LEARNING PREFERENCES OF STRATEGIC USERS IN SEQUENTIAL DELEGATED MENU DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

AI assistants often learn users' preferences from their choices and use this feedback to shape future recommendations. Strategic users, however, may choose to influence what the assistant offers next. We study this feedback loop in sequential delegated menu design, where an assistant repeatedly reveals options, learns from user choices, and maximizes its own payoff while protecting user welfare. We show that treating strategic choices as truthful can lead to learning incorrect preferences, whereas accounting for strategic behavior enables learning when feedback sufficiently distinguishes possible preferences. Yet, equilibrium feedback need not be informative: there can be an equilibrium in which all users provide the same feedback, leaving the assistant with only its best value without feedback. Moreover, when the assistant's behavior determines a terminal item for each possible preference profile, strategic feedback cannot outperform truthful feedback, even if it reveals the user's preferences. This conclusion has a sharp boundary: strategic feedback can outperform truthful feedback if the assistant observes feedback before randomizing future recommendations. If it randomizes before the user responds, we characterize exactly when truthful and strategic feedback have the same value. Experiments on synthetic interactions and preferences estimated from real-world datasets support these learning and value results.

1 INTRODUCTION

AI assistants often personalize what they show users by learning their preferences from their choices. A shopping assistant may reveal products over several rounds, a travel assistant may refine itineraries, and a recommender may use earlier choices to determine what to display next. In such interactions, the item a user chooses from the available options provides feedback from which the assistant can learn the user's preferences. Real-world preference data illustrate this learning problem: the Coat shopping dataset (Schnabel et al., 2016), for example, records user preferences over items encountered while navigating an online store. However, when users know that their choices affect future recommendations, this feedback need not simply reveal their current preferences. A user may select a less-preferred option to steer the assistant toward more desirable future alternatives.

This creates a feedback loop absent under truthful feedback, where choices reflect the user's current preferences. We study it as *sequential delegated menu design with strategic users*. Over a finite horizon, the assistant reveals items that remain available and learns from the user's choices to guide future recommendations. Yet, the user may choose strategically to improve their final choice. The assistant seeks to maximize its own payoff, such as engagement or revenue, while ensuring that the user's final choice is close, in expectation, to what they would choose if all options were available.

The central question is therefore not only what the assistant can learn from strategic feedback, but how much such feedback is ultimately worth. These questions are distinct: feedback may be informative without being correctly interpreted or improving the assistant's equilibrium payoff. Our main contributions clarify this relationship as follows.

(1) Strategic incentives and equilibrium (Section 3). We formulate strategic feedback in sequential delegated menu design, characterize the user's dynamic best response to a fixed assistant policy, and establish equilibrium existence when both parties act strategically.

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.

(2) Learning from strategic feedback (Section 4). We show that interpreting strategic choices as truthful can drive the assistant's posterior toward an incorrect preference profile, whereas accounting for strategic behavior enables finite-sample learning of the true preferences when feedback sufficiently distinguishes possible profiles.

(3) Equilibrium information and the value of truthful feedback (Sections 5.1–5.2). We show that equilibrium feedback need not be informative: there exists an equilibrium in which the same feedback is provided regardless of the user's preferences, and the assistant obtains exactly its best non-adaptive value, which is also its worst-equilibrium value.

(4) When strategic information has value (Section 5.3). We show that when each possible preference profile receives the same terminal item across its possible equilibrium histories, every such equilibrium attains the best non-adaptive value. This includes deterministic assistant behavior. If the assistant can instead observe the user's feedback and then randomize its future recommendations, strategic feedback can even outperform truthful feedback. When this randomization occurs *before* the user responds, we characterize exactly when truthful and strategic feedback have the same value.

(5) Empirical evaluation (Section 6). Synthetic experiments show that accounting for strategic behavior improves preference learning, while more informative equilibrium feedback need not yield higher value. Experiments using preferences estimated from real-world datasets show that uninformative feedback can attain the best non-adaptive value while truthful feedback can strictly improve on it, and scalability experiments characterize the scope of exact analysis.

Related work. We discuss the closest work here and defer broader comparisons to Appendix A.

Strategic users and learning from endogenous feedback. Recent work studies users who adapt their behavior to influence learning systems, including recommenders (Haupt et al., 2023; Cen et al., 2024a;b). Most closely, Haupt et al. (2023) study strategic consumption in a two-stage recommender system, Cen et al. (2024a) provide experimental evidence that users alter engagement to shape future content, and Cen et al. (2024b) study how such behavior affects platform learning. Strategic classification instead studies agents who manipulate observable features to obtain favorable classification outcomes (Hardt et al., 2016; Dong et al., 2018). In contrast, our setting follows a persistent user over an arbitrary finite horizon, where strategic feedback repeatedly affects beliefs and future recommendations while revealed items accumulate into the user's terminal menu. We characterize both what the assistant can learn from this feedback and its equilibrium value.

Preference learning and adaptive elicitation. IRL, cooperative IRL, learning from human preferences, recommender systems, and preference bandits infer preferences from observed behavior (Ng & Russell, 2000; Hadfield-Menell et al., 2016; Christiano et al., 2017; Ricci et al., 2011; Yue et al., 2012), typically treating feedback as generated from user preferences rather than chosen to influence future interactions. In contrast, our user anticipates how their choices affect the assistant's beliefs and future recommendations. We show that treating such strategic feedback as truthful can lead to incorrect learning, while accounting for strategic behavior enables learning when feedback sufficiently distinguishes preferences. Incentive-compatible active learning recognizes strategic responses to adaptive preference queries and designs them to induce truthfulness (Echenique & Prasad, 2020). Unlike such queries, our recommendations are not payoff-neutral queries that can be withdrawn: revealed items remain in the user's menu and may affect their final choice. Thus, eliciting information also changes the user's strategic incentives and eventual choice set.

Delegation, persuasion, and strategic recommendation. Delegation and Bayesian persuasion shape an agent's decision by restricting feasible actions or designing information (Holmström, 1984; Kamenica & Gentzkow, 2011; Kolotilin & Zapechelnyuk, 2025), including dynamic and online variants (Castiglioni et al., 2020; Gan et al., 2022; Bernasconi et al., 2023; Bacchiocchi et al., 2024a). Unlike these models, our assistant neither merely restricts a known user's choices nor designs signals about a payoff-relevant state: it learns private preferences from strategic feedback while revealing actual alternatives that persist in the user's terminal menu. Work on multistakeholder and profit-aware recommendation instead studies platforms whose objectives may conflict with user welfare (Burke et al., 2016; Hosanagar et al., 2008; Zhang et al., 2021; Bourreau & Gaudin, 2022). We additionally model the user's strategic response to such a platform: their choices both reveal preferences and influence which alternatives the assistant reveals next.

2 PRELIMINARIES

We study a repeated interaction between an assistant a and a user u over a finite horizon of T rounds. In each round, the assistant reveals items and the user chooses from the accumulated menu. The assistant initially does not know the user's preferences and learns from the user's choices, while steering the final outcome toward items that are good for the assistant but acceptable to the user.

Items. To model preferences in a structured way, we encode items by a collection of interpretable features. An item has m features (e.g., price and engine type for a car). Feature $j \in \{1, \dots, m\}$ takes values in a finite set V_j, and feasible items are tuples $\mathbf{x} = (x_1, \dots, x_m)$ in a subset $X \subseteq V_1 \times \cdots \times V_m$; not every combination of feature values needs to be available.

Feature-value utilities. We model preferences over structured items using additive feature-value utilities, a standard framework for capturing preference tradeoffs (Keeney & Raiffa, 1976; Bohanec & Rajkovic, 1988; Boutilier et al., 2004). For each agent $i \in \{a, u\}$, preferences over items are represented by feature weights $\mathbf{w}^i = (w_1^i, \dots, w_m^i) \in \Delta_m$, where $\Delta_m := \{\mathbf{w} \in \mathbb{R}_+^m : \sum_{j=1}^m w_j = 1\}$ is the m-dimensional simplex and the coefficient w_j^i measures the importance that agent i assigns to feature j. For each feature j, agent i also has a value-score function $r_j^i : V_j \to [0, 1]$, where the quantity $r_j^i(v)$ is the score assigned by agent i to value $v \in V_j$ (e.g., a user may prefer a low to a high price). The utility of agent i from item $\mathbf{x}$ is then $U_i(\mathbf{x}) := \sum_{j=1}^m w_j^i r_j^i(x_j)$, and we call $\eta_i = (\mathbf{w}^i, r_1^i, \dots, r_m^i)$ the *type of agent* i. Normalizing each positive-weight feature's scores to span $[0, 1]$ makes w_j^i the range of feature j's utility contribution (see Appendix B.1).

Information structure. The assistant knows its own type η_a but not the user's realized type η_u. It starts from a prior $p \in \Delta(\mathcal{E})$ over a finite hypothesis space $\mathcal{E}$ of candidate user types. Here, $\Delta(\mathcal{E})$ denotes the set of probability distributions over $\mathcal{E}$. For $\eta \in \mathcal{E}$, let U_η denote the corresponding utility.

Interaction. The assistant learns the user's type only from the user's observed choices. At round t, its information is summarized by (A_t, q_t), where $A_t \subseteq X$ is the accumulated menu and $q_t \in \Delta(\mathcal{E})$ the posterior over user types, with $(A_1, q_1) = (\varnothing, p)$. Under the *canonical slate correspondence*, $\mathcal{S}_t(A) := \{S \subseteq X \setminus A : |S| \le k,\ A \cup S \neq \varnothing\}$. After the assistant chooses $S_t \in \mathcal{S}_t(A_t)$, the menu becomes $A_{t+1} = A_t \cup S_t$. Thus, the first-round slate is nonempty, though later slates may be empty. Previously revealed items remain available, so the user chooses from the accumulated menu, capturing settings with persistent alternatives (e.g., itineraries) while abstracting from attention costs and expiration. Failure to switch can itself reveal preferences when a new item does not displace a previous favorite. These dynamics also determine which terminal menus are horizon-reachable. A nonempty menu $A^\star$ is *horizon-reachable* if $A_{T+1} = A^\star$ under some feasible slate sequence. Under the canonical correspondence, this is equivalent to $|A^\star| \le kT$; additional slate restrictions may further restrict reachability (Appendix B.4).

User feedback. The assistant learns the user's preferences from their choices from the accumulated menu. Under *truthful feedback*, choices follow a known type-dependent likelihood. Our canonical model is deterministic truthful choice: the user reports $\mathbf{y}_t \in \arg\max_{\mathbf{x} \in A_{t+1}} U_{\eta_u}(\mathbf{x})$, with fixed tie-breaking. We also consider multinomial-logit feedback as a stochastic truthful-choice model (Appendix B.5). After observing $\mathbf{y}_t$, the assistant updates its posterior by Bayes' rule. As the user's type is fixed, (A_t, q_t) is a sufficient belief state for planning (see Appendix B.5).

Strategic users. Truthful feedback describes myopic choice: the user chooses according to their preferences over the current menu. However, if the user anticipates that their choice affects the assistant's posterior and future slates, they may choose strategically rather than report their currently preferred item. Their choice then serves both as a signal of preference and as a message that can influence future recommendations. We model this behavior by a *signaling rule* σ_t: for each state (A, q), new menu B, and true user type η, $\sigma_t(\cdot \mid A, q, B, \eta) \in \Delta(B)$ gives the distribution over reports. Deterministic truthful choice is the special case that places all probability on a maximizer of U_η over B. If the assistant accounts for strategic behavior, Bayes' rule updates the posterior as $\tau_t^\sigma(q, A, B, \mathbf{y})(\eta) = \frac{q(\eta)\sigma_t(\mathbf{y}|A,q,B,\eta)}{\sum_{\eta' \in \mathcal{E}} q(\eta')\sigma_t(\mathbf{y}|A,q,B,\eta')}$ whenever the denominator is positive. Thus, the strategic signaling probability $\sigma_t(\mathbf{y} \mid A, q, B, \eta)$ replaces the truthful-choice likelihood $\mathbb{P}(\mathbf{y} \mid B, \eta)$; its predictive probability for observing $\mathbf{y}$ is $P_q^\sigma(\mathbf{y} \mid A, B) = \sum_{\eta \in \mathcal{E}} q(\eta)\sigma_t(\mathbf{y} \mid A, q, B, \eta)$.

Terminal decision and objective. After T rounds, the user chooses their preferred item from the terminal menu, $\mathbf{x}_T^\star \in \arg\max_{\mathbf{x}\in A_{T+1}} U_{\eta_u}(\mathbf{x})$. This terminal choice rule is common to all feedback models, which differ only in how intermediate feedback $\mathbf{y}_t$ is generated: under deterministic truthful feedback, $\mathbf{x}_T^\star = \mathbf{y}_T$; under multinomial-logit feedback, $\mathbf{y}_t$ is a noisy preference signal; and under strategic feedback, it may influence the assistant's beliefs and future recommendations. Thus, $\mathbf{x}_T^\star$ is always the user's final utility-maximizing choice from the revealed menu. For any type η, let $\bar{\mathbf{x}}(\eta) \in \arg\max_{\mathbf{x}\in X} U_\eta(\mathbf{x})$ denote a globally optimal item and $\mathbf{x}^\star(A,\eta) \in \arg\max_{\mathbf{x}\in A} U_\eta(\mathbf{x})$ the user's preferred item in nonempty menu A. The assistant payoff and user-welfare loss are $G_a(A,\eta) := \tilde{U}_a(\mathbf{x}^\star(A,\eta)) \in [0,1]$ and $G_u(A,\eta) := U_\eta(\bar{\mathbf{x}}(\eta)) - U_\eta(\mathbf{x}^\star(A,\eta)) \in [0,1]$, where $\tilde{U}_a$ is a positive affine rescaling of U_a to $[0,1]$ (Appendix B.6).

Under a specified truthful feedback model, a randomized Markov assistant policy $\pi = \{\pi_t\}_{t=1}^T$ assigns at each belief state (A,q) a distribution $\pi_t(\cdot \mid A,q)$ over feasible assistant slates $\mathcal{S}_t(A)$ at round t. The posterior evolves according to the corresponding truthful feedback likelihood, and expectations are over $\eta_u \sim p$, feedback randomness, and the assistant's randomization. The assistant maximizes expected terminal payoff subject to an ex-ante bound ϵ on expected user-welfare loss. We call the resulting optimum the *constrained Bayes value* $\mathrm{OPT}(\epsilon) := \sup_\pi\{\mathbb{E}^\pi[G_a(A_{T+1},\eta_u)] : \mathbb{E}^\pi[G_u(A_{T+1},\eta_u)] \le \epsilon\}$, where the supremum is over randomized Markov policies on the belief state and the expectation is taken under the specified truthful feedback model. We consider only values of ϵ for which the constrained optimization problem is feasible (Appendix B.8).

Delegated choice. The assistant controls the revealed menu, but the user chooses the terminal item, making this a *delegated* decision problem. Appendix C compares this problem with a first-best benchmark in which the assistant may directly assign type-specific outcomes.

3 Strategic User Incentives and Equilibrium Existence

We first characterize the user's best response to a fixed assistant policy (Section 3.1) and then establish equilibrium existence when both parties act strategically (Section 3.2). Throughout this section, we adopt a *common-knowledge convention*: the assistant's type η_a, feasible slates, policy class, objective, welfare constraint, and updating rule are known to the user. Only the user's type $\eta \in \mathcal{E}$ is private, while the assistant maintains the public posterior $q \in \Delta(\mathcal{E})$. This isolates one-sided learning while allowing the user to anticipate how feedback affects the assistant's beliefs and future decisions. Appendix D.4 gives a two-sided private information extension.

3.1 User Best Response to a Fixed Assistant Policy

We first characterize a strategic user's best response to a fixed assistant policy π and posterior-update rule $\bar{\tau} = \{\bar{\tau}_t\}_{t=1}^T$. Given state (A,q), menu B, and report $\mathbf{y} \in B$, $\bar{\tau}_t(q,A,B,\mathbf{y}) \in \Delta(\mathcal{E})$ specifies the next posterior, including after zero-probability reports. Under the common-knowledge convention above, the user anticipates how reports affect the posterior and thus the assistant's future slates. Since the user's payoff comes from their terminal choice, they report to maximize expected continuation utility. Let $U_t^{\pi,\bar{\tau}}(A,q,\eta)$ denote the type-η user's optimal continuation utility. Their dynamic best response is $U_t^{\pi,\bar{\tau}}(A,q,\eta) = \sum_S \pi_t(S \mid A,q) \max_{\mathbf{y}\in A\cup S} U_{t+1}^{\pi,\bar{\tau}}\big(A\cup S, \bar{\tau}_t(q,A,A\cup S,\mathbf{y}),\eta\big)$, with $U_{T+1}^{\pi,\bar{\tau}}(A,q,\eta) = \max_{\mathbf{x}\in A} U_\eta(\mathbf{x})$. A deviation changes only the report, not $\bar{\tau}$, so an optimal report may differ from the current utility maximizer to affect future slates.

We next ask when truthful feedback is dynamically optimal, so that the user has no incentive to manipulate the assistant's beliefs through intermediate choices. We call this property *dynamic incentive compatibility*; Appendix D.1 gives its continuation-value formulation.

Definition 3.1 (Dynamic incentive compatibility)**.** *An assistant policy π and posterior-update rule $\bar{\tau}$ are* dynamically incentive compatible *(DIC) if, at every state reachable under truthful reporting and after every feasible assistant slate, each type's utility-maximizing items in the resulting menu also maximize its continuation utility.*

When $\bar{\tau}$ is clear from the context, we simply call π DIC. DIC holds whenever feedback cannot affect future slates, including in the final round. For adaptive policies, however, strategic reporting may improve future recommendations and the user's terminal utility; Appendix D.2 bounds this gain.

3.2 SEQUENTIAL-EQUILIBRIUM EXISTENCE

We now consider the interaction in which both parties optimize strategically. Under the common-knowledge convention, this is a finite-horizon extensive-form Bayesian game with private user type η. Unlike in Section 3.1, the assistant's strategy must also be optimal given the user's reporting strategy. An equilibrium specifies both strategies and beliefs after every history, including zero-probability reports, which we call *off-path reports*. We represent beliefs after such reports by an *off-path posterior selector* $\mu_t(\cdot \mid A, q, B, \mathbf{y}) \in \Delta(\mathcal{E})$. For a signaling rule σ, let $\tau_t^{\sigma,\mu}(q, A, B, \mathbf{y}) = \tau_t^{\sigma}(q, A, B, \mathbf{y})$ when $P_q^{\sigma}(\mathbf{y} \mid A, B) > 0$, and $\tau_t^{\sigma,\mu}(q, A, B, \mathbf{y}) = \mu_t(\cdot \mid A, q, B, \mathbf{y})$ otherwise.

We use sequential equilibrium, which requires both players to act optimally given their information, continuation play, and beliefs. We use *full-history* to emphasize that strategies may condition on the entire observed history rather than only on the public state (A, q). Appendix D.3 gives the formal formulation and conditions for a finite-state Markov representation.

Theorem 3.1 (Sequential-equilibrium existence)**.** *Suppose that X, $\mathcal{E}$, the feasible slate sets, the report sets, and T are finite, and utilities are bounded. Then, the finite extensive-form Bayesian game on the active type space $\mathcal{E}_+ = \mathrm{supp}(p)$ admits a full-history sequential equilibrium.*

Proof sketch. Restricting the initial type draw to $\mathcal{E}_+$ removes only zero-probability types, yielding a finite extensive-form game with perfect recall and a full-support prior. Standard completely mixed perturbations and a vanishing-tremble argument yield a sequential equilibrium; see Appendix D.3.

4 MISSPECIFIED VERSUS EQUILIBRIUM-AWARE LEARNING

We now ask what the assistant can learn from strategic feedback. We compare a *misspecified assistant* that updates its beliefs as if choices followed the truthful-choice model with an *equilibrium-aware assistant* that uses the strategic signalling rule generating them. Both belief updates are applied to the same strategic interaction and observed reports. We show that the misspecified assistant can shift belief toward a false user type, whereas the equilibrium-aware assistant learns the true type when strategic feedback sufficiently separates types. When the signalling rule arises in equilibrium, its likelihood can in principle be obtained from the equilibrium strategies.

Fix the strategic process generating the interaction. For each candidate type η, the *strategic likelihood* is its report distribution under the fixed signalling rule, while the *truthful-choice likelihood* is its distribution under the specified truthful model. Applied to the same reports, they yield the *strategic-likelihood posterior* $q_t^\star$ and *truthful-likelihood posterior* $\widetilde{q}_t$. We assume uniform full support: at every positive-probability history, both likelihoods assign each feasible report probability at least $\alpha > 0$ for every candidate type. Our results below use *cumulative separation*, measuring evidence through round t for distinguishing types. Under the misspecified likelihood, this evidence may favor a false type that better explains the true type's strategic reports. Under the strategic likelihood, it reflects differences between types' strategic reporting distributions. Individual rounds need not be informative if sufficient evidence accumulates. Both results below are finite-sample, not asymptotic. Formal definitions and conditions are in Appendices E.1–E.3.

The first result concerns the misspecified assistant. The true type's strategic reports may resemble the truthful-choice behavior of a false type more closely than that of the true type. If this advantage accumulates across rounds, the truthful-likelihood posterior can shift toward the false type.

Theorem 4.1 (Strategic decoy under the truthful-choice likelihood)**.** *Suppose that, through round t, the strategic reports generated by the true type are cumulatively better explained under the truthful-choice likelihood by some false type $\zeta \neq \eta_0$ than by the true type itself, with an advantage that grows proportionally with the number of observed rounds. Then, with high probability, the truthful-likelihood posterior odds of ζ against η_0 increase according to this accumulated advantage, up to a finite-sample deviation term. In particular, once the accumulated advantage is large enough to overcome the initial posterior odds and the deviation term, the posterior favors ζ over the true type.*

Proof sketch. Log-posterior odds accumulate evidence across rounds. Under the stated condition, their expected drift favors the false type, while full support controls fluctuations. Concentration yields the result; see Appendix E.2 for the full statement and proof.

This failure reflects the wrong likelihood, not uninformative strategic feedback. When strategic reporting distributions sufficiently separate the true type from false types, the strategic-likelihood posterior instead moves toward the true type.

Theorem 4.2 (Finite-sample concentration of the strategic-likelihood posterior)**.** *Suppose that, through round t, the accumulated separation between the strategic reporting distributions of the true type and every false type grows proportionally with the number of observed rounds. Then, with high probability, the strategic-likelihood posterior mass assigned to false types decreases exponentially in this accumulated separation, up to a finite-sample deviation term.*

Proof sketch. Under the separation condition, log-posterior odds drift toward the true type, while full support controls their fluctuations. Concentration and a union bound over false types yield the result; see Appendix E.3 for the full statement and proof.

Together, Theorems 4.1 and 4.2 separate likelihood specification from feedback informativeness. A truthful-choice likelihood can shift belief toward a false type, whereas an equilibrium-aware assistant can learn the true type when strategic feedback is sufficiently informative.

5 Equilibrium Collapse and the Value of Truthful Feedback

We now turn from learning to the value of strategic feedback. We first ask whether equilibrium feedback must be informative and characterize its payoff floor (Section 5.1). We then compare truthful feedback with the worst strategic equilibrium (Section 5.2). Finally, we identify when equilibrium value collapses to the best non-adaptive value and thus cannot exceed truthful feedback (Section 5.3).

5.1 Strategic Manipulation and the Limits of Identification

Section 4 shows that the assistant can learn when strategic feedback sufficiently distinguishes types, but leaves open whether equilibrium provides such information. We show that it need not: all user types can report identically, leaving the assistant's posterior at its prior. Intermediate reports affect the user's payoff only through future recommendations, so report-independent recommendations give no incentive to reveal their type. The reports are therefore analogous to cheap-talk messages (Crawford & Sobel, 1982), although here they can affect the user's eventual payoff by changing the future menu. Following cheap-talk terminology, we call an equilibrium with type-independent reports a *babbling equilibrium*. In our setting, this equilibrium attains exactly the best value without feedback and hence the assistant's worst-equilibrium value under the welfare constraint. We next formalize the equilibria used for our worst- and best-case comparisons.

Definition 5.1 (Welfare-feasible strategic assessment)**.** *Fix $\epsilon \geq 0$. A* welfare-feasible strategic assessment *is a triple (π, σ, μ) of a behavioral Markov assistant policy π, a possibly type-dependent and mixed Markov user signalling rule σ, and an off-path posterior selector μ as in Section 3, such that:* **(i)** (welfare feasibility) $\mathbb{E}^{(\pi,\sigma)}[G_u(A_{T+1}, \eta_u)] \leq \epsilon$*;* **(ii)** (assistant optimality) *holding (σ, μ) fixed, no other behavioral Markov policy satisfying the welfare constraint gives the assistant higher expected terminal payoff;* **(iii)** (user optimality) *at every reporting decision, including after zero-probability events, each type assigns positive probability only to reports maximizing expected continuation utility when future reporting follows σ; and* **(iv)** (belief consistency) *beliefs follow Bayes' rule after positive-probability reports and μ otherwise.*

Let $\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon)$ and $\mathrm{OPT}^{\max}_{\mathrm{strat}}(\epsilon)$ denote the infimum and supremum, respectively, of the assistant's expected terminal payoff over these assessments.

Assistant optimality considers complete alternative policies while holding the user's signalling rule and off-path beliefs fixed and preserving welfare feasibility, whereas user optimality applies at every reporting decision, including after zero-probability reports. Appendices F.1 and F.2 formalize these conditions and their computation. Appendix F.2 also gives an entropy-regularized self-play procedure for generating candidate assessments. We benchmark equilibrium value against what the assistant can secure without feedback.

Definition 5.2 (Non-adaptive value)**.** *A* non-adaptive control *is a family $\kappa = \{\kappa_t(\cdot \mid A)\}_{t=1}^{T}$ of distributions over feasible slates. It may depend on the accumulated menu A and may randomize, but it does not depend on the assistant's posterior or the user's feedback. Let $\mathcal{K}_{\mathrm{nf}}$ denote the set*

of such controls and D_κ the resulting distribution over terminal menus. The non-adaptive value *is* $\mathrm{OPT}_{\mathrm{nf}}(\epsilon) := \max_{\kappa \in \mathcal{K}_{\mathrm{nf}}} \{\sum_\eta p(\eta)\, \mathbb{E}_{A \sim D_\kappa}[G_a(A,\eta)] \ : \ \sum_\eta p(\eta)\, \mathbb{E}_{A \sim D_\kappa}[G_u(A,\eta)] \le \epsilon\}$.

Equivalently, $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ is the best welfare-feasible lottery over horizon-reachable terminal menus (Appendix F.1). Because a non-adaptive control ignores reports, its terminal-menu distribution, expected payoff, and welfare loss are independent of the user's signalling rule. Hence, it lower-bounds the assistant's worst-equilibrium value $\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon)$ of Definition 5.1, i.e., $\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon) \ge \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. Appendix F.1 gives the formal argument; the next result shows equality whenever a welfare-feasible non-adaptive control exists.

Theorem 5.1 (Babbling equilibrium and exact strategic maximin)**.** *Suppose that X, $\mathcal{E}$, the feasible slate sets, the report sets, and the horizon T are finite, and utilities are bounded. Fix $\epsilon \ge 0$ for which a welfare-feasible non-adaptive control exists. Then, there exists a welfare-feasible strategic assessment $(\pi^\circ, \sigma^\circ, \mu^\circ)$ in which all user types follow the same reporting rule. Consequently, the reports reveal no information about the user's type and the assistant's posterior remains equal to the prior, $q_t = p$ for every round t. This babbling equilibrium gives the assistant value* $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$*, and therefore* $\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon) = \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$.

Proof sketch. Construct a type-independent reporting rule and pair it with an optimal welfare-feasible non-adaptive control. Since future slates ignore reports, no type benefits from revealing itself. Conversely, any assistant policy induces a type-independent terminal-menu distribution reproducible non-adaptively, yielding both equilibrium and the value identity. See Appendix F.3.

Together, Theorems 4.1, 4.2 and 5.1 identify two ways learning can fail: the assistant may interpret strategic reports using the wrong likelihood, while even an equilibrium-aware assistant cannot identify the user's type when equilibrium behavior provides no separating information.

5.2 The Value of Truthful Feedback

Theorem 5.1 shows that strategic behavior can leave the assistant with only its best non-adaptive value. We now ask how costly this can be relative to truthful feedback. We prove that truthful feedback is always weakly more valuable than the worst strategic equilibrium, but its relative advantage can be arbitrarily large (Theorem 5.2). We quantify this gap by the following ratio.

Definition 5.3 (Price of Truthful Feedback)**.** *Let* $\mathrm{OPT}_{\mathrm{truthful}}(\epsilon) := \mathrm{OPT}(\epsilon)$ *denote the constrained Bayes value under truthful feedback (Section 2), and let the assistant's worst-equilibrium value* $\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon)$ *be as defined in Definition 5.1. When* $\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon) > 0$*, define the* Price of Truthful Feedback *by* $\mathrm{PoTF}(\epsilon) := \mathrm{OPT}_{\mathrm{truthful}}(\epsilon) / \mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon)$.

Theorem 5.2 (Bounds on the value of truthful feedback)**.** *Fix $\epsilon \ge 0$ for which a welfare-feasible non-adaptive control exists. Then, the assistant's best non-adaptive value* $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ *satisfies* $\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon) = \mathrm{OPT}_{\mathrm{nf}}(\epsilon) \le \mathrm{OPT}_{\mathrm{truthful}}(\epsilon)$*. Hence, whenever it is defined,* $\mathrm{PoTF}(\epsilon) \ge 1$*, and the additive value of truthful feedback is* $\mathrm{OPT}_{\mathrm{truthful}}(\epsilon) - \mathrm{OPT}_{\mathrm{nf}}(\epsilon) \ge 0$*. Moreover, the Price of Truthful Feedback is unbounded across instances and welfare budgets for which* $\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon) > 0$*.*

Proof sketch. By Theorem 5.1, the assistant's worst-equilibrium value equals the non-adaptive value ($\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon) = \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$). The truthful value is at least as large, as every welfare-feasible non-adaptive control is feasible under truthful feedback ($\mathrm{OPT}_{\mathrm{nf}}(\epsilon) \le \mathrm{OPT}_{\mathrm{truthful}}(\epsilon)$). Appendix F.4 proves the result, gives conditions for $\mathrm{PoTF}(\epsilon) = 1$, and shows that PoTF can be unbounded.

5.3 Best-Equilibrium Comparison: When Truthful Feedback Dominates

Section 5.2 compares truthful feedback with the worst strategic equilibrium. We now ask whether the *best* strategic equilibrium can outperform it. Informativeness alone does not determine the answer, since strategic feedback may separate types that truthful feedback pools. Instead, the key restriction comes from incentives over terminal outcomes. We show that if each type receives the same terminal item across its possible equilibrium histories, every welfare-feasible equilibrium attains the best non-adaptive value and cannot outperform truthful feedback. Throughout this subsection, we impose:

(A1) *Strict preferences*: $U_\eta(\mathbf{x}) \ne U_\eta(\mathbf{x}')$ for any type $\eta \in \mathcal{E}$ and distinct items $\mathbf{x}, \mathbf{x}' \in X$, i.e., any nonempty menu has a unique utility-maximizing item for each type (tie-breaking is irrelevant).

(A2) *Non-binding reachability*: the slate correspondence is canonical, $\mathcal{S}_t(A) = \{S \subseteq X \setminus A : |S| \leq k,\ A \cup S \neq \varnothing\}$, and $k \geq |X|$.

(A3) *Deterministic assistant behavior*: at every state at which the assistant may act, it chooses a single feasible slate with probability one, so its behavioral Markov policy is degenerate. The user's signalling rule remains unrestricted and may be type-dependent and mixed.

Assumption (A1) translates equality of continuation utility across histories into equality of the terminal item, while Assumption (A2) makes every nonempty menu horizon-reachable. Assumption (A3) is the substantive equilibrium restriction, ruling out assistant randomization but not user signalling. Appendices F.6, F.10 and F.12 formalize these conditions and their limitations.

The key restriction is self-selection: for the type-specific terminal items to be reproduced by a common menu, each type must prefer its own item to those of the other types.

Definition 5.4 (Self-selecting assignment)**.** *An assignment $(\mathbf{x}_\eta)_{\eta \in S}$ is* self-selecting *if its image menu $A := \{\mathbf{x}_\eta : \eta \in S\}$ satisfies $\mathbf{x}^\star(A, \eta) = \mathbf{x}_\eta$ for any $\eta \in S$, i.e., from the common menu of all assigned items, each user type selects its assigned item rather than one assigned to another type.*

Equilibrium incentives imply self-selection. A type-η user can imitate the reporting strategy of another type η'. This admissible deviation induces the same distribution over public states, and hence over terminal menus, that type η' faces in equilibrium. Since it cannot improve type η's terminal utility, each type must weakly prefer its own equilibrium item to that of any other type. Under Assumption (A1), Appendix F.7 shows that the equilibrium items are thus self-selecting. Placing all these items in a single menu reproduces the equilibrium outcomes. Under Assumption (A2), this menu is horizon-reachable and can be revealed non-adaptively. Together with Theorem 5.1, this yields the collapse result below. More generally, deterministic assistant behavior is unnecessary: it suffices that every equilibrium history for a given type leads to the same terminal item. We call such an assessment *outcome-determinate*; see Definition F.6 in Appendix F.6.

Theorem 5.3 (Equilibrium collapse under outcome determinacy)**.** *Fix ϵ for which a welfare-feasible non-adaptive control exists and assume (A1). Let (π, σ, μ) be an outcome-determinate welfare-feasible strategic assessment, and suppose its image menu is horizon-reachable, which is automatic under Assumption (A2). Then, $\mathbb{E}^{(\pi,\sigma)}\big[\tilde{U}_a(\mathbf{x}_T^\star)\big] = \mathrm{OPT}_{\mathrm{nf}}(\epsilon) \leq \mathrm{OPT}(\epsilon)$. Assumption (A3) implies outcome determinacy, so the result holds even with type-dependent and mixed user signalling.*

Proof sketch. By the imitation argument above, outcome determinacy makes the equilibrium outcomes reproducible by a non-adaptive menu, giving the upper bound $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. Theorem 5.1 gives the reverse inequality. See Appendix F.8 for the full proof.

