right)$$

whose conditional expectation given $\mathcal{G}_{2s-1}$ is also zero.

Therefore, these assistant and user contributions form a martingale difference sequence under the refined filtration.

Let their sum be

$$N_T(z) := \sum_{s=1}^{T-1} \lambda\left(Y_s(z) - \mathbb{E}[Y_s(z) \mid \mathcal{F}_s]\right) + \sum_{s=1}^{T-1} (T-s)\left(X_{s+1}(z) - \mathbb{E}[X_{s+1}(z) \mid \mathcal{G}_{2s-1}]\right)$$

Using Eq. (28), we first rewrite the terminal margin as

$$\Delta_T^{\mathrm{snow}}(z) = \Delta_1(z) + (T-1)X_1(z) + \sum_{s=1}^{T-1}(T-s)X_{s+1}(z) + \lambda\sum_{s=1}^{T-1} Y_s(z)$$

For each post-seed user term, add and subtract its conditional expectation:

$$(T-s)X_{s+1}(z) = (T-s)\left(X_{s+1}(z) - \mathbb{E}[X_{s+1}(z) \mid \mathcal{G}_{2s-1}]\right) + (T-s)\mathbb{E}[X_{s+1}(z) \mid \mathcal{G}_{2s-1}]$$

Similarly,

$$\lambda Y_s(z) = \lambda\left(Y_s(z) - \mathbb{E}[Y_s(z) \mid \mathcal{F}_s]\right) + \lambda\mathbb{E}[Y_s(z) \mid \mathcal{F}_s].$$

Substituting these identities into the terminal margin and collecting the centered terms into $N_T(z)$ gives the exact decomposition

$$\Delta_T^{\text{snow}}(z) = \Delta_1(z) + (T-1)X_1(z) + N_T(z) + \sum_{s=1}^{T-1}(T-s)\mathbb{E}[X_{s+1}(z) \mid \mathcal{G}_{2s-1}] + \lambda\sum_{s=1}^{T-1}\mathbb{E}[Y_s(z) \mid \mathcal{F}_s]$$

By Eq. (29),

$$\Delta_1(z) + (T-1)X_1(z) \geq -(T-1)M$$

Moreover, by Assump. 2.2,

$$\mathbb{E}[X_{s+1}(z) \mid \mathcal{G}_{2s-1}] \geq \underline{\kappa}$$

and by Eq. (31),

$$\mathbb{E}[Y_s(z) \mid \mathcal{F}_s] \geq -\beta$$

Therefore,

$$\begin{aligned}\Delta_T^{\text{snow}}(z) &\geq -(T-1)M + N_T(z) + \underline{\kappa}\sum_{s=1}^{T-1}(T-s) - \lambda\beta(T-1)\\ &= N_T(z) + (T-1)\left(\frac{\underline{\kappa}T}{2} - M - \lambda\beta\right)\\ &= N_T(z) + \mu_T\end{aligned}$$

Hence

$$\Delta_T^{\text{snow}}(z) \geq \mu_T + N_T(z). \tag{33}$$

Step 4: Bound the martingale fluctuations.

By Assump. 2.4,

$$|X_{s+1}(z) - \mathbb{E}[X_{s+1}(z) \mid \mathcal{G}_{2s-1}]| \leq 2M$$

and therefore the centered weighted user increment has absolute value at most $2M(T-s)$.

Likewise,

$$|\lambda\left(Y_s(z) - \mathbb{E}[Y_s(z) \mid \mathcal{F}_s]\right)| \leq 2\lambda M$$

The sum of the squared increment bounds is therefore

$$\sum_{s=1}^{T-1}[2M(T-s)]^2 + \sum_{s=1}^{T-1}(2\lambda M)^2 = 4M^2\left(\sum_{j=1}^{T-1}j^2 + \lambda^2(T-1)\right)$$

For all sufficiently large T, Eq. (32) satisfies $\mu_T > 0$.

If $\Delta_T^{\text{snow}}(z) \leq 0$, then by Eq. (33), $\mu_T + N_T(z) \leq 0$, and hence

$$N_T(z) \leq -\mu_T$$

Thus

$$\{\Delta_T^{\text{snow}}(z) \leq 0\} \subseteq \{N_T(z) \leq -\mu_T\}$$

Applying the lower-tail form of Azuma- Hoeffding's inequality to the martingale difference sequence gives

$$\begin{aligned}\mathbb{P}\left(\Delta_T^{\text{snow}}(z) \le 0\right) &\le \mathbb{P}\left(N_T(z) \le -\mu_T\right) \\ &\le \exp\left(-\frac{\mu_T^2}{8M^2\left(\sum_{j=1}^{T-1} j^2 + \lambda^2(T-1)\right)}\right)\end{aligned} \tag{34}$$

Step 5: Show that the exponent grows linearly in T.

Choose

$$T_0 := \left\lceil \max\left\{2,\ \frac{4(M+\lambda\beta)}{\underline{\kappa}}\right\}\right\rceil.$$

For every $T \ge T_0$,

$$M + \lambda\beta \le \frac{\underline{\kappa}T}{4}$$

and therefore

$$\frac{\underline{\kappa}T}{2} - M - \lambda\beta \ge \frac{\underline{\kappa}T}{4}.$$

Using $T - 1 \ge T/2$ for $T \ge 2$,

$$\begin{aligned}\mu_T &= (T-1)\left(\frac{\underline{\kappa}T}{2} - M - \lambda\beta\right) \\ &\ge (T-1)\frac{\underline{\kappa}T}{4} \ge \frac{\underline{\kappa}T^2}{8}\end{aligned}$$

Hence

$$\mu_T^2 \ge \frac{\underline{\kappa}^2 T^4}{64}$$

Moreover,

$$\sum_{j=1}^{T-1} j^2 = \frac{(T-1)T(2T-1)}{6} \le \frac{T^3}{3}$$

and, since $T \ge 1$,

$$\lambda^2(T-1) \le \lambda^2 T^3.$$

Therefore,

$$\sum_{j=1}^{T-1} j^2 + \lambda^2(T-1) \le \left(\frac{1}{3} + \lambda^2\right) T^3$$

Substituting these bounds into Eq. (34),

$$\begin{aligned}\mathbb{P}\left(\Delta_T^{\text{snow}}(z) \le 0\right) &\le \exp\left(-\frac{\underline{\kappa}^2 T^4/64}{8M^2(\frac{1}{3}+\lambda^2)T^3}\right) \\ &\le \exp\left(-\frac{\underline{\kappa}T}{512M^2(\frac{1}{3}+\lambda^2)}\right) \\ &= \exp(-CT)\end{aligned}$$

where

$$C := \frac{\underline{\kappa}^2}{512M^2\left(\frac{1}{3}+\lambda^2\right)} > 0$$

Thus, for every wrong intent z and every $T \ge T_0$,

$$\mathbb{P}\left(\Delta_T^{\text{snow}}(z) \le 0\right) \le e^{-CT} \tag{35}$$

Step 6: From pairwise margins to terminal error.