The theorem is conditional: an outcome-determinate equilibrium requires the deterministic and randomized non-adaptive values to coincide (Corollary F.3). Deterministic assistant behavior is an important special case, while user signalling may remain type-dependent and mixed. Assistant randomization is essential: the collapse holds when it is resolved and observed before the user reports, but can fail when it remains unresolved (Appendices F.11 and F.12). The roles of reachability and strict preferences are discussed in Appendix F.10.

This motivates a public-randomization variant in which the assistant's random seed is realized and observed *before* the user reports (Appendix F.11). Under Assumptions (A1) and (A2), every equilibrium in this variant attains the non-adaptive value $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ (Proposition F.4), allowing us to characterize exactly when it attains the truthful-feedback value.

Theorem 5.4 (When truthful feedback is optimal)**.** *Assume (A1) and (A2), and fix ϵ for which a welfare-feasible non-adaptive control exists. In the public-randomization variant, the following are equivalent:* **(i)** *every welfare-feasible equilibrium has value $\mathrm{OPT}(\epsilon)$;* **(ii)** $\mathrm{OPT}(\epsilon) = \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$*;* **(iii)** *some truthful-optimal policy and off-path beliefs make truthful reporting optimal after every possible report, including zero-probability ones.*

Proof sketch. Conditioning on the public seed reduces the problem to deterministic assistant behavior, so every equilibrium attains $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$, giving **(i)**⇔**(ii)**. Under **(ii)**, an optimal non-adaptive menu lottery makes reports irrelevant and hence truthful reporting optimal; conversely, **(iii)** yields an equilibrium attaining $\mathrm{OPT}(\epsilon)$. See Appendix F.13 for a full proof.

Thus, truthful feedback strictly outperforms every public-randomization equilibrium exactly when truthful adaptation improves on the best non-adaptive control (Proposition F.7). If the assistant

instead randomizes after observing the report, the collapse can fail and strategic feedback can outperform truthful feedback; Appendix F.12 gives an exact counterexample. The worst-equilibrium result of Theorem 5.1 remains unchanged.

6 EXPERIMENTS

We test the main learning and equilibrium results in controlled experiments and with preferences estimated from real choice data. Full protocols and results are in Appendices G–M.

Accounting for strategic behavior changes what is learned. We test Theorems 4.1 and 4.2 by interpreting the same strategic feedback as truthful or strategic. Across $12{,}900$ interactions from 129 reporting behaviors, a false type ends favored in 67.8% of interactions under the truthful interpretation, versus 27.4% under the strategic one; the latter also assigns a higher mean terminal posterior to the true type for every reporting behavior (Appendix L.1). Extending the comparison from $T = 2$ to $T = 4$ strengthens this contrast: false-type dominance under the truthful interpretation rises from 56.0% to 78.6%, while the strategic posterior remains close to the prior (Appendix L.4). Thus, the gap widens mainly because treating strategic feedback as truthful becomes increasingly misleading, rather than because strategic inference improves. This qualitative comparison is robust to the alternative continuation-value convention (Appendix M.2).

More informative feedback need not be more valuable. We construct and verify the uninformative equilibrium of Theorem 5.1 on 600 independently generated synthetic instances across six configurations. It certifies on every instance, with zero informativeness and assistant value equal to the best non-adaptive value (Appendix M.3). In a separate three-instance study, equilibria with different informativeness can have the same assistant value; in one instance, completely uninformative and perfectly informative equilibria attain the same value (Appendix M.4). Thus, informativeness alone is not a proxy for equilibrium value. Exact evaluation of the constructions establishing Theorem 5.2, Proposition F.5 and Proposition F.6 recovers all three comparisons between strategic and truthful feedback: strategic feedback can yield lower, equal, or higher assistant value (Appendix M.5).

The value gap persists with preferences estimated from real data. We fit preference types from Coat, a 300-item shopping-rating dataset (Schnabel et al., 2016), and SUSHI, where $5{,}000$ respondents rated items from a 100-item catalogue (Kamishima et al., 2010); see Appendix H. These datasets provide item structure and preference heterogeneity rather than strategic behavior; the strategic interaction itself is simulated. On every one of the 40 Coat and 20 SUSHI instances, we certify an equilibrium in which feedback reveals no information and the assistant obtains only its best non-adaptive value. Yet, truthful feedback strictly improves on this value on all 20 SUSHI instances and 29 of 40 Coat instances. Thus, even with preferences estimated from real choice data, strategic behavior can eliminate the value of feedback when truthful adaptation would be valuable.

Exact analysis scales to all instances used in our experiments. A scaling study varies catalogue size, preference types, horizon, and slate size (Appendix M.9). Runtime grows most when large catalogues and slates are combined, driven by equilibrium verification and the user's optimal strategic response. With 14 catalogue items and up to 14 items per slate, runtime exceeds three hours and neither step finishes within an hour. All other tested configurations finish within an hour per step.

7 CONCLUSION AND FUTURE WORK

We studied sequential recommendation when an assistant learns a user's private preferences from choices the user may strategically manipulate to influence future recommendations. We showed that treating strategic choices as truthful can lead to incorrect learning, whereas accounting for strategic behavior enables learning when feedback sufficiently separates types. Yet, informative feedback need not have value in equilibrium: uninformative feedback can leave the assistant at its best non-adaptive value, while the timing of assistant randomization determines whether strategic feedback can outperform truthful feedback. Our experiments supported these learning and value results on controlled instances and preferences estimated from real choice data. Future work could allow users to exit early, requiring the assistant to balance information acquisition, recommendation quality, and retention; study which equilibria arise when informative and uninformative equilibria coexist; and allow noisy strategic choices, where feedback reflects both strategic incentives and uncertainty.

AI USE STATEMENT

Generative AI tools were used to assist with writing and presentation of the paper, identifying potentially relevant related work, generating synthetic experimental instances, and generating and editing experimental code. The controlled instance used in the learning experiments and the randomly generated synthetic instances used in the other experiments are described in Appendix G and Appendix G.6, respectively. The authors consulted the original sources and independently verified all citations and characterizations of prior work. All AI-assisted text was reviewed and revised by the authors, and all AI-assisted instances, code, and reported results were independently checked.

Generative AI was not used to develop the theoretical results or proofs. The learning experiments use the controlled instance in Appendix G to study the mechanisms established by Theorems E.1 and E.2; the theoretical results themselves are independent of the experimental instance. All theoretical results and proofs were developed and independently verified by the authors. All empirical analyses, interpretations, and scientific conclusions are the authors' own, and the authors take full responsibility for the content of the paper.

ETHICS STATEMENT

This work is theoretical and computational. It involves no experiments with human participants or deployment of an AI assistant. Our synthetic experiments use generated preference instances. We also construct model instances from the public Coat and SUSHI datasets to incorporate preference heterogeneity and item structure; these datasets are not used as evidence of strategic behavior, and the strategic interactions in our experiments are simulated.

The paper studies assistants that learn user preferences from choices while pursuing an objective that may differ from the user's. This setting raises questions about user welfare and the incentives created when feedback affects future recommendations. We address these concerns explicitly by constraining the expected welfare loss of the user's final choice and by analyzing how strategic behavior affects both preference learning and the value of feedback. The resulting framework provides a formal way to study these tradeoffs, but real deployments may introduce additional considerations, such as privacy, fairness, autonomy, changing preferences, and application-specific legal or institutional requirements. Applying the framework in practice would therefore require evaluating these considerations in the relevant deployment context.

REPRODUCIBILITY STATEMENT

The assumptions, definitions, theorem statements, and proof sketches are given in the main text, with complete proofs and supporting results in Appendices B–F. The experimental instances and data-generating procedures are specified in Appendices G and H; the computational procedures and evaluation metrics are given in Appendices I and J. Detailed empirical results and diagnostics are reported in Appendices L and M, including the numerical-tolerance audit in Appendix M.1 and the computational scaling study in Appendix M.9.

Appendix K specifies the random seeds and validation checks used in the learning experiments. We retain the experimental configurations, deterministic seeding rules, generated instances, fitted preference instances, trajectory-level and intermediate outputs, implementation and validation code, and scripts used to produce the reported tables and figures. Posterior sequences are independently reconstructed from stored interactions, and reported aggregates are recomputed from trajectory-level outputs. Together with the documented procedures, these materials permit reproduction of the reported synthetic and real-data-based analyses.

REFERENCES

Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, pp. 1–8. ACM, 2004.

Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

Guy Aridor and Duarte Gonçalves. Recommenders' originals: The welfare effects of the dual role of platforms as producers and recommender systems. *International Journal of Industrial Organization*, 83:102845, 2022.

Francesco Bacchiocchi, Matteo Bollini, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Online Bayesian persuasion without a clue. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:76404–76449, 2024a.

Francesco Bacchiocchi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Learning optimal contracts: How to exploit small action spaces. In *International Conference on Learning Representations (ICLR)*, volume 2024, pp. 11944–11970, 2024b.

Martino Bernasconi, Matteo Castiglioni, Andrea Celli, Alberto Marchesi, Francesco Trovò, and Nicola Gatti. Optimal rates and efficient algorithms for online Bayesian persuasion. In *International Conference on Machine Learning (ICML)*, pp. 2164–2183. PMLR, 2023.

Marko Bohanec and Vladislav Rajkovic. Knowledge acquisition and explanation for multi-attribute decision making. In *International Workshop on Expert Systems and their Applications*, pp. 59–78. Avignon France, 1988.

Matteo Bollini, Matteo Castiglioni, and Alberto Marchesi. Online contract design with unknown technology. In *International Conference on Machine Learning (ICML)*, 2026.

Marc Bourreau and Germain Gaudin. Streaming platform and strategic recommendation bias. *Journal of Economics & Management Strategy*, 31(1):25–47, 2022.

Craig Boutilier, Ronen I Brafman, Carmel Domshlak, Holger H Hoos, and David Poole. CP-nets: A tool for representing and reasoning withconditional ceteris paribus preference statements. *Journal of artificial intelligence research*, 21:135–191, 2004.

Robin Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.

Robin D Burke, Himan Abdollahpouri, Bamshad Mobasher, and Trinadh Gupta. Towards multi-stakeholder utility evaluation of recommender systems. *UMAP (Extended Proceedings)*, 750, 2016.

Matteo Castiglioni, Andrea Celli, Alberto Marchesi, and Nicola Gatti. Online Bayesian persuasion. *Advances in neural information processing systems (NeurIPS)*, 33:16188–16198, 2020.

Matteo Castiglioni, Alberto Marchesi, Andrea Celli, and Nicola Gatti. Multi-receiver online Bayesian persuasion. In *International Conference on Machine Learning (ICML)*, pp. 1314–1323. PMLR, 2021.

Sarah H. Cen, Andrew Ilyas, Jennifer Allen, Hannah Li, and Aleksander Madry. Measuring strategization in recommendation: Users adapt their behavior to shape future content. In *ACM Conference on Economics and Computation (EC)*, pp. 203–204, 2024a.

Sarah H. Cen, Andrew Ilyas, and Aleksander Madry. User strategization and trustworthy algorithms. In *ACM Conference on Economics and Computation (EC)*, pp. 202, 2024b.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Vincent P Crawford and Joel Sobel. Strategic information transmission. *Econometrica: Journal of the Econometric Society*, pp. 1431–1451, 1982.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *ACM Conference on Economics and Computation (EC)*, pp. 55–70, 2018.

Federico Echenique and Siddharth Prasad. Incentive Compatible Active Learning. In *Innovations in Theoretical Computer Science Conference (ITCS)*, volume 151 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 67:1–67:20, 2020.

Jiarui Gan, Rupak Majumdar, Goran Radanovic, and Adish Singla. Bayesian persuasion in sequential decision-making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5025–5033, 2022.

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, pp. 3909–3917, 2016.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *ACM conference on Innovations in Theoretical Computer Science (ITCS)*, pp. 111–122, 2016.

Andreas Haupt, Dylan Hadfield-Menell, and Chara Podimata. Recommending to strategic users. *arXiv preprint arXiv:2302.06559*, 2023.

Bengt Holmström. On the theory of delegation. In *Bayesian Models in Economic Theory*, pp. 115–141. Elsevier Science, New York, 1984.

Kartik Hosanagar, Ramayya Krishnan, and Liye Ma. Recommended for you: The impact of profit incentives on the relevance of online recommendations. *International Conference on Information Systems (ICIS)*, pp. 31, 2008.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101 (6):2590–2615, 2011.

Toshihiro Kamishima, Hideto Kazawa, and Shotaro Akaho. A survey and empirical comparison of object ranking methods. In *Preference learning*, pp. 181–201. Springer, 2010.

Ralph L. Keeney and Howard Raiffa. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Wiley, New York, 1976.

Anton Kolotilin and Andriy Zapechelnyuk. Persuasion meets delegation. *Econometrica*, 93(1): 195–228, 2025.

Joseph A. Konstan, Bradley N. Miller, David Maltz, Jonathan L. Herlocker, Lee R. Gordon, and John Riedl. GroupLens: Applying collaborative filtering to Usenet news. *Communications of the ACM*, 40(3):77–87, 1997.

Fei Li and Peter Norman. Sequential persuasion. *Theoretical Economics*, 16(2):639–675, 2021.

Kaisa Miettinen. *Nonlinear Multiobjective Optimization*, volume 12 of *International Series in Operations Research & Management Science*. Kluwer Academic Publishers, Boston, MA, 1999.

Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, pp. 663–670. Morgan Kaufmann, 2000.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 27730–27744, 2022.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning (ICML)*, pp. 7599–7609. PMLR, 2020.

Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor (eds.). *Recommender Systems Handbook*. Springer, New York, NY, 2011.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *International Conference on Machine Learning (ICML)*, volume 48, pp. 1670–1679. PMLR, 2016.

Dandan Wang and Yan Chen. Multiobjective and multistakeholder recommender systems. In *International Conference on Intelligent Computing and Signal Processing (ICSP)*, pp. 1221–1224. IEEE, 2021.

Jibang Wu, Zixuan Zhang, Zhe Feng, Zhaoran Wang, Zhuoran Yang, Michael I Jordan, and Haifeng Xu. Sequential information design: Markov persuasion process and its efficient reinforcement learning. In *ACM Conference on Economics and Computation (EC)*, 2022.

Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Annual International Conference on Machine Learning (ICML)*, pp. 1201–1208. ACM, 2009.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Xiaochen Zhang, Pedro Ferreira, Miguel Godinho de Matos, and Rodrigo Belo. Welfare properties of profit maximizing recommender systems: Theory and results from a randomized experiment. *MIS Quarterly*, 45(1):1–43, 2021.

Banghua Zhu, Stephen Bates, Zhuoran Yang, Yixin Wang, Jiantao Jiao, and Michael I Jordan. The sample complexity of online contract design. In *ACM Conference on Economics and Computation (EC)*, pp. 1188–1188, 2023.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Omitted Related Work

This appendix expands the discussion in Section 1. Table 1 summarizes the closest conceptual connections before we discuss them in more detail.

Strategic users, recommendation, and endogenous feedback. The closest work recognizes that behavior observed by a recommendation system may itself be chosen to influence what the system does next. Haupt et al. (2023) model users who strategically consume content during a cold-start phase to affect later recommendations, using a two-stage signaling game, and also provide survey evidence of such behavior. Cen et al. (2024a) experimentally show that users alter engagement when they understand that it affects future content, while Cen et al. (2024b) model repeated interaction between a user and a data-driven platform and study how strategization affects the data available for subsequent decisions. These works share our departure from exogenous feedback: observed behavior can depend on how users expect the system to react.

Our setting differs in the interaction between strategic feedback, learning, and delegated menu construction. We consider a persistent user over an arbitrary finite horizon, with feedback affecting the assistant's beliefs and future recommendations at multiple rounds. Moreover, recommendations are persistent: revealed items accumulate into the menu from which the user ultimately chooses. The assistant must therefore learn the same private preferences that determine both the user's strategic reports and their terminal choice. We characterize when strategic feedback can be learned from, when equilibrium behavior instead reveals no preference information, and when such information has value to an assistant whose objective may differ from the user's.

Our babbling result also connects to cheap-talk games (Crawford & Sobel, 1982). In canonical cheap talk, a sender sends a costless message that can influence the receiver's subsequent action but does not itself enter either player's payoff. Intermediate reports in our model share this indirect-payoff feature, but the strategic environment is different: rather than inducing a single receiver action, a report can change the assistant's beliefs and thereby its future recommendations, which accumulate into the menu from which the user ultimately chooses. The assistant is thus simultaneously learning the user's private preferences and shaping the user's future opportunity set, subject to a welfare constraint. Our babbling result exploits this sequential menu structure: we construct a type-independent reporting rule that, paired with an optimal welfare-feasible non-adaptive control, yields an uninformative equilibrium and exactly identifies the assistant's worst-equilibrium value.

Canonical problem	Key distinction from our model
Strategic recommendation	Models users influencing future recommendations through their behavior, but typically in more restricted interaction structures; unlike such works, our model follows a persistent user over an arbitrary finite horizon while jointly learning preferences and constructing an accumulated terminal menu.
Strategic classification / performative learning	Models observations that respond to a deployed decision rule, but not strategic preference feedback that shapes a persistent user's future choice set. In contrast, our feedback affects both future recommendations and the accumulated terminal menu.
Incentive-compatible elicitation	Accounts for strategic responses to adaptive queries, but those queries need not alter the user's eventual choice set. Unlike such studies, our recommendations are not payoff-neutral queries that can be withdrawn, since revealed items remain available for the terminal choice.
Preference learning / IRL	Infers latent preferences from behavior, but typically does not model feedback as deliberately chosen to influence the learner's future actions. Instead, our user can strategically choose feedback to shape subsequent recommendations.
Bayesian persuasion / delegation	Designs information or restricts feasible actions rather than learning private preferences from strategic feedback while constructing the user's choice set. In contrast, our assistant does both sequentially, with revealed alternatives persisting in the terminal menu.
Principal-agent learning	Learns under strategic interaction and private information through contracts or payments rather than recommendations. Our assistant learns from strategic choices while its recommendations accumulate into the user's terminal menu.

Table 1: Closest canonical problem classes and their main distinctions from our strategic sequential menu-design setting.

Strategic learning and incentive-compatible elicitation. Strategic classification studies prediction when individuals modify observable features in response to the classifier (Hardt et al., 2016; Dong et al., 2018), while performative prediction studies the broader feedback effect in which deploying a predictor changes the distribution on which subsequent predictions are evaluated (Perdomo et al., 2020). These literatures make observed data endogenous to the learning or decision rule. Our endogeneity has a different structure: a persistent user strategically selects intermediate feedback to change the assistant's beliefs and thus influence future recommendations. The latent object being learned is that same user's preference profile, and the assistant's actions also determine the alternatives available for the user's final choice. We also compare learning under a misspecified truthful-feedback model with learning that accounts for the strategic process generating the observations.

A related issue arises in incentive-compatible active learning. Echenique & Prasad (2020) consider adaptive preference elicitation in which subjects may answer strategically to steer later questions and design procedures that preserve incentives for truthful responses. Our recommendations, however, are not payoff-neutral queries that can be withdrawn. Every revealed item remains in the user's accumulated menu and may affect their terminal choice. Thus, elicitation changes not only future queries and strategic incentives but also the feasible set over which the user's final payoff is determined.

Preference learning, inverse reinforcement learning, and human feedback. Inverse reinforcement learning and apprenticeship learning recover latent reward functions from observed behavior (Ng & Russell, 2000; Abbeel & Ng, 2004); cooperative IRL models interaction under uncertainty about a human's reward (Hadfield-Menell et al., 2016); and learning from human feedback estimates reward models from comparisons, rankings, or related signals (Christiano et al., 2017; Ziegler et al., 2019; Ouyang et al., 2022). These approaches motivate our inference problem, but typically model feedback through a specified relationship between latent preferences and observed behavior rather than as a signal deliberately chosen to influence the learner's future actions. In our model,

however, the user anticipates how feedback changes the assistant's beliefs and recommendations. Consequently, the observation process itself depends on strategic incentives. We show that treating such observations as truthful preference evidence can lead to incorrect learning, while accounting for the strategic feedback process permits learning when different preferences generate sufficiently distinct behavior. The assistant also has its own objective and constructs the user's terminal menu, rather than merely estimating or optimizing the user's reward.

Recommendation and preference bandits. Classical recommender systems personalize items under incomplete information about user preferences (Konstan et al., 1997; Adomavicius & Tuzhilin, 2005; Burke, 2002; Ricci et al., 2011), while preference and dueling bandits learn from comparative feedback (Yue & Joachims, 2009; Yue et al., 2012). These settings connect naturally to our sequential preference-learning problem, but our feedback and action dynamics differ. The user strategically chooses from the accumulated revealed menu rather than simply supplying feedback about a current recommendation or comparison. Previously revealed items remain available, so revealing an item both creates a new source of preference information and permanently changes the user's future choice set. Moreover, because the user anticipates subsequent recommendations, intermediate choices need not coincide with the choices that would reveal their current preferences truthfully.

Strategic and multistakeholder platform objectives. A separate recommendation literature studies conflicts among the objectives of platforms, users, providers, and other stakeholders (Burke et al., 2016; Wang & Chen, 2021). Economic work likewise examines profit-aware recommendation, self-preferencing, and strategic recommendation bias, where a platform may favor alternatives that advance its own objective at a cost to relevance or user welfare (Hosanagar et al., 2008; Bourreau & Gaudin, 2022; Zhang et al., 2021; Aridor & Gonçalves, 2022). This literature motivates the separation between assistant and user utilities in our model, but its strategic actor is primarily the platform. Our focus is the resulting response of a strategic user whose private preferences the platform is trying to learn. The assistant chooses recommendations to advance its own payoff subject to a user-welfare constraint, while the user can manipulate the feedback used to choose later recommendations. Our focus is the resulting response of a strategic user whose private preferences the platform is trying to learn. Thus, the assistant must learn preferences from feedback that users may strategically choose to influence its future recommendations.

Bayesian persuasion, delegation, and sequential information design. Bayesian persuasion changes a receiver's decision by designing information, whereas delegation changes the actions available to an agent (Kamenica & Gentzkow, 2011; Holmström, 1984); both are closely related as alternative ways of controlling information and discretion (Kolotilin & Zapechelnyuk, 2025). Sequential and dynamic extensions study repeated disclosure and information design when current signals or actions affect continuation behavior (Li & Norman, 2021; Gan et al., 2022; Wu et al., 2022). Their central instrument nevertheless remains information about a payoff-relevant state. In contrast, our assistant does not privately observe such a state or choose a Bayes-plausible signal. It instead learns the user's private preferences from strategic choices while selecting actual alternatives to reveal. Those alternatives persist, so the assistant simultaneously changes what it knows about the user and the menu from which the user will ultimately choose. The terminal decision is delegated to the user, which further restricts which preference-dependent outcomes can be implemented.

Online Bayesian persuasion introduces learning into information design: signaling schemes are selected repeatedly when receiver utilities, types, or priors are initially unknown (Castiglioni et al., 2020; 2021; Bernasconi et al., 2023; Bacchiocchi et al., 2024a). The learning problem is different from ours. Online-persuasion models generally learn a signaling policy across a sequence of interactions, whereas our model follows one persistent private preference profile within a finite-horizon interaction. More importantly, our control variable is a growing menu rather than a signaling scheme, and the observations used for learning may themselves be strategically chosen to influence that menu.

Online principal-agent learning and contract design. Online principal-agent models study repeated interaction in which a principal faces uncertainty about an agent or its available technology and learns how to design effective contracts from observed outcomes (Zhu et al., 2023; Bacchiocchi et al., 2024b; Bollini et al., 2026). Strategic responses and private information are therefore central in both settings, but the mechanism and learning signal differ. These models use payments or contracts to induce hidden actions and learn from their outcomes. However, our assistant reveals alternatives and learns from the user's explicit but potentially strategic choices. Those choices affect

subsequent recommendations, and the recommendations themselves accumulate into the terminal action set. Thus, unlike a contract that primarily changes incentives, the assistant's learning actions also irreversibly modify the delegated decision problem.

Multi-attribute utility and multi-objective optimization. Our additive feature-value representation follows multi-attribute utility models in which utility is decomposed across attributes (Keeney & Raiffa, 1976; Bohanec & Rajkovic, 1988). The welfare constraint also connects the assistant's optimization problem to the classical ϵ-constraint approach to multi-objective optimization (Miettinen, 1999). These are modeling and optimization connections rather than strategic-learning models: they do not capture a private user who selects feedback to influence subsequent decisions. In our setting, the utility representation defines the latent preferences being learned and the welfare consequences of the terminal menu, while the central difficulty comes from learning those preferences through endogenous strategic feedback as the feasible menu evolves.

B Additional Model Details

B.1 Normalization and Interpretation of Weights

The simplex condition $\mathbf{w} \in \Delta_m$, i.e. $w_j \geq 0$ for every j and $\sum_{j=1}^m w_j = 1$, fixes a convenient scale for the feature weights, but without an additional within-feature normalization it does not in general uniquely determine how each absorbed contribution is split between w_j and $r_j(\cdot)$.

To see this while remaining inside the parameter class $r_j(v) \in [0,1]$, write $\theta_{jv} := w_j r_j(v)$, so that $\theta_{jv} \in [0, w_j]$ and $\sum_j \max_{v \in V_j} \theta_{jv} \leq \sum_j w_j = 1$. Any alternative vector $\mathbf{w}' \in \Delta_m$ satisfying $w'_j \geq \max_{v \in V_j} \theta_{jv}$ for every j induces admissible scores

$$r'_j(v) := \begin{cases} \theta_{jv}/w'_j, & w'_j > 0, \\ 0, & w'_j = 0, \end{cases}$$

where $w'_j = 0$ can occur only when $\theta_{jv} = 0$ for all $v \in V_j$. Then $0 \leq r'_j(v) \leq 1$ and $w'_j r'_j(v) = \theta_{jv}$, so

$$\sum_{j=1}^m w'_j r'_j(x_j) = \sum_{j=1}^m w_j r_j(x_j) \qquad \forall \mathbf{x} \in X.$$

The original $\mathbf{w}$ always satisfies the displayed inequalities. Whenever the feasible set

$$\left\{ \mathbf{w}' \in \Delta_m : \; w'_j \geq \max_{v \in V_j} \theta_{jv} \text{ for every } j \right\}$$

contains more than one vector, the same utility function therefore admits multiple admissible weight–score decompositions. For example, if $m \geq 2$ and $\sum_{j=1}^m \max_{v \in V_j} \theta_{jv} < 1$, then the feasible set contains more than one such vector: writing $a_j := \max_{v \in V_j} \theta_{jv}$ and $\delta := 1 - \sum_j a_j > 0$, moving the slack δ onto one feature or onto a different feature gives two distinct admissible weight vectors.

For this reason, we impose the following within-feature normalization. For every feature j with nonzero weight, the value scores satisfy $\min_{v \in V_j} r_j(v) = 0$ and $\max_{v \in V_j} r_j(v) = 1$. Equivalently, each feature-specific value scale runs from a worst value of 0 to a best value of 1. Under this normalization, $w_j = \max_{v \in V_j} w_j r_j(v) - \min_{v \in V_j} w_j r_j(v)$, so w_j is the range of the latent utility contribution associated with feature j. If the relevant counterfactual tuples belong to X, this is also the attainable utility gain from moving feature j from its worst to its best value while holding all other feature values fixed. When $X \subsetneq \prod_j V_j$, that counterfactual interpretation need not correspond to a feasible pair of items, although the feature-contribution interpretation remains valid.

Under this normalization, w_j equals the range of the feature-j contribution and $r_j(v)$ records the location of value v on that normalized within-feature scale. The normalization is a parameterization convention: it resolves the weight–score scaling ambiguity of a given additive utility representation. It does *not* by itself imply that all such parameters are statistically identified from an arbitrary choice dataset. Identification from observations additionally depends on the available item set, the menus that are observed, and the choice model.

B.2 Feature-Value Indicator Representation

Several proofs are simplified by representing items through feature-value indicator vectors rather than explicit feature weights and value scores; we introduce that notation here. Let $d := \sum_{j=1}^{m} |V_j|$ be the total number of feature-value pairs across all features. For each feature-value pair (j, v), define $\phi_{jv}(\mathbf{x}) := \mathbf{1}\{x_j = v\}$, which equals one exactly when item x realizes value v on feature j. Collecting these indicators gives the feature-value representation $\phi(\mathbf{x}) := (\phi_{jv}(\mathbf{x}))_{j,v} \in \{0,1\}^d$.

B.3 Absorbed Feature-Value Coefficients

Throughout this appendix, we use the feature-value indicator notation introduced in Appendix B.2, together with the absorbed coefficients θ.

Although the feature-value utility model is written in terms of weights and scores, it is often convenient to introduce the absorbed feature-value coefficients, where each feature-value contribution is represented by a single coefficient. For agent $i \in \{a, u\}$, feature j, and value $v \in V_j$, define the absorbed coefficient $\theta^i_{jv} := w^i_j r^i_j(v)$. The coefficient θ^i_{jv} combines the importance of feature j with the attractiveness of value v for agent i. In this notation, the utility of item $\mathbf{x}$ is $U_i(\mathbf{x}) = \sum_{j=1}^{m} \theta^i_{j,x_j} = (\theta^i)^\top \phi(\mathbf{x})$.

The explicit type parameter $\eta = (\mathbf{w}, r_1, \ldots, r_m)$ is used when we discuss feature weights, value scores, and feature-level beliefs. The absorbed coefficient notation θ is used throughout the appendix for algebraic derivations, posterior likelihoods, and norm bounds.

Once the within-feature range normalization of Appendix B.1 is imposed, the normalized weight–score representative associated with a given absorbed coefficient array is recovered feature by feature. Specifically, the weight of feature j is

$$w_j = \max_{v \in V_j} \theta_{jv} - \min_{v \in V_j} \theta_{jv}.$$

If $w_j > 0$, the corresponding normalized value score is

$$r_j(v) = \frac{\theta_{jv} - \min_{\mathbf{z} \in V_j} \theta_{jz}}{w_j}.$$

If $w_j = 0$, feature j has no effect on utility and the value-score function $r_j(\cdot)$ is not pinned down by the representation. These identities concern the normalized parameterization of the additive utility function; they should not be read as a claim that an arbitrary choice dataset identifies every coordinate of θ, since observational identification depends on the menus and feasible items that are observed.

Example B.1. *Consider two features, price and quality. The price values are* $V_{\text{price}} = \{\text{cheap}, \text{moderate}, \text{expensive}\}$ *and the quality values are* $V_{\text{quality}} = \{\text{basic}, \text{standard}, \text{premium}\}$. *Suppose a user has weights* $w^u_{\text{price}} = 0.6$ *and* $w^u_{\text{quality}} = 0.4$. *Thus, price is more important to this user than quality, but quality still has a substantial effect on utility.*

Suppose the user's normalized value scores for price are $r^u_{\text{price}}(\text{cheap}) = 1, r^u_{\text{price}}(\text{moderate}) = 0.55$ *and* $r^u_{\text{price}}(\text{expensive}) = 0$. *The user's normalized value scores for quality are* $r^u_{\text{quality}}(\text{basic}) = 0, r^u_{\text{quality}}(\text{standard}) = 0.65$ *and* $r^u_{\text{quality}}(\text{premium}) = 1$. *The user utility for an item* $\mathbf{x} = (x_{\text{price}}, x_{\text{quality}})$ *is therefore*

$$U_u(\mathbf{x}) = 0.6\, r^u_{\text{price}}(x_{\text{price}}) + 0.4\, r^u_{\text{quality}}(x_{\text{quality}}).$$

The corresponding absorbed coefficients are

$$\theta^u_{\text{price,cheap}} = 0.6, \theta^u_{\text{price,moderate}} = 0.33, \theta^u_{\text{price,expensive}} = 0,\ \theta^u_{\text{quality,basic}} = 0,$$
$$\theta^u_{\text{quality,standard}} = 0.26,\ \theta^u_{\text{quality,premium}} = 0.4.$$

Now, consider the three items $\mathbf{x}^1 = (\text{cheap}, \text{basic}), \mathbf{x}^2 = (\text{moderate}, \text{standard})$ *and* $\mathbf{x}^3 = (\text{expensive}, \text{premium})$. *Their user utilities are* $U_u(\mathbf{x}^1) = 0.6 \cdot 1 + 0.4 \cdot 0 = 0.6, U_u(\mathbf{x}^2) = 0.6 \cdot 0.55 + 0.4 \cdot 0.65 = 0.59, U_u(x^3) = 0.6 \cdot 0 + 0.4 \cdot 1 = 0.4$. *Thus, the user ranks these items as* $\mathbf{x}^1 \succ_u \mathbf{x}^2 \succ_u \mathbf{x}^3$.

The example illustrates the distinction between within-feature value scores and cross-feature tradeoffs. The user assigns the highest within-feature quality score to premium quality, but an expensive premium item is still unattractive because the price feature has a larger weight. Thus, the same preference model can express both the ordering of values within a feature, captured by $r_j(\cdot)$, and the strength of tradeoffs across features, captured by w_j.

B.4 Horizon Reachability

Recall that a nonempty menu $A^\star \subseteq X$ is horizon-reachable if it can arise as A_{T+1} after T feasible rounds. The following observation specializes this notion to the canonical slate correspondence.

Remark B.1 (Reachability under the canonical slate correspondence)**.** *Suppose*

$$\mathcal{S}_t(A) = \big\{S \subseteq X \setminus A : |S| \le k,\ A \cup S \neq \varnothing\big\}$$

for every round t and accumulated menu A. Then a menu $A^\star \subseteq X$ is horizon-reachable if and only if $1 \le |A^\star| \le \min\{kT, |X|\}$. In particular, if $kT \ge |X|$, every nonempty menu is horizon-reachable. With additional restrictions on feasible slates, the cardinality bound $|A^\star| \le kT$ remains necessary but need not be sufficient.

B.5 Sufficiency of the Belief State

In this section, we introduce a *multinomial-logit model* that allows for stochastic choices. Under a user type η, recall that the probability of selecting item $\mathbf{x} \in A_{t+1}$ is

$$\mathbb{P}(\mathbf{y}_t = \mathbf{x} \mid A_{t+1}, \eta) = \frac{\exp(\beta U_\eta(\mathbf{x}))}{\sum_{\mathbf{z} \in A_{t+1}} \exp(\beta U_\eta(\mathbf{z}))},$$

where $\beta > 0$ is an inverse-temperature parameter controlling the degree of choice noise.

Using the absorbed feature-value coefficients introduced in Appendix B.3, the same probability can be written as

$$\mathbb{P}(\mathbf{y}_t = \mathbf{x} \mid A_{t+1}, \eta) = \frac{\exp(\beta \theta_\eta^\top \phi(\mathbf{x}))}{\sum_{\mathbf{z} \in A_{t+1}} \exp(\beta \theta_\eta^\top \phi(\mathbf{z}))}.$$

This form makes explicit that the model is a standard multinomial-logit model with feature vector $\phi(\mathbf{x})$ and parameter vector θ_η.

After observing the user's choice, the assistant updates its belief about the user's type using Bayes' rule. Given the newly formed menu $A_{t+1} = A_t \cup S_t$ and the observed choice $\mathbf{y}_t$, the posterior becomes

$$q_{t+1}(\eta) = \frac{q_t(\eta)\mathbb{P}(\mathbf{y}_t \mid A_{t+1}, \eta)}{\sum_{\eta' \in \mathcal{E}} q_t(\eta')\mathbb{P}(\mathbf{y}_t \mid A_{t+1}, \eta')}.$$

The corresponding predictive probability of observing a choice $y \in A_{t+1}$ under the current belief is

$$\mathbb{P}_t(y \mid A_{t+1}) = \sum_{\eta \in \mathcal{E}} q_t(\eta)\mathbb{P}(y \mid A_{t+1}, \eta).$$

The state therefore evolves from (A_t, q_t) to (A_{t+1}, q_{t+1}). The posterior q_t summarizes all information about the user's preferences that is contained in the observed choices. In particular, it captures uncertainty about both feature importance weights and preferences among values within a feature. As a result, the assistant can reason separately about which features matter to the user and which values of those features are preferred.

Under the fixed-type multinomial-logit model, conditional on the round t, accumulated menu A_t, and posterior q_t, future observations under any slate sequence are independent of the earlier public history. Terminal rewards depend only on the terminal menu and user type, while feasible slates depend on (t, A_t). Hence, (t, A_t, q_t) is a sufficient finite-horizon planning state. Since policies are indexed by t, we write the state as (A_t, q_t). This sufficiency is relative to the specified type space and observation model and does not imply statistical identification.

Remark B.2 (Local-comparison feedback)**.** *An alternative protocol asks the user to choose only from the newly revealed slate S_t, providing direct comparisons among new items. Our model instead keeps previously revealed items available, so the user chooses from the accumulated menu A_{t+1}. This captures persistent candidate sets (e.g., shopping carts or saved itineraries) while isolating the interaction between preference learning and menu construction. Expiring menus and explicit attention costs are natural extensions.*

B.6 Rescaling the Assistant's Utility

Recall that the assistant's raw utility is $U_a(\mathbf{x}) = \sum_{j=1}^{m} w_j^a r_j^a(x_j)$. For the planning problem and all regret quantities defined later in Section 2, we use a fixed normalization of the assistant payoff. Let $U_a^{\min} := \min_{\mathbf{x}\in X} U_a(\mathbf{x})$ and $U_a^{\max} := \max_{\mathbf{x}\in X} U_a(\mathbf{x})$, and assume $U_a^{\max} > U_a^{\min}$. Define $R_a := U_a^{\max} - U_a^{\min}$ and

$$\tilde{U}_a(\mathbf{x}) := \frac{U_a(\mathbf{x}) - U_a^{\min}}{R_a} \in [0,1]. \tag{1}$$

The user-welfare loss is not normalized; it remains in the user's utility units. However, under the utility model of Section 2, user utilities are already bounded. Indeed, because the feature weights satisfy $w_j \geq 0$ and $\sum_{j=1}^{m} w_j = 1$, and the value scores satisfy $r_j(v) \in [0,1]$, it follows that $0 \leq U_\eta(\mathbf{x}) \leq 1$ for every item $\mathbf{x}$ and every user type η. Consequently, user-regret quantities such as $U_\eta(\bar{\mathbf{x}}(\eta)) - U_\eta(\mathbf{x}_T^\star)$ also lie in $[0,1]$.

Using $\tilde{U}_a$ in (1) instead of U_a does not change which policies or oracle items are optimal, because $\tilde{U}_a$ is a positive affine transformation of U_a. The normalization only fixes the scale of assistant payoffs and the units of the Lagrange multiplier used later. The welfare tolerance ϵ remains expressed in the user's utility units and is therefore instance dependent. Consequently, identical numerical values of ϵ need not correspond to the same feasible region across different instances. Throughout the experiments, ϵ should therefore be interpreted as a problem-specific welfare budget rather than a universally comparable quantity.

B.7 Monotonicity of User Utility under Deterministic Revealed-Menu Choice

Proposition B.1 (Monotonicity of user utility under deterministic revealed-menu choice)**.** *Fix a user type η and suppose feedback is deterministic, so that $y_t \in \arg\max_{x\in A_{t+1}} U_\eta(\mathbf{x})$. Then, the user's achieved utility is nondecreasing over time: $U_\eta(y_{t+1}) \geq U_\eta(y_t)$ for any $t \in \{1,\dots,T-1\}$. Equivalently, the user's regret relative to the global optimum is nonincreasing: $U_\eta(\bar{\mathbf{x}}(\eta)) - U_\eta(y_{t+1}) \leq U_\eta(\bar{\mathbf{x}}(\eta)) - U_\eta(y_t)$.*

Proof. Since $A_{t+1} \subseteq A_{t+2}$, maximizing U_η over A_{t+2} cannot yield a lower value than maximizing over A_{t+1}. Therefore, $\max_{x\in A_{t+2}} U_\eta(\mathbf{x}) \geq \max_{x\in A_{t+1}} U_\eta(\mathbf{x})$. The deterministic choices y_t and y_{t+1} attain these two maxima. □

B.8 Feasibility of the Constrained Bayes Value

Recall that an assistant's policy π maps belief states (A_t, q_t) to distributions over slates. As noted in our definition for the constrained Bayes value in Section 2, the assistant solves

$$\begin{aligned} \max_{\pi} \quad & \mathbb{E}^{\pi}_{\eta\sim p}\left[\tilde{U}_a(\mathbf{x}_T^\star)\right] \\ \text{s.t.} \quad & \mathbb{E}^{\pi}_{\eta\sim p}\left[U_\eta(\bar{\mathbf{x}}(\eta)) - U_\eta(\mathbf{x}_T^\star)\right] \leq \epsilon. \end{aligned} \tag{2}$$

The constraint requires the final user-selected item to be close, in expectation, to the user's true best item. For any policy, the unnormalized objective equals $U_a^{\min} + R_a \mathbb{E}^{\pi}[\tilde{U}_a(\mathbf{x}_T^\star)]$, where $R_a = U_a^{\max} - U_a^{\min}$ as defined in Appendix B.6. Hence, the normalized and unnormalized formulations induce the same optimal policies but different multiplier units.

The feasibility and strength of the constraint in (2) depend on the prior p, the hypothesis space $\mathcal{E}$, and the tolerance parameter ϵ. If ϵ is very small, the feasible set may be empty or highly restrictive. Conversely, if ϵ is large, the constraint may become nearly nonbinding. The constrained Bayes value

is therefore understood only for values of ϵ for which the corresponding constrained optimization problem is feasible.

Definition B.1 (Horizon-reachable menu)**.** *A menu $A^\star \subseteq X$ is* horizon-reachable *if there exists a deterministic sequence $S_t \in \mathcal{S}_t(A_t)$, $A_{t+1} = A_t \cup S_t$, $A_1 = \varnothing$, for $t = 1, \dots, T$, such that $A_{T+1} = A^\star$.*

Example B.2 (Expected ϵ-feasibility is not worst-case protection)**.** *The constraint in the main optimization problem is an expectation constraint. It therefore need not imply that every realized user type receives regret at most ϵ. Fix any $\epsilon \in (0, 1)$ and consider a one-round instance with $T = 1$, $k \geq 1$, whose feasible first-round slate correspondence contains the singleton $\{\mathbf{x}^A\}$; this holds, for example, under the canonical slate correspondence of Remark B.1. There are two possible user types, η_C and η_R, with prior probabilities*

$$p(\eta_C) = 1 - \epsilon, \qquad p(\eta_R) = \epsilon.$$

There are two items, $\mathbf{x}^A$ and $\mathbf{x}^R$. The assistant prefers $\mathbf{x}^A$:

$$U_a(\mathbf{x}^A) = 1, \qquad U_a(\mathbf{x}^R) = 0.$$

The common type η_C also prefers $\mathbf{x}^A$, while the rare type η_R strictly prefers $\mathbf{x}^R$:

type	$U_\eta(\mathbf{x}^A)$	$U_\eta(\mathbf{x}^R)$	$\bar{\mathbf{x}}(\eta)$
η_C	*1*	*0*	$\mathbf{x}^A$
η_R	*0*	*1*	$\mathbf{x}^R$

Consider the deterministic policy $S_1 = \{\mathbf{x}^A\}$, which is admissible by construction; its terminal menu is $A_2 = \{\mathbf{x}^A\}$. Both types then choose $\mathbf{x}^A$ from the revealed menu. The expected user-regret is

$$\mathbb{E}\big[U_\eta(\bar{\mathbf{x}}(\eta)) - U_\eta(\mathbf{x}_T^\star)\big] = (1 - \epsilon) \cdot 0 + \epsilon \cdot 1 = \epsilon.$$

Thus, the policy satisfies the expected welfare constraint exactly. However, if the realized type is η_R, then the terminal regret is

$$U_{\eta_R}(\mathbf{x}^R) - U_{\eta_R}(\mathbf{x}^A) = 1 > \epsilon.$$

So a user can end up choosing an item whose welfare loss is larger than ϵ, even though the policy is feasible in expectation. Thus expected feasibility does not imply pathwise welfare protection. A genuine almost-sure version of the welfare constraint would require

$$\mathbb{P}^\pi(G_u(A_{T+1}, \eta_u) \leq \epsilon) = 1,$$

where the probability includes the draw of the user type, feedback randomness, and assistant randomization. Other stronger notions include a chance constraint or a max-regret constraint.

C The Price of Delegated Choice

Before returning to the equilibrium implications of strategic behavior, we isolate a structural feature that is independent of how intermediate feedback is generated. The assistant determines which items are revealed, but the user ultimately chooses its preferred item from the resulting menu. Thus, the assistant influences the terminal outcome indirectly through menu design. We refer to this as a *delegated* decision problem. We next show that delegation is costless when the user's type is known but can be arbitrarily costly under type uncertainty. The key issue is whether a type-contingent assignment can be implemented through a common menu while inducing each type to choose its intended item. We call this property *self-selection*. It also provides the key link to the equilibrium-collapse result in Section 5.3: incentive compatibility forces equilibrium terminal outcomes to be self-selecting and hence reproducible by a single non-adaptive menu.

We compare two constrained benchmarks at the same belief q and under the same ex-ante welfare budget. Under *delegated choice*, the assistant selects a single menu and each user type chooses its preferred item from that menu. Under *direct choice*, the assistant may instead assign a type-specific terminal item. The direct benchmark therefore removes the common-menu constraint and provides a first-best benchmark for measuring the cost of delegation. To formalize the comparison, fix a belief $q \in \Delta(\mathcal{E})$ and welfare tolerance $\epsilon \geq 0$. Let $r_\eta(\mathbf{x}) := U_\eta(\bar{\mathbf{x}}(\eta)) - U_\eta(\mathbf{x}) \geq 0$ denote the regret of

item $\mathbf{x}$ for type η, and recall that $\mathbf{x}^\star(A,\eta)$ denotes the user's tie-broken favorite item in a nonempty menu A. Throughout this section, the fixed tie-breaking rule is induced by a total priority order $\succ$ on X, so $\mathbf{x}^\star(A,\eta)$ is the $\succ$-greatest element of $\arg\max_{\mathbf{x}\in A} U_\eta(\mathbf{x})$.

Definition C.1 (Delegated and direct values)**.** *Let $S := \mathrm{supp}(q) \subseteq \mathcal{E}$. The* delegated value *at belief q is*

$$\mathrm{OPT}_{\mathrm{del}}(q,\epsilon) := \max_{\varnothing \neq A \subseteq X} \left\{ \sum_{\eta\in S} q(\eta)\tilde{U}_a(\mathbf{x}^\star(A,\eta)) : \sum_{\eta\in S} q(\eta) r_\eta(\mathbf{x}^\star(A,\eta)) \le \epsilon \right\}.$$

The direct value *at belief q is*

$$\mathrm{OPT}_{\mathrm{dir}}(q,\epsilon) := \max_{(\mathbf{x}_\eta)_{\eta\in S}\in X^S} \left\{ \sum_{\eta\in S} q(\eta)\tilde{U}_a(\mathbf{x}_\eta) : \sum_{\eta\in S} q(\eta) r_\eta(\mathbf{x}_\eta) \le \epsilon \right\}.$$

When $\mathrm{OPT}_{\mathrm{dir}}(q,\epsilon) > 0$, the Price of Delegated Choice *is* $\mathrm{PoD}(q,\epsilon) := \frac{\mathrm{OPT}_{\mathrm{del}}(q,\epsilon)}{\mathrm{OPT}_{\mathrm{dir}}(q,\epsilon)}$.

This appendix treats delegated choice as a static menu benchmark: $\mathrm{OPT}_{\mathrm{del}}$ ranges over all nonempty menus $A \subseteq X$ and is not restricted by the sequential horizon or by $\mathcal{S}_t(A)$. Whenever we compare this benchmark with the sequential model, reachability is imposed separately.

Definition C.2 (Randomized direct value)**.** *Let $S := \mathrm{supp}(q)$. The* randomized direct value *at belief q is*

$$\mathrm{OPT}^{\mathrm{rand}}_{\mathrm{dir}}(q,\epsilon) := \max_{(\rho_\eta)_{\eta\in S},\ \rho_\eta\in\Delta(X)} \sum_{\eta\in S} q(\eta)\Big\langle \rho_\eta, \tilde{U}_a \Big\rangle \quad \textit{s.t.} \quad \sum_{\eta\in S} q(\eta)\Big\langle \rho_\eta, r_\eta \Big\rangle \le \epsilon,$$

where $\langle \rho, f\rangle := \sum_{\mathbf{x}\in X} \rho(\mathbf{x}) f(\mathbf{x})$. It relaxes Definition C.1 by allowing the assistant to assign each type a lottery *over items rather than a single item, so $\mathrm{OPT}_{\mathrm{dir}}(q,\epsilon) \le \mathrm{OPT}^{\mathrm{rand}}_{\mathrm{dir}}(q,\epsilon)$, with strict inequality in general.*

Proposition C.1 (Randomized first-best upper bound)**.** *For every $\epsilon \ge 0$, $\mathrm{OPT}(\epsilon) \le \mathrm{OPT}^{\mathrm{rand}}_{\mathrm{dir}}(p,\epsilon)$. More generally, the assistant value of every welfare-feasible strategic assessment of Definition 5.1 is at most $\mathrm{OPT}^{\mathrm{rand}}_{\mathrm{dir}}(p,\epsilon)$.*

Proof. Let π be any welfare-feasible policy (respectively, let (π,σ,μ) be any welfare-feasible strategic assessment), and for each $\eta \in \mathrm{supp}(p)$ let $\rho_\eta \in \Delta(X)$ be the conditional law of the terminal item $\mathbf{x}^\star_T$ given $\eta_u = \eta$. Then

$$\sum_\eta p(\eta) \sum_{\mathbf{x}} \rho_\eta(\mathbf{x})\tilde{U}_a(\mathbf{x}) = \mathbb{E}\big[\tilde{U}_a(\mathbf{x}^\star_T)\big], \qquad \sum_\eta p(\eta) \sum_{\mathbf{x}} \rho_\eta(\mathbf{x}) r_\eta(\mathbf{x}) = \mathbb{E}\big[G_u(A_{T+1},\eta_u)\big] \le \epsilon,$$

where the second equality uses $G_u(A_{T+1},\eta) = r_\eta(\mathbf{x}^\star(A_{T+1},\eta))$ and Proposition F.1. Hence $(\rho_\eta)_\eta$ is feasible for Definition C.2 and attains the same value. □

The delegated value corresponds to the terminal mechanics of our model when the assistant reveals a common menu A and each type self-selects $\mathbf{x}^\star(A,\eta)$. The direct value instead allows the assistant to assign a type-specific item $\mathbf{x}_\eta$, providing a first-best benchmark without the self-selection constraint imposed by a common menu. Both problems use the same ex-ante welfare criterion and are feasible for every $\epsilon \ge 0$. For the direct problem, assigning $\mathbf{x}_\eta = \bar{\mathbf{x}}(\eta)$ for every $\eta \in \mathrm{supp}(q)$ gives zero regret. For the delegated problem, revealing the common menu $A = X$ also gives zero regret, because the fixed tie-breaking rule selects $\mathbf{x}^\star(X,\eta) = \bar{\mathbf{x}}(\eta)$ for each type.

The gap between the two benchmarks depends on whether a type-contingent assignment can be implemented through a common menu while preserving each type's incentive to choose its assigned item. This motivates the following notion.

Definition C.3 (Self-selecting assignment)**.** *An assignment $(\mathbf{x}_\eta)_{\eta\in S}$ is* self-selecting *if, letting $A := \{\mathbf{x}_\eta : \eta \in S\}$ denote its image menu, $\mathbf{x}^\star(A,\eta) = \mathbf{x}_\eta$ for every $\eta \in S$. In other words, if the assistant places all assigned items in a common menu, each user type voluntarily selects the item assigned to it rather than an item assigned to another type.*

The following theorem characterizes the cost of delegated choice. Delegation can never improve the assistant's value, and it is costless exactly when a direct-optimal assignment can be implemented through a self-selecting menu.

Theorem C.1 (Delegation gap and its exact characterization)**.** *For every belief $q \in \Delta(\mathcal{E})$ and every $\epsilon \geq 0$, $\mathrm{OPT}_{\mathrm{del}}(q,\epsilon) \leq \mathrm{OPT}_{\mathrm{dir}}(q,\epsilon)$. Moreover, equality holds if and only if there exists a direct-optimal assignment that is self-selecting. In particular, delegation is costless, i.e., $\mathrm{OPT}_{\mathrm{del}}(q,\epsilon) = \mathrm{OPT}_{\mathrm{dir}}(q,\epsilon)$ (and $\mathrm{PoD}(q,\epsilon) = 1$ whenever $\mathrm{OPT}_{\mathrm{dir}}(q,\epsilon) > 0$), in either of the following cases:* **(i)** *the user's type is known (q is degenerate), in which case the direct optimum can be implemented by a singleton menu; or* **(ii)** *there exists an assistant-optimal item $\mathbf{x}^\circ \in X$ that is welfare-feasible for every type, i.e., $r_\eta(\mathbf{x}^\circ) \leq \epsilon$ for every $\eta \in S$ and $\tilde{U}_a(\mathbf{x}^\circ) = \max_{\mathbf{x}\in X} \tilde{U}_a(\mathbf{x})$, in which case the singleton menu $\{\mathbf{x}^\circ\}$ implements the direct optimum.*

Proof. **Inequality.** Let A be any feasible menu for the delegated problem and define the induced assignment $\mathbf{x}_\eta := \mathbf{x}^\star(A,\eta)$ for each $\eta \in S$. The assignment $(\mathbf{x}_\eta)_{\eta\in S}$ is feasible for the direct problem because its expected regret is exactly the delegated regret under A, and hence is at most ϵ. Moreover, its objective value $\sum_{\eta\in S} q(\eta)\tilde{U}_a(\mathbf{x}_\eta)$ is exactly the delegated objective attained by A. Thus, every feasible delegated menu induces a feasible direct assignment with the same assistant value. Taking the maximum over feasible menus yields $\mathrm{OPT}_{\mathrm{del}}(q,\epsilon) \leq \mathrm{OPT}_{\mathrm{dir}}(q,\epsilon)$.

Sufficiency of a self-selecting direct optimum. Suppose that $(\mathbf{x}_\eta)_{\eta\in S}$ is a direct-optimal self-selecting assignment, and let $A := \{\mathbf{x}_\eta : \eta \in S\}$ be its image menu. By self-selection, $\mathbf{x}^\star(A,\eta) = \mathbf{x}_\eta$ for every $\eta \in S$. Hence, when the assistant reveals A, each type chooses exactly the item assigned to it under the direct solution. It follows that the delegated menu A induces the same expected regret and the same assistant value as the direct assignment. Since the latter is feasible and optimal, A is delegated-feasible and attains $\mathrm{OPT}_{\mathrm{dir}}(q,\epsilon)$. Therefore, $\mathrm{OPT}_{\mathrm{del}}(q,\epsilon) \geq \mathrm{OPT}_{\mathrm{dir}}(q,\epsilon)$. Together with the inequality above, this gives $\mathrm{OPT}_{\mathrm{del}}(q,\epsilon) = \mathrm{OPT}_{\mathrm{dir}}(q,\epsilon)$.

Necessity. Suppose now that $\mathrm{OPT}_{\mathrm{del}}(q,\epsilon) = \mathrm{OPT}_{\mathrm{dir}}(q,\epsilon)$, and let $A^\star$ be a menu attaining the delegated optimum. For each $\eta \in S$, define $\mathbf{x}_\eta := \mathbf{x}^\star(A^\star,\eta)$. By the argument used for the inequality, $(\mathbf{x}_\eta)_{\eta\in S}$ is feasible for the direct problem and attains the same value as $A^\star$. Since $\mathrm{OPT}_{\mathrm{del}}(q,\epsilon) = \mathrm{OPT}_{\mathrm{dir}}(q,\epsilon)$, the induced assignment is therefore direct-optimal.

It remains to show that this assignment is self-selecting. Let $A := \{\mathbf{x}_\eta : \eta \in S\}$ denote its image menu. By construction, $A \subseteq A^\star$. Fix any $\eta \in S$. Since $\mathbf{x}_\eta = \mathbf{x}^\star(A^\star,\eta)$, the item $\mathbf{x}_\eta$ maximizes U_η over $A^\star$. Because $\mathbf{x}_\eta \in A \subseteq A^\star$, it also maximizes U_η over the smaller menu A. It remains only to verify that the fixed tie-breaking rule continues to select $\mathbf{x}_\eta$. By definition, $\mathbf{x}_\eta$ is the $\succ$-greatest U_η-maximizer in $A^\star$. The set of U_η-maximizers in A is a subset of the maximizers in $A^\star$ and contains $\mathbf{x}_\eta$. Therefore, $\mathbf{x}_\eta$ is also the $\succ$-greatest maximizer in A. Hence $\mathbf{x}^\star(A,\eta) = \mathbf{x}_\eta$. Since this holds for every $\eta \in S$, the assignment $(\mathbf{x}_\eta)_{\eta\in S}$ is self-selecting.

Special cases. For (i), suppose $q = \delta_{\eta_0}$. The direct problem reduces to $\max\{\tilde{U}_a(\mathbf{x}) : r_{\eta_0}(\mathbf{x}) \leq \epsilon\}$. Let $\mathbf{x}_0$ be an optimizer. The corresponding assignment has the singleton image menu $\{\mathbf{x}_0\}$, from which type η_0 necessarily selects $\mathbf{x}_0$. It is therefore self-selecting, and the equality characterization above gives $\mathrm{OPT}_{\mathrm{del}}(q,\epsilon) = \mathrm{OPT}_{\mathrm{dir}}(q,\epsilon)$ and $\mathrm{PoD}(q,\epsilon) = 1$.

For (ii), suppose there exists $\mathbf{x}^\circ$ satisfying the conditions in the theorem. Assign $\mathbf{x}_\eta := \mathbf{x}^\circ$ to every $\eta \in S$. The assignment is direct-feasible because $r_\eta(\mathbf{x}^\circ) \leq \epsilon$ for every type, and it is direct-optimal because $\mathbf{x}^\circ$ maximizes $\tilde{U}_a$ over X. Its image menu is the singleton $\{\mathbf{x}^\circ\}$ and is therefore self-selecting. The equality characterization again yields $\mathrm{PoD}(q,\epsilon) = 1$. □

Theorem C.1 formalizes Example C.1 below: with a known type, the assistant attains the direct optimum by revealing a singleton menu, and the welfare constraint is identical in both problems. Delegation is therefore not the source of any loss when the type is known.

Example C.1 (Strategic revelation can improve the assistant's payoff)**.** *This example isolates the role of the revealed menu. Suppose the assistant has already learned the user type, so the posterior is degenerate. Let $T = 2$, $k = 1$, and $\epsilon = 0.1$. There are three items:*

item	$U_u(\mathbf{x})$	$U_a(\mathbf{x})$
$\mathbf{x}^U$	*1.0*	*0.0*
$\mathbf{x}^A$	*0.9*	*1.0*
$\mathbf{x}^0$	*0.4*	*0.2*

The user's global optimum is $\bar{\mathbf{x}} = \mathbf{x}^U$. If the assistant reveals both $\mathbf{x}^U$ and $\mathbf{x}^A$, then the final revealed-menu choice is $\mathbf{x}^U$, because $U_u(\mathbf{x}^U) > U_u(\mathbf{x}^A)$. The assistant then receives payoff 0.0.

By contrast, under a slate correspondence permitting these actions — for instance the canonical one of Remark B.1 with $T = 2$, $k \geq 1$ — a strategic assistant can reveal $\mathbf{x}^A$ in the first round and either reveal x^0 or reveal nothing in the second, the empty second slate being feasible because the accumulated menu is already nonempty. The user's final choice is then $\mathbf{x}^A$. The user-regret is

$$U_u(x^U) - U_u(\mathbf{x}^A) = 1.0 - 0.9 = 0.1 = \epsilon,$$

so the welfare constraint is exactly satisfied, while the assistant receives payoff 1.0. *Thus, even with a known user utility, the assistant may benefit from being strategic about what to reveal: revealing the user's true optimum can permanently remove the possibility that the user chooses an assistant-preferred near-optimal item.*

Theorem C.1 shows that delegation is costless when the user's type is known. Under type uncertainty, however, a common menu must simultaneously induce appropriate choices across multiple types. This self-selection constraint can be maximally costly, as the next result shows.

Theorem C.2 (Delegation can be arbitrarily costly)**.** *There exist instances with $|\mathcal{E}| = 2$, a full product item space, and $\epsilon \in (0, \frac{1}{4})$ for which $\mathrm{PoD}(p, \epsilon) = 0$. Consequently, $\inf \mathrm{PoD}(q, \epsilon) = 0$, where the infimum is over all instances and beliefs with $\mathrm{OPT}_{\mathrm{dir}}(q, \epsilon) > 0$: the delegated value can be zero even when the direct value is strictly positive.*

Proof. Fix $\epsilon \in (0, \frac{1}{4})$ and consider the following decoy-anchor instance. Take a single feature $m = 1$ with $w_1 = 1$ and $X = V_1 = \{b, g, p_1, p_2, o\}$, so that $U_\eta(\mathbf{x}) = r_1^\eta(\mathbf{x})$ for each type. There are two user types $\mathcal{E} = \{\eta_1, \eta_2\}$ with $p(\eta_1) = p(\eta_2) = \frac{1}{2}$ and value scores

	b	g	p_1	p_2	o
η_1	1	$1-\epsilon$	0.5	0.1	0
η_2	1	0	0.1	0.5	0

Note that each row has minimum 0 and maximum 1, consistent with the normalization of Section B.1. The assistant's normalized utility satisfies $\tilde{U}_a(g) = 1$ and $\tilde{U}_a(b) = \tilde{U}_a(p_1) = \tilde{U}_a(p_2) = \tilde{U}_a(o) = 0$. Thus, $\bar{\mathbf{x}}(\eta_1) = \bar{\mathbf{x}}(\eta_2) = b$, $r_{\eta_1}(g) = \epsilon$, $r_{\eta_2}(g) = 1$, and $r_{\eta_i}(b) = 0$.

Consider the decoy-anchor construction above at the prior $q = p$.

Direct value. Consider the assignment $\mathbf{x}_{\eta_1} = g$ and $\mathbf{x}_{\eta_2} = b$. Its expected regret is $\frac{1}{2} r_{\eta_1}(g) + \frac{1}{2} r_{\eta_2}(b) = \frac{1}{2}\epsilon \leq \epsilon$, so the assignment is feasible. Its assistant value is $\frac{1}{2}\tilde{U}_a(g) + \frac{1}{2}\tilde{U}_a(b) = \frac{1}{2}$, and hence $\mathrm{OPT}_{\mathrm{dir}}(p, \epsilon) \geq \frac{1}{2}$.

To obtain the reverse inequality, recall that g is the only item with positive assistant utility and that $r_{\eta_2}(g) = 1$. Any assignment with $\mathbf{x}_{\eta_2} = g$ therefore incurs expected regret at least $\frac{1}{2} r_{\eta_2}(g) = \frac{1}{2} > \epsilon$ and is infeasible. Thus, every feasible assignment satisfies $\tilde{U}_a(\mathbf{x}_{\eta_2}) = 0$, so its assistant value is at most $\frac{1}{2}\tilde{U}_a(\mathbf{x}_{\eta_1}) \leq \frac{1}{2}$. Therefore, $\mathrm{OPT}_{\mathrm{dir}}(p, \epsilon) = \frac{1}{2}$.

Delegated value. Let A be any feasible menu for the delegated problem. Feasibility requires $\frac{1}{2} r_{\eta_2}(\mathbf{x}^\star(A, \eta_2)) \leq \epsilon$. Since $\epsilon < \frac{1}{4}$, this implies $r_{\eta_2}(\mathbf{x}^\star(A, \eta_2)) < \frac{1}{2}$, or equivalently $U_{\eta_2}(\mathbf{x}^\star(A, \eta_2)) > \frac{1}{2}$. By construction, b is the only item with utility strictly greater than $\frac{1}{2}$ for type η_2. Hence, every delegated-feasible menu must contain b, and type η_2 selects it: $\mathbf{x}^\star(A, \eta_2) = b$.

But b is also the unique utility-maximizing item for type η_1, since $U_{\eta_1}(b) = 1 > 1 - \epsilon = U_{\eta_1}(g)$ and all other items have still lower utility. Thus, whenever $b \in A$, $\mathbf{x}^\star(A, \eta_1) = b$ as well, so $\tilde{U}_a(\mathbf{x}^\star(A, \eta_1)) = 0$. Likewise, $\mathbf{x}^\star(A, \eta_2) = b$, so $\tilde{U}_a(\mathbf{x}^\star(A, \eta_2)) = 0$. Since $\tilde{U}_a(b) = 0$, both types therefore select an item with zero assistant utility from every feasible menu. It follows that $\mathrm{OPT}_{\mathrm{del}}(p, \epsilon) = 0$.

Consequently, $\mathrm{PoD}(p,\epsilon) = \mathrm{OPT}_{\mathrm{del}}(p,\epsilon)/\,\mathrm{OPT}_{\mathrm{dir}}(p,\epsilon) = 0$. □

The construction also illustrates the mechanism behind the delegation gap. With a known type, a type-specific singleton menu can implement the direct optimum. Under uncertainty, however, the assistant must offer a common menu to all types. In the construction above, welfare feasibility forces the anchor item b into this menu, but b is also preferred by η_1 to the assistant-preferred item g. Thus, the direct-optimal assignment $(\mathbf{x}_{\eta_1}, \mathbf{x}_{\eta_2}) = (g, b)$ cannot be implemented through a common menu: in the image menu $\{g, b\}$, type η_1 selects b rather than g. By Theorem C.1, this failure of self-selection is exactly what makes delegation costly.

Together, Theorems C.1 and C.2 identify self-selection as the key constraint imposed by delegated choice. When a direct optimum is self-selecting, delegation is costless; otherwise, the loss can be maximal. This property will play a central role in the strategic equilibrium analysis that follows.

D Omitted Proofs and Additional Results for Section 3

D.1 Continuation-Value Formulation of Dynamic Incentive Compatibility

Definition D.1 (Continuation-value formulation of dynamic incentive compatibility)**.** *Fix an assistant policy π, a posterior-update rule $\bar{\tau}$, and the deterministic truthful signaling rule σ^{tr} for revealed-menu choice. A public state (A, q) at the beginning of round t is* truthfully reachable under $(\pi, \bar{\tau})$ *if it has positive probability starting from $(A_1, q_1) = (\varnothing, p)$ when the assistant follows π, the user follows σ^{tr}, and posteriors evolve according to $\bar{\tau}$. The pair $(\pi, \bar{\tau})$ is* dynamically incentive compatible *for truthful revealed-menu choice if, for every truthfully reachable state (A, q) at round t, every type η, and every feasible assistant slate S with resulting menu $B = A \cup S$, every myopic maximizer $\mathbf{y}^\star \in \arg\max_{\mathbf{x}\in B} U_\eta(\mathbf{x})$ also maximizes continuation utility, i.e., $\mathbf{y}^\star \in \arg\max_{\mathbf{y}\in B} U_{t+1}^{\pi,\bar{\tau}}(B, \bar{\tau}_t(q, A, B, \mathbf{y}), \eta)$.*

D.2 Strategic Gain is Bounded by Posterior Sensitivity

Theorem D.1 (Strategic gain is bounded by posterior sensitivity)**.** *Fix an assistant policy π, a posterior-update rule $\bar{\tau}$, a round t, a truthfully reachable state (A, q) under $(\pi, \bar{\tau})$ in the sense of Definition D.1, and a user type η. Suppose π selects a deterministic slate S at (A, q), and let $B = A \cup S$ be the resulting menu. Define the one-step strategic signaling gain relative to truthful myopic feedback by*

$$\begin{aligned}\Delta_t^{\pi,\bar{\tau}}(A, q, \eta) := \max_{\mathbf{y}\in B} U_{t+1}^{\pi,\bar{\tau}}\big(B,\ \bar{\tau}_t(q, A, B, \mathbf{y}),\ \eta\big) \\ - \max_{\mathbf{y}^\star \in \arg\max_{\mathbf{x}\in B} U_\eta(\mathbf{x})} U_{t+1}^{\pi,\bar{\tau}}\big(B,\ \bar{\tau}_t(q, A, B, \mathbf{y}^\star),\ \eta\big).\end{aligned}$$

Assume that $U_{t+1}^{\pi,\bar{\tau}}(B, \cdot, \eta)$ is L-Lipschitz in total variation distance. Then

$$0 \le \Delta_t^{\pi,\bar{\tau}}(A, q, \eta) \le L \max_{\substack{\mathbf{y}\in B \\ \mathbf{y}^\star \in \arg\max_{\mathbf{x}\in B} U_\eta(\mathbf{x})}} \big\|\bar{\tau}_t(q, A, B, \mathbf{y}) - \bar{\tau}_t(q, A, B, \mathbf{y}^\star)\big\|_{\mathrm{TV}}.$$

In particular, if the user's continuation utility is constant over the posteriors induced by the available reports in B, then $\Delta_t^{\pi,\bar{\tau}}(A, q, \eta) = 0$.