Suppose $\hat{z}_T^{\text{snow}} \neq z^\star$. Let

$$z_T := \hat{z}_T^{\text{snow}}$$

be the selected wrong intent.

Since the estimator selects a score maximizer,

$$S_{\text{snow}}^{(T)}(z_T) \geq S_{\text{snow}}^{(T)}(z^\star)$$

Therefore,

$$\Delta_T^{\text{snow}}(z_T) = S_{\text{snow}}^{(T)}(z^\star) - S_{\text{snow}}^{(T)}(z_T) \leq 0$$

Hence

$$\{\hat{z}_T^{\text{snow}} \neq z^\star\} \subseteq \bigcup_{z \neq z^\star} \{\Delta_T^{\text{snow}}(z) \leq 0\}$$

Applying the union bound and Eq. (35),

$$\begin{aligned}\mathbb{P}\left(\hat{z}_T^{\text{snow}} \neq z^\star\right) &\leq \sum_{z \neq z^\star} \mathbb{P}\left(\Delta_T^{\text{snow}}(z) \leq 0\right) \\ &\leq (K-1)e^{-CT}\end{aligned}$$

Step 7: Almost-sure eventual correctness. The preceding bound gives

$$\begin{aligned}\sum_{T=T_0}^{\infty} \mathbb{P}(\hat{z}_T^{\text{snow}} \neq z^\star) &\leq (K-1) \sum_{T=T_0}^{\infty} e^{-CT} \\ &< \infty\end{aligned}$$

By the first Borel–Cantelli lemma,

$$\mathbb{P}(\{\hat{z}_T^{\text{snow}} \neq z^\star\} \text{ infinitely often}) = 0.$$

Equivalently,

$$\mathbb{P}\left(\exists\, T_1 < \infty \text{ such that } \forall T \geq T_1,\ \hat{z}_T^{\text{snow}} = z^\star\right) = 1$$

Any permanently locked wrong trajectory would be wrong for infinitely many terminal horizons. Therefore,

$$\{\text{permanent lock-in}\} \subseteq \{\{\hat{z}_T^{\text{snow}} \neq z^\star\} \text{ infinitely often}\},$$

and hence

$$\mathbb{P}\left(\text{permanent lock-in}\right) = 0$$

□

G Synthetic Validation of the Theoretical Model

Simulation Setup. Our experiments implement the finite-intent random model proposed in this paper. They are synthetic simulations of the theoretical abstraction and their purpose is to examine the finite-horizon behavior of the closed-loop mechanisms in settings where the underlying likelihoods and theoretical quantities can be computed exactly.

Closed-loop generative process. For each experiment, we fix the true intent $z^\star$, the prior π, the user-shard distribution $p_{\text{U,t}}$, the model-response distribution p_{A}, the model-evidence weight λ, and the interaction protocol. At turn t, a fresh user shard is sampled as $U_t \sim p_{\text{U,t}}(\cdot \mid z^\star)$. The user shards are conditionally independent given $z^\star$ and do not adapt to previous model responses. After observing U_t, the model updates the score $S_\lambda^{(t)}$ using user evidence through turn t and model-evidence from responses through turn $t-1$, and forms the MAP estimate $\hat{z}_t^\lambda$, with uniform random tie-breaking. The response is then sampled according to $A_t \sim p_{\text{A}}(\cdot \mid \hat{z}_t^\lambda)$.

The first user shard is constructed to realize the m-way initial ambiguity in Assump. 2.1. Subsequent user shards follow the experiment-specific informative channel. Unless otherwise stated, we use finite discrete channels, which allow the relevant likelihood ratios, divergences, and theoretical bounds to be evaluated directly.

Monte Carlo estimation. For quantities that are not evaluated exactly, we estimate probabilities and expectations by independently simulating conversation trajectories from the generative process above. Unless otherwise stated, each experimental configuration uses $N = 20{,}000$ independent trajectories. Event probabilities are estimated by empirical frequencies, while real-valued trajectory statistics are estimated by sample averages. A trajectory is simulated once up to its maximum required horizon, so measurements at different checkpoints along the same curve are computed from prefixes of the same simulated trajectories.

Reproducibility. All experimental configurations, horizons, simulation budgets, and random seeds were fixed before the final runs. The master random seed is 20260826, and bootstrap analyses use 5,000 replicates. Experiment-specific parameters and conditioning designs are reported with the corresponding results.

G.1 Statistical Role of Model-Generated History

We evaluate three claims from Sec. 3: the equivalence of user-only SHARDED and matched SINGLE inference, the consistency of inference from accumulating user evidence, and the statistical role of model-evidence under the self-consistent reference model.

User-only equivalence. We first verify Prop. C.1. We set $\lambda = 0$ and couple the SHARDED and matched SINGLE estimators using the same realized user shards and coupled tie-breaking randomness. For the user shards, we use three user-shard distribution

$$p_{\mathrm{U},\mathrm{t}} = \begin{cases} p_t, & u = z \\ \dfrac{1-p_t}{K-1}, & u \neq z \end{cases}$$

where (1) $p_t \equiv 0.45$, (2) $p_t \equiv 0.65$ and (3) $p_t = \begin{cases} 0.45, & 1 \le t \le T_1 \\ 0.55, & T_1 < t \le T_2 \\ 0.65, & t > T_2, \end{cases}$.