Proof. The gain is nonnegative because the maximization over all $\mathbf{y} \in B$ includes the truthful reports $\mathbf{y}^\star \in \arg\max_{\mathbf{x}\in B} U_\eta(\mathbf{x})$. Let $\hat{\mathbf{y}}$ attain the first maximum and let $\hat{\mathbf{y}}^\star$ attain the second maximum. Write $q_{\mathbf{y}}^+ := \bar{\tau}_t(q, A, B, \mathbf{y})$. By the Lipschitz condition, $|U_{t+1}^{\pi,\bar{\tau}}(B, q', \eta) - U_{t+1}^{\pi,\bar{\tau}}(B, q'', \eta)| \le L\|q' - q''\|_{\mathrm{TV}}$, from which we infer

$$\begin{aligned}\Delta_t^{\pi,\bar{\tau}}(A, q, \eta) &= U_{t+1}^{\pi,\bar{\tau}}(B, q_{\hat{\mathbf{y}}}^+, \eta) - U_{t+1}^{\pi,\bar{\tau}}(B, q_{\hat{\mathbf{y}}^\star}^+, \eta) \\ &\le L\,\|q_{\hat{\mathbf{y}}}^+ - q_{\hat{\mathbf{y}}^\star}^+\|_{\mathrm{TV}} \\ &\le L \max_{\substack{\mathbf{y}\in B \\ \mathbf{y}^\star \in \arg\max_{\mathbf{x}\in B} U_\eta(\mathbf{x})}} \big\|\bar{\tau}_t(q, A, B, \mathbf{y}) - \bar{\tau}_t(q, A, B, \mathbf{y}^\star)\big\|_{\mathrm{TV}}.\end{aligned}$$

This proves the stated bound. If $U_{t+1}^{\pi,\bar{\tau}}(B, q', \eta)$ is constant over the posteriors induced by the available reports in B, then its Lipschitz constant on this set can be taken to be zero, and hence $\Delta_t^{\pi,\bar{\tau}}(A, q, \eta) = 0$. □

D.3 Proof of Theorem 3.1

The equilibrium-existence result below is formulated in the finite extensive-form game with behavioral strategies. In the baseline model, the assistant uses a behavioral Markov policy: at each round t and public belief state (A_t, q_t), $\pi_t(\cdot \mid A_t, q_t)$ is a distribution over feasible slates, and the realized slate is publicly observed. The strategic user likewise uses a behavioral Markov signalling rule: after state (A, q) and revealed menu $B = A \cup S$, a type-η user reports according to $\sigma_t(\cdot \mid A, q, B, \eta) \in \Delta(B)$, allowing type-dependent mixed signalling. Thus, the baseline game contains only the behavioral randomization generated by these kernels; there is no persistent hidden assistant randomization.

We now give the finite-state Markov formulation used in the representation part of Theorem 3.1.

Definition D.2 (Finite-state Markov perfect Bayesian assessment)**.** *Fix a finite collection of public belief states $\mathcal{Z}_t \subseteq 2^X \times \Delta(\mathcal{E})$, $t = 1, \dots, T+1$. A Markov assessment $(\pi^\star, \sigma^\star, \mu^\star)$ is a* Markov perfect Bayesian assessment *on $\{\mathcal{Z}_t\}_{t=1}^{T+1}$ if, at any state $(A, q) \in \mathcal{Z}_t$ and any round t:*

(i) Assistant optimality. *Given $(\sigma^\star, \mu^\star)$, the assistant policy $\pi_t^\star(\cdot \mid A, q)$ maximizes its expected continuation value over all feasible mixed slate choices;*

(ii) User optimality. *Given $(\pi^\star, \mu^\star)$, after any assistant slate S and revealed menu $B = A \cup S$, the user signaling rule $\sigma_t^\star(\cdot \mid A, q, B, \eta)$ maximizes the type-η user's expected continuation value for any $\eta \in \mathcal{E}$; and*

(iii) Belief consistency. *After observing $\mathbf{y} \in B$, the next posterior is $q^+ = \tau_t^{\sigma^\star, \mu^\star}(q, A, B, \mathbf{y})$, so beliefs are updated by Bayes' rule following positive-probability signals and by the off-path posterior selector $\mu^\star$ following zero-probability signals.*

Since the horizon, action sets, and type space are finite, the interaction defines a finite extensive-form Bayesian game with full-history sequential equilibrium as its natural equilibrium notion. Such an equilibrium need not be Markov in (A, q), since the same public state can arise after histories with different continuation play. We therefore state the condition under which a finite-state Markov representation is valid.

Assumption D.1 (Markov sufficiency: state consistency of an assessment)**.** *A full-history sequential assessment is* state-consistent *if, whenever two public histories at the same round induce the same public state (A, q), they induce the same continuation assistant strategy, the same continuation reporting strategy for every active user type, and the same continuation belief system.*

The existence result in Theorem 3.1 uses sequential equilibrium for the underlying full-history game. We state the formal requirements here before proving the result. A full-history *assessment* $\mathbf{a} = (s, \nu)$ consists of a behavioral-strategy profile s for the assistant and the user, specified at every information set at which they may act, together with a belief system ν over the user's type at every information set where such beliefs are needed to evaluate continuation play. Sequential equilibrium imposes two requirements on such an assessment: behavior must remain optimal as the interaction unfolds, and the beliefs used to evaluate continuation play must be consistent with the strategies.

The first requirement is sequential rationality. Optimality is imposed not only from the beginning of the interaction, but at every subsequent decision point: conditional on the information available there, the beliefs specified by the assessment, and the continuation play prescribed by the assessment, the acting player must have no profitable deviation.

Definition D.3 (Sequential rationality)**.** *A full-history assessment $\mathbf{a} = (s, \nu)$ is* sequentially rational *if, at every information set of each player, the continuation strategy prescribed by s maximizes that player's expected continuation utility given the information available at that information set, the beliefs specified by ν, and the other player's continuation strategy.*

Sequential rationality alone does not determine how beliefs should be formed after histories that have zero probability under the prescribed strategies. Sequential equilibrium therefore also requires

consistency of beliefs. To define this requirement, consider completely mixed behavioral-strategy profiles, in which every feasible action at every information set is assigned positive probability and beliefs are determined by Bayes' rule. The beliefs specified by the sequential equilibrium, including those following zero-probability histories, must arise as limits of these Bayesian beliefs as the completely mixed strategy profiles converge to the equilibrium strategy profile.

Definition D.4 (Sequential equilibrium)**.** *A full-history assessment* $\mathbf{a} = (s, \nu)$ *is a* sequential equilibrium *if:*

(i) Sequential rationality. *The assessment* $\mathbf{a}$ *is sequentially rational in the sense of Definition D.3; and*

(ii) Belief consistency. *There exists a sequence* $\{(s^n, \nu^n)\}_{n=1}^{\infty}$ *such that each* s^n *is a completely mixed behavioral-strategy profile,* ν^n *is the belief system obtained from* s^n *by Bayes' rule, and* $s^n \to s$ *and* $\nu^n \to \nu$ *as* $n \to \infty$*.*

For the existence result, let $\mathcal{E}_+ := \operatorname{supp}(p)$. Types outside $\mathcal{E}_+$ occur with zero probability, so restricting the initial type draw to $\mathcal{E}_+$ does not change any positive-probability history. We identify posteriors on $\mathcal{E}_+$ with elements of $\Delta(\mathcal{E})$ assigning zero probability to $\mathcal{E} \setminus \mathcal{E}_+$.

Theorem D.2 (Full statement of Theorem 3.1: Sequential equilibrium and finite-state Markov representation)**.** *Suppose that* X*,* $\mathcal{E}$*, the feasible slate sets, the report sets, and the horizon* T *are finite, and utilities are bounded. Then, the finite extensive-form Bayesian game on the active type space* $\mathcal{E}_+ = \operatorname{supp}(p)$ *admits a full-history sequential equilibrium.*

Moreover, let $\mathbf{a}^\star$ *be any full-history sequential-equilibrium assessment of this reduced game that is state-consistent in the sense of Assumption D.1. For each round* t*, define*

$$\mathcal{Z}_t(\mathbf{a}^\star) := \{(A_t(h_t), q_t(h_t)) : h_t \text{ is a public history at round } t\},$$

where $q_t(h_t)$ *is the public posterior specified by the equilibrium belief system at* h_t*. Then, each* $\mathcal{Z}_t(\mathbf{a}^\star)$ *is finite, and* $\mathbf{a}^\star$ *admits a Markov perfect Bayesian representation on* $\{\mathcal{Z}_t(\mathbf{a}^\star)\}_{t=1}^{T+1}$ *in the sense of Definition D.2, with the type space restricted to* $\mathcal{E}_+$*. The Markov representation preserves the continuation strategies and beliefs of* $\mathbf{a}^\star$ *at every public history.*

Proof. We prove the two claims separately.

Full-history sequential-equilibrium existence. Restrict the initial type draw to $\mathcal{E}_+ = \operatorname{supp}(p)$. Every type in this reduced game therefore has strictly positive prior probability. The resulting interaction is a finite extensive-form game with perfect recall because the horizon, type space, feasible slate sets, and report sets are all finite.

For sufficiently small $\delta > 0$, perturb the game by requiring every feasible behavioral action at every information set to receive probability at least δ. The perturbed game is finite and admits an equilibrium in behavioral strategies. Because every active type has positive prior probability and every action is completely mixed, every feasible information set is reached with positive probability in the perturbed game, so Bayes' rule determines beliefs at every such information set.

Let $\delta_n \downarrow 0$. The behavioral-strategy spaces and belief spaces are finite-dimensional and compact, so a subsequence of the corresponding equilibrium assessments converges to an assessment $\mathbf{a}^\star$. By construction, the perturbed strategy profiles are completely mixed, their beliefs are obtained by Bayes' rule, and both the strategies and beliefs converge to those of $\mathbf{a}^\star$. Hence the limiting assessment satisfies the belief-consistency condition for sequential equilibrium from Definition D.4. Sequential rationality follows from the vanishing-tremble argument. If an action prescribed with positive probability in the limit were strictly suboptimal at some information set, the same strict improvement would persist for all sufficiently small perturbations, contradicting optimality in the perturbed equilibria. Hence $\mathbf{a}^\star$ is sequentially rational in the sense of Definition D.3. Together with belief consistency, Definition D.4 implies that $\mathbf{a}^\star$ is a full-history sequential equilibrium of the reduced game.

Finite-state Markov representation under state consistency. Now, let $\mathbf{a}^\star$ be any full-history sequential-equilibrium assessment of the reduced game satisfying Assumption D.1. This need not be the equilibrium obtained in the first part.

For each round t, define $\mathcal{Z}_t(\mathbf{a}^\star) := \{(A_t(h_t), q_t(h_t)) : h_t \text{ is a public history at round } t\}$, where $q_t(h_t)$ is the public posterior specified by the belief system of $\mathbf{a}^\star$ at h_t. The extensive-form game has finitely many public histories, so each $\mathcal{Z}_t(\mathbf{a}^\star)$ is finite.

Fix $(A, q) \in \mathcal{Z}_t(\mathbf{a}^\star)$ and choose any public history h_t inducing this state. Define $\pi_t^\star(\cdot \mid A, q)$ to be the assistant continuation strategy prescribed by $\mathbf{a}^\star$ at h_t. For any feasible slate S, let $B = A \cup S$ and define $\sigma_t^\star(\cdot \mid A, q, B, \eta)$ to be the continuation reporting strategy prescribed by $\mathbf{a}^\star$ for type $\eta \in \mathcal{E}_+$ at the corresponding user information set. Finally, for a report $\mathbf{y} \in B$ having zero probability under the prescribed signaling rule, define $\mu_t^\star(\cdot \mid A, q, B, \mathbf{y})$ to be the corresponding off-path posterior prescribed by the continuation belief system of $\mathbf{a}^\star$.

These objects are well defined and do not depend on the choice of the public history h_t. Indeed, by Assumption D.1, any two public histories at the same round that induce the same state (A, q) prescribe the same continuation assistant strategy, the same continuation reporting strategy for every active type, and the same continuation belief system. Hence, $(\pi^\star, \sigma^\star, \mu^\star)$ is well defined as a collection of Markov kernels on $\{\mathcal{Z}_t(\mathbf{a}^\star)\}_{t=1}^{T+1}$, which do not depend on the choice of h_t.

It remains to verify the three conditions of Definition D.2. First, sequential rationality of the full-history equilibrium $\mathbf{a}^\star$ implies that, at every public history inducing (A, q), the assistant's prescribed mixed slate choice maximizes its expected continuation value against the continuation assessment prescribed by $\mathbf{a}^\star$. By Assumption D.1, all public histories at the same round that induce (A, q) have the same continuation assessment. Hence the same optimality condition depends only on (A, q) and gives assistant optimality for $\pi_t^\star(\cdot \mid A, q)$.

Second, after every feasible slate and for every active type $\eta \in \mathcal{E}_+$, sequential rationality of the full-history equilibrium implies that the prescribed reporting strategy assigns positive probability only to reports that maximize the type's expected continuation utility. State consistency again makes this prescription depend only on (A, q, B, η), so $\sigma^\star$ satisfies the user-optimality condition.

Third, after every positive-probability report, the belief system of $\mathbf{a}^\star$ follows Bayes' rule. After a zero-probability report, the posterior is, by construction, the off-path belief inherited from $\mathbf{a}^\star$. Therefore, the Markov representation uses the posterior update

$$q^+ = \tau_t^{\sigma^\star, \mu^\star}(q, A, B, \mathbf{y}),$$

with the Bayesian posterior after positive-probability reports and $\mu^\star$ after off-path reports. Hence, the representation preserves the beliefs of $\mathbf{a}^\star$ and satisfies the belief-updating condition of Definition D.2.

Consequently, $(\pi^\star, \sigma^\star, \mu^\star)$ is a Markov perfect Bayesian representation of $\mathbf{a}^\star$ on $\{\mathcal{Z}_t(\mathbf{a}^\star)\}_{t=1}^{T+1}$, with the type space restricted to $\mathcal{E}_+$. The representation is sequen- tially rational at every public history, follows Bayes' rule after every positive-probability report, and retains the equilibrium's off-path beliefs after zero-probability reports. □

Types in $\mathcal{E} \setminus \mathcal{E}_+$ have zero probability under the prior and therefore do not affect the existence result or the finite-state representation above. The theorem therefore restricts the equilibrium type space to $\mathcal{E}_+$. If a later definition requires strategies to be specified for all candidate types in $\mathcal{E}$, prescriptions for types outside $\mathcal{E}_+$ can be supplied separately; this is the case, for example, in Definition 5.1. The completely mixed perturbations used in the existence proof apply to players' actions and do not alter the initial type distribution.

The finite-state conclusion in Theorem D.2 is conditional on state consistency. The existence of a full-history sequential equilibrium does not by itself imply the existence of an equilibrium that is Markov in (A, q). Rather, if a full-history sequential equilibrium satisfies Assumption D.1, its continuation strategies and beliefs are the same at public histories that induce the same state (A, q) and can therefore be represented on the finite state domain constructed above.

D.4 Assistant-Side Uncertainty

The strategic equilibrium model of Section 3.2 assumes that the assistant-side primitives relevant to the interaction are common knowledge. These include the assistant's objective, utility function, feasible-slate correspondence, and information. We now relax this common-knowledge assumption

by allowing the assistant to have privately observed information about these primitives. Let Ω be a finite set of assistant-side models, where an element $\omega \in \Omega$ specifies the assistant's objective, utility function, information, and feasible-slate correspondence.

Let $\lambda \in \Delta(\mathcal{E} \times \Omega)$ be a common prior over the user's type and the assistant-side model. At the beginning of the interaction, a pair (η, ω) is drawn according to λ. The user observes η, the assistant observes ω, and both observe the public interaction history. Thus, each party may have private information that is relevant to continuation play. Let $\Theta_+ := \mathrm{supp}(\lambda) \subseteq \mathcal{E} \times \Omega$ denote the active joint type space.

In this Bayesian extension, posterior beliefs are induced by the common prior, the players' strategies, and their observations. In particular, an arbitrary exogenous non-Bayesian updating algorithm is not treated as a component of ω. Allowing the assistant to have a privately known misspecified or otherwise constrained updating rule would instead define a different subjective or behavioral model, for which the assistant's decision problem and the user's beliefs about that problem would have to be specified separately. The result below concerns the Bayesian extension described above.

Proposition D.1 (Equilibrium with assistant-side uncertainty)**.** *Suppose that X, $\mathcal{E}$, Ω, the feasible action sets, the report sets, and the horizon T are finite, and utilities are bounded. Then, the finite extensive-form Bayesian game on the active joint type space $\Theta_+ = \mathrm{supp}(\lambda)$ admits a full-history sequential equilibrium.*

Proof. Restricting the initial draw to $\Theta_+ = \mathrm{supp}(\lambda)$ removes only joint types that occur with zero probability and gives a finite extensive-form game with perfect recall and a full-support prior on the active joint type space. The standard completely mixed perturbation and vanishing-tremble argument used in the proof of Theorem 3.1 therefore yields a full-history sequential equilibrium. □

The existence result is stated for full-history strategies. Obtaining a Markov representation requires additional care because the two players now have different private information. In particular, the user's posterior about ω, conditional on its privately observed type η, is not generally a public belief shared with the assistant. Likewise, the assistant may form beliefs about η using both the public history and its privately observed ω. Thus, separate player-specific posteriors cannot simply be appended to (A, q) and declared to form a public state.

A common public belief can instead be obtained from the history observed by both players. Fix a full-history sequential assessment $\mathbf{a}$. At a positive-probability public history H_t^{pub}, define the *public-history posterior*

$$\beta_t^{\mathbf{a}} := \mathbb{P}_{\lambda,\mathbf{a}}\Big((\eta, \omega) \in \cdot \;\Big|\; H_t^{\mathrm{pub}}\Big) \in \Delta(\Theta_+).$$

Because both players observe the same public history, this posterior is common to them. Each player can then combine it with its own private information when evaluating continuation play. At public histories that occur with zero probability, the corresponding beliefs are supplied by the consistent off-path beliefs of the sequential assessment.

This public-history posterior suggests an augmented state such as $\psi_t = (A_t, \beta_t^{\mathbf{a}})$, but it does not by itself imply that equilibrium play is Markov in that state. As in the baseline model, two different public histories may generate the same proposed state while prescribing different continuation behavior. We therefore require the following state-consistency condition.

Assumption D.2 (Augmented state consistency)**.** *Fix a full-history sequential assessment and an explicitly specified augmented public state ψ_t. Whenever two public histories at the same round induce the same ψ_t, the assessment prescribes the same continuation assistant strategy, the same continuation reporting strategy for every private type, and the same continuation beliefs.*

If a full-history sequential assessment satisfies Assumption D.2, continuation play depends only on the augmented state. In particular, if $\psi_t = (A_t, \beta_t^{\mathbf{a}})$ is sufficient for a given assessment, the common continuation behavior at histories inducing the same ψ_t defines a Markov representation on the finite collection of augmented states generated by that assessment's public-history tree. This is a representation of a full-history equilibrium that already satisfies state consistency; the existence result above does not assert that every equilibrium, or some equilibrium without further conditions, must be Markov in $(A_t, \beta_t^{\mathbf{a}})$.

E Omitted Proofs and Additional Results for Section 4

E.1 Formal Setup for Misspecified and Equilibrium-Aware Learning

Definition E.1 (Truthful-choice and strategic likelihoods)**.** *Fix a strategic data-generating process $(\pi^\dagger, \sigma^\dagger, \mu^\dagger)$, consisting of an assistant policy $\pi^\dagger$, a user signalling rule $\sigma^\dagger$, and an off-path belief selector $\mu^\dagger$, and fix a true type $\eta_0 \in \mathcal{E}$. Let H_s denote the public history before the round-s report and let $B_s = A_s \cup S_s$ denote the revealed menu. Starting from $(A_1, q_1^\dagger) = (\varnothing, p)$, the process evolves according to $S_s \sim \pi_s^\dagger(\cdot \mid A_s, q_s^\dagger)$, $B_s = A_s \cup S_s$, $Y_s \sim \sigma_s^\dagger(\cdot \mid A_s, q_s^\dagger, B_s, \eta_0)$, $q_{s+1}^\dagger = \tau_s^{\sigma^\dagger, \mu^\dagger}(q_s^\dagger, A_s, B_s, Y_s)$, and $A_{s+1} = B_s$. Thus, $q_s^\dagger$ is the belief state carried by the fixed data-generating process. For each candidate type η, the* strategic likelihood *is $M_s^\eta(\mathbf{y} \mid H_s) := \sigma_s^\dagger(\mathbf{y} \mid A_s, q_s^\dagger, B_s, \eta)$, so that $Y_s \sim M_s^{\eta_0}(\cdot \mid H_s)$. Let $L_s^\eta(\cdot \mid H_s)$ be a specified* truthful-choice likelihood*, representing how a type-η user would choose from the revealed menu under the specified truthful feedback model, for example, the finite-temperature multinomial-logit model of Section B.5. On the same realized history, the* truthful-likelihood posterior $\widetilde{q}_t$ *is obtained by updating with $\{L_s^\eta\}_{s<t}$, while the* strategic-likelihood posterior $q_t^\star$ *is obtained by updating with $\{M_s^\eta\}_{s<t}$. Both posterior sequences start from full-support distributions on $\mathcal{E}$.*

Throughout the learning results below, probability and almost-sure statements are taken under the law of the fixed strategic data-generating process in Definition E.1, conditional on $\eta_u = \eta_0$. If $q_1^\star = p$, then the strategic-likelihood posterior coincides with the belief state of the data-generating process along every positive-probability history, so $q_s^\star = q_s^\dagger$.

The two learning results below assume uniform full support: for some $\alpha \in (0, 1)$, $L_s^\eta(\mathbf{y} \mid H_s) \geq \alpha$ and $M_s^\eta(\mathbf{y} \mid H_s) \geq \alpha$ at every positive-probability history H_s, every $\eta \in \mathcal{E}$, and every $\mathbf{y} \in B_s$. This holds, for example, for the finite-temperature multinomial-logit specification of the truthful-choice likelihood in Section B.5. By contrast, the canonical deterministic tie-broken truthful rule of Section 2 is a point mass and generally does not satisfy full support. Both results are finite-sample statements for a single interaction of fixed horizon T: each concerns a round $t \leq T$, and no limit in t is taken.

The two learning results compare the truthful-choice and strategic likelihoods on the same realized feedback stream generated by the fixed strategic process. They do not compare interactions generated by assistants whose different interpretations of feedback lead them to choose different future slates. In an endogenous adaptive interaction, changing how the assistant interprets feedback may change subsequent slates, the user's incentives, and hence the feedback generated later in the interaction. The results therefore apply to the fixed strategic process specified above, whenever their stated conditions hold.

E.2 Proof of Theorem 4.1

The separation hypotheses in the two results are cumulative. Writing δ_s for the round-s conditional drift of the relevant log-likelihood ratio, the condition requires the first t increments to sum to at least $t\gamma$, rather than requiring every increment separately to exceed γ. The pointwise condition $\delta_s \geq \gamma$ for every $s \leq t$ implies the cumulative condition, while the cumulative formulation allows individual rounds to contribute little or no separation.

The round-T report illustrates why this distinction matters. Since $A_{T+1} = B_T$ is already determined when $\mathbf{y}_T$ is submitted, every report is a continuation-best response at $s = T$, and sequential rationality alone does not determine $\sigma_T^\dagger$. Whether the final round separates types therefore depends on how this indifference is resolved. A rule based only on continuation payoffs is type-independent at $s = T$ and contributes no separation, whereas a rule that breaks ties using the type's utility U_η may continue to separate types. The cumulative formulation accommodates both cases.

We first consider the consequences of interpreting strategic feedback through the truthful-choice likelihood. If the reports generated by the true type are cumulatively better explained, under this likelihood, by a false type than by the true type itself, the resulting posterior odds can systematically move toward that false type. The following result makes this statement precise.

Theorem E.1 (Full statement of Theorem 4.1: Strategic decoy under the truthful-choice likelihood)**.** *Under the setup of Definition E.1, fix a round $t \leq T$ and suppose there exist a false type $\zeta \neq \eta_0$ and*

a constant $\gamma > 0$ such that

$$\sum_{s=1}^{t}\Big[D_{\mathrm{KL}}\big(M_s^{\eta_0}(\cdot \mid H_s)\,\big\|\,L_s^{\eta_0}(\cdot \mid H_s)\big) - D_{\mathrm{KL}}\big(M_s^{\eta_0}(\cdot \mid H_s)\,\big\|\,L_s^{\zeta}(\cdot \mid H_s)\big)\Big] \;\geq\; t\gamma \quad \textit{almost surely.}$$

For every $\rho \in (0,1)$, let $c_t(\rho) := \log(1/\alpha)\sqrt{2t\log(1/\rho)}$. Then, $\log \frac{\widetilde{q}_{t+1}(\zeta)}{\widetilde{q}_{t+1}(\eta_0)} \geq \log \frac{\widetilde{q}_1(\zeta)}{\widetilde{q}_1(\eta_0)} + t\gamma - c_t(\rho)$ with probability at least $1-\rho$. Hence, if $t\gamma > c_t(\rho) - \log \frac{\widetilde{q}_1(\zeta)}{\widetilde{q}_1(\eta_0)}$, the truthful-likelihood posterior favors the false type ζ over the true type η_0. Thus, strategic feedback can make a misspecified assistant favor a decoy type that better explains the user's strategic signals.

Proof. All expectations and probabilities in this proof are under the data-generating law fixed in Definition E.1, conditional on $\eta_u = \eta_0$. By Bayes' rule for the misspecified reference-likelihood update,

$$\frac{\widetilde{q}_{t+1}(\zeta)}{\widetilde{q}_{t+1}(\eta_0)} = \frac{\widetilde{q}_1(\zeta)}{\widetilde{q}_1(\eta_0)} \prod_{s=1}^{t} \frac{L_s^{\zeta}(Y_s \mid H_s)}{L_s^{\eta_0}(Y_s \mid H_s)},$$

so that, writing $Z_s := \log\big[L_s^{\zeta}(Y_s \mid H_s)/L_s^{\eta_0}(Y_s \mid H_s)\big]$,

$$\log \frac{\widetilde{q}_{t+1}(\zeta)}{\widetilde{q}_{t+1}(\eta_0)} = \log \frac{\widetilde{q}_1(\zeta)}{\widetilde{q}_1(\eta_0)} + \sum_{s=1}^{t} Z_s.$$

Conditional drift. Since $Y_s \sim M_s^{\eta_0}(\cdot \mid H_s)$, write $P_s := M_s^{\eta_0}(\cdot \mid H_s)$ and $L_s^{\eta} := L_s^{\eta}(\cdot \mid H_s)$. Then

$$\delta_s := \mathbb{E}[Z_s \mid H_s] = \sum_{\mathbf{y} \in B_s} P_s(\mathbf{y}) \log \frac{L_s^{\zeta}(\mathbf{y})}{L_s^{\eta_0}(\mathbf{y})} = D_{\mathrm{KL}}(P_s \| L_s^{\eta_0}) - D_{\mathrm{KL}}(P_s \| L_s^{\zeta}),$$

each δ_s being H_s-measurable. The hypothesis states exactly that $\sum_{s=1}^{t} \delta_s \geq t\gamma$ almost surely. This is the only place the separation hypothesis is used, and it is used only through the sum: no individual δ_s needs to be positive.

Boundedness. Since B_s is H_s-measurable, conditionally on H_s the variable Z_s takes values in the finite set $\{\log[L_s^{\zeta}(\mathbf{y})/L_s^{\eta_0}(\mathbf{y})] : \mathbf{y} \in B_s\}$. Full support gives $\alpha \leq L_s^{\eta}(\mathbf{y} \mid H_s) \leq 1$ for every η and every $\mathbf{y} \in B_s$, hence $L_s^{\zeta}/L_s^{\eta_0} \in [\alpha, 1/\alpha]$ and every such value lies in $[-\log(1/\alpha), \log(1/\alpha)]$. Writing $a_s := \min_{\mathbf{y} \in B_s} \log[L_s^{\zeta}/L_s^{\eta_0}]$ and b_s for the corresponding maximum — both H_s-measurable — we have $a_s \leq Z_s \leq b_s$ with $b_s - a_s \leq 2\log(1/\alpha)$. Since $\delta_s = \mathbb{E}[Z_s \mid H_s] \in [a_s, b_s]$, the martingale difference $D_s := Z_s - \delta_s$ lies in $[a_s - \delta_s,\, b_s - \delta_s]$, an interval of the *same* width: centring shifts a range without widening it.

Concentration. The sequence $\{D_s\}_{s \leq t}$ is a martingale difference sequence with respect to the filtration generated by the public histories. Applying Hoeffding's lemma to the regular conditional law of D_s given H_s — which has mean zero and is supported on an interval of width at most $2\log(1/\alpha)$ — gives, for every $u \in \mathbb{R}$,

$$\mathbb{E}\big[e^{uD_s} \mid H_s\big] \leq \exp\Big(\tfrac{1}{8}u^2(b_s - a_s)^2\Big) \leq \exp\Big(\tfrac{1}{2}u^2 \log^2(1/\alpha)\Big),$$

a *deterministic* bound, so iterating the tower property over $s = t, \ldots, 1$ yields $\mathbb{E}[\exp(-u\sum_{s \leq t} D_s)] \leq \exp(\tfrac{1}{2}tu^2\log^2(1/\alpha))$ for $u > 0$. Markov's inequality then gives

$$\mathbb{P}\Big[\sum_{s=1}^{t} D_s \leq -\lambda\Big] \leq \exp\Big(-u\lambda + \tfrac{1}{2}tu^2\log^2(1/\alpha)\Big) = \exp\Big(-\frac{\lambda^2}{2t\log^2(1/\alpha)}\Big)$$

at the optimal $u = \lambda/(t\log^2(1/\alpha))$. Setting the right-hand side equal to ρ and solving for λ gives $\lambda = \log(1/\alpha)\sqrt{2t\log(1/\rho)} = c_t(\rho)$. Hence, with probability at least $1-\rho$, $\sum_{s=1}^{t} D_s \geq -c_t(\rho)$.

Combination. On that event, $\sum_{s=1}^{t} Z_s = \sum_{s=1}^{t} \delta_s + \sum_{s=1}^{t} D_s \geq t\gamma - c_t(\rho)$. Substituting into the log-odds identity proves the claim. □

E.3 Proof of Theorem 4.2

The preceding failure stems from interpreting strategic feedback through the wrong likelihood, rather than from strategic feedback being intrinsically uninformative. We next consider the strategic-likelihood posterior. If the strategic reporting distributions of the true and false types accumulate sufficient separation, the posterior instead moves toward the true type. The following result gives the corresponding finite-sample guarantee.

Theorem E.2 (Full statement of Theorem 4.2: Finite-sample concentration of the strategic-likelihood posterior under separation)**.** *Assume $|\mathcal{E}| \geq 2$. Under the setup of Definition E.1, fix a round $t \leq T$ and suppose that the strategic-likelihood posterior is updated with M_s^{η} and that there is a constant $\gamma_\star > 0$ such that, almost surely,*

$$\sum_{s=1}^{t} D_{\mathrm{KL}}\big(M_s^{\eta_0}(\cdot \mid H_s) \,\big\|\, M_s^{\zeta}(\cdot \mid H_s)\big) \;\geq\; t\gamma_\star \qquad \text{for every false type } \zeta \neq \eta_0.$$

Then, for every $\rho \in (0,1)$, $1 - q_{t+1}^\star(\eta_0) \leq \frac{1-q_1^\star(\eta_0)}{q_1^\star(\eta_0)} \exp(-t\gamma_\star + c_t^\star(\rho))$ with probability at least $1-\rho$, where $c_t^\star(\rho) := \log(1/\alpha)\sqrt{2t\log((|\mathcal{E}|-1)/\rho)}$. If $|\mathcal{E}| = 1$ the posterior is identically one on the unique type, so the conclusion is trivial and no union-bound constant is needed.

Proof. All expectations and probabilities in this proof are under the data-generating law fixed in Definition E.1, conditional on $\eta_u = \eta_0$. Fix a false type $\zeta \neq \eta_0$. Bayes' rule for the strategic-likelihood update gives

$$\log \frac{q_{t+1}^\star(\eta_0)}{q_{t+1}^\star(\zeta)} = \log \frac{q_1^\star(\eta_0)}{q_1^\star(\zeta)} + \sum_{s=1}^{t} W_s^\zeta, \qquad W_s^\zeta := \log \frac{M_s^{\eta_0}(Y_s \mid H_s)}{M_s^\zeta(Y_s \mid H_s)}.$$

Conditional drift. Letting $P_s := M_s^{\eta_0}(\cdot \mid H_s)$ and $R_s^\zeta := M_s^\zeta(\cdot \mid H_s)$, and since $Y_s \sim P_s$, we have

$$\mathbb{E}[W_s^\zeta \mid H_s] = \sum_{\mathbf{y} \in B_s} P_s(\mathbf{y}) \log \frac{P_s(\mathbf{y})}{R_s^\zeta(\mathbf{y})} = D_{\mathrm{KL}}(P_s \| R_s^\zeta) \;\geq\; 0,$$

so every conditional increment is non-negative, and the hypothesis states that the first t of them sum to at least $t\gamma_\star$ almost surely. Rounds in which the two strategic likelihoods coincide contribute zero and are therefore harmless.

Boundedness. Conditionally on H_s, the variable W_s^ζ takes values in the finite set $\{\log[M_s^{\eta_0}(\mathbf{y})/M_s^\zeta(\mathbf{y})] : \mathbf{y} \in B_s\}$, and full support gives $M_s^{\eta_0}/M_s^\zeta \in [\alpha, 1/\alpha]$, so those values lie in $[-\log(1/\alpha), \log(1/\alpha)]$, an interval of width $2\log(1/\alpha)$. Exactly as in the proof of Theorem E.1, centring leaves the width unchanged, so the martingale difference $D_s^\zeta := W_s^\zeta - \mathbb{E}[W_s^\zeta \mid H_s]$ is supported, conditionally on H_s, on an interval of width at most $2\log(1/\alpha)$.

Concentration. Applying the conditional Hoeffding bound exactly as in the proof of Theorem E.1, but at failure level $\rho/(|\mathcal{E}|-1)$,

$$\sum_{s=1}^{t} D_s^\zeta \;\geq\; -\log(1/\alpha)\sqrt{2t\log\big((|\mathcal{E}|-1)/\rho\big)} = -c_t^\star(\rho)$$

with probability at least $1 - \rho/(|\mathcal{E}|-1)$, whence $\sum_{s=1}^t W_s^\zeta \geq t\gamma_\star - c_t^\star(\rho)$ on that event. A union bound over the $|\mathcal{E}|-1$ false types makes the inequality hold simultaneously for all of them with probability at least $1-\rho$.

Aggregation. On that event, for every $\zeta \neq \eta_0$,

$$\frac{q_{t+1}^\star(\zeta)}{q_{t+1}^\star(\eta_0)} \leq \frac{q_1^\star(\zeta)}{q_1^\star(\eta_0)} \exp\big(-t\gamma_\star + c_t^\star(\rho)\big),$$

and summing over the false types,

$$\frac{1 - q_{t+1}^\star(\eta_0)}{q_{t+1}^\star(\eta_0)} = \sum_{\zeta \neq \eta_0} \frac{q_{t+1}^\star(\zeta)}{q_{t+1}^\star(\eta_0)} \leq \frac{1 - q_1^\star(\eta_0)}{q_1^\star(\eta_0)} \exp\big(-t\gamma_\star + c_t^\star(\rho)\big).$$

Since $1 - q_{t+1}^\star(\eta_0) \leq \big[1 - q_{t+1}^\star(\eta_0)\big]/q_{t+1}^\star(\eta_0)$, the stated bound follows. □

The cumulative separation condition allows only a subset of rounds to be informative. Each term $D_{\mathrm{KL}}(M_s^{\eta_0} \| M_s^{\zeta})$ is non-negative, so rounds that do not distinguish η_0 from ζ do not offset the separation accumulated in other rounds. In particular, suppose that $D_{\mathrm{KL}}(M_s^{\eta_0} \| M_s^{\zeta}) \geq \gamma'$ for every $s \in I \subseteq \{1, \ldots, t\}$, while the remaining rounds provide no guaranteed separation. Then the hypothesis of Theorem E.2 holds with $\gamma_\star = \gamma'|I|/t$, and the exponential term in the bound becomes $\exp(-\gamma'|I| + c_t^\star(\rho))$. Thus, the amount of accumulated separation depends on the informative rounds rather than on every round being informative.

Uninformative rounds are nevertheless not costless in the finite-sample bound. The deviation term $c_t^\star(\rho)$ grows as $\sqrt{t}$ regardless of how many rounds separate the types, so the accumulated divergence must be large enough to dominate this term before the bound becomes informative. More generally, the bound in Theorem E.1 becomes informative once its drift dominates its deviation term, which requires $t > 2\log^2(1/\alpha)\log(1/\rho)/\gamma^2$. The corresponding condition for Theorem E.2 is $t > 2\log^2(1/\alpha)\log((|\mathcal{E}|-1)/\rho)/\gamma_\star^2$.

These are finite-sample conditions at a fixed round $t \leq T$, not asymptotic consistency statements. In particular, $\gamma_\star$ need not remain bounded away from zero as the horizon changes, so Theorem E.2 should not be interpreted as an asymptotic concentration result without additional assumptions.

When strategic incentives uniquely sustain truthful reporting, the two likelihoods coincide. DIC alone does not ensure this selection: it makes truthful feedback a continuation-best response where the condition applies, but other equilibrium reports may remain optimal (Definition 3.1).

F Omitted Proofs and Additional Results for Section 5

F.1 Terminal Choice and the Assistant's Security Value

We establish here the non-adaptive benchmark underlying the equilibrium-collapse results and show that it is a security value for the assistant. We first isolate the role of the terminal menu and characterize the menu lotteries that can be generated without adapting to feedback. We then embed such non-adaptive controls into the strategic model by defining the domain on which complete assistant-policy deviations are evaluated. This allows any feasible non-adaptive control to be used as a deviation against a strategic assessment, yielding the security-value bound.

Proposition F.1 (Terminal selection is non-strategic)**.** *For every signaling rule σ and every realized terminal menu A_{T+1}, the strategic user's terminal choice is $\mathbf{x}^\star(A_{T+1}, \eta_u)$, and the realized user regret $r_{\eta_u}(\mathbf{x}^\star(A_{T+1}, \eta_u))$ depends on σ only through A_{T+1}. In particular, for a non-adaptive control the expected regret is determined by its terminal-menu law and the prior alone.*

Proof. At the terminal round T there is no continuation, so in the user best-response recursion of Section 3.1 the terminal condition is $U_{T+1}^{\pi,\tau}(A, q, \eta) = \max_{\mathbf{x}\in A} U_\eta(\mathbf{x})$, attained by $\mathbf{x}^\star(A, \eta)$. Thus the user's terminal item is its tie-broken U_η-maximizer in A_{T+1}, irrespective of any earlier signals; this is the content of the equation defining $\mathbf{x}_T^\star$ in Section 2. Hence the realized regret is a function only of (A_{T+1}, η_u). Now fix a non-adaptive control κ. Under its lift, the accumulated-menu process obeys the autonomous recursion $A_1 = \varnothing$, $S_t \sim \kappa_t(\cdot \mid A_t)$, $A_{t+1} = A_t \cup S_t$, which contains neither the posterior nor the user's report; by Lemma F.1 its terminal-menu *marginal law* is the same distribution D_κ under every signalling rule to whose assessment domain the control is separately lifted, and $\mathbb{P}(A_{T+1} = A \mid \eta_u = \eta) = D_\kappa(A)$ for every supported type η. The resulting expected regret therefore depends only on κ and the prior. This is an invariance-of-marginal-law statement; it does not assert statistical independence between A_{T+1} and the realized report sequence. □

Proposition F.1 shows that, conditional on the terminal menu, the user's final choice depends only on its type and not on its earlier reports. This makes it useful to isolate what the assistant can achieve by constructing the terminal menu without conditioning its future slates on feedback. We formalize such behavior as a non-adaptive control.

Definition F.1 (Non-adaptive control)**.** *A* non-adaptive control *is a collection $\kappa = \{\kappa_t(\cdot \mid A)\}_{t=1}^T$ such that, for every round t and every accumulated menu A feasible at that round, $\kappa_t(\cdot \mid A)$ is a distribution over feasible slates at A. The control may randomize and may condition on the assistant's own accumulated menu, but it has no posterior or feedback argument. Write $\mathcal{K}_{\mathrm{nf}}$ for the set of non-adaptive controls.*

A non-adaptive control is defined independently of the strategic assessment, and the non-adaptive benchmark obtained by optimizing over such controls therefore does not depend on a strategic signalling rule or an assessment domain. To use a non-adaptive control as a complete assistant-policy deviation in Definition 5.1, however, it must be specified at every public state that can arise under the deviation while the user's signalling rule σ and off-path belief selector μ remain fixed. This requires a domain larger than the states reached under the incumbent policy. We therefore define the following deviation-closed domain.

Definition F.2 (Deviation-closed Markov domain)**.** *Fix a Markov signalling rule σ and an off-path belief selector μ. Define finite sets $\mathcal{D}_t(\sigma,\mu) \subseteq 2^X \times \Delta(\mathcal{E})$ and information-set collections $\mathcal{I}_t(\sigma,\mu)$ by the forward recursion*

$$\mathcal{D}_1 := \{(\varnothing, p)\}, \qquad \mathcal{I}_t := \big\{(A,q,B) : (A,q) \in \mathcal{D}_t,\ B = A \cup S \text{ for a feasible slate } S \text{ at } A\big\},$$

$$\mathcal{D}_{t+1} := \big\{\big(B,\ \tau_t^{\sigma,\mu}(q,A,B,\mathbf{y})\big) : (A,q,B) \in \mathcal{I}_t,\ \mathbf{y} \in B\big\}, \qquad t = 1,\dots,T,$$

where $\sigma_t(\cdot \mid A,q,B,\eta)$ and $\mu_t(\cdot \mid A,q,B,\mathbf{y})$ are specified at every $(A,q,B) \in \mathcal{I}_t$ and every $\mathbf{y} \in B$, and $\tau_t^{\sigma,\mu}$ is the update of Section 3.1: Bayes' rule when $P_q^\sigma(\mathbf{y} \mid A,B) > 0$ and μ_t otherwise. We call $\mathcal{D}(\sigma,\mu) = \{\mathcal{D}_t(\sigma,\mu)\}_{t=1}^{T+1}$ the deviation-closed domain generated by (σ,μ).

The domain is finite because the horizon, item set, and feasible action and report sets are finite. Its construction includes every feasible assistant slate and every report in the resulting menu, including slates not chosen by a given assistant policy and reports that may occur with zero probability. Hence, for fixed signalling and belief rules (σ,μ), $\mathcal{D}(\sigma,\mu)$ contains every public state that can arise under any feasible behavioral Markov assistant policy. The domain therefore provides a common state space on which complete assistant-policy deviations can be compared while (σ,μ) remain fixed. Replacing σ or μ, by contrast, generally generates a different deviation-closed domain.

The deviation-closed domain serves a different purpose from the finite-state Markov representation in Theorem 3.1. The latter represents a specified state-consistent full-history equilibrium using only the states generated by its public-history tree. By contrast, $\mathcal{D}(\sigma,\mu)$ includes the states reached after every feasible slate and report under fixed (σ,μ) so that complete assistant-policy deviations can be evaluated on a common domain. It is therefore a bookkeeping device for the deviation problem, not an additional equilibrium-existence result. We reserve $\mathcal{Z}_t$ for the equilibrium representation of Section 3 and $\mathcal{D}_t$ for the deviation-closed domain used in Section 5.

A non-adaptive control is defined only in terms of the accumulated menu, while a strategic assessment is specified on belief states (A,q). To compare the two, we lift a control to the strategic domain by prescribing the same slate distribution at every state with the same accumulated menu.

Definition F.3 (Lift of a non-adaptive control)**.** *Fix an assessment with deviation domain $\mathcal{D}(\sigma,\mu)$. The* lift *of a non-adaptive control κ to this domain is the behavioral Markov assistant policy π^κ given by*

$$\pi_t^\kappa(\cdot \mid A,q) := \kappa_t(\cdot \mid A) \qquad \text{for every } (A,q) \in \mathcal{D}_t(\sigma,\mu).$$

A policy obtained in this way is called non-adaptive on the incumbent domain.

The lift is defined on the entire deviation-closed domain $\mathcal{D}(\sigma,\mu)$, including states that are off path under the incumbent assessment. Thus, $\pi_t^\kappa(\cdot \mid A,q) = \kappa_t(\cdot \mid A)$ at every state in the domain, as required when evaluating the complete assistant-policy deviations in Definition 5.1. If a different pair (σ',μ') is considered, the same control κ is lifted separately to its corresponding domain $\mathcal{D}(\sigma',\mu')$.

The following result records the properties of this lift that will be used to establish the assistant's security value.

Lemma F.1 (Consequences of a non-adaptive control)**.** *Fix a Markov signalling rule σ, an off-path selector μ, and the deviation domain $\mathcal{D}(\sigma,\mu)$. Let κ be a non-adaptive control and π^κ its lift to this domain. Then*

(i) *the marginal law of the accumulated-menu process under (π^κ,σ) is generated by*

$$A_1 = \varnothing, \qquad S_t \sim \kappa_t(\cdot \mid A_t), \qquad A_{t+1} = A_t \cup S_t;$$

(ii) *this recursion contains neither the posterior nor the user's type, so writing D_κ for the induced terminal-menu law, $\mathbb{P}^{(\pi^\kappa,\sigma)}(A_{T+1} = A \mid \eta_u = \eta) = D_\kappa(A)$ for every $\eta \in \operatorname{supp}(p)$, equivalently $\mathbb{P}^{(\pi^\kappa,\sigma)}(A_{T+1} = A,\ \eta_u = \eta) = D_\kappa(A)\,p(\eta)$.*

Proof. Conditional on any public history in $\mathcal{D}(\sigma, \mu)$ with accumulated menu $A_t = A$, Definition F.3 gives $\mathbb{P}(S_t = S \mid H_t) = \kappa_t(S \mid A)$. Hence, the marginal transition kernel of the accumulated-menu process is $\mathbb{P}(A_{t+1} = A \cup S \mid A_t = A) = \kappa_t(S \mid A)$, and induction from $A_1 = \varnothing$ proves ***(i)***. The same recursion holds after conditioning on $\eta_u = \eta$, because neither the initial menu nor any kernel $\kappa_t(\cdot \mid A)$ depends on η; thus the conditional law of A_{T+1} is D_κ for every type, proving ***(ii)***. □

The same conclusion holds when κ is lifted separately to a different assessment domain $\mathcal{D}(\sigma', \mu')$: the accumulated-menu recursion, and hence the terminal-menu law D_κ and the resulting payoff-regret pair, depend only on κ and the prior, not on the signalling rule. This does not require the two assessment domains to coincide. Nor does non-adaptivity imply statistical independence between the menu and the user's reports: reports generally respond to the revealed menu. Rather, non-adaptivity means that the assistant's slate distribution depends on the accumulated menu but not on those reports.

In particular, whenever the non-adaptive problem is feasible, lifting a welfare-feasible non-adaptive control gives a feasible complete assistant deviation in Definition 5.1**(ii)**. Hence the assistant's constrained deviation set is nonempty. Moreover, because the behavioral-policy space is compact and the objective and welfare constraint are continuous, the assistant's constrained best-response problem attains its maximum. The terminal-menu law of the lifted control is unaffected by the signalling rule and, by Proposition F.1, its payoff is determined by that law and the prior. Constrained assistant optimality therefore implies that every welfare-feasible strategic assessment gives the assistant at least the non-adaptive benchmark.

Proposition F.2 (Assistant security value)**.** *Fix $\epsilon \geq 0$ for which the non-adaptive problem is feasible. In every welfare-feasible strategic assessment (π, σ, μ), $\mathbb{E}^{(\pi,\sigma)}[\tilde{U}_a(\mathbf{x}^\star_T)] \geq \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$, and therefore $\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon) \geq \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$.*

Proof. Let $\kappa^\star \in \mathcal{K}_{\mathrm{nf}}$ attain $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ in Definition 5.2, and let $\pi^{\kappa^\star}$ be its lift to the incumbent deviation domain:

$$\pi^{\kappa^\star}_t(\cdot \mid A, q) = \kappa^\star_t(\cdot \mid A) \qquad \text{for every } (A, q) \in \mathcal{D}_t(\sigma, \mu).$$

This is a complete behavioral Markov policy on the incumbent domain. By Lemma F.1,

$$\mathbb{P}^{(\pi^{\kappa^\star}, \sigma)}\big(A_{T+1} = A,\ \eta_u = \eta\big) = D_{\kappa^\star}(A)\, p(\eta),$$

and by Proposition F.1 the terminal item is $\mathbf{x}^\star(A_{T+1}, \eta_u)$ irrespective of the intermediate reports. Hence

$$\mathbb{E}^{(\pi^{\kappa^\star}, \sigma)}\big[\tilde{U}_a(\mathbf{x}^\star_T)\big] = \sum_{A,\eta} D_{\kappa^\star}(A) p(\eta) \tilde{U}_a(\mathbf{x}^\star(A, \eta)) = \mathrm{OPT}_{\mathrm{nf}}(\epsilon),$$

and likewise $\mathbb{E}^{(\pi^{\kappa^\star}, \sigma)}\big[G_u(A_{T+1}, \eta_u)\big] \leq \epsilon$. Thus $\pi^{\kappa^\star}$ is a complete admissible assistant deviation against the incumbent σ in Definition 5.1**(ii)**, and constrained assistant optimality gives $\mathbb{E}^{(\pi,\sigma)}[\tilde{U}_a(\mathbf{x}^\star_T)] \geq \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. □

The security bound above is stated in terms of the non-adaptive benchmark of Definition 5.2. This benchmark can also be characterized directly in terms of terminal-menu lotteries, provided that every lottery supported on horizon-reachable menus can be generated by a non-adaptive control. The following lemma establishes this implementation result and hence characterizes exactly the attainable terminal-menu laws.

Lemma F.2 (Implementation of non-adaptive menu lotteries)**.** *Let D be a distribution with finite support such that every $A \in \mathrm{supp}(D)$ is horizon-reachable. Then there exists a non-adaptive control κ whose induced terminal-menu law D_κ equals D. Consequently, the attainable terminal-menu laws $\{D_\kappa : \kappa \in \mathcal{K}_{\mathrm{nf}}\}$ are exactly the distributions supported on horizon-reachable menus.*

Proof. **Paths.** For each $A^\star \in \mathrm{supp}(D)$ fix, by Definition B.1, one feasible deterministic slate sequence $S^{A^\star}_1, \ldots, S^{A^\star}_T$ of length T whose accumulated terminal menu is $A^\star$, and write $A^{A^\star}_t := \bigcup_{s<t} S^{A^\star}_s$.

Flows. In the finite acyclic graph with nodes (t, A), $1 \le t \le T+1$, and edges $(t, A) \to (t+1, A \cup S)$ for feasible slates S at A, send flow $D(A^\star)$ along the path of $A^\star$ and aggregate:

$$f_t(A, S) := \sum_{A^\star:\, A_t^{A^\star} = A,\; S_t^{A^\star} = S} D(A^\star), \quad m_t(A) := \sum_S f_t(A, S), \quad m_{T+1}(A) := \sum_{A^\star:\, A^\star = A} D(A^\star).$$

Each path contributes its mass to exactly one edge out of each node it visits, so $m_1(\varnothing) = 1$ and $m_{t+1}(A') = \sum_{(A,S):A\cup S = A'} f_t(A, S)$.

The control. Set $\kappa_t(S \mid A) := f_t(A, S)/m_t(A)$ when $m_t(A) > 0$ and any fixed feasible distribution otherwise.

Correctness. The probability of reaching (t, A) under the recursion of Lemma F.1**(i)** is $m_t(A)$: true at $t = 1$, and if true at t then for any A' it is $\sum_{(A,S):A\cup S = A'} m_t(A) \frac{f_t(A,S)}{m_t(A)} = m_{t+1}(A')$, nodes with $m_t(A) = 0$ contributing nothing. At the terminal layer $m_{T+1}(A) = \sum_{A^\star} D(A^\star)\mathbf{1}\{A^\star = A\} = D(A)$, so $D_\kappa = D$. The argument reproduces the *aggregated flow*; a realized trajectory need not coincide with one of the initially selected paths, since where paths share a node the normalized edge-flow control may combine one path's prefix with another's suffix. Conversely, every D_κ is supported on horizon-reachable menus, since every realization of the recursion follows a feasible deterministic slate sequence.

A non-adaptive control. The normalized edge flows $\kappa_t(S \mid A)$ have no posterior argument, so they define a non-adaptive control in the sense of Definition F.1. For any strategic assessment (π, σ, μ), lifting this same control to $\mathcal{D}(\sigma, \mu)$ via Definition F.3 yields a complete behavioral Markov assistant policy whose terminal-menu law is again D, by Lemma F.1. □

Consequently, $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ is equivalently the best welfare-feasible lottery over horizon-reachable terminal menus. We write $\mathrm{OPT}_{\mathrm{nf}}^{\mathrm{det}}(\epsilon) \le \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ for the restriction to a single committed menu; the inequality can be strict. These benchmarks depend only on the model primitives and feasible-slate correspondence, not on the user's signalling rule. We use them only when a welfare-feasible non-adaptive control exists.

F.2 Equilibrium Computation by Regularized Self-Play and Verification

This appendix describes the entropy-regularized self-play procedure referenced in Section 4. When the strategic signalling rule arises from equilibrium behavior, computing an equilibrium provides the assistant policy and user signalling rule needed to construct the corresponding strategic likelihood and simulate equilibrium play. As we show later in Section 5.3 and Appendix F.8, equilibrium computation is not always needed to determine the assistant's equilibrium value: under the conditions studied there, a broad class of equilibria all yield the same non-adaptive value even when their signalling behavior differs. Computing the equilibrium strategies remains useful when the strategies themselves are of interest and outside the classes for which the equilibrium value is characterized analytically.

For computation and simulation, we use entropy-regularized self-play on the belief-state game. Starting from an assistant policy and a user signalling rule, the procedure repeatedly evaluates continuation payoffs and updates both policies toward soft best responses. Entropy regularization smooths the best-response maps and can make the iteration more stable.

Fix a temperature parameter $\lambda > 0$, and let (π^n, σ^n) denote the policies at iteration n. At each round t and state (A, q), let $Q_t^{a,n}(A, q, S)$ denote the assistant's continuation value from choosing a feasible slate $S \in \mathcal{S}_t(A)$, and let $Q_t^{u,n}(A, q, B, \eta, y)$ denote the continuation value of a type-η user from reporting $y \in B$. The next policies are the corresponding soft best responses:

$$\pi_t^{n+1}(S \mid A, q) = \frac{\exp\left(Q_t^{a,n}(A, q, S)/\lambda\right)}{\sum_{S' \in \mathcal{S}_t(A)} \exp\left(Q_t^{a,n}(A, q, S')/\lambda\right)}, \qquad S \in \mathcal{S}_t(A),$$

and

$$\sigma_t^{n+1}(y \mid A, q, B, \eta) := \frac{\exp\left(Q_t^{u,n}(A, q, B, \eta, y)/\lambda\right)}{Z_t^{u,n}(A, q, B, \eta)},$$
$$Z_t^{u,n}(A, q, B, \eta) := \sum_{z \in B} \exp\left(Q_t^{u,n}(A, q, B, \eta, z)/\lambda\right).$$

After each update, the continuation values are recomputed by backward induction using the posterior update $\tau^{\sigma^{n+1},\mu}$; the selector μ determines beliefs only after reports having zero probability under the updated signalling rule.

This procedure is a computational method rather than a general convergence result. For a fixed temperature $\lambda > 0$ and finite state domain, if the resulting entropy-regularized best-response operator is a contraction under some specified complete metric, then Banach's fixed-point theorem implies convergence to a unique regularized fixed point. We do not assume that this contraction property holds generally: whether it holds depends on the particular instance, temperature, state domain, and metric. Accordingly, the self-play iteration by itself should not be interpreted as verifying an equilibrium.

When an equilibrium candidate is obtained numerically, its optimality conditions can instead be checked directly. For this purpose, it is useful to distinguish the user's continuation value under a prescribed signalling rule from the best-response value used in Section 3.1. Once an assessment is fixed, the former is determined by a direct backward recursion.

Definition F.4 (User continuation values under a prescribed signalling rule)**.** *Fix a behavioral Markov assistant policy π, a Markov signalling rule σ, and a belief selector μ, all specified on the deviation-closed Markov domain $\mathcal{D}(\sigma, \mu)$ of Definition F.2. For a type η, define by backward recursion*

$$U_{T+1}^{\pi,\sigma}(A, q, \eta) := \max_{\mathbf{x} \in A} U_\eta(\mathbf{x}), \qquad (A, q) \in \mathcal{D}_{T+1}(\sigma, \mu),$$

and, for $t = T, \ldots, 1$,

$$W_t^{\pi,\sigma}(A, q, B, \eta) := \sum_{\mathbf{y} \in B} \sigma_t(\mathbf{y} \mid A, q, B, \eta)\, U_{t+1}^{\pi,\sigma}\big(B, \tau_t^{\sigma,\mu}(q, A, B, \mathbf{y}), \eta\big)$$

for $(A, q, B) \in \mathcal{I}_t(\sigma, \mu)$, and

$$U_t^{\pi,\sigma}(A, q, \eta) := \sum_S \pi_t(S \mid A, q)\, W_t^{\pi,\sigma}(A, q, A \cup S, \eta), \qquad (A, q) \in \mathcal{D}_t(\sigma, \mu).$$

Here, $W_t^{\pi,\sigma}$ is the type-η user's continuation value at the round-t reporting decision, after the menu B has been revealed, whereas $U_t^{\pi,\sigma}$ is its continuation value before the assistant chooses the round-t slate. Replacing the expectation over reports in $W_t^{\pi,\sigma}$ by a maximization over $\mathbf{y} \in B$ gives the best-response recursion of Section 3.1, with $\bar{\tau}_t = \tau_t^{\sigma,\mu}$.

This recursion is well defined for every triple (π, σ, μ) on $\mathcal{D}(\sigma, \mu)$, whether or not σ is optimal: the horizon and the state and information-set domains are finite. It therefore defines the continuation values against which user optimality can be stated without presupposing that the prescribed signalling rule is itself optimal.

Definition F.5 (Full statement of Definition 5.1: welfare-feasible strategic assessment)**.** *Fix $\epsilon \geq 0$. A* welfare-feasible strategic assessment *is a triple (π, σ, μ): a behavioral Markov user signalling rule $\sigma_t(\cdot \mid A, q, B, \eta) \in \Delta(B)$, possibly type-dependent and mixed, and an off-path belief selector μ, together generating the deviation-closed domain $\mathcal{D}(\sigma, \mu)$ of Definition F.2, both specified at every $(A, q, B) \in \mathcal{I}_t(\sigma, \mu)$ and every $\mathbf{y} \in B$; and a behavioral Markov assistant policy $\pi_t(\cdot \mid A, q)$ specified at every $(A, q) \in \mathcal{D}_t(\sigma, \mu)$; such that:*

(i) (ex-ante welfare feasibility) $\mathbb{E}^{(\pi,\sigma)}[G_u(A_{T+1}, \eta_u)] \leq \epsilon$;

(ii) (ex-ante constrained assistant optimality) *given (σ, μ), the policy π maximizes $\mathbb{E}^{(\pi',\sigma)}[\tilde{U}_a(\mathbf{x}_T^\star)]$ over* all *behavioral Markov policies π' on $\mathcal{D}(\sigma, \mu)$ satisfying $\mathbb{E}^{(\pi',\sigma)}[G_u(A_{T+1}, \eta_u)] \leq \epsilon$;*

(iii) (sequential Markov user optimality) *given (π, μ), for every round t, every type $\eta \in \mathcal{E}$ and every $(A, q, B) \in \mathcal{I}_t(\sigma, \mu)$,*

$$\operatorname{supp} \sigma_t(\cdot \mid A, q, B, \eta) \;\subseteq\; \arg\max_{\mathbf{y} \in B} U_{t+1}^{\pi,\sigma}\big(B,\ \tau_t^{\sigma,\mu}(q, A, B, \mathbf{y}),\ \eta\big),$$

with $U^{\pi,\sigma}$ as in Definition F.4: the prescribed mixed report is supported on reports maximizing the type-η user's expected continuation utility conditional on the current Markov information state, when future reporting follows σ;

(iv) (belief consistency) *beliefs follow Bayes' rule after every positive-probability report and are given by μ after every off-path report; equivalently, after observing $\mathbf{y} \in B$, the posterior is $\tau_t^{\sigma,\mu}(q, A, B, \mathbf{y})$.*

Let $\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon)$ and $\mathrm{OPT}^{\max}_{\mathrm{strat}}(\epsilon)$ denote the infimum and supremum of $\mathbb{E}^{(\pi,\sigma)}[\tilde{U}_a(\mathbf{x}_T^\star)]$ over such assessments.

The two optimality requirements have different scopes. Assistant optimality in Definition F.5**(ii)** is an ex-ante, policy-level constrained best response: because welfare feasibility concerns the entire induced interaction, a deviation replaces the complete assistant policy π by another behavioral Markov policy π' while holding (σ, μ) fixed. The deviation-closed domain $\mathcal{D}(\sigma, \mu)$ ensures that the prescribed and deviating policies are evaluated on the same state space, including states not reached under π.

User optimality in Definition F.5**(iii)** is instead sequential. At every reporting information set in the deviation-closed domain, including those reached only after zero-probability reports, the prescribed signalling rule must assign positive probability only to reports that maximize the continuation values of Definition F.4.

Without a verified contraction condition, entropy-regularized self-play provides candidate assessments rather than an equilibrium certificate. A candidate must therefore be checked directly against Definition F.5. The two optimality conditions admit different verification procedures.

User optimality can be checked exactly by backward induction. Given the continuation values of Definition F.4, replacing the prescribed report distribution by a maximization over reports gives the Markov best-response problem of Section 3.1. The difference between this best-response value and the value under σ is therefore the user's exploitability within the behavioral Markov strategy class of Definition F.5; Lemma F.5 establishes the corresponding optimality against complete behavioral Markov reporting deviations.

Assistant optimality requires solving the corresponding ex-ante constrained policy problem. Holding (σ, μ) fixed, the assistant's exploitability is

$$\sup_{\pi'}\Big\{\mathbb{E}^{(\pi',\sigma)}\big[\tilde{U}_a(\mathbf{x}_T^\star)\big] \;:\; \mathbb{E}^{(\pi',\sigma)}\big[G_u(A_{T+1}, \eta_u)\big] \le \epsilon\Big\} - \mathbb{E}^{(\pi,\sigma)}\big[\tilde{U}_a(\mathbf{x}_T^\star)\big],$$

where the supremum ranges over all behavioral Markov assistant policies on $\mathcal{D}(\sigma, \mu)$. To evaluate this problem as a finite Markov control problem, we first verify that the posterior carried by the deviation-domain state remains the correct conditional distribution of the user type along every positive-probability history generated by a deviation π'.

Lemma F.3 (Belief correctness under assistant-policy deviations)**.** *Fix (σ, μ) and any behavioral Markov assistant policy π' on $\mathcal{D}(\sigma, \mu)$. At every public history H_t with positive probability under (π', σ), the posterior q_t carried by the public state equals the conditional distribution of the user type:*

$$q_t(\eta) = \Pr^{(\pi',\sigma)}\big(\eta_u = \eta \mid H_t\big), \qquad \eta \in \mathcal{E}.$$

Consequently,

$$\mathbb{E}^{(\pi',\sigma)}[\tilde{U}_a(\mathbf{x}_T^\star)] = \mathbb{E}^{(\pi',\sigma)}\left[\sum_{\eta\in\mathcal{E}} q_{T+1}(\eta)\tilde{U}_a\big(\mathbf{x}^\star(A_{T+1}, \eta)\big)\right],$$

with the analogous identity for G_u.

Proof. The claim follows by induction on t. At $t = 1$, $q_1 = p$, so the result holds. Suppose it holds at a positive-probability public history H_t with state (A, q). Conditional on H_t, the assistant draws its slate according to $\pi'_t(\cdot \mid A, q)$, which does not depend on η_u. Hence observing the realized slate does not change the conditional distribution of η_u. After the resulting menu B is revealed, a report $\mathbf{y}$ with $P_q^\sigma(\mathbf{y} \mid A, B) > 0$ updates this distribution by Bayes' rule to $\tau_t^\sigma(q, A, B, \mathbf{y})$, which is exactly

the posterior carried by the next public state. Reports with $P_q^\sigma(\mathbf{y} \mid A, B) = 0$ occur with zero probability under (π', σ) and therefore do not affect on-path expectations; μ only specifies beliefs at those off-path information sets. □

This belief correctness reduces the assistant's deviation problem to a finite Markov control problem on $\mathcal{D}(\sigma, \mu)$. The next lemma shows that allowing behavioral randomization convexifies, but does not otherwise enlarge, the set of attainable ex-ante payoff-regret pairs.

Lemma F.4 (Convexification of behavioral Markov values)**.** *Fix a Markov signalling rule σ, an off-path selector μ, and the deviation domain $\mathcal{D}(\sigma, \mu)$ they generate. Let $\mathcal{V}_{\mathrm{beh}}$ and $\mathcal{V}_{\mathrm{det}}$ be the sets of ex-ante pairs $(\mathbb{E}^{(\pi,\sigma)}[\tilde{U}_a(\mathbf{x}_T^\star)], \mathbb{E}^{(\pi,\sigma)}[G_u(A_{T+1}, \eta_u)])$ achievable by behavioral Markov and deterministic Markov assistant policies on $\mathcal{D}(\sigma, \mu)$, respectively. Then, $\mathcal{V}_{\mathrm{beh}} = \mathrm{conv}(\mathcal{V}_{\mathrm{det}})$.*

Proof. By Lemma F.3, the public state is a finite Markov chain on $\mathcal{D}(\sigma, \mu)$ along positive-probability histories, with transition kernel determined by (π, σ), and both coordinates are ex-ante expectations of terminal functions of that state. They are therefore linear in the state–action occupancy measure.