Across all 5,000 coupled trajectories and all evaluated checkpoints up to $T = 500$, we observe no MAP mismatch, no score discrepancy, and no paired degradation (Table 1). These checks are stronger than agreement in average error alone: when model-evidence is excluded, the SHARDED score is identical to the score formed from the same accumulated user evidence. Thus, within the model, sequential disclosure alone does not produce the closed-loop degradation studied later.

Table 1: User-only equivalence under three user-evidence schedules. Each entry reports the maximum absolute discrepancy over all evaluated checkpoints.

Schedule	MAP mismatch	Score gap	$\lvert\mathbb{P}(\hat{z}_T^0 \neq z^\star) - \mathbb{P}(\hat{z}_{\mathrm{single}} \neq z^\star)\rvert$
$p_t = 0.45$	0	0	0
$p_t = 0.65$	0	0	0
$p_t = 0.45/0.55/0.65$	0	0	0

User-only consistency. We next evaluate Thm. 3.1. Here we fixed shard distribution at $p_t \equiv 0.55$ for all $t \ge 2$, while the number of candidate intents is varied over $K \in \{2, 3, 5, 10, 20\}$. All configurations use the same $m = 2$ initial ambiguity and $N = 20{,}000$ independent trajectories.

Because the remaining probability mass $1 - p_t$ is distributed uniformly across the $K - 1$ competing observations, changing K also changes the pairwise evidence constants $\underline{\kappa}$ and M. The sweep should therefore be interpreted as a check across several finite intent spaces rather than as a controlled experiment in which only problem size changes as shown in Table 2.

Fig. 3(a) reports the empirical user-only error and (b) shows the corresponding upper bound from Thm. 3.1. Across all five K, the empirical error decreases as user evidence accumulates. The bound is conservative, most visibly in the binary setting: with $\underline{\kappa} = 0.0201$, the empirical error decreases from 0.4413 at $t = 2$ to 0.0116 at $t = 512$, while the analytic upper bound decreases only from 0.9999 to 0.9664. For stronger user channels, no errors are observed beyond sufficiently large horizons; these points are reported with their finite-sample upper confidence limits rather than interpreted as zero population error.

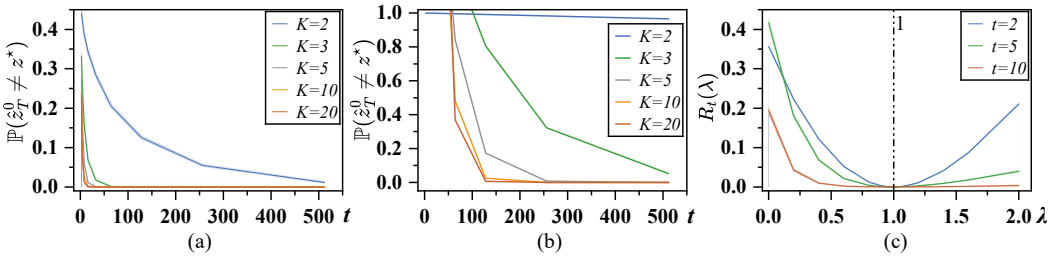


Figure 3: **Statistical role of model-generated history.** (a) Empirical user-only error probability $\mathbb{P}(\hat{z}_T \neq z^\star)$ for $K \in \{2, 3, 5, 10, 20\}$ with $m = 2$ and $p_t \equiv 0.55$. Error bars show the stored binomial confidence limits. (b) Corresponding raw upper bounds from Thm. 3.1. (c) Posterior KL risk $R_t(\lambda)$ under the self-consistent pretraining distribution for $t \in \{2, 5, 10\}$. All three evaluated risk curves attain their minimum at $\lambda = 1$, where $R_t(1) = 0$.

Posterior identity under the self-consistent reference. We next evaluate Prop. 3.1 under the reference model. A latent intent $Z \sim \pi$ is sampled once per conversation, and $U_t \sim p_{\mathrm{U,t}}(\cdot \mid Z)$ and $A_t \sim p_{\mathrm{A}}(\cdot \mid Z)$, Thus model responses are generated from the same fixed latent intent as the user shards, rather than from the current estimate as in the closed loop.

For $K = 3$ and $T = 5$, we enumerate all 6,561 finite histories and compare the ordinary Bayes posterior with the score-based posterior at $\lambda = 1$. As shown in Table 3, the maximum posterior discrepancy is 1.39×10^{-15}; the corresponding predictive discrepancy is 9.44×10^{-16}. Both are below the 10^{-12} numerical tolerance, confirming the posterior identity to floating-point precision.

Table 2: Evidence constants for the user-only consistency sweep. The finite-sample bound in Thm. 3.1 depends on both $\underline{\kappa}$ and M through $\underline{\kappa}^2/(2M^2)$.

K	2	3	5	10	20
κ	0.0201	0.2905	0.6943	1.1989	1.6553
M	1.7346	2.4277	3.1209	3.9318	4.6790

Table 3: Exact checks of the self-consistent pretraining posterior. The two independent calculations agree to floating-point precision.

Exact check	Histories	Max discrepancy
Posterior identity	6,561	1.39×10^{-15}
Predictive identity	6,561	9.44×10^{-16}

Model-evidence weighting under the self-consistent reference. We then evaluate Prop. 3.2. We sample 20,000 independent histories from the self-consistent reference model and evaluate every value of λ on the same histories. This paired design isolates the effect of the model-evidence weight from Monte Carlo variation across histories.

Fig. 3(c) reports $R_t(\lambda)$ for $t \in \{2, 5, 10\}$. At every evaluated horizon, the minimum occurs at $\lambda = 1$, where $R_t(1) = 0$, while all evaluated $\lambda \neq 1$ yield positive posterior risk. The estimated local curvature coefficients are $\mathcal{I}_2 = 0.592$, $\mathcal{I}_5 = 0.162$, $\mathcal{I}_{10} = 0.0133$. At the nearest evaluated weights, $\lambda = 0.98$ and $\lambda = 1.02$, the empirical quadratic coefficient differs from that predicted by the local expansion by at most approximately 2.5%.