Let $\mathcal{O}$ be the polytope of occupancy measures defined by the flow constraints

$$\omega_1(z) = \mathbf{1}\{z = z_1\}, \qquad \sum_S \omega_t(z, S) = \omega_t(z),$$

$$\omega_{t+1}(z') = \sum_{z,S} \omega_t(z, S) P_t(z' \mid z, S), \qquad \omega \geq 0.$$

Every behavioral Markov policy induces a point of $\mathcal{O}$, and every $\omega \in \mathcal{O}$ is induced by the behavioral Markov policy

$$\pi_t(S \mid z) = \frac{\omega_t(z, S)}{\omega_t(z)}$$

whenever $\omega_t(z) > 0$. Each extreme point of this finite-horizon occupancy polytope is induced by a deterministic Markov policy. Writing L for the linear map from occupancies to payoff-regret pairs,

$$\mathcal{V}_{\mathrm{beh}} = L(\mathcal{O}) = \mathrm{conv}(\mathcal{V}_{\mathrm{det}}).$$

The same argument applies to the non-adaptive subclass. □

Holding σ fixed, Lemma F.4 implies that the assistant's constrained best-response value is the upper concave envelope, at ϵ, of the payoff-regret pairs achievable by deterministic Markov policies on $\mathcal{D}(\sigma, \mu)$. It can therefore be computed by a two-objective dynamic program. Local one-shot assistant deviations do not by themselves verify Definition F.5**(ii)**.

The self-play procedure should therefore be viewed as a method for generating candidate assessments rather than as an equilibrium solver. At positive temperature, its fixed points are entropy-regularized and need not be exact equilibria of the unregularized game; without a verified approximation bound, they should not be interpreted as such. Likewise, no-regret methods generally target solution concepts such as coarse correlated equilibrium rather than the assessment of Definition 5.1. In either case, satisfying assistant optimality ***(ii)*** requires the global verification described above.

F.3 Proof of Theorem 5.1

Theorem 5.1 (Babbling equilibrium and exact strategic maximin)**.** *Suppose that X, $\mathcal{E}$, the feasible slate sets, the report sets, and the horizon T are finite, and utilities are bounded. Fix $\epsilon \geq 0$ for which a welfare-feasible non-adaptive control exists. Then, there exists a welfare-feasible strategic assessment $(\pi^\circ, \sigma^\circ, \mu^\circ)$ in which all user types follow the same reporting rule. Consequently, the reports reveal no information about the user's type and the assistant's posterior remains equal to the prior, $q_t = p$ for every round t. This babbling equilibrium gives the assistant value $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$, and therefore $\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon) = \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$.*

Proof. We first construct a welfare-feasible babbling assessment with assistant value $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. Let σ° place, at every $(A, q, B) \in \mathcal{I}_t(\sigma^\circ, \mu^\circ)$ and for every user type, all probability on the $\succ$-greatest element of B. For every report, set $\mu_t^\circ(\cdot \mid A, q, B, \mathbf{y}) := q$. Both σ° and μ° are thus

specified at every information set and after every report, as required by Definition F.5. Because σ° is type-independent, Bayes' rule leaves the posterior unchanged after every positive-probability report, while μ° leaves it unchanged after every zero-probability report. Hence, $\tau_t^{\sigma^\circ,\mu^\circ}(q, A, B, \mathbf{y}) = q$ for every report $\mathbf{y}$. The resulting deviation-closed domain is therefore $\mathcal{D}_t(\sigma^\circ, \mu^\circ) = \{(A, p) : A \text{ is accumulated in } t-1 \text{ feasible rounds}\}$, so every state in the domain carries the prior. Since $q_1 = p$, it follows that $q_t = p$ at every round.

By the hypothesis of Theorem 5.1, the non-adaptive feasible set is nonempty. Because the horizon, accumulated-menu state space, and feasible slate sets are finite, the set of behavioral non-adaptive controls is compact. The ex-ante welfare constraint defines a closed subset, and expected assistant payoff is continuous. Hence, there exists a welfare-feasible non-adaptive control κ° attaining $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. Let $\pi^\circ := \pi^{\kappa^\circ}$ be its lift to $\mathcal{D}(\sigma^\circ, \mu^\circ)$. We next verify the four conditions of Definition F.5.

(i) Welfare feasibility. The control κ° was chosen to satisfy the non-adaptive welfare constraint. By Lemma F.1, its lift to $\mathcal{D}(\sigma^\circ, \mu^\circ)$ has the same expected regret and is therefore welfare-feasible.

(ii) Assistant optimality.

Step 1 (type independence). Fix any behavioral Markov policy π' on $\mathcal{D}(\sigma^\circ, \mu^\circ)$. Under the type-independent reporting rule σ°, the conditional distribution of the report given the public history is the same for every user type, while the assistant's policy depends only on the public state. An induction on t therefore shows that the distribution of the public history under (π', σ°) does not depend on η_u. Consequently, neither does the induced distribution D' of the terminal menu A_{T+1}.

Step 2 (implementation). Every menu in the support of D' is horizon-reachable, so Lemma F.2 implies that there exists a non-adaptive control κ' with $D_{\kappa'} = D'$. Lift κ' to the same domain $\mathcal{D}(\sigma^\circ, \mu^\circ)$. By Lemma F.1, this lift induces the joint distribution $D' \otimes p$ over (A_{T+1}, η_u). By Step 1, (π', σ°) induces the same joint distribution. Hence, by Proposition F.1, π' and the lift of κ' induce the same assistant value and expected regret. It follows that every welfare-feasible π' on this domain has value at most $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$, which π° attains. Thus π° satisfies assistant optimality.

(iii) User optimality. Because π° is the lift of a non-adaptive control, its action kernels are unchanged when the user changes its reports: conditional on the accumulated menu, no current report alters the distribution of future slates. This remains true after every unilateral report deviation, including reports that have zero probability under σ°, because the lift is q-invariant at every state of $\mathcal{D}(\sigma^\circ, \mu^\circ)$ and the domain is deviation-closed. When evaluating such a deviation, the prescribed belief rule remains fixed; no alternative signalling or belief rule is introduced.

Hence, by Proposition F.1, a type-η user's terminal utility is $\max_{\mathbf{x} \in A_{T+1}} U_\eta(\mathbf{x})$ regardless of its intermediate reports. Every report therefore attains the same continuation value at every information set in $\mathcal{I}_t(\sigma^\circ, \mu^\circ)$, so every report is optimal there and, in particular, σ° is sequentially optimal. This argument does *not* require the terminal menu and the reports to be statistically independent; they need not be.

(iv) Belief consistency. Because σ° is type-independent, Bayes' rule leaves the posterior unchanged after every positive-probability report. After every zero-probability report, μ° likewise leaves the posterior unchanged. Thus, the prescribed beliefs satisfy the required updating rule. In particular, as established in the construction above, $q_t = p$ at every round.

We have therefore constructed a welfare-feasible babbling assessment $(\pi^\circ, \sigma^\circ, \mu^\circ)$ with $q_t = p$ for every t and assistant value $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. Hence, $\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon) \le \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. Conversely, Proposition F.2 gives $\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon) \ge \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. Therefore, $\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon) = \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. □

We next quantify the cost of this identification failure using the decoy-anchor instance from Theorem C.2, with enough rounds and slate capacity for truthful feedback to identify the user type.

Corollary F.1 (Strategic non-identification can cause a persistent payoff loss)**.** *Consider the decoy-anchor construction of Theorem C.2 with $\epsilon \in (0, \frac{1}{4})$, and suppose that either $k \ge 2$ and $T \ge 2$, or*

$k \geq 1$ *and* $T \geq 3$*. Then,*

$$\mathrm{OPT}_{\mathrm{nf}}(\epsilon) = \mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon) = \frac{2\epsilon}{1+\epsilon}, \qquad \mathrm{OPT}(\epsilon) = \frac{1+\epsilon}{2}.$$

Consequently,

$$\mathrm{OPT}(\epsilon) - \mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon) = \frac{(1-\epsilon)^2}{2(1+\epsilon)} > 0, \qquad \mathrm{PoTF}(\epsilon) = \frac{(1+\epsilon)^2}{4\epsilon}.$$

At the babbling equilibrium of Theorem 5.1, the posterior therefore remains equal to the prior and the user type is never identified, while the additive loss relative to the truthful optimum is at least $\frac{9}{40}$ *for every* $\epsilon \in (0, \frac{1}{4})$*, uniformly in the horizon* T*. Moreover,* $\mathrm{PoTF}(\epsilon) \to +\infty$ *as* $\epsilon \downarrow 0$*, while* $\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon) > 0$ *for every* $\epsilon > 0$*.*

Proof. Recall the decoy-anchor construction. There is a single feature and $X = \{b, g, p_1, p_2, o\}$, with $r^{\eta_1} = (b{:}1,\ g{:}1-\epsilon,\ p_1{:}0.5,\ p_2{:}0.1,\ o{:}0)$ and $r^{\eta_2} = (b{:}1,\ g{:}0,\ p_1{:}0.1,\ p_2{:}0.5,\ o{:}0)$. The prior is uniform, $\tilde{U}_a(g) = 1$, and the assistant receives zero utility from every other item. Both types have $\bar{\mathbf{x}}(\eta_i) = b$, so $r_{\eta_1}(g) = \epsilon$, $r_{\eta_2}(g) = 1$, and $r_{\eta_i}(b) = 0$. For any menu A, define its expected assistant payoff and user regret by $P(A) := \sum_i \frac{1}{2}\tilde{U}_a(\mathbf{x}^\star(A, \eta_i))$ and $R(A) := \sum_i \frac{1}{2} r_{\eta_i}(\mathbf{x}^\star(A, \eta_i))$.

Step 1: A feasible non-adaptive control attains $\frac{2\epsilon}{1+\epsilon}$**.** Consider the non-adaptive control that reveals the singleton menu $\{g\}$ with probability $\lambda := \frac{2\epsilon}{1+\epsilon}$ and $\{b\}$ otherwise. Both menus are horizon-reachable, so such a control exists by Lemma F.2. From $\{b\}$, both types choose b, giving $P = R = 0$. From $\{g\}$, both choose g, giving $P = 1$ and $R = \frac{1}{2}\epsilon + \frac{1}{2}\cdot 1 = \frac{1+\epsilon}{2}$. The control therefore has value λ and expected regret $\lambda\frac{1+\epsilon}{2} = \epsilon$, and is thus welfare-feasible. Moreover, $\lambda < \frac{1}{2} < 1$ because $\epsilon < \frac{1}{4}$. Hence, $\mathrm{OPT}_{\mathrm{nf}}(\epsilon) \geq \frac{2\epsilon}{1+\epsilon}$.

Step 2: Every menu satisfies $P(A) \leq \frac{2}{1+\epsilon}R(A)$**.** If $b \in A$, both types choose b, which is their unique utility maximizer, so $P(A) = R(A) = 0$. Now suppose $b \notin A$. If $P(A) = 0$, the inequality is immediate, so consider $P(A) > 0$, which requires $g \in A$.

Since $1 - \epsilon > \frac{1}{2} > \frac{1}{10} > 0$ for $\epsilon < \frac{1}{4}$, item g is the unique maximizer for type η_1 in A. This type therefore contributes $\frac{1}{2}$ to $P(A)$ and $\frac{\epsilon}{2}$ to $R(A)$. For type η_2, the ranking over $A \subseteq \{g, p_1, p_2, o\}$ is $p_2 \succ_{\eta_2} p_1 \succ_{\eta_2} g \sim_{\eta_2} o$. We distinguish between three cases:

- If $p_2 \in A$, type η_2 chooses p_2, so $P(A) = \frac{1}{2}$ and $R(A) = \frac{\epsilon}{2} + \frac{1}{4}$. Hence, $\frac{2}{1+\epsilon}R(A) = \frac{\epsilon+\frac{1}{2}}{1+\epsilon} \geq \frac{1}{2} = P(A)$, since this equivalent to:

  $$\tfrac{1+\epsilon}{2} \leq \epsilon + \tfrac{1}{2} \iff \tfrac{\epsilon}{2} \leq \epsilon,$$

  which indeed holds.
- If $p_2 \notin A$ but $p_1 \in A$, type η_2 chooses p_1, so $P(A) = \frac{1}{2}$ and $R(A) = \frac{\epsilon}{2} + \frac{9}{20}$. Hence, $\frac{2}{1+\epsilon}R(A) = \frac{\epsilon+\frac{9}{10}}{1+\epsilon} \geq \frac{1}{2} = P(A)$, as this is equivalent $\epsilon \geq -\frac{4}{5}$, which indeed holds.
- Otherwise $A \subseteq \{g, o\}$. Type η_2 then obtains utility zero, so $R(A) = \frac{1+\epsilon}{2}$, while $P(A) \leq 1$. Therefore, $\frac{2}{1+\epsilon}R(A) = 1 \geq P(A)$, with equality when $A = \{g\}$.

Thus, $P(A) \leq \frac{2}{1+\epsilon}R(A)$ for every menu A. The conclusion does not depend on how the tie between g and o is resolved for type η_2.

Step 3: The non-adaptive value is $\frac{2\epsilon}{1+\epsilon}$**.** By Definition 5.2, a non-adaptive control induces a distribution D_κ over horizon-reachable terminal menus. By applying the inequality from Step 2 and taking expectations, welfare feasibility then gives $\mathbb{E}_{A\sim D_\kappa}[P(A)] \leq \frac{2}{1+\epsilon}\mathbb{E}_{A\sim D_\kappa}[R(A)] \leq \frac{2\epsilon}{1+\epsilon}$. Together with Step 1, this yields $\mathrm{OPT}_{\mathrm{nf}}(\epsilon) = \frac{2\epsilon}{1+\epsilon}$. By Theorem 5.1, $\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon) = \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$.

Step 4: The truthful value is at most $\frac{1+\epsilon}{2}$**.** Let π be any welfare-feasible policy and define $a_i := \mathbb{P}^\pi[\mathbf{x}^\star_T = g \mid \eta_u = \eta_i]$. Because g is the only item with positive assistant utility, $\mathbb{E}^\pi[\tilde{U}_a(\mathbf{x}^\star_T)] =$

$\frac{a_1+a_2}{2}$. All regrets are nonnegative, while $r_{\eta_1}(g) = \epsilon$ and $r_{\eta_2}(g) = 1$, so welfare feasibility implies $\epsilon \geq \mathbb{E}^{\pi}[G_u] \geq \frac{1}{2}\epsilon a_1 + \frac{1}{2}a_2$, or equivalently $\epsilon a_1 + a_2 \leq 2\epsilon$.

It remains to maximize $\frac{a_1+a_2}{2}$ over $a_1, a_2 \in [0,1]$ subject to the inequality $\epsilon a_1 + a_2 \leq 2\epsilon$. Since assigning probability to a_1 costs only ϵ units of regret per unit of objective probability, whereas assigning it to a_2 costs one, the optimum first sets $a_1 = 1$ and then $a_2 = 2\epsilon - \epsilon = \epsilon \leq 1$ because $\epsilon < \frac{1}{4} < 1$. Thus every welfare-feasible truthful policy has value at most $\frac{1+\epsilon}{2}$.

Step 5: A truthful policy attains $\frac{1+\epsilon}{2}$. Let π^{id} be the deterministic identifying policy: it reveals the probes p_1, p_2 together when $k \geq 2, T \geq 2$, or sequentially when $k \geq 1, T \geq 3$. Under deterministic truthful feedback, type η_1 chooses p_1 and type η_2 chooses p_2. The assistant then reveals g on the η_1 branch and b on the η_2 branch. This policy has value $\frac{1}{2}$ and expected regret $\frac{\epsilon}{2}$.

Let π^g be the non-adaptive policy that reveals only $\{g\}$. It has value 1 and expected regret $\frac{1+\epsilon}{2}$. Randomize at the root, using π^g with probability θ and π^{id} otherwise. The resulting value is $\frac{1}{2} + \frac{\theta}{2}$, and its expected regret is $(1-\theta)\frac{\epsilon}{2} + \theta\frac{1+\epsilon}{2} = \frac{\epsilon}{2} + \frac{\theta}{2}$. Setting $\theta = \epsilon$ therefore gives expected regret exactly ϵ and value $\frac{1+\epsilon}{2}$. This mixture is a randomized Markov policy admissible in the constrained Bayes value defined in Section 2, so together with Step 4, $\mathrm{OPT}(\epsilon) = \frac{1+\epsilon}{2}$.

Step 6: Payoff gap and Price of Truthful Feedback. Combining the preceding steps gives:

$$\mathrm{OPT}(\epsilon) - \mathrm{OPT}_{\mathrm{nf}}(\epsilon) = \tfrac{1+\epsilon}{2} - \tfrac{2\epsilon}{1+\epsilon} = \tfrac{(1+\epsilon)^2-4\epsilon}{2(1+\epsilon)} = \tfrac{(1-\epsilon)^2}{2(1+\epsilon)},$$

and $\mathrm{PoTF}(\epsilon) = \frac{(1+\epsilon)/2}{2\epsilon/(1+\epsilon)} = \frac{(1+\epsilon)^2}{4\epsilon}$. Both quantities are independent of T. The additive gap is decreasing in ϵ on $(0, \frac{1}{4})$ and is therefore at least its value at $\epsilon = \frac{1}{4}$, namely $\frac{9}{40}$. Finally, $\mathrm{PoTF}(\epsilon) \to +\infty$ as $\epsilon \downarrow 0$, while $\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon) = \frac{2\epsilon}{1+\epsilon} > 0$ for every $\epsilon > 0$. □

For comparison, the corresponding deterministic benchmark is $\mathrm{OPT}^{\mathrm{det}}_{\mathrm{nf}}(\epsilon) = 0$, where recall that we write $\mathrm{OPT}^{\mathrm{det}}_{\mathrm{nf}}(\epsilon) \leq \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ for the same benchmark $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ when the assistant is restricted to a single committed menu. Furthermore, $\lim_{\epsilon\downarrow 0} \mathrm{OPT}(\epsilon) = \frac{1}{2}$. Under a deterministic non-adaptive policy, welfare feasibility forces the committed menu to contain b, after which both types choose b and the assistant receives $\tilde{U}_a(b) = 0$. The contrast with $\mathrm{OPT}_{\mathrm{nf}}(\epsilon) = \frac{2\epsilon}{1+\epsilon} > 0$ illustrates why the benchmark's randomization class must be specified explicitly.

F.4 Proof of Theorem 5.2

Theorem F.1 (Full statement of Theorem 5.2: Bounds on the value of truthful feedback)**.** *Fix $\epsilon \geq 0$ for which a welfare-feasible non-adaptive control exists.*

(i) $\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon) = \mathrm{OPT}_{\mathrm{nf}}(\epsilon) \leq \mathrm{OPT}(\epsilon) = \mathrm{OPT}_{\mathrm{truthful}}(\epsilon)$. *Hence* $\mathrm{PoTF}(\epsilon) \geq 1$ *whenever it is defined, and*

$$\mathrm{OPT}_{\mathrm{truthful}}(\epsilon) - \mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon) = \mathrm{OPT}(\epsilon) - \mathrm{OPT}_{\mathrm{nf}}(\epsilon) \geq 0.$$

(ii) (Coincidence.) $\mathrm{OPT}(\epsilon) = \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$*, and hence* $\mathrm{PoTF}(\epsilon) = 1$ *whenever it is defined, in each of the following cases:*

- *(a) the prior p is degenerate;*
- *(b) the welfare constraint is non-binding, $\epsilon \geq \max_{\eta\in\mathrm{supp}(p)} \max_{\mathbf{x}\in X} r_\eta(\mathbf{x})$, and there exists an assistant-optimal item $\mathbf{x}^a \in \arg\max_{\mathbf{x}\in X} \tilde{U}_a(\mathbf{x})$ whose singleton menu $\{\mathbf{x}^a\}$ is horizon-reachable;*
- *(c) there exists a distribution over horizon-reachable* self-selecting *assignments (Definition 5.4) that attains the randomized direct benchmark $\mathrm{OPT}^{\mathrm{rand}}_{\mathrm{dir}}(p,\epsilon)$. This benchmark is the assistant's first-best value when it may assign each user type a lottery over terminal items, subject to the same ex-ante welfare constraint (Definition C.2).*

(iii) (Unboundedness.) $\sup\{\mathrm{OPT}_{\mathrm{truthful}}(\epsilon)/\,\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon)\} = +\infty$*, where the supremum is over instances and tolerances $\epsilon > 0$ for which* $\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon) > 0$.

Proof. **Part (i).** By Theorem 5.1, $\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon) = \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. A welfare-feasible non-adaptive control, when used under truthful feedback, is a randomized Markov policy whose slate choices do not depend on the observed feedback and is therefore admissible in the constrained Bayes value defined in Section 2. By Lemma F.1 and proposition F.1, its value and expected regret are exactly those induced by its terminal-menu law. Hence, $\mathrm{OPT}_{\mathrm{nf}}(\epsilon) \le \mathrm{OPT}(\epsilon)$. The additive identity and $\mathrm{PoTF}(\epsilon) \ge 1$ follow immediately.

Part (ii)(a). Let $p = \delta_{\eta_0}$. Under either truthful feedback model, the posterior remains δ_{η_0} at every history because $\mathrm{supp}(p) = \{\eta_0\}$. Fix any welfare-feasible policy π, and let $D \in \Delta(2^X)$ denote its induced distribution over terminal menus. Every menu in the support of D is horizon-reachable. By Lemma F.2, there exists a non-adaptive control κ with $D_\kappa = D$. It therefore induces the same terminal-menu law of A_{T+1} and, by Proposition F.1, the same assistant value $\sum_A D(A)\tilde{U}_a(\mathbf{x}^\star(A, \eta_0))$ and the same expected regret. Thus, every welfare-feasible truthful policy can be matched by a welfare-feasible non-adaptive control with the same value, giving $\mathrm{OPT}(\epsilon) \le \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. Together with part (i), this yields equality. The construction preserves randomization; it does not replace the original policy by a deterministic one.

Part (ii)(b). Let $\mathbf{x}^a \in \arg\max_{\mathbf{x}\in X} \tilde{U}_a(\mathbf{x})$ be an assistant-optimal item whose singleton menu $\{\mathbf{x}^a\}$ is horizon-reachable. Because the welfare constraint is non-binding, the deterministic non-adaptive control that reveals $\{\mathbf{x}^a\}$ is welfare-feasible. It gives the assistant $\max_{\mathbf{x}\in X} \tilde{U}_a(\mathbf{x})$, which is an upper bound on the payoff attainable under any policy. Therefore, $\mathrm{OPT}_{\mathrm{nf}}(\epsilon) = \mathrm{OPT}(\epsilon) = \max_{\mathbf{x}\in X} \tilde{U}_a(\mathbf{x})$.

Part (ii)(c). Let ν be a distribution over horizon-reachable self-selecting assignments that attains $\mathrm{OPT}^{\mathrm{rand}}_{\mathrm{dir}}(p, \epsilon)$ (recall Definition C.3 for the definition of self-selecting assignments). Consider the non-adaptive control that draws an assignment $(\mathbf{x}_\eta)_\eta \sim \nu$ and reveals its image menu $\hat{A} = \{\mathbf{x}_\eta : \eta \in \mathrm{supp}(p)\}$. Because every assignment in the support of ν is self-selecting, each type η chooses its assigned item $\mathbf{x}_\eta$ from $\hat{A}$ by self-selection. The resulting non-adaptive control therefore reproduces exactly the expected assistant payoff and expected regret of the randomized direct solution. It is welfare-feasible and attains $\mathrm{OPT}^{\mathrm{rand}}_{\mathrm{dir}}(p, \epsilon)$, so $\mathrm{OPT}_{\mathrm{nf}}(\epsilon) \ge \mathrm{OPT}^{\mathrm{rand}}_{\mathrm{dir}}(p, \epsilon)$.

Conversely, Proposition C.1 gives $\mathrm{OPT}(\epsilon) \le \mathrm{OPT}^{\mathrm{rand}}_{\mathrm{dir}}(p, \epsilon)$, while part (i) gives $\mathrm{OPT}_{\mathrm{nf}}(\epsilon) \le \mathrm{OPT}(\epsilon)$. Hence, $\mathrm{OPT}^{\mathrm{rand}}_{\mathrm{dir}}(p, \epsilon) \le \mathrm{OPT}_{\mathrm{nf}}(\epsilon) \le \mathrm{OPT}(\epsilon) \le \mathrm{OPT}^{\mathrm{rand}}_{\mathrm{dir}}(p, \epsilon)$, and equality holds throughout.

Part (iii). Apply Corollary F.1 to the decoy-anchor family. For every $\epsilon \in (0, \frac{1}{4})$, $\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon) = \frac{2\epsilon}{1+\epsilon} > 0$ and $\mathrm{OPT}_{\mathrm{truthful}}(\epsilon) = \frac{1+\epsilon}{2}$, so $\mathrm{PoTF}(\epsilon) = \frac{(1+\epsilon)^2}{4\epsilon} \to +\infty$ as $\epsilon \downarrow 0$. Since the denominator remains strictly positive for every $\epsilon > 0$, the supremum over instances and tolerances with positive worst-equilibrium value is $+\infty$. □

F.5 Relation among the Direct, Truthful, and Strategic benchmarks

We can now relate the direct-choice benchmark of Appendix C to the truthful and strategic values of Section 5.2. Under direct choice, the assistant may assign a different terminal item to each user type, with $\mathrm{OPT}_{\mathrm{dir}}(p, \epsilon)$ denoting the optimal value under the same ex-ante welfare constraint. A randomized truthful policy, however, can induce a lottery over terminal items for each type. The corresponding comparison therefore allows the assistant to assign a type-specific lottery directly, rather than a single type-specific item. This is the randomized direct value $\mathrm{OPT}^{\mathrm{rand}}_{\mathrm{dir}}(p, \epsilon)$ of Definition C.2.

Combining Proposition C.1 with Theorem 5.2(i) and Theorem 5.1, for every ϵ for which a welfare-feasible non-adaptive control exists,

$$\underbrace{\mathrm{OPT}^{\mathrm{rand}}_{\mathrm{dir}}(p, \epsilon)}_{\text{randomized direct}} \ge \underbrace{\mathrm{OPT}(\epsilon)}_{\text{truthful}} \ge \underbrace{\mathrm{OPT}_{\mathrm{nf}}(\epsilon) = \mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon)}_{\text{worst equilibrium}}.$$

The first gap captures the loss from requiring type-specific randomized outcomes to arise through the truthful-feedback interaction. The second captures the value of adapting to truthful feedback relative to the best possibly randomized non-adaptive control.

In the decoy-anchor instance of Corollary F.1, the first gap vanishes: $\mathrm{OPT}^{\mathrm{rand}}_{\mathrm{dir}}(p,\epsilon) = \mathrm{OPT}(\epsilon) = \frac{1+\epsilon}{2}$, whereas $\mathrm{OPT}_{\mathrm{nf}}(\epsilon) = \frac{2\epsilon}{1+\epsilon}$. Thus, strategic non-identification eliminates the value of adaptation without eliminating the assistant's payoff altogether. The distinction between deterministic and randomized benchmarks is important here: $\mathrm{OPT}_{\mathrm{dir}}(p,\epsilon) = \frac{1}{2} < \mathrm{OPT}(\epsilon)$ for $\epsilon > 0$. A deterministic direct assignment is therefore not an upper bound on the value of a randomized adaptive policy, which is why the randomized direct benchmark is required in the comparison above.

F.6 Deterministic Assistant Behavior and Outcome Determinacy

The collapse result in Theorem 5.3 does not require the assistant's policy itself to be deterministic. Its key requirement is instead that, conditional on the user's type, equilibrium play leads to a uniquely determined terminal item. We formalize this property as outcome determinacy and then show that Assumption (A3) is a sufficient condition for it.

Definition F.6 (Outcome-determinate assessment)**.** *A welfare-feasible strategic assessment (π,σ,μ) is* outcome-determinate *if, for every $\eta \in \mathrm{supp}(p)$, there exists an item $\phi(\eta) \in X$ such that*

$$\mathbb{P}^{(\pi,\sigma)}[\mathbf{x}^\star(A_{T+1},\eta) = \phi(\eta) \mid \eta_u = \eta] = 1.$$

We call $(\phi(\eta))_\eta$ the induced assignment *and $\hat{A} := \{\phi(\eta) : \eta \in \mathrm{supp}(p)\}$ its* image menu.

Condition ***(iii)*** of Definition 5.1 imposes user optimality separately at every reporting decision. The imitation argument used later, however, compares the prescribed signalling rule with an alternative reporting strategy followed throughout the remaining interaction. The next lemma connects these two scopes: in the finite-horizon Markov problem, local sequential optimality at every information set implies optimality against any complete Markov reporting deviation. The update rule and assistant policy are held fixed when evaluating such a deviation.

Lemma F.5 (Sequential optimality against Markov deviations)**.** *Fix a behavioral Markov assistant policy π, a belief selector μ, and the deviation domain $\mathcal{D}(\sigma,\mu)$ of Definition F.2. Fix a type η and suppose that Definition F.5**(iii)** holds for η at every $(A,q,B) \in \mathcal{I}_t(\sigma,\mu)$, $t = 1,\dots,T$.*

For any alternative behavioral Markov reporting strategy $\widetilde{\sigma} = \{\widetilde{\sigma}_t(\cdot \mid A,q,B)\}$ for type η on the same domain, let $U^{\pi,\widetilde{\sigma}}$ denote the continuation values obtained when $\widetilde{\sigma}$ replaces the type-η component of σ, while π and $\tau^{\sigma,\mu}$ remain fixed. Then, $U^{\pi,\widetilde{\sigma}}_t(A,q,\eta) \le U^{\pi,\sigma}_t(A,q,\eta)$ for every t and every $(A,q) \in \mathcal{D}_t(\sigma,\mu)$. Hence, the prescribed reporting strategy for type η is optimal, from every state in the deviation domain, against all behavioral Markov reporting deviations on that domain.

Proof. Only type η deviates. The deviation is unilateral and unobserved: the assistant's kernels $\pi_t(\cdot \mid A,q)$ and the assessment's update rule $\tau^{\sigma,\mu}_t(q,A,B,\mathbf{y})$ depend only on the public state and the observed report, and therefore remain unchanged. Consequently, $\widetilde{\sigma}$ induces a process on the same deviation domain $\mathcal{D}(\sigma,\mu)$, so the prescribed and deviating continuation values below are defined on the same finite state sets.

We specialize the best-response recursion of Section 3.1 to the assessment's own update rule, setting $\bar{\tau}_t = \tau^{\sigma,\mu}_t$ and writing $U^\pi_t := U^{\pi,\bar{\tau}}_t$. Thus,

$$U^\pi_{T+1}(A,q,\eta) = \max_{\mathbf{x}\in A} U_\eta(\mathbf{x}),$$

and

$$U^\pi_t(A,q,\eta) = \sum_S \pi_t(S \mid A,q) \max_{\mathbf{y}\in A\cup S} U^\pi_{t+1}\big(A\cup S, \tau^{\sigma,\mu}_t(q,A,A\cup S,\mathbf{y}),\eta\big).$$

The recursion first averages over the assistant's round-t slate and then maximizes over the user's report from the resulting menu. This matches the order of moves: the assistant chooses the slate, the user observes the resulting menu and reports, the report determines the next posterior, and the assistant chooses its next slate in the following round.

We first show that $U^{\pi,\widetilde{\sigma}}_t \le U^\pi_t$ pointwise for every Markov $\widetilde{\sigma}$. We use backward induction on t. At $t = T+1$, the two values coincide, both being $\max_{\mathbf{x}\in A} U_\eta(\mathbf{x})$. Suppose that $U^{\pi,\widetilde{\sigma}}_{t+1} \le U^\pi_{t+1}$ pointwise on $\mathcal{D}_{t+1}(\sigma,\mu)$. For every $(A,q,B) \in \mathcal{I}_t(\sigma,\mu)$,

$$W^{\pi,\widetilde{\sigma}}_t(A,q,B,\eta) = \sum_{\mathbf{y}\in B} \widetilde{\sigma}_t(\mathbf{y} \mid A,q,B)\, U^{\pi,\widetilde{\sigma}}_{t+1}\big(B, \tau^{\sigma,\mu}_t(q,A,B,\mathbf{y}),\eta\big)$$

and hence

$$W_t^{\pi,\widetilde{\sigma}}(A,q,B,\eta) \le \max_{\mathbf{y}\in B} U_{t+1}^{\pi}\big(B,\tau_t^{\sigma,\mu}(q,A,B,\mathbf{y}),\eta\big).$$

The inequality follows by applying the induction hypothesis to each report and then bounding the resulting average by its maximum. Averaging over the assistant's slate therefore gives $U_t^{\pi,\widetilde{\sigma}} \le U_t^{\pi}$ pointwise on $\mathcal{D}_t(\sigma,\mu)$.

We next show, again by backward induction, that the prescribed strategy attains this bound, i.e., $U_t^{\pi,\sigma} = U_t^{\pi}$ pointwise under the hypothesis on σ. The equality is immediate at $t = T+1$. Suppose that $U_{t+1}^{\pi,\sigma} = U_{t+1}^{\pi}$ pointwise on $\mathcal{D}_{t+1}(\sigma,\mu)$. For any $(A,q,B) \in \mathcal{I}_t(\sigma,\mu)$, the two maps $\mathbf{y} \mapsto U_{t+1}^{\pi,\sigma}\big(B,\tau_t^{\sigma,\mu}(q,A,B,\mathbf{y}),\eta\big)$ and $\mathbf{y} \mapsto U_{t+1}^{\pi}\big(B,\tau_t^{\sigma,\mu}(q,A,B,\mathbf{y}),\eta\big)$ coincide by the induction hypothesis and therefore have the same maximizers. By the assumed user optimality condition, $\sigma_t(\cdot \mid A,q,B,\eta)$ is supported on these maximizing reports. It follows that

$$W_t^{\pi,\sigma}(A,q,B,\eta) = \max_{\mathbf{y}\in B} U_{t+1}^{\pi}\big(B,\tau_t^{\sigma,\mu}(q,A,B,\mathbf{y}),\eta\big).$$

Averaging over the assistant's slate gives $U_t^{\pi,\sigma} = U_t^{\pi}$ pointwise on $\mathcal{D}_t(\sigma,\mu)$. Crucially, the user-optimality hypothesis applies at *every* information set, including those reached with zero probability under π and σ. This is what makes the equality $U_t^{\pi,\sigma} = U_t^{\pi}$ hold pointwise on the entire deviation domain $\mathcal{D}_t(\sigma,\mu)$, rather than only along equilibrium play.