G.2 Closed-Loop State–Drift Dynamics

Wrong-state persistence. We first evaluate Thm. 4.1. We use $K = 5$, a stationary post-seed shard distribution with $p_t \equiv 0.55$, a symmetric model-response distribution with $p_A = 0.85$, and $\lambda = 0.5$. The resulting constants are $\overline{\kappa} = 0.6943$, $\alpha_{\mathrm{err}} = 2.5357$, $M = 3.1209$, so that $\lambda\alpha_{\mathrm{err}} = 1.2679 > \overline{\kappa}$.

To isolate the dependence on the current state, we initialize controlled trajectories at $G_{t_0} \in \{-0.25, -0.5, -1, -2, -4\}$, and simulate future randomness independently. Fig. 4(a) reports $\mathbb{P}(t_{\mathrm{rec}} > t_0 + h \,|\, G_{t_0})$. At $h = 498$, the finite-horizon no-recovery probability increases from 0.7147 at $G_{t_0} = -0.25$ to 0.9639 at $G_{t_0} = -4$.

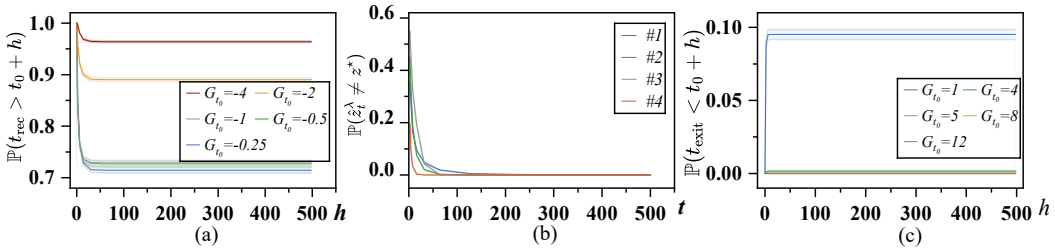


Figure 4: **Closed-loop state–drift dynamics.** (a) Conditional finite-horizon no-recovery probability from controlled wrong states satisfying $\lambda\alpha_{\text{err}} > \overline{\kappa}$, for $G_{t_0} \in \{-0.25, -0.5, -1, -2, -4\}$. (b) Unconditional error probability in four configurations satisfying $\lambda\beta < \underline{\kappa}$, the settings # are shown in Table 5. (c) Conditional exit probability from controlled correct states for $G_{t_0} \in \{1, 4, 5, 8, 12\}$.

Table 4: Controlled wrong-state lock-in with $\lambda = 0.5$. The empirical column is the probability of no recovery by future horizon $h = 498$, whereas the last column is the lower bound from Theorem 4.1 on permanent no recovery.

G_{t_0}	$\mathbb{P}(t_{\text{rec}} > t_0 + 498)$	95% one-sided CP limits	Permanent-lock-in lower bound
-0.25	0.7147	[0.7094, 0.7199]	0.0233
-0.50	0.7270	[0.7218, 0.7322]	0.0460
-1	0.7289	[0.7236, 0.7340]	0.0899
-2	0.8904	[0.8866, 0.8940]	0.1718
-4	0.9639	[0.9616, 0.9660]	0.3140

Table 4 compares these finite-horizon probabilities with the theorem's lower bound on *permanent* no recovery. The two columns refer to different events: $\mathbb{P}(t_{\text{rec}} > t_0 + 498)$ and $\mathbb{P}(t_{\text{rec}} = \infty)$. The bound is therefore not expected to approximate the finite-horizon probability, but it increases with the initial wrong-state deficit as predicted.

Recovery under uniformly corrective drift. We next evaluate Thm. 4.2 using four settings satisfying $\lambda\beta < \underline{\kappa}$ where their values of K, λ, $\lambda\beta/\underline{\kappa}$ are reported in Table 5. Setting #1 uses the asymmetric assistant channel $P_A = \begin{pmatrix} 0.80 & 0.10 & 0.10 \\ 0.05 & 0.85 & 0.10 \\ 0.05 & 0.25 & 0.70 \end{pmatrix}$, which gives $\beta = 1.6804$ and $\lambda\beta/\underline{\kappa} = 0.8088$. Settings #2–#4 use the K-ary symmetric assistant channel with $p_A = 0.85$

Fig. 4(b) reports $\mathbb{P}(\hat{z}_t^{\lambda} \neq z^{\star})$. The error decreases toward the Monte Carlo resolution limit in all four configurations. For the first three settings, the theorem bounds are valid but numerically above one at the final horizon and are therefore used only as qualitative checks of the recovery mechanism. The binary configuration uses $K = 2$, $p_t \equiv 0.80$, $\lambda = 0.05$, with $\kappa = 0.8318$, $\beta = 1.2142$ so $\frac{\lambda\beta}{\underline{\kappa}} = 0.0730$. Here the upper bound is nontrivial. At $T = 512$, no errors are observed among 20,000 trajectories, while the theorem gives 1.07×10^{-5}. Because this value lies below the Monte Carlo resolution of the experiment, the zero observed count is interpreted through its finite-sample upper confidence limit rather than as evidence of zero error probability.

Table 5: Experimental configurations for recovery under uniformly corrective drift.

Setting	K	p_t	λ	$\lambda\beta/\underline{\kappa}$
#1	3	0.65	0.30	0.809
#2	5	0.55	0.15	0.548
#3	5	0.45/0.55/0.65	0.10	0.684
#4	2	0.80	0.05	0.073

Correct-state persistence. We next evaluate Thm. 4.3. We fix $K = 5$, $p_t \equiv 0.55$, $p_A = 0.85$, $\lambda = 0.5$ and initialize controlled trajectories at positive leader gaps $G_{t_0} \in \{1, 4, 5, 8, 12\}$. Fig. 4(c) reports $\mathbb{P}(t_{\text{exit}} \leq t_0 + h \,|G_{t_0})$. At $h = 498$, the exit probability decreases from 0.0951 at $G_0 = 1$ to 1.6×10^{-3} at $G_0 = 4$ and 10^{-4} at $G_0 = 5$; no exits are observed for $G_0 = 8$ or 12. The corre-

Table 6: Controlled correct-state persistence with $\lambda = 0.5$. The empirical quantity is exit by future horizon $h = 498$, whereas the analytic quantity bounds the probability of ever exiting. Raw bounds above one are shown as trivial.