Combining the two inductions above gives

$$U_t^{\pi,\widetilde{\sigma}}(A,q,\eta) \le U_t^{\pi}(A,q,\eta) = U_t^{\pi,\sigma}(A,q,\eta)$$

for every t and every $(A,q) \in \mathcal{D}_t(\sigma,\mu)$. Evaluating at $t = 1$ and $(A,q) = (\varnothing,p)$ gives the corresponding ex-ante optimality statement. □

Thus, the local user-optimality conditions in Definition 5.1**(iii)** also rule out complete Markov reporting deviations on the same deviation-closed domain. This is the global user-deviation property needed below; in particular, the imitation deviation used later in Proposition F.3 belongs to this strategy class.

The scope of Lemma F.5 is deliberately limited to Markov reporting strategies. Its proof is a finite-horizon dynamic-programming argument on the Markov state, rather than an application of the one-shot-deviation principle. It does not assert optimality against arbitrary full-history continuation strategies of the underlying extensive-form game, nor is such a result needed below. The proof also identifies the recursion from Section 3.1, specialized to $\bar{\tau}_t = \tau_t^{\sigma,\mu}$, with the user's behavioral Markov best-response problem against π. Thus, the value U^{π} is the best value attainable by a Markov reporting strategy, and the corresponding dynamic program provides an exact test of user optimality within this class. The update rule remains fixed throughout: a deviation changes the user's reports, not the rule by which those reports update the assistant's beliefs.

A related distinction arises when dynamic incentive compatibility is used to establish the user-optimality condition of Definition 5.1**(iii)**. Because that condition also covers off-path reporting decisions, the incentive condition must have the same scope. We formalize this stronger, deviation-complete notion later in Appendix F.9.

Assumption (A3) requires deterministic assistant behavior at every state at which the assistant may act. Formally, this means that $\pi_t(\cdot \mid A,q)$ is degenerate at every $(A,q) \in \mathcal{D}_t(\sigma,\mu)$, where $\mathcal{D}(\sigma,\mu)$ is the deviation-closed domain of Definition F.2. Thus, the restriction applies throughout the domain on which assistant deviations are evaluated, not only along equilibrium play. It places no restriction on the user's signalling rule, which may remain type-dependent and mixed.

The purpose of this restriction is to rule out assistant randomization whose realization is still unknown when the user reports. To distinguish this case from randomization realized beforehand, we use the following terminology.

Definition F.7 (Resolved and unresolved assistant randomization)**.** *At a user reporting decision, assistant randomization relevant to future play is* resolved and public *if its realization is observed by the user before the report is chosen. It is* unresolved *if the user's continuation utility from a report averages over assistant randomization that has not yet been realized or observed.*

In the baseline behavioral model, assistant randomization occurring after a report is unresolved at that reporting decision, whereas a deterministic assistant policy has no such continuation randomization. In the public-randomization variant introduced later in Appendix F.11, the relevant random seed is instead realized and publicly revealed before the user reports.

We now connect deterministic assistant behavior to the outcome-determinacy condition in Definition F.6. Determinism alone does not prevent the user's signalling rule from being type-dependent or mixed, so different reports may still generate different public histories. The key point is that sequential user optimality forces all histories followed with positive probability by a given type to deliver the same terminal utility. Under strict preferences, equality of terminal utility then implies equality of the terminal item.

Lemma F.6 (Deterministic assistant behavior implies outcome determinacy)**.** *Assume (A1). Let (π, σ, μ) be a welfare-feasible strategic assessment such that $\pi_t(\cdot \mid A, q)$ is degenerate at every $(A, q) \in \mathcal{D}_t(\sigma, \mu)$. The signalling rule σ may be type-dependent and mixed. Then, (π, σ, μ) is outcome-determinate.*

Proof. Fix $\eta \in \operatorname{supp}(p)$. Because the public history determines (t, A_t, q_t) and π is degenerate at every such state, the slate chosen at each round, and hence the terminal menu, is determined by the public history. For a user information set H reached with positive probability under type η's equilibrium play, let $W_\eta(H)$ denote that type's expected continuation utility under (π, σ), where the only remaining randomness comes from its future reports. At a terminal history,

$$W_\eta(H) = \max_{\mathbf{x} \in A(H)} U_\eta(\mathbf{x}).$$

Because π and σ are Markov, $W_\eta(H)$ depends on H only through its stage and public state and coincides with the corresponding continuation value of Definition F.4. Thus, Definition 5.1**(iii)** applies directly to the continuation comparisons below.

We prove by backward induction on the reporting stage that, at every such information set H, every terminal history reached with positive probability under the continuation of σ yields terminal utility exactly $W_\eta(H)$.

At the final reporting stage $t = T$, the terminal menu has already been determined by the public history. By Proposition F.1, the user's terminal item is $\mathbf{x}^\star(A_{T+1}, \eta)$ regardless of its report. Hence, every continuation yields terminal utility $\max_{\mathbf{x} \in A(H)} U_\eta(\mathbf{x}) = W_\eta(H)$.

Now, suppose the claim holds at every supported user information set at stage $t+1$, and consider a supported information set H at stage $t < T$, with state (A, q) and menu B. By user optimality, Definition 5.1**(iii)**, $\sigma_t(\cdot \mid A, q, B, \eta)$ is supported on reports maximizing $\mathbf{y} \mapsto U_{t+1}^{\pi,\sigma}\big(B, \tau_t^{\sigma,\mu}(q, A, B, \mathbf{y}), \eta\big)$. Therefore, $W_\eta(H, \mathbf{y}) = W_\eta(H)$ for every $\mathbf{y} \in \operatorname{supp} \sigma_t(\cdot \mid A, q, B, \eta)$.

Fix any such report $\mathbf{y}$. Since π is degenerate, the assistant's next slate is determined by the public history $(H, \mathbf{y})$, so there is a unique successor user information set H'. Moreover,

$$W_\eta(H') = W_\eta(H, \mathbf{y}) = W_\eta(H).$$

By the induction hypothesis at H', every terminal history reached with positive probability below H' yields terminal utility $W_\eta(H)$. This holds for every supported report $\mathbf{y}$, and every terminal history reached with positive probability below H follows one of these reports. The claim therefore holds at H.

Finally, because π is degenerate, the first slate is determined and there is a unique initial user information set H_1. The preceding argument shows that every terminal history reached with positive probability under type η's equilibrium play yields the same terminal utility $U_\eta^\star := W_\eta(H_1)$. Thus, $\max_{\mathbf{x} \in A(H)} U_\eta(\mathbf{x}) = U_\eta^\star$ at every supported terminal history H. Under Assumption (A1), distinct items have distinct utilities for type η, so there is at most one item with utility $U_\eta^\star$. Hence, $\mathbf{x}^\star(A(H), \eta)$ is the same item at every supported terminal history. Denote it by $\phi(\eta)$. Since the argument applies to every $\eta \in \operatorname{supp}(p)$, the assessment is outcome-determinate. □

Degeneracy of π is essential to this argument: it makes both the terminal menu a function of the public history and the successor information set unique. With unresolved continuation random-

ization, the continuation value $W_\eta(H, \mathbf{y})$ instead averages over future realizations, so equality of continuation values need not imply equality of realized terminal utilities.

F.7 Self-Selection of Equilibrium Terminal Items

Outcome determinacy assigns a terminal item $\phi(\eta)$ to each user type, but the collapse argument requires more: these items must be jointly implementable through their common image menu. We now show that equilibrium incentives provide exactly this property. The argument lets one type imitate another type's reporting strategy and uses Lemma F.5 to show that such imitation cannot improve its utility.

Proposition F.3 (Equilibrium terminal items are self-selecting)**.** *Assume (A1). Let (π, σ, μ) be an outcome-determinate welfare-feasible strategic assessment with induced assignment ϕ and image menu $\hat{A} := \{\phi(\eta) : \eta \in \mathrm{supp}(p)\}$. Then, the induced assignment is self-selecting: $\mathbf{x}^\star(\hat{A}, \eta) = \phi(\eta)$ for every $\eta \in \mathrm{supp}(p)$.*

Proof. Fix $\eta, \eta' \in \mathrm{supp}(p)$. Consider the deviation in which type η imitates the reporting strategy of type η' at every information set $(A, q, B) \in \mathcal{I}_t(\sigma, \mu)$:

$$\widehat{\sigma}_{\eta,t}(\cdot \mid A, q, B) := \sigma_t(\cdot \mid A, q, B, \eta').$$

This is an admissible behavioral Markov deviation: the user's true type is private, and the strategy class permits replacing the type argument of the reporting kernel.

We first show that, conditional on $\eta_u = \eta$, the public-state process under $(\pi, \widehat{\sigma}_\eta)$ has the same law as the equilibrium public-state process conditional on $\eta_u = \eta'$ under (π, σ). We prove this by induction on t. Both processes start from $(A_1, q_1) = (\varnothing, p)$. Suppose they have the same law at the beginning of round t. Conditional on a common state (A, q), the assistant draws its slate from the same kernel $\pi_t(\cdot \mid A, q)$, so the resulting menu B has the same conditional law. Conditional on B, both users report according to the same distribution $\sigma_t(\cdot \mid A, q, B, \eta')$. After any realized report $\mathbf{y}$, both processes therefore move to the same posterior $q^+ = \tau_t^{\sigma,\mu}(q, A, B, \mathbf{y})$. Indeed, the assessment's update rule depends on the observed public state and report, not on the deviating user's unobserved true type. Hence, the next public state has the same law in both processes, and induction establishes the claim.

The beliefs in this comparison are those prescribed by the fixed assessment (π, σ, μ) after the observed public state; a unilateral deviation does not cause the assistant to condition on the deviator's unobserved true type. Moreover, because $\eta' \in \mathrm{supp}(p)$, every state in the support of the imitated process has positive ex-ante probability under (π, σ). Beliefs along this process therefore follow Bayes' rule, rather than the off-path selector μ.

By outcome determinacy for type η', every terminal menu in the support of this process contains $\phi(\eta')$, since $\mathbf{x}^\star(A_{T+1}, \eta') = \phi(\eta')$. Here, recall that ϕ is the assignment induced by the assessment (π, σ, μ) and $\hat{A} = \{\phi(\eta) : \eta \in \mathrm{supp}(p)\}$ is ϕ's image menu. The deviating type-η user instead makes its own terminal choice from that menu, so Proposition F.1 gives, pathwise,

$$\max_{\mathbf{x} \in A_{T+1}} U_\eta(\mathbf{x}) \geq U_\eta(\phi(\eta')).$$

Thus, imitating type η' gives type η expected utility at least $U_\eta(\phi(\eta'))$.

The imitation rule $\widehat{\sigma}_\eta$ is a complete behavioral Markov reporting strategy on the incumbent deviation domain $\mathcal{D}(\sigma, \mu)$: it is specified at every information set in $\mathcal{I}_t(\sigma, \mu)$ and depends only on the Markov state (A, q, B). Hence, Lemma F.5 implies that type η cannot gain from this imitation deviation. Outcome determinacy gives its equilibrium utility as $U_\eta(\phi(\eta))$, while the imitation deviation yields at least $U_\eta(\phi(\eta'))$. Therefore, $U_\eta(\phi(\eta)) \geq U_\eta(\phi(\eta'))$ for any $\eta, \eta' \in \mathrm{supp}(p)$. Thus, $\phi(\eta)$ maximizes U_η over the image menu $\hat{A}$. Under Assumption (A1), this maximizer is unique, so $\mathbf{x}^\star(\hat{A}, \eta) = \phi(\eta)$ for every $\eta \in \mathrm{supp}(p)$. □

F.8 Proof of Theorem 5.3

We now combine outcome determinacy with the self-selection result of Proposition F.3. The argument has two parts. First, the type-specific equilibrium outcomes can be reproduced by revealing

their common image menu non-adaptively, which bounds the equilibrium value from above. Second, the assistant can always secure the non-adaptive value against strategic feedback. We then specialize the result to deterministic assistant behavior and record the resulting restriction on equilibrium existence.

The first result shows that every outcome-determinate equilibrium with a reachable image menu gives the assistant exactly its non-adaptive value $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$, defined in Definition 5.2 as the best welfare-feasible value attainable without conditioning future slates on feedback.

Theorem F.2 (Collapse of outcome-determinate equilibria)**.** *Fix $\epsilon \geq 0$ for which a welfare-feasible non-adaptive control exists, and assume (A1). Let (π, σ, μ) be an outcome-determinate welfare-feasible strategic assessment with induced assignment ϕ and horizon-reachable image menu $\hat{A} = \{\phi(\eta) : \eta \in \mathrm{supp}(p)\}$. Then,*

$$\mathbb{E}^{(\pi,\sigma)}\left[\tilde{U}_a(\mathbf{x}_T^\star)\right] = \mathrm{OPT}_{\mathrm{nf}}(\epsilon).$$

Proof. Write $V := \mathbb{E}^{(\pi,\sigma)}[\tilde{U}_a(\mathbf{x}_T^\star)]$. By outcome determinacy, the terminal item of type η is $\phi(\eta)$ almost surely, so

$$V = \sum_\eta p(\eta)\tilde{U}_a(\phi(\eta)), \qquad \sum_\eta p(\eta) r_\eta(\phi(\eta)) \leq \epsilon,$$

where the second inequality follows from welfare feasibility due to Definition 5.1**(i)**.

Upper bound. By Proposition F.3, the assignment ϕ is self-selecting. Hence, revealing its image menu $\hat{A}$ makes every type η select $\mathbf{x}^\star(\hat{A}, \eta) = \phi(\eta)$; see Proposition F.1. Because $\hat{A}$ is horizon-reachable, the deterministic non-adaptive control revealing this menu is admissible in Definition 5.2. Its terminal-menu law $\delta_{\hat{A}}$ reproduces the assessment's terminal item for every type, and therefore has the same assistant value V and the same expected regret. It is thus welfare-feasible, giving

$$V \leq \mathrm{OPT}_{\mathrm{nf}}^{\mathrm{det}}(\epsilon) \leq \mathrm{OPT}_{\mathrm{nf}}(\epsilon).$$

Lower bound. By Proposition F.2, the assistant can secure $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ against strategic feedback, so $V \geq \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. Combining the two bounds gives $V = \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. □

Theorem F.2 is stated directly in terms of outcome determinacy. The deterministic-assistant condition used in Assumption (A3) is a sufficient condition for outcome determinacy, while Assumption (A2) ensures that the resulting image menu is reachable. This gives the following specialization.

Corollary F.2 (Deterministic assistant behavior)**.** *Assume (A1) and (A2). Every welfare-feasible strategic assessment of the baseline game whose assistant policy is degenerate at every state of $\mathcal{D}_t(\sigma, \mu)$ has assistant value exactly $\mathrm{OPT}_{\mathrm{nf}}(\epsilon) \leq \mathrm{OPT}(\epsilon)$, regardless of the user's Markov signalling rule, which may be type-dependent and mixed.*

Proof. By Lemma F.6, the assessment is outcome-determinate, and Proposition F.3 makes its induced assignment self-selecting. Assumption (A2) makes the induced image menu $\hat{A}$ horizon-reachable. Revealing $\hat{A}$ non-adaptively reproduces the assessment's typewise terminal items and expected regret, so welfare feasibility of the assessment implies welfare feasibility of this non-adaptive control. Thus, the feasibility requirement of Theorem F.2 is satisfied, and that theorem gives the value $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. Finally, Theorem 5.2(i) gives $\mathrm{OPT}_{\mathrm{nf}}(\epsilon) \leq \mathrm{OPT}(\epsilon)$. □

The proof also reveals an existence restriction. An outcome-determinate equilibrium is reproduced by a *deterministic* non-adaptive control, yet assistant optimality compares it with randomized policies as well. Thus, such an equilibrium can exist only when randomization does not improve the non-adaptive benchmark.

Corollary F.3 (Necessary condition for outcome-determinate equilibrium)**.** *Recall that $\mathrm{OPT}_{\mathrm{nf}}^{\mathrm{det}}(\epsilon)$ is the non-adaptive value restricted to a single committed menu. If an outcome-determinate welfare-feasible strategic assessment exists, then $\mathrm{OPT}_{\mathrm{nf}}^{\mathrm{det}}(\epsilon) = \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. Consequently, if $\mathrm{OPT}_{\mathrm{nf}}^{\mathrm{det}}(\epsilon) < \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$, no outcome-determinate welfare-feasible strategic assessment exists.*

Proof. For an outcome-determinate assessment with value V, the upper-bound argument in Theorem F.2 gives $V \leq \mathrm{OPT}_{\mathrm{nf}}^{\mathrm{det}}(\epsilon)$, while Proposition F.2 gives $V \geq \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. Since $\mathrm{OPT}_{\mathrm{nf}}^{\mathrm{det}}(\epsilon) \leq \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ by definition, equality follows. □

However, this condition can fail. In the decoy-anchor instance of Corollary F.1, $\mathrm{OPT}_{\mathrm{nf}}^{\mathrm{det}}(\epsilon) = 0$ while $\mathrm{OPT}_{\mathrm{nf}}(\epsilon) = 2\epsilon/(1+\epsilon) > 0$, so no outcome-determinate welfare-feasible equilibrium exists. This does not reflect a restriction on assistant deviations: under Definition 5.1**(ii)**, a deterministic equilibrium policy must still be optimal against all admissible behavioral policy deviations, including randomized ones.

We can now prove Theorem 5.3. The first part follows directly from the outcome-determinate collapse, while the deterministic-assistant statement follows from the preceding corollary.

Theorem 5.3 (Equilibrium collapse under outcome determinacy)**.** *Fix ϵ for which a welfare-feasible non-adaptive control exists and assume (A1). Let (π, σ, μ) be an outcome-determinate welfare-feasible strategic assessment, and suppose its image menu is horizon-reachable, which is automatic under Assumption (A2). Then, $\mathbb{E}^{(\pi,\sigma)}\big[\tilde{U}_a(\mathbf{x}_T^\star)\big] = \mathrm{OPT}_{\mathrm{nf}}(\epsilon) \leq \mathrm{OPT}(\epsilon)$. Assumption (A3) implies outcome determinacy, so the result holds even with type-dependent and mixed user signalling.*

Proof. By Theorem F.2, every assessment satisfying the stated conditions has value $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$, and Theorem 5.2(i) gives $\mathrm{OPT}_{\mathrm{nf}}(\epsilon) \leq \mathrm{OPT}(\epsilon)$. Under Assumption (A3), Lemma F.6 gives outcome determinacy, so Corollary F.2 yields the final statement. □

F.9 Dynamic Incentive Compatibility beyond Truthful Play

Dynamic incentive compatibility in Definition D.1 requires truthful reporting to be continuation-optimal at states reached under truthful play. For a strategic assessment, however, user optimality has a broader scope: Definition 5.1**(iii)** must hold at every information set in the deviation-closed domain, including those reached only after a zero-probability report. We therefore strengthen DIC by imposing the same continuation-optimality condition throughout that domain.

Definition F.8 (Deviation-complete dynamic incentive compatibility)**.** *Fix the truthful Markov signalling rule σ^{tr} and an off-path selector μ. Let $\mathcal{D}(\sigma^{\mathrm{tr}}, \mu)$ be the associated deviation-closed domain of Definition F.2, and write $\bar{\tau}_t^{\mathrm{tr},\mu} := \tau_t^{\sigma^{\mathrm{tr}},\mu}$ for its posterior-update rule. An assistant policy π is* deviation-complete dynamically incentive compatible *relative to μ if, for every round t, every type η, every information set $(A, q, B) \in \mathcal{I}_t(\sigma^{\mathrm{tr}}, \mu)$, and every myopically optimal report $\mathbf{y}^\star \in \arg\max_{\mathbf{x}\in B} U_\eta(\mathbf{x})$,*

$$\mathbf{y}^\star \in \arg\max_{\mathbf{y}\in B} U_{t+1}^{\pi,\bar{\tau}^{\mathrm{tr},\mu}}\big(B, \bar{\tau}_t^{\mathrm{tr},\mu}(q, A, B, \mathbf{y}), \eta\big).$$

Thus, truthful choice must remain continuation-optimal at every information set needed to evaluate report deviations, not only those reached with positive probability under truthful play.

Deviation-complete DIC therefore strengthens ordinary DIC from Definition D.1. Ordinary DIC imposes continuation-optimality only at states reached under truthful reporting for the fixed pair $(\pi, \bar{\tau})$, whereas Definition F.8 imposes the same condition at every information set of the deviation-closed domain $\mathcal{I}_t(\sigma^{\mathrm{tr}}, \mu)$, including information sets reached only after zero-probability reports. Hence, deviation-complete DIC implies ordinary DIC, but the converse need not hold. As the counterexamples in Appendix F.12 show, a policy can satisfy ordinary DIC throughout truthful play while truthful signalling fails the user optimality condition of Definition 5.1**(iii)** off path.

The next lemma makes the connection to strategic assessments precise. Under strict preferences, truthful reporting selects a unique myopically optimal item. Deviation-complete DIC then guarantees that this report also maximizes continuation utility at every information set in the deviation-closed domain.

Lemma F.7 (Deviation-complete DIC implies truthful sequential optimality)**.** *Assume (A1). Fix an assistant policy π, truthful Markov signalling σ^{tr}, and an off-path selector μ on $\mathcal{D}(\sigma^{\mathrm{tr}}, \mu)$. If π is deviation-complete dynamically incentive compatible relative to μ, then σ^{tr} satisfies Definition 5.1**(iii)** at every information set in $\mathcal{I}_t(\sigma^{\mathrm{tr}}, \mu)$.*

Proof. Under Assumption (A1), U_η is injective on X. Hence, at every $(A, q, B) \in \mathcal{I}_t(\sigma^{\mathrm{tr}}, \mu)$, the myopic maximizer $\mathbf{y}^{\mathrm{tr}} = \arg\max_{\mathbf{x}\in B} U_\eta(\mathbf{x})$ is unique, and $\sigma_t^{\mathrm{tr}}(\cdot \mid A, q, B, \eta)$ is the point mass at $\mathbf{y}^{\mathrm{tr}}$.

We first show that $U_s^{\pi,\sigma^{\mathrm{tr}}} = U_s^{\pi,\bar{\tau}^{\mathrm{tr},\mu}}$ at every state of $\mathcal{D}_s(\sigma^{\mathrm{tr}}, \mu)$, for every s and every type. We proceed by backward induction on s. At $s = T + 1$, both values equal $\max_{\mathbf{x}\in A} U_\eta(\mathbf{x})$. Suppose the equality holds at $s + 1$, and fix $(A, q, B) \in \mathcal{I}_s(\sigma^{\mathrm{tr}}, \mu)$. By the induction hypothesis, the two maps $\mathbf{y} \mapsto U_{s+1}^{\pi,\sigma^{\mathrm{tr}}}(B, \bar{\tau}_s^{\mathrm{tr},\mu}(q, A, B, \mathbf{y}), \eta)$ and $\mathbf{y} \mapsto U_{s+1}^{\pi,\bar{\tau}^{\mathrm{tr},\mu}}(B, \bar{\tau}_s^{\mathrm{tr},\mu}(q, A, B, \mathbf{y}), \eta)$ coincide and therefore have the same maximizers. By Definition F.8, applied at this information set, $\mathbf{y}^{\mathrm{tr}}$ maximizes the second map and hence also the first. Since σ^{tr} places probability one on $\mathbf{y}^{\mathrm{tr}}$, averaging under it returns exactly this maximum:

$$W_s^{\pi,\sigma^{\mathrm{tr}}}(A, q, B, \eta) = \max_{\mathbf{y}\in B} U_{s+1}^{\pi,\bar{\tau}^{\mathrm{tr},\mu}}\big(B, \bar{\tau}_s^{\mathrm{tr},\mu}(q, A, B, \mathbf{y}), \eta\big).$$

Averaging over the assistant's slate then gives $U_s^{\pi,\sigma^{\mathrm{tr}}} = U_s^{\pi,\bar{\tau}^{\mathrm{tr},\mu}}$ on $\mathcal{D}_s(\sigma^{\mathrm{tr}}, \mu)$. Because Definition F.8 is imposed at *every* information set in the deviation-closed domain, this induction holds throughout the domain, not only along positive-probability play.

Now, fix any $(A, q, B) \in \mathcal{I}_t(\sigma^{\mathrm{tr}}, \mu)$ and any type η. By Definition F.8 and the equality just established, $\mathbf{y}^{\mathrm{tr}}$ maximizes $\mathbf{y} \mapsto U_{t+1}^{\pi,\sigma^{\mathrm{tr}}}(B, \bar{\tau}_t^{\mathrm{tr},\mu}(q, A, B, \mathbf{y}), \eta)$, i.e.,

$$\mathbf{y}^{\mathrm{tr}} \in \arg\max_{\mathbf{y}\in B} U_{t+1}^{\pi,\bar{\tau}^{\mathrm{tr},\mu}}\big(B, \bar{\tau}_t^{\mathrm{tr},\mu}(q, A, B, \mathbf{y}), \eta\big) = \arg\max_{\mathbf{y}\in B} U_{t+1}^{\pi,\sigma^{\mathrm{tr}}}\big(B, \bar{\tau}_t^{\mathrm{tr},\mu}(q, A, B, \mathbf{y}), \eta\big),$$

Since $\operatorname{supp}\sigma_t^{\mathrm{tr}}(\cdot \mid A, q, B, \eta) = \{\mathbf{y}^{\mathrm{tr}}\}$, the prescribed truthful report is therefore supported on continuation-optimal reports. This is precisely Definition 5.1**(iii)** at the given information set. Since the information set and type were arbitrary, the condition holds throughout $\mathcal{I}_t(\sigma^{\mathrm{tr}}, \mu)$. □

F.10 Reachability, Tie-Breaking, and Existence

We collect here two qualifications to the collapse argument that are separated from its main proof: reachability of the image menu and uniqueness of each type's preferred item. Both are needed to reproduce the equilibrium terminal outcomes with a single non-adaptive menu.

Assumption (A2) guarantees that the image menu $\hat{A}$ formed from the equilibrium terminal items is horizon-reachable. Without this assumption, the menus arising for individual types may each be reachable even though their union $\hat{A}$ is not. The equilibrium value may then exceed the horizon-restricted non-adaptive value $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. The collapse argument nevertheless applies whenever the particular image menu $\hat{A}$ induced by the equilibrium is horizon-reachable, which is the weaker condition imposed in Theorem 5.3.

Assumption (A1) serves a different purpose: it ensures that a type's terminal utility uniquely determines its terminal item. Without strict preferences, revealing the common image menu may preserve each type's terminal utility, and hence user welfare, while tie-breaking selects a different utility-maximizing item. Because the assistant's payoff depends on the selected item, its value need not be preserved. A fixed strict perturbation of user utilities restores the comparison, but we make no collapse claim without Assumption (A1).

F.11 Public Randomization and Equilibrium Collapse

We now formalize the public-randomization variant described in Section 5.3. Its purpose is to separate randomization itself from its timing. Unlike in the baseline game, any assistant randomization relevant to future play is realized and publicly observed *before* the user reports. Conditional on that realization, the assistant therefore behaves deterministically. We first define this augmented game and its belief updates, and then show that each realized seed reduces to a baseline deterministic problem, allowing the collapse argument to be applied seed by seed.

To represent public randomization without restricting the assistant to an arbitrarily small collection of deterministic policies, fix a common finite set of seed labels. Let N_{plan} be the number of complete deterministic assistant contingent plans on the finite public-history tree generated by the primitive

action and report sets. This number is finite because the horizon and all primitive sets are finite. Fix a seed-label set $\mathcal{C}$ with $|\mathcal{C}| \geq N_{\text{plan}}$. Every deterministic Markov assistant policy on any finite assessment-relative domain induces one of these contingent plans, so $\mathcal{C}$ is large enough to represent any finite lottery over such policies. The labels themselves carry no intrinsic meaning.

Definition F.9 (Public-randomization game)**.** *A* public-randomization assistant policy *is a pair $\pi^{\text{pub}} = (\nu, (\pi^c)_{c\in\mathcal{C}})$, where $\nu \in \Delta(\mathcal{C})$ and each π^c is a deterministic Markov assistant policy for every $c \in \mathcal{C}$. Before round 1, the assistant draws $C \sim \nu$ independently of η_u and publicly reveals its realization. Conditional on $C = c$, it follows π^c. The public state is therefore (A_t, q_t, C), and the user's behavioral Markov signalling rule is specified for every seed: $\sigma_t(\cdot \mid A, q, B, c, \eta) \in \Delta(B)$.*

Assistant deviations may change both ν and the family $(\pi^c)_{c\in\mathcal{C}}$, but use the same seed-label set $\mathcal{C}$.

Using a common master seed set is important when evaluating assistant optimality: it allows the user's strategy and beliefs to remain fixed while the assistant changes its entire public-randomization policy. Moreover, every horizon-reachable terminal menu is generated by some deterministic contingent plan. Hence, $|\mathcal{C}| \geq |\mathcal{A}_{\text{H}}|$, and we may fix an injection $\iota : \mathcal{A}_{\text{H}} \hookrightarrow \mathcal{C}$ for the menu-lottery construction used below.

The public-randomization game is a separate variant of the baseline model, not an additional property imposed on a baseline assessment. In particular, assistant optimality here is evaluated only against policies whose randomization is publicly resolved in this way. An assessment that is optimal against these public-randomization deviations need not remain optimal when the unrestricted behavioral deviations of the baseline game are restored.

Because the realized seed is part of the public information, beliefs and off-path updates must also condition on it. The next definition extends the baseline Bayes/off-path update rule to this augmented public state.

Definition F.10 (Belief update under public randomization)**.** *Fix an augmented Markov signalling rule $\sigma_t(\cdot \mid A, q, B, c, \eta) \in \Delta(B)$ and an augmented off-path selector $\mu_t^{\text{pub}}(\cdot \mid A, q, B, c, \mathbf{y}) \in \Delta(\mathcal{E})$. For (A, q, B, c) and a report $\mathbf{y} \in B$, define*

$$P_q^\sigma(\mathbf{y} \mid A, B, c) := \sum_{\eta\in\mathcal{E}} q(\eta)\sigma_t(\mathbf{y} \mid A, q, B, c, \eta).$$

If $P_q^\sigma(\mathbf{y} \mid A, B, c) > 0$, set

$$\tau_t^{\sigma,\mu^{\text{pub}},\text{pub}}(q, A, B, c, \mathbf{y})(\eta) := \frac{q(\eta)\sigma_t(\mathbf{y} \mid A, q, B, c, \eta)}{P_q^\sigma(\mathbf{y} \mid A, B, c)};$$

otherwise, set $\tau_t^{\sigma,\mu^{\text{pub}},\text{pub}}(q, A, B, c, \mathbf{y}) := \mu_t^{\text{pub}}(\cdot \mid A, q, B, c, \mathbf{y})$.

Thus, the belief update is the same Bayes/off-path rule as in the baseline game, except that it conditions additionally on the publicly observed seed. This observation lets us analyze the augmented game one seed at a time.

In the public-randomization variant, the revealed seed is part of the public state. We therefore augment the deviation-closed domain of Definition F.2 by carrying this seed through the interaction. Formally, $\mathcal{D}_t^{\text{pub}}(\sigma, \mu^{\text{pub}}) \subseteq 2^X \times \Delta(\mathcal{E}) \times \mathcal{C}$, with $\mathcal{D}_1^{\text{pub}} = \{(\varnothing, p, c) : c \in \mathcal{C}\}$. The corresponding information sets are $(A, q, B, c) \in \mathcal{I}_t^{\text{pub}}(\sigma, \mu^{\text{pub}})$, and after report $\mathbf{y}$ the successor state is $\big(B, \tau_t^{\sigma,\mu^{\text{pub}},\text{pub}}(q, A, B, c, \mathbf{y}), c\big)$. Thus, the seed never changes along a continuation.

For any augmented signalling rule σ and augmented off-path selector μ^{pub}, define the rules associated with seed $c \in \mathcal{C}$ by

$$\sigma_t^c(\cdot \mid A, q, B, \eta) := \sigma_t(\cdot \mid A, q, B, c, \eta), \qquad \mu_t^c(\cdot \mid A, q, B, \mathbf{y}) := \mu_t^{\text{pub}}(\cdot \mid A, q, B, c, \mathbf{y}).$$

These are the *seed-c sections* of the augmented signalling and belief rules. Comparing Definition F.10 with the baseline update shows that, for every $c \in \mathcal{C}$,

$$\tau_t^{\sigma^c,\mu^c}(q, A, B, \mathbf{y}) = \tau_t^{\sigma,\mu^{\text{pub}},\text{pub}}(q, A, B, c, \mathbf{y}).$$

Indeed, both sides use the same Bayes ratio after a positive-probability report and the same off-path selector otherwise. Consequently, the seed-c section of the augmented deviation domain is exactly the corresponding baseline deviation domain:

$$\big\{(A,q) : (A,q,c) \in \mathcal{D}_t^{\mathrm{pub}}(\sigma,\mu^{\mathrm{pub}})\big\} = \mathcal{D}_t(\sigma^c,\mu^c).$$

with c retained as a fixed public label. We write $\bar{\tau}^c$ for this common update rule.

Fixing a seed c therefore reduces the augmented problem to the baseline problem associated with (π^c,σ^c,μ^c). By the sectional identities above, its domain is $\mathcal{D}(\sigma^c,\mu^c)$ and its posterior update is $\tau^{\sigma^c,\mu^c} = \bar{\tau}^c$, with c retained as a fixed public label. Hence, the baseline Markov-deviation, outcome-determinacy, and self-selection arguments apply separately to each seed section whenever their respective hypotheses hold. The corresponding user state in the augmented game is (t,A,q,B,c,η).