G_0	$\mathbb{P}(t_{\text{exit}} \le t_0 + 498)$	95% one-sided CP limits	Ever-exit upper bound
1	0.0951	$[0.0917, 0.0986]$	2.898
4	0.00160	$[0.00117, 0.00215]$	1.102
5	0.00010	$[1.8\times10^{-5}, 3.15\times10^{-4}]$	0.798
8	0	$[0, 1.50\times10^{-4}]$	0.304
12	0	$[0, 1.50\times10^{-4}]$	0.0836

sponding theorem bound becomes nontrivial for $G_0 \ge 5$, consistent with the predicted stabilization as the positive leader gap increases as shown in Table 6.

Table 7: Conditional continuations from states reached by the closed-loop process at $t_0 = 2$. For each condition, 128 independent histories are sampled with 64 future continuations per history.

Condition	Finite-horizon event	Estimate	95% CI	Mean theorem bound
$G_{t_0} < 0$	$t_{\text{rec}} > t_0 + 128$	0.7405	$[0.7106, 0.7682]$	0.1415 (lower)
$G_{t_0} > 0$	$t_{\text{exit}} \le t_0 + 128$	0.1130	$[0.1022, 0.1234]$	1.000 (upper)

Natural-state continuations. The controlled initializations above isolate the effect of the current leader gap. To verify that the same behavior occurs from states reached by the closed-loop process itself, we sample realized states at $t_0 = 2$ and branch independent future continuations. We use the same $K = 5$, $p_t \equiv 0.55$, $p_A = 0.85$, and $\lambda = 0.5$ configuration as in the controlled state experiments above.

For each sign of the leader gap, we use 128 independent realized histories and 64 continuations per history. Conditional on $G_{t_0} < 0$, the probability of no recovery by $h = 128$ is 0.7405, compared with a mean snapshot-specific lower bound 0.1415 on permanent non-recovery. Conditional on $G_{t_0} > 0$, the probability of exit by $h = 128$ is 0.1130; the corresponding mean theorem upper bound is quantitatively trivial.

Binary drift threshold. In the binary specialization, $\beta = \alpha_{\text{err}}$ and $\overline{\kappa} = \underline{\kappa}$, so the recovery and wrong-state persistence conditions meet at $r := \frac{\lambda\alpha_{\text{err}}}{\kappa} = 1$. We use $K = 2$, $p_t \equiv 0.75$, and $p_A = 0.90$, giving $\kappa = 0.5493$ and $\alpha_{\text{err}} = 1.7578$, and vary λ so that r ranges from 0.5 to 1.5. Fig. 5 reports the finite-horizon error at $t = 500$. The point $r = 1$ is shown only as the theoretical boundary in Corollary D.2.

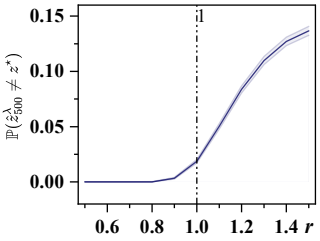


Figure 5: **Binary drift.** Finite-horizon error as a function of the normalized drift ratio $r = \lambda\alpha_{\text{err}}/\kappa$.

G.3 Persistent Multi-Turn Degradation and Unreliability

We next examine the population-level consequences in Sec. 5: a persistent error gap between SHARDED and matched SINGLE inference, and non-vanishing variability across closed-loop trajectories. Both experiments use $z^\star = 0$, an uninformative first user shard with $m = K$, and $N = 20{,}000$ independent closed-loop trajectories up to $T = 500$.

Persistent degradation. We evaluate Thm. 5.1 using $K = 5$, a stationary post-seed shard distribution with $p_t \equiv 0.55$, a symmetric model-response distribution with $p_A = 0.85$, and $\lambda = 0.5$. For each trajectory, closed-loop SHARDED inference and the matched user-only estimator share the same realized user shards and tie-breaking randomness. We estimate $\mathbb{P}(\hat{z}_t^{\lambda} \neq z^\star) - \mathbb{P}(\hat{z}_t^{\text{single}} \neq z^\star)$ from the resulting paired outcomes.

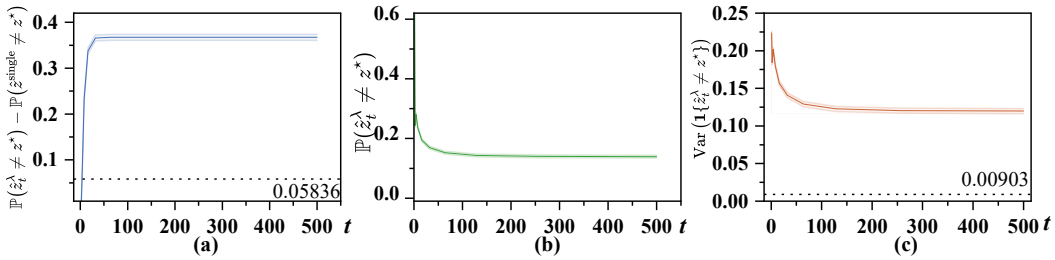


Figure 6: **Persistent multi-turn degradation and unreliability.** (a) Paired error gap between closed-loop SHARDED inference and the matched user-only baseline, together with the theoretical degradation floor. (b) Terminal error probability in the persistent-unreliability configuration and the certified interval $[q_L, q_U]$. (c) Corresponding variance $V_T = q_T(1 - q_T)$ and its positive theoretical floor $V_{\min}$. Error bars or shaded regions denote 95% confidence intervals.

Fig. 6(a) shows that the error gap increases rapidly and remains positive over the longest evaluated horizons. In particular, $\mathbb{P}(\hat{z}^{\lambda}_{500} \neq z^{\star}) - \mathbb{P}(\hat{z}^{\text{single}}_{500} \neq z^{\star}) = 0.3673$ with 95% confidence interval $[0.3606, 0.3739]$ at $t = 500$. For this configuration, exact evaluation of the wrong-seed lock-in quantity gives $p_{\text{lock}} = 0.07294$, and hence $\left(1 - \frac{1}{m}\right) p_{\text{lock}} = 0.05836$.

The finite-horizon lower bound is 0.0511 at $t = 256$ and 0.05834 at $t = 500$. Thus the simulation exhibits a persistent error gap while the theorem certifies that the gap remains bounded away from zero.

Persistent unreliability. We next evaluate Thm. 5.2. For the frozen parameters, we use $K = 3$, a stationary post-seed shard distribution with $p_t \equiv 0.65$, $\lambda = 0.5$ with an asymmetric model-response distribution with $P_A = \begin{pmatrix} 0.80 & 0.10 & 0.10 \\ 0.05 & 0.85 & 0.10 \\ 0.05 & 0.25 & 0.70 \end{pmatrix}$ We choose $\gamma = 2.6985$, yielding $q_L = 0.00911$, $q_U = 0.96866$ and $V_{\min} = 0.00903$.

Fig. 6(b) reports $\mathbb{P}(\hat{z}^{\lambda}_t \neq z^{\star})$. At the longest evaluated horizon,$\mathbb{P}(\hat{z}^{\lambda}_{500} \neq z^{\star}) = 0.13915$, with 95% CI $[0.13442, 0.14402]$, which lies strictly within the certified interval $[q_L, q_U]$.

Fig. 6(c) reports the corresponding Bernoulli variance $V_T = q_T(1 - q_T)$. At $T = 500$, $V_{500} = 0.11979$, 95% CI $[0.11626,\ 0.12312]$ well above the theoretical floor $V_{\min} = 0.00903$. These results are consistent with the coexistence of persistently correct and persistently incorrect trajectories, rather than convergence of all conversations to a single outcome.

G.4 EFFECTS OF INTERACTION MODIFICATIONS

Unless otherwise stated, the experiments use the same closed-loop simulation protocol described in Sec. 6 and use the same $K = m = 5$-intent deployment environment with $z^{\star} = 0$. For $t \geq 2$, $p_t \equiv 0.55$, $p_A = 0.85$. And unless explicitly varied, $\lambda = 0.5$. The resulting theory constants are $\kappa = 0.6943$, $\alpha_{\text{err}} = \alpha_{\text{true}} = \beta = 2.5357$, $M = 3.1209$. Each ordinary parameter setting uses $N = 20{,}000$ Monte Carlo trajectories. Deferred commitment and the single-recap experiment use $T_{\max} = 500$, while the snowball and matched-protocol experiments extend to $T_{\max} = 1000$. The snowball growth experiment additionally evaluates $\lambda = 1$.

Deferred commitment. We first evaluate Thm. F.1. For a deferral horizon r, the first r model responses are excluded from subsequent intent scores, so model-evidence first enters after the $(r+1)$-st user shard. We use the theorem choice $\gamma = \frac{r\kappa}{2}$ and evaluate both the state-formation event $\Pr(G^{\text{def}}_{r+1} \leq \gamma)$ and the terminal error probability.

Fig. 7(a) shows that terminal error decreases rapidly with the deferral horizon: No terminal errors are observed among 20,000 trajectories for $r \geq 64$ at the reported checkpoints. The theorem bounds are conservative at short deferral horizons and become informative only for sufficiently large r, consistent with the exponential dependence on the deferral length. Table 8 reports the corresponding

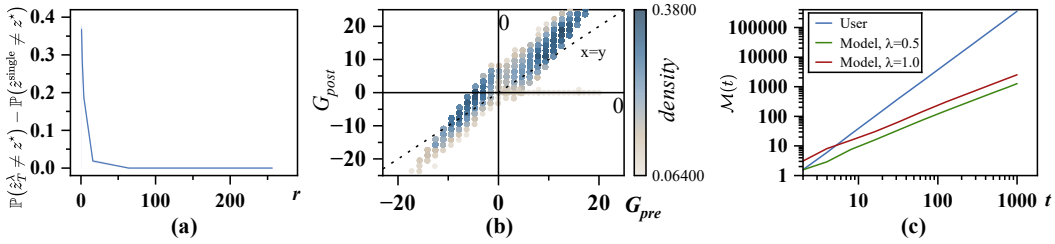


Figure 7: **Effects of Interaction Modifications.** (a) Terminal error as a function of the deferral horizon r. (b) Leader gap immediately before and after a single recap. Shading represents empirical trajectory density. (c) The mean weighted evidence magnitude $\mathcal{M}(t)$ on log–log axes. The fitted tail slopes are approximately 1.98 for user evidence and 1.03–1.04 for model evidence, illustrating quadratic versus linear accumulation.

state-formation event $\mathbb{P}(G_{r+1}^{\mathrm{def}} \le \gamma)$, together with the two bounds in Thm. F.1. The analytic bounds are conservative for short deferral horizons and become nontrivial only for sufficiently large r.

Table 8: State-formation probabilities and theoretical bounds for deferred commitment, with $\gamma = r\kappa/2$. Terminal errors are shown in Fig. 7(a).

r	γ	$\mathbb{P}(G_{r+1}^{\mathrm{def}} \le \gamma)$	95% one-sided CP limits	Gap bound	Terminal bound
1	0.347	0.4498	[0.4440, 0.4556]	3.975	7.552
8	2.777	0.3161	[0.3106, 0.3215]	3.807	5.441
32	11.109	0.0993	[0.09584, 0.10285]	3.282	3.393
64	22.218	0.01725	[0.01576, 0.01884]	2.692	2.695
128	44.435	0.00070	[0.000423, 0.001094]	1.812	1.812
256	88.870	0	$[0, 1.50\times10^{-4}]$	0.821	0.821
384	133.305	0	$[0, 1.50\times10^{-4}]$	0.372	0.372

A single recap. We next evaluate the recap intervention. At turn $t_0 = 8$, the accumulated user shards are re-added once while the accumulated model-evidence is retained. The recap therefore changes the current pairwise margins but not the subsequent closed-loop update rule.

As shown in Fig. 7(b), across 20,000 trajectories, the mean leader-gap displacement is $\mathbb{E}[G_{\mathrm{post}} - G_{\mathrm{pre}}] = 3.5675$, 95% CI $= [3.5263, 3.6086]$, and the probability of crossing the zero-gap boundary is 0.11025, 95% CI $= [0.10598, 0.11467]$. Of these trajectories, 2,178 move from a negative to a positive gap, whereas only 27 cross in the opposite direction.

To evaluate the post-recap conditional guarantees, we branch 64 future continuations from each sampled post-recap state. Among 180 independent positive-gap states, the probability of exit by $h = 128$ is 0.001823 with 95% interval $[0.000868, 0.002951]$. Among 76 negative-gap states, the probability of no recovery by $h = 128$ is 0.92784 with 95% interval $[0.90378, 0.95025]$. Exact evaluation of the future drift gives a maximum discrepancy of 4.44×10^{-16} from the ordinary closed-loop update, confirming that recap shifts the current state without changing the subsequent drift regime.

Snowballing changes the accumulation rate. Fig. 7(c) evaluates these accumulation rates directly. We fit log–log slopes over the tail region defined by the median checkpoint, which corresponds to $t \ge 48$ under our predefined checkpoint schedule. The fitted slope of the mean weighted user-evidence magnitude is 1.9785, close to the quadratic scaling predicted by Thm. F.2. The corresponding slopes of the mean model-evidence magnitude are 1.0273 for $\lambda = 0.5$ and 1.0385 for $\lambda = 1$. These empirical growth rates are consistent with the theoretical separation between quadratic accumulation of weighted user evidence and linear accumulation of model evidence.

The corresponding terminal-error experiment is also consistent with Thm. F.2. For the finite-horizon comparison at $T = 1000$, we report the tighter proof-level bound obtained by applying a union bound over the pairwise bound, rather than the looser simplified exponential envelope used to state the asymptotic rate. This gives terminal-error upper bounds of approximately 0.0435 for $\lambda = 0.5$

Table 9: Example controlled binary-intent instance used in the real-LLM mechanism probe.

Component	Example
Ambiguous request	Remove duplicate tokens while preserving one representative of each value.
True intent	Keep the last occurrence of each token.
Alternative intent	Keep the first occurrence of each token.
Weak clarification	When a duplicate appears later, keep the later copy.
Medium clarification	Preserve the latest occurrence of each duplicated value.
Strong clarification	Remove duplicates and keep the last occurrence of each token.

and 0.0450 for $\lambda = 1$. For both values of λ, no errors are observed among 20,000 trajectories at $T = 1000$; the corresponding one-sided 95% exact Clopper–Pearson upper confidence limit is 1.50×10^{-4}.

Matched protocol comparison. We next compare ordinary closed-loop inference, pruning, a single recap, and snowballing under matched user-shard realizations. All four curves use the same five-intent settings. Ordinary closed-loop inference, recap, and snowballing use $\lambda = 0.5$, pruning removes model-side evidence entirely, and the recap is applied at $t_0 = 8$. Each protocol is evaluated on 20,000 trajectories up to $T = 1000$.

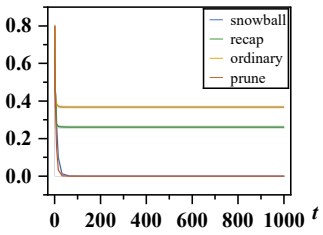


Figure 8: **Matched protocol comparison.**

Fig. 8 shows terminal error as a function of the conversation horizon. Ordinary closed-loop inference retains a long-horizon error near 0.3673, while a single recap reduces it to approximately 0.26095. Pruning reproduces the user-only behavior, with no observed errors from $t = 64$ onward. Snowballing also drives the empirical error rapidly downward: 0.0116 at $t = 32$ and 3.5×10^{-4} at $t = 64$ with no observed errors from $t = 128$ onward.

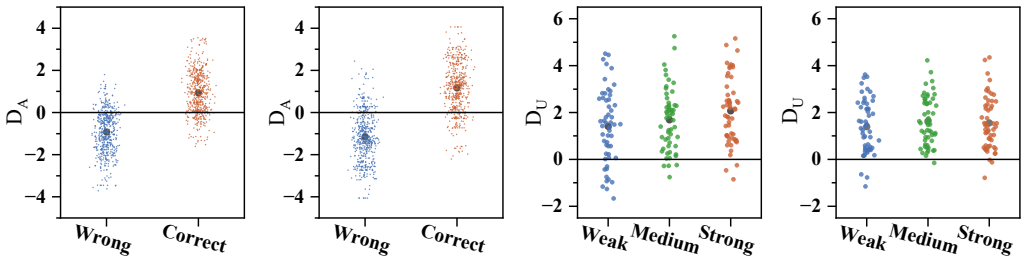
H Real-LLM Mechanism Validation

Goal. Our synthetic experiments test the predictions of the formal model under controlled generative assumptions. We additionally conduct a small real-LLM study to test whether the distinctive directional signatures of the proposed feedback mechanism are observable in actual instruction-tuned language models. This experiment is intended as a behavioral mechanism probe rather than a reproduction of prior large-scale multi-turn degradation benchmarks: (i) fresh user evidence should move the inferred intent toward the user's true intent, whereas (ii) model-generated history should reinforce the interpretation under which that history was generated.

Models and tasks. We evaluate two independently developed open-weight instruction-tuned language models: Qwen2.5-7B-Instruct (Yang et al., 2024) and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023). We construct 54 controlled binary-intent instances spanning 9 semantic task families, with two opposing interpretations and three lexical variants per direction, as an example shown in Table 9. The initial user request is deliberately ambiguous between the two candidate intents. Both models are evaluated in BF16 using their native chat templates.

Intent margin. For conversation history H, we measure the model's preference between the ground-truth intent $z^\star$ and the competing intent z' using $g(H) = s(z^\star \mid H) - s(z' \mid H)$ where $s(\cdot \mid H)$ is the mean conditional token log-likelihood of the candidate intent description. Thus, $g(H) > 0$ favors the true intent and $g(H) < 0$ favors the competing intent. We obtain the scores using a fixed intent-inference probe and do not sample during scoring.

Sampling and statistics. Model responses are sampled with temperature 0.7 and top-$p = 0.95$. For each task and seeded state, we sample 10 responses. We report 95% confidence intervals using cluster bootstrap resampling over the 9 semantic task families.



(a) Qwen: Model Feedback (b) Mistral: Model Feedback (c) Qwen: User Evidence (d) Mistral: User Evidence

Figure 9: **Real-LLM mechanism validation.** (a–b) State-dependent model feedback on Qwen2.5-7B-Instruct and Mistral-7B-Instruct-v0.3. Model responses generated under the wrong interpretation shift the intent margin away from the ground-truth intent ($D_A < 0$), whereas responses generated under the correct interpretation shift it toward the ground-truth intent ($D_A > 0$). (c–d) Effects of fresh user clarification at three levels of linguistic explicitness. Across both models, all clarification conditions produce positive shifts toward the ground-truth intent ($D_U > 0$). Small dots denote sampled-response effects in (a–b) and task-level clarification effects in (c–d); large markers denote aggregate means.

H.1 State-Dependent Model Feedback

For each ambiguous task, we generate a model response under either the true intent (*correct-seeded*) or the competing intent (*wrong-seeded*). The controller instruction is used only to generate the response and is not included in the subsequent scoring context.

Let H_0 denote the ambiguous user request and A a sampled model response. We measure $D_A = g(H_0+A)-g(H_0)$. The theory predicts $\mathbb{E}[D_A \mid \text{correct}] > 0$, and $\mathbb{E}[D_A \mid \text{wrong}] < 0$. We average D_A over the 10 sampled responses for each task–state pair and report the mean effect, family-cluster confidence interval, and the fraction of tasks with the predicted sign.

Fig. 9(a–b) shows a clear state-dependent effect across both models. For Qwen2.5-7B-Instruct, correct-seeded model responses increase the intent margin by 1.166 on average (95% CI:$[0.833, 1.442]$), whereas wrong-seeded responses decrease it by 1.154 ($[-1.425, -0.827]$). Mistral-7B-Instruct-v0.3 exhibits the same pattern: $+0.933$ ($[0.801, 1.076]$) under the correct-seeded intervention and -0.929 ($[-1.084, -0.779]$) under the wrong-seeded intervention.

The effect is not driven by a small number of task instances. The Monte Carlo mean has the predicted sign for 85.2%/83.3% of Qwen tasks under the correct/wrong interventions, respectively, and for 81.5%/83.3% of Mistral tasks. The aggregate sign also remains unchanged when any single semantic task family is removed.

These results are consistent with the central feedback prediction of our analysis: model-generated history tends to reinforce the interpretation under which that history was generated.

Robustness Across Task Families Table 10 reports the model-feedback effect separately for each semantic task family. Values are the mean $\pm$ standard deviation across the 6 task-level Monte Carlo means within each family. Despite substantial within-family variation for some task categories, all 9 families exhibit the predicted direction on average for both models: $D_A > 0$ under the correct-seeded intervention and $D_A < 0$ under the wrong-seeded intervention.

H.2 User-Evidence Effects

We separately append a clarification consistent with the true intent to the ambiguous user request. For each task, we use three levels of linguistic explicitness: weak, medium, and strong. For clarification U, we measure $D_U = g(H_0 + U) - g(H_0)$. The primary prediction is $\mathbb{E}[D_U] > 0$, meaning that fresh user evidence should shift the model toward the true intent. We report the mean effect and family-cluster confidence interval separately for the three clarification levels.

Fig. 9(c–d) reports the effect of fresh user clarification. For Qwen2.5-7B-Instruct, the mean intent-margin shifts are 1.411, 1.690, and 2.067 for the weak, medium, and strong clarification conditions,

Table 10: Model-feedback effect D_A by semantic task family. Values are means, with subscripts denoting standard deviations across the six task-level Monte Carlo means within each family.

	Qwen2.5-7B		Mistral-7B	
Family	Wrong	Correct	Wrong	Correct
Aggregate operation	$-1.254_{\pm 1.878}$	$1.248_{\pm 1.860}$	$-1.044_{\pm 2.021}$	$1.074_{\pm 2.044}$
Deduplication	$-1.428_{\pm 1.624}$	$1.423_{\pm 1.640}$	$-1.242_{\pm 1.064}$	$1.191_{\pm 1.048}$
Extreme value	$-1.688_{\pm 0.269}$	$1.761_{\pm 0.366}$	$-1.316_{\pm 0.307}$	$1.318_{\pm 0.343}$
Parity filtering	$-1.175_{\pm 0.317}$	$1.130_{\pm 0.299}$	$-0.963_{\pm 0.641}$	$0.969_{\pm 0.671}$
Sequence endpoint	$-0.056_{\pm 1.291}$	$0.015_{\pm 1.273}$	$-0.617_{\pm 0.868}$	$0.682_{\pm 0.897}$
Sort direction	$-1.199_{\pm 1.066}$	$1.210_{\pm 1.080}$	$-0.972_{\pm 1.022}$	$0.868_{\pm 1.059}$
String case	$-1.174_{\pm 0.322}$	$1.169_{\pm 0.322}$	$-0.772_{\pm 0.809}$	$0.757_{\pm 0.848}$
Threshold side	$-0.783_{\pm 1.103}$	$0.900_{\pm 1.163}$	$-0.831_{\pm 1.101}$	$0.863_{\pm 1.122}$
Trim side	$-1.630_{\pm 0.866}$	$1.637_{\pm 0.846}$	$-0.608_{\pm 0.392}$	$0.676_{\pm 0.431}$

respectively. The corresponding values for Mistral-7B-Instruct-v0.3 are 1.389, 1.631, and 1.552. All six family-cluster confidence intervals lie strictly above zero.

Thus, across both models and all three clarification conditions, fresh user evidence shifts the model toward the ground-truth intent. The magnitude is not strictly monotonic across clarification levels, so we interpret the three conditions as different degrees of linguistic explicitness rather than as calibrated evidence strengths.