Non-adaptive menu lotteries are still defined from the model primitives, independently of an incumbent assessment. A public non-adaptive lottery specifies a distribution over $\mathcal{C}$ and a deterministic control for each seed. When evaluated as an assistant deviation, it is lifted to the incumbent augmented domain $\mathcal{D}^{\mathrm{pub}}(\sigma,\mu^{\mathrm{pub}})$, just as in Definition F.3. This does not assert that the same augmented domain would remain valid if the user's signalling rule were changed.

We can now define the equilibrium analogue of Definition F.5 for the public-randomization variant. The four requirements are the same in substance: welfare feasibility, assistant optimality, user optimality, and belief consistency. The difference is that policies, reports, and beliefs condition on the publicly observed seed, and assistant optimality is evaluated within the public-randomization policy class.

Definition F.11 (Welfare-feasible public strategic assessment)**.** *Fix $\epsilon \geq 0$ and the master seed space $\mathcal{C}$. A* welfare-feasible public strategic assessment *is a triple $(\pi^{\mathrm{pub}},\sigma,\mu^{\mathrm{pub}})$, where $\pi^{\mathrm{pub}} = (\nu,(\pi^c)_{c\in\mathcal{C}})$, σ is an augmented behavioral Markov signalling rule $\sigma_t(\cdot \mid A,q,B,c,\eta) \in \Delta(B)$, and μ^{pub} is an augmented off-path selector $\mu_t^{\mathrm{pub}}(\cdot \mid A,q,B,c,\mathbf{y}) \in \Delta(\mathcal{E})$, all specified on $\mathcal{D}^{\mathrm{pub}}(\sigma,\mu^{\mathrm{pub}})$, such that:*

(i) (aggregate ex-ante welfare feasibility) $\mathbb{E}^{(\pi^{\mathrm{pub}},\sigma)}[G_u(A_{T+1},\eta_u)] \leq \epsilon$*;*

(ii) (complete public-policy assistant optimality) *holding $(\sigma,\mu^{\mathrm{pub}})$ fixed, π^{pub} maximizes $\mathbb{E}^{(\widetilde{\pi},\sigma)}[\tilde{U}_a(\mathbf{x}_T^\star)]$ over every public assistant policy $\widetilde{\pi} = (\widetilde{\nu},(\widetilde{\pi}^c)_{c\in\mathcal{C}})$ on the same master seed set $\mathcal{C}$, with each $\widetilde{\pi}^c$ deterministic Markov on the c-section of the incumbent augmented domain, subject to $\mathbb{E}^{(\widetilde{\pi},\sigma)}[G_u(A_{T+1},\eta_u)] \leq \epsilon$;*

(iii) (sequential augmented-Markov user optimality) *for every seed $c \in \mathcal{C}$, round t, type $\eta \in \mathcal{E}$, and $(A,q,B,c) \in \mathcal{I}_t^{\mathrm{pub}}(\sigma,\mu^{\mathrm{pub}})$,*

$$\operatorname{supp}\sigma_t^c(\cdot \mid A,q,B,\eta) \subseteq \arg\max_{\mathbf{y}\in B} U_{t+1}^{\pi^c,\sigma^c}\big(B,\tau_t^{\sigma^c,\mu^c}(q,A,B,\mathbf{y}),\eta\big),$$

where U^{π^c,σ^c} is the prescribed-strategy continuation value of Definition F.4;

(iv) (belief consistency) *beliefs evolve according to Definition F.10.*

Let $\mathrm{OPT}^{\min}_{\mathrm{strat,pub}}(\epsilon)$ and $\mathrm{OPT}^{\max}_{\mathrm{strat,pub}}(\epsilon)$ denote, respectively, the infimum and supremum of the assistant's expected terminal payoff over welfare-feasible public strategic assessments. Condition **(iii)** applies at every seed label, including labels assigned zero probability by the incumbent seed distribution: an assistant deviation may assign positive probability to such a label, so the user's strategy must already be sequentially optimal there. By contrast, welfare feasibility in **(i)** is imposed only in expectation over the seed distribution, not separately for each seed.

The sectional representation also transfers the global user-optimality result of Lemma F.5 to the public-randomization game. Since the seed is observed before reporting and remains fixed thereafter, a Markov reporting deviation may depend on the seed but can be evaluated separately within each seed section.

Corollary F.4 (Optimality against seed-dependent Markov deviations)**.** *Suppose that, for a type η, the prescribed reporting rule is supported on continuation-optimal reports at every augmented information set $(A,q,B,c) \in \mathcal{I}_t^{\mathrm{pub}}(\sigma,\mu^{\mathrm{pub}})$ and for every $c \in \mathcal{C}$, including seed labels of zero*

incumbent probability. Then, no behavioral Markov reporting strategy in the augmented game, including one that conditions on the revealed seed, gives type η higher expected utility from any augmented state, and hence none does so ex ante.

Proof. Because the seed is drawn once before round 1, publicly observed, and fixed thereafter, the augmented game decomposes into $|\mathcal{C}|$ finite-horizon Markov sections, one per seed label. Its deviation-closed domain $\mathcal{D}^{\mathrm{pub}}(\sigma, \mu^{\mathrm{pub}})$ is correspondingly the disjoint union of the seed-sectional domains. For each $c \in \mathcal{C}$, the identities following Definition F.10 identify the c-section with the baseline problem (π^c, σ^c, μ^c) on $\mathcal{D}(\sigma^c, \mu^c)$. Thus, Lemma F.5 applies within every seed section and shows that no alternative behavioral Markov reporting rule on that c-section improves type η's continuation utility. An augmented behavioral Markov reporting strategy is precisely a collection of such seed-dependent Markov rules, one for each $c \in \mathcal{C}$. The conclusion therefore holds pointwise at every augmented state. Averaging over the incumbent seed distribution ν gives the ex-ante conclusion. □

The preceding sectional reduction gives the key implication of public randomization. Conditional on each revealed seed, the assistant behaves deterministically, so the outcome-determinacy and self-selection arguments apply within that seed section. Averaging the resulting image menus over the public seed then produces a non-adaptive menu lottery with the same value and welfare loss. Conversely, the public policy class contains every non-adaptive menu lottery. These two observations imply that the entire public-randomization equilibrium class collapses to the non-adaptive value.

Proposition F.4 (Equilibrium collapse under public randomization)**.** *Assume (A1) and (A2), and fix $\epsilon \geq 0$ for which a welfare-feasible non-adaptive control exists. The public-randomization game admits an equilibrium, and every equilibrium has assistant value exactly $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. Consequently,*

$$\mathrm{OPT}^{\min}_{\mathrm{strat,pub}}(\epsilon) = \mathrm{OPT}^{\max}_{\mathrm{strat,pub}}(\epsilon) = \mathrm{OPT}_{\mathrm{nf}}(\epsilon) \leq \mathrm{OPT}(\epsilon). \tag{3}$$

Proof. We first show that the public policy class contains an optimal policy attaining the non-adaptive value. Let $D^\star$ attain $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ in Definition 5.2. Its support consists of nonempty horizon-reachable menus and therefore lies in $\mathcal{A}_{\mathrm{H}}$. Using the injection $\iota : \mathcal{A}_{\mathrm{H}} \hookrightarrow \mathcal{C}$ introduced above immediately after Definition F.9, set $\nu^\star(\iota(A)) := D^\star(A)$ for $A \in \mathrm{supp}(D^\star)$ and $\nu^\star(c) := 0$ on the unused labels. For each $A \in \mathrm{supp}(D^\star)$, let $\pi^{\iota(A)}$ follow a fixed deterministic feasible slate sequence reaching A; under Assumption (A2), A can be revealed in the first round. On unused labels, let π^c follow any fixed deterministic non-adaptive feasible control. This defines $\pi^{\mathrm{nf}}_{\mathrm{pub}} = (\nu^\star, (\pi^c)_{c\in\mathcal{C}})$ on the same master seed space, with terminal-menu law $D^\star$. Since its slates do not depend on the posterior or the user's reports, Lemma F.1 and Proposition F.1 imply that its value is $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ and its expected regret is at most ϵ.

We next use this policy to establish nonemptiness of the equilibrium class. For every seed $c \in \mathcal{C}$, every augmented information set (A, q, B, c), and every type η, let $\sigma^\circ_t(\cdot \mid A, q, B, c, \eta)$ be any fixed reporting distribution independent of η, and set $\mu^{\mathrm{pub},\circ}_t(\cdot \mid A, q, B, c, \mathbf{y}) := q$ at every $(A, q, B, c, \mathbf{y})$ in the augmented domain. The corresponding sectional selector is therefore $\mu^{\circ,c}_t(\cdot \mid A, q, B, \mathbf{y}) = q$.

Because σ° is type-independent, every positive-probability report leaves the posterior unchanged by Bayes' rule, while the chosen off-path selector does the same after zero-probability reports. Thus, every seed section carries the prior p throughout. Conditional on any seed, $\pi^{\mathrm{nf}}_{\mathrm{pub}}$ follows a deterministic non-adaptive control, so future slates depend on neither beliefs nor the user's reports. Every report therefore gives the user the same continuation utility at every augmented information set, including zero-probability ones, and Definition F.11**(iii)** holds.

Against this fixed type-independent signalling rule, any complete public assistant deviation induces a terminal-menu law independent of the user's type. By Lemma F.2, that law can be implemented by a non-adaptive menu lottery. Hence, no welfare-feasible public deviation can attain value above $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. It follows that $(\pi^{\mathrm{nf}}_{\mathrm{pub}}, \sigma^\circ, \mu^{\mathrm{pub},\circ})$ is a welfare-feasible public strategic assessment with value $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$.

Now, we proceed to proving the upper bound in (3). Let $(\pi^{\mathrm{pub}}, \sigma, \mu^{\mathrm{pub}})$ be any welfare-feasible public strategic assessment, and write its assistant value as V. Conditional on $C = c$, the corresponding section is the baseline-style assessment (π^c, σ^c, μ^c) on $\mathcal{D}(\sigma^c, \mu^c)$, with posterior c-sectional update

$\bar{\tau}_t^c(q, A, B, \mathbf{y}) := \tau_t^{\sigma,\mu^{\mathrm{pub}},\mathrm{pub}}(q, A, B, c, \mathbf{y})$. Since π^c is deterministic and Definition F.11**(iii)** supplies sequential Markov user optimality in that section, Lemma F.6 and Proposition F.3 apply. Thus, each seed c induces a self-selecting assignment ϕ_c with image menu $\hat{A}_c$, which is horizon-reachable under Assumption (A2).

Let $V_c := \sum_\eta p(\eta)\tilde{U}_a(\phi_c(\eta))$ and $W_c := \sum_\eta p(\eta) r_\eta(\phi_c(\eta))$. Then, $V = \sum_c \nu(c) V_c$ and $\sum_c \nu(c) W_c \le \epsilon$. The distribution of $\hat{A}_C$ is supported on horizon-reachable menus and is independent of η_u. By Lemma F.2, it can therefore be implemented by a non-adaptive control. Self-selection ensures that this control gives type η the item $\phi_c(\eta)$ when menu $\hat{A}_c$ is drawn, so it reproduces both V and the assessment's expected regret. Consequently, $V \le \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$.

For the reverse inequality, recall that public assistant deviations may change both the seed distribution and the seed-conditional deterministic policies while keeping the master seed set $\mathcal{C}$ fixed. Hence, $\pi^{\mathrm{nf}}_{\mathrm{pub}}$ constructed above is an admissible complete deviation under Definition F.11**(ii)** and is welfare-feasible. Assistant optimality therefore gives $V \ge \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. Combining the two inequalities yields $V = \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ for every welfare-feasible public strategic assessment. The final inequality $\mathrm{OPT}_{\mathrm{nf}}(\epsilon) \le \mathrm{OPT}(\epsilon)$ follows from Theorem 5.2**(i)**.

Notice that the lower bound does not invoke Proposition F.2, whose assistant-deviation class is the unrestricted baseline class. Here, the deviation class is restricted to public-randomization policies; the required lower bound follows instead because this restricted class still contains a policy attaining $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. □

F.12 Why Unresolved Randomization Breaks the Collapse

Appendix F.11 showed that the collapse survives when the assistant's randomization is realized and observed *before* the user reports. We now explain why the timing of that randomization matters. If the assistant can randomize its continuation only *after* observing the user's report, the user chooses its report without knowing which continuation will be realized. Equilibrium incentives then compare expected continuation utilities rather than realized terminal outcomes, and the pathwise argument behind Lemma F.6 no longer applies.

To see the issue, recall the imitation argument from Proposition F.3. A type-η user can imitate the reporting strategy prescribed for another type η', and equilibrium optimality requires that this deviation not increase type η's expected utility. When continuation outcomes are randomized, let Z_η denote type η's equilibrium terminal item and $Z_{\eta'}$ the terminal item generated by imitating type η'. The argument then gives only $\mathbb{E}[U_\eta(Z_\eta)] \ge \mathbb{E}[U_\eta(Z_{\eta'})]$. This comparison is in expectation and does not imply that type η prefers its own realized terminal item on every path. For example, $U_\eta(Z_\eta)$ may equal 0 or 1 with equal probability while $U_\eta(Z_{\eta'}) = \frac{1}{2}$ with probability one. Thus, equilibrium optimality no longer yields the pathwise comparison needed for self-selection.

The economic consequence is a budget-allocation effect. The ex-ante welfare constraint provides a single welfare budget, but the regret cost of an assistant-preferred outcome may differ across types. Once feedback separates the types, the assistant can use a type-contingent continuation lottery to spend more of this budget on the type for which the preferred outcome has the lowest regret cost, while compensating that type through the lottery's other realization. A type-independent non-adaptive menu lottery cannot allocate the budget in this way: because its menu distribution cannot depend on the user's unknown type, it obtains the assistant-preferred outcome at its population-average regret cost.

Outcome determinacy rules out this channel by assigning each type a fixed terminal item. Likewise, when the assistant's randomization is realized and publicly observed before the relevant reports, the pathwise logic underlying the collapse is restored. Unresolved continuation randomization does not provide this restriction. The counterexamples in Propositions F.5 and F.6 show that the distinction matters: unresolved randomization can raise equilibrium value above $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ and can even allow strategic feedback to outperform truthful feedback.

The following construction makes this failure concrete. It separates the two types with an initial report and then uses unresolved continuation randomization to allocate the welfare budget asymmetrically across them. The resulting equilibrium attains the truthful-feedback optimum but strictly exceeds the best non-adaptive value.

Proposition F.5 (Unresolved randomization can break the collapse)**.** *Let $X = \{g, b_1, b_2, p_1, p_2\}$ be a single-feature item space, $\mathcal{E} = \{\eta_1, \eta_2\}$, $p = (\frac{1}{2}, \frac{1}{2})$, $T = 2$, $k \geq 3$, $\epsilon = \frac{1}{5}$, and*

	g	b_1	b_2	p_1	p_2
U_{η_1}	$\frac{1}{5}$	1	$\frac{1}{2}$	$\frac{3}{20}$	$\frac{1}{20}$
U_{η_2}	0	$\frac{3}{10}$	1	$\frac{1}{20}$	$\frac{3}{20}$
$\tilde{U}_a$	1	0	0	0	0

Both types have strict preferences. Then, $\mathrm{OPT}_{\mathrm{nf}}(\frac{1}{5}) = \frac{2}{9}$ and $\mathrm{OPT}(\frac{1}{5}) = \frac{1}{4}$. Moreover, there exists a welfare-feasible strategic assessment in the sense of Definition 5.1 with assistant value $\frac{1}{4} > \frac{2}{9} = \mathrm{OPT}_{\mathrm{nf}}(\frac{1}{5})$. Its assistant policy is truthful-optimal and deviation-complete dynamically incentive compatible in the sense of Definition F.8.

Proof. The relevant regrets are $r_{\eta_1}(g) = \frac{4}{5}$, $r_{\eta_1}(b_1) = 0$, $r_{\eta_2}(g) = 1$, and $r_{\eta_2}(b_2) = 0$.

Universal upper bound. We first show that every welfare-feasible policy has value at most $\frac{1}{4}$. Let $a_i := \mathbb{P}[\mathbf{x}^\star_T = g \mid \eta_u = \eta_i]$ under any policy and any user behavior. Since g is the only item with positive assistant utility, the assistant's value is $\frac{a_1+a_2}{2}$. Nonnegativity of regret and welfare feasibility imply $\frac{1}{5} \geq \mathbb{E}[G_u] \geq \frac{1}{2}(\frac{4}{5}a_1) + \frac{1}{2}a_2$, or equivalently $\frac{4}{5}a_1 + a_2 \leq \frac{2}{5}$. Maximizing $\frac{a_1+a_2}{2}$ subject to this inequality and $a_i \in [0, 1]$ allocates the entire budget to the cheaper coordinate a_1, whose regret cost is $\frac{4}{5} < 1$. Thus, $a_1 = \frac{1}{2}$, $a_2 = 0$, and the maximum value is $\frac{1}{4}$.

This bound applies both to $\mathrm{OPT}(\frac{1}{5})$ and to every welfare-feasible strategic assessment, each evaluated on its own domain. It uses only the definition of a_i and nonnegativity of regret.

The non-adaptive value. We next show that $\mathrm{OPT}_{\mathrm{nf}}(\frac{1}{5}) = \frac{2}{9}$. For a menu A, define $P(A) := \sum_i \frac{1}{2}\tilde{U}_a(\mathbf{x}^\star(A, \eta_i))$ and $R(A) := \sum_i \frac{1}{2} r_{\eta_i}(\mathbf{x}^\star(A, \eta_i))$. Type η_1 ranks the items as $b_1 \succ b_2 \succ g \succ p_1 \succ p_2$, while type η_2 ranks them as $b_2 \succ b_1 \succ p_2 \succ p_1 \succ g$. Hence, η_1 selects g exactly when $g \in A$ and $A \cap \{b_1, b_2\} = \varnothing$, whereas η_2 selects g exactly when $A = \{g\}$, since g is its strict minimum.

The menus with $P(A) > 0$ are therefore

$$A = \{g\} : \ (P, R) = \big(1, \tfrac{9}{10}\big); \quad A \in \{\{g, p_2\}, \{g, p_1, p_2\}\} : \big(\tfrac{1}{2}, \tfrac{33}{40}\big); \quad A = \{g, p_1\} : \ \big(\tfrac{1}{2}, \tfrac{7}{8}\big),$$

and $P(A) = 0$ for every other menu. In each of these four cases, $P(A) \leq \frac{10}{9}R(A)$, with equality only for $A = \{g\}$; when $P(A) = 0$, the inequality is immediate. Averaging over any menu lottery, every welfare-feasible non-adaptive control therefore has value at most $\frac{10}{9} \cdot \frac{1}{5} = \frac{2}{9}$.

This bound is attained by revealing $\{g\}$ with probability $\frac{2}{9}$ and $\{b_1, b_2\}$ otherwise. Under the latter menu, η_1 selects b_1 and η_2 selects b_2, so $P = R = 0$. The resulting expected regret is $\frac{2}{9} \cdot \frac{9}{10} = \frac{1}{5}$, and the assistant value is $\frac{2}{9}$. Hence, $\mathrm{OPT}_{\mathrm{nf}}(\frac{1}{5}) = \frac{2}{9}$.

Construction of the strategic assessment. Let σ be truthful at every user information set: each type reports a myopic utility maximizer from the revealed menu. Let π the behavioral Markov policy that reveals $S_1 = \{p_1, p_2\}$ in round 1. After report p_1, which induces posterior δ_{η_1}, it reveals $\{g\}$ with probability $\frac{1}{2}$ and $\{b_1\}$ with probability $\frac{1}{2}$. After report p_2, which induces posterior δ_{η_2}, it reveals $\{b_2\}$ deterministically. At every other state, it reveals nothing.

The realization of the round-2 random slate is not known when the user chooses its round-1 report. Thus, by Definition F.7, this continuation randomization is unresolved, and the user evaluates the continuation lottery in expectation. Under σ, type η_1 reports p_1, since $\frac{3}{20} > \frac{1}{20}$, and type η_2 reports p_2. Both reports are therefore on path, so no off-path beliefs are needed in round 1.

For type η_1, the terminal menu is $\{p_1, p_2, g\}$ with probability $\frac{1}{2}$ and $\{p_1, p_2, b_1\}$ with probability $\frac{1}{2}$, yielding terminal items g and b_1, respectively. For type η_2, the terminal menu is $\{p_1, p_2, b_2\}$ and the terminal item is b_2. Hence,

$$V = \tfrac{1}{2}\big(\tfrac{1}{2} \cdot 1 + \tfrac{1}{2} \cdot 0\big) + \tfrac{1}{2} \cdot 0 = \tfrac{1}{4}, \qquad W = \tfrac{1}{2}\big(\tfrac{1}{2} \cdot \tfrac{4}{5} + \tfrac{1}{2} \cdot 0\big) + \tfrac{1}{2} \cdot 0 = \tfrac{1}{5} = \epsilon.$$

Verification of the strategic-assessment conditions in Definition F.5. Condition **(i)** of Definition F.5 holds because $W = \epsilon$.

For condition **(ii)**, the universal bound above shows that no welfare-feasible policy on this domain can attain value above $\frac{1}{4}$ against σ, while π attains $\frac{1}{4}$. Thus, π is an ex-ante constrained best response to σ, holding (σ, μ) fixed as required by Definition 5.1.

For condition **(iii)**, the only user information set at which the report affects continuation play is the round-1 menu $\{p_1, p_2\}$. Type η_1 obtains $\frac{1}{2} \cdot \frac{1}{5} + \frac{1}{2} \cdot 1 = \frac{3}{5}$ from reporting p_1 and $\max\{\frac{3}{20}, \frac{1}{20}, \frac{1}{2}\} = \frac{1}{2}$ from reporting p_2, so p_1 is strictly optimal. Type η_2 obtains 1 from reporting p_2 and $\frac{1}{2} \cdot \frac{3}{20} + \frac{1}{2} \cdot \frac{3}{10} = \frac{9}{40}$ from reporting p_1, so p_2 is strictly optimal. Round-2 reports do not affect the terminal choice, so every report is optimal there.

Finally, condition **(iv)** holds because both round-1 reports occur with positive probability and Bayes' rule yields the degenerate posteriors used above.

Truthful optimality and deviation-complete DIC. Because users follow truthful reporting under σ, the policy π is admissible for the constrained Bayes value defined in Section 2 and attains value $\frac{1}{4}$. Together with the universal upper bound, this gives $\mathrm{OPT}(\frac{1}{5}) = \frac{1}{4}$, so π is truthful-optimal.

It remains to verify that π is deviation-complete dynamically incentive compatible in the sense of Definition F.8, rather than merely dynamically incentive compatible in the sense of Definition D.1. There are three kinds of information sets in $\mathcal{I}_t(\sigma^{\mathrm{tr}}, \mu)$:

1. At the probe $(\varnothing, p, \{p_1, p_2\})$, the truthful myopic reports are strict continuation best responses by the computation above, with margins $\frac{1}{10}$ for η_1 and $\frac{31}{40}$ for η_2. Both reports are on path, so μ is not invoked.

2. At every other round-1 information set $(\varnothing, p, B)$, where the accumulated menu B after round 1 satisfies $B \neq \{p_1, p_2\}$, the policy π reveals nothing in round 2. Hence, $U_2^{\pi, \bar{\tau}^{\mathrm{tr}, \mu}}(B, q', \eta) = \max_{\mathbf{x} \in B} U_\eta(\mathbf{x})$ for every posterior q'. The continuation value is therefore independent of the report, so every report, and in particular every myopic maximizer, is continuation-optimal regardless of the posterior assigned by the off-path selector.

3. At every round-2 information set, there is no further slate. Thus, $U_{T+1}^{\pi, \bar{\tau}^{\mathrm{tr}, \mu}}(B, \cdot, \eta) = \max_{\mathbf{x} \in B} U_\eta(\mathbf{x})$ is again independent of the report. In the latter two cases, μ never affects the relevant continuation comparison, so the conclusion is independent of the off-path selector.

We have therefore constructed a welfare-feasible strategic assessment with value $\frac{1}{4} > \frac{2}{9} = \mathrm{OPT}_{\mathrm{nf}}(\frac{1}{5})$, proving that unresolved assistant randomization can break the collapse. □

The counterexample has several consequences for the scope of the collapse result. First, equilibrium value is no longer unique. By Theorem 5.1, $\mathrm{OPT}^{\min}_{\mathrm{strat}}(\frac{1}{5}) = \mathrm{OPT}_{\mathrm{nf}}(\frac{1}{5}) = \frac{2}{9}$, whereas the construction above, together with its universal upper bound, gives $\mathrm{OPT}^{\max}_{\mathrm{strat}}(\frac{1}{5}) = \frac{1}{4}$. Thus, without Assumption (A3), equilibrium values need not collapse to $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. The example also shows why the characterization in Theorem 5.4 is restricted to the public-randomization variant. In the baseline game, the implication **(iii)⇒(ii)** fails: the policy constructed above is truthful-optimal and deviation-complete dynamically incentive compatible, yet $\mathrm{OPT}(\frac{1}{5}) = \frac{1}{4} \neq \frac{2}{9} = \mathrm{OPT}_{\mathrm{nf}}(\frac{1}{5})$. Finally, the example is consistent with the existence restriction in Corollary F.3. Here, $\mathrm{OPT}^{\mathrm{det}}_{\mathrm{nf}}(\frac{1}{5}) = 0 < \frac{2}{9} = \mathrm{OPT}_{\mathrm{nf}}(\frac{1}{5})$, so no outcome-determinate welfare-feasible equilibrium exists.

The preceding example shows that unresolved continuation randomization can raise equilibrium value above the non-adaptive benchmark. The next example shows that the effect can be stronger: strategic feedback can give the assistant strictly higher value than truthful feedback. The construction again uses feedback to separate the types, after which the assistant can allocate the ex-ante welfare budget differently across the resulting continuations.

Proposition F.6 (Strategic feedback can strictly outperform truthful feedback)**.** *Let* $X = \{A, B, M, Z\}$ *be a single-feature item space,* $\mathcal{E} = \{\eta_1, \eta_2\}$*,* $p = (\frac{1}{2}, \frac{1}{2})$*,* $T = 2$*,* $k = |X| = 4$*,*

$\epsilon = \frac{11}{80}$, *and*

	A	B	M	Z
U_{η_1}	$\frac{3}{5}$	1	$\frac{1}{2}$	0
U_{η_2}	1	$\frac{1}{2}$	$\frac{9}{20}$	0
$\tilde{U}_a$	0	$\frac{3}{5}$	1	0

Both types have strict preferences. Then, $\mathrm{OPT}(\frac{11}{80}) = \mathrm{OPT}_{\mathrm{nf}}(\frac{11}{80}) = \frac{29}{60}$*, while there exists a welfare-feasible strategic assessment with assistant value* $\frac{11}{20} > \frac{29}{60}$*. Thus, without Assumption (A3), strategic feedback can give the assistant strictly higher value than truthful feedback.*

Proof. The regrets are $r_{\eta_1} = (A{:}\frac{2}{5}, B{:}0, M{:}\frac{1}{2}, Z{:}1)$ and $r_{\eta_2} = (A{:}0, B{:}\frac{1}{2}, M{:}\frac{11}{20}, Z{:}1)$. Type η_1 ranks the items as $B \succ A \succ M \succ Z$, while type η_2 ranks them as $A \succ B \succ M \succ Z$. Thus, their rankings differ only in the relative order of A and B. Under Assumption (A2), every menu is reachable in one round.

The non-adaptive value. We first show that the non-adaptive value is $\mathrm{OPT}_{\mathrm{nf}}(\frac{11}{80}) = \frac{29}{60}$. Let (R, P) denote the prior-averaged regret and assistant payoff induced by a menu. The distinct menu outcomes are

$$\{A, B\}{:}\,\big(0, \tfrac{3}{10}\big), \quad \{A\}{:}\,\big(\tfrac{1}{5}, 0\big), \quad \{B\}{:}\,\big(\tfrac{1}{4}, \tfrac{3}{5}\big), \quad \{M\}{:}\,\big(\tfrac{21}{40}, 1\big), \quad \{Z\}{:}\,(1, 0),$$

with every other menu duplicating one of these outcomes. The upper concave envelope through $(0, \frac{3}{10})$ has maximal initial slope $(1 - \frac{3}{10})/(\frac{21}{40}) = \frac{4}{3}$, attained by the chord to $(\frac{21}{40}, 1)$. Hence, $\mathrm{OPT}_{\mathrm{nf}}(\rho) = \frac{3}{10} + \frac{4}{3}\rho$ for $\rho \le \frac{21}{40}$. At $\rho = \frac{11}{80}$, this gives $\mathrm{OPT}_{\mathrm{nf}}(\frac{11}{80}) = \frac{3}{10} + \frac{11}{60} = \frac{29}{60}$.

The truthful value. We next show that truthful feedback cannot improve on this value, i.e., the truthful value is also $\frac{29}{60}$. Under deterministic truthful feedback, the two types choose different items from a menu S if and only if $\{A, B\} \subseteq S$. If the round-1 slate S_1 does not contain both A and B, the round-1 report is therefore uninformative and the posterior remains p. Since $T = 2$, the round-2 slate then depends only on a type-independent public history. The resulting terminal-menu law is type-independent and is therefore outcome-equivalent to a non-adaptive control, so its value is at most $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$.

If instead $\{A, B\} \subseteq S_1$, both items remain in the terminal menu regardless of the round-2 action. Type η_1 then terminates at B and type η_2 at A, their respective global optima, yielding regret zero and assistant value $\frac{3}{10}$. This is exactly the non-adaptive outcome associated with the menu $\{A, B\}$. Randomization over the round-1 slate only mixes these two cases. Hence, the achievable payoff-regret set under truthful feedback is contained in the non-adaptive one, giving $\mathrm{OPT}(\frac{11}{80}) \le \mathrm{OPT}_{\mathrm{nf}}(\frac{11}{80})$. The reverse inequality follows from Theorem 5.2(i), so $\mathrm{OPT}(\frac{11}{80}) = \mathrm{OPT}_{\mathrm{nf}}(\frac{11}{80}) = \frac{29}{60}$.

Construction of the strategic assessment. Define the Markov signalling rule σ as follows. At the round-1 menu $\{M, Z\}$, type η_1 reports M and type η_2 reports Z. At every other round-1 menu, each type reports its myopic utility maximizer, and round-2 reports are arbitrary. Thus, σ is not truthful: at the probe $\{M, Z\}$, type η_2 reports the dominated item Z. At every non-probe menu, however, the assistant subsequently reveals nothing, so continuation utility is independent of the report and any prescribed report is sequentially optimal.

Define the behavioral Markov policy π that reveals $S_1 = \{M, Z\}$ in round 1. After report M, which induces posterior δ_{η_1}, it reveals $\{B\}$. After report Z, which induces posterior δ_{η_2}, the behavioral Markov policy π reveals nothing with probability $\frac{1}{2}$ and $\{A\}$ with probability $\frac{1}{2}$. At every state whose accumulated menu differs from $\{M, Z\}$, it reveals nothing. Both round-1 reports are on path, so no off-path beliefs are needed there. Since σ is Markov, Lemma F.3 supplies the hypothesis of Lemma F.4; the reachable posteriors are p, δ_{η_1}, and δ_{η_2}.

Under this assessment, type η_1 faces terminal menu $\{B, M, Z\}$ and selects B, yielding assistant payoff $\frac{3}{5}$ and regret zero. Type η_2 faces $\{M, Z\}$ with probability $\frac{1}{2}$, selects M, and yields assistant payoff 1 and regret $\frac{11}{20}$; with probability $\frac{1}{2}$, it faces $\{A, M, Z\}$ and selects A, yielding assistant payoff and regret both equal to zero. Therefore,

$$V = \tfrac{1}{2} \cdot \tfrac{3}{5} + \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot 1 = \tfrac{11}{20}, \qquad W = \tfrac{1}{2} \cdot 0 + \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot \tfrac{11}{20} = \tfrac{11}{80} = \epsilon.$$

User sequential optimality. We verify condition **(iii)** of Definition 5.1. At the probe menu $\{M, Z\}$, type η_1 obtains $\max\{1, \frac{1}{2}, 0\} = 1$ from reporting M and $\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{3}{5} = \frac{11}{20}$ from reporting Z, so M is strictly optimal. Type η_2 obtains $\max\{\frac{1}{2}, \frac{9}{20}, 0\} = \frac{1}{2}$ from reporting M and $\frac{1}{2} \cdot \frac{9}{20} + \frac{1}{2} \cdot 1 = \frac{29}{40}$ from reporting Z, so Z is strictly optimal. At every other user information set, π reveals nothing thereafter, making continuation utility independent of the report. Every report is therefore optimal there, and round-2 reports never affect the terminal choice.

Assistant ex-ante constrained optimality. It remains to verify condition **(ii)** of Definition 5.1. We show that, against the fixed signalling rule σ, every behavioral Markov assistant policy with expected regret at most $\epsilon = \frac{11}{80}$ has value at most $\frac{11}{20}$.

It suffices to condition first on the round-1 slate. Behavioral randomization over different round-1 slates produces convex combinations of the corresponding regret-value pairs, while the upper bound derived below is concave. There are three cases.

Case (a): $S_1 \neq \{M, Z\}$ and $\{A, B\} \not\subseteq S_1$. At such a round-1 menu, both types send the same report. Their preference rankings differ only in the relative order of A and B, so any menu that does not contain both has the same unique maximizer for both types. The round-1 report is therefore type-independent. Consequently, the public history after round 1, and hence the terminal-menu law under any continuation policy, is independent of the latent type. As in the babbling argument in Theorem 5.1, this terminal-menu law can be reproduced by a non-adaptive control. Every regret-value pair in this case therefore lies below the non-adaptive upper boundary

$$g(\rho) = \frac{3}{10} + \frac{4}{3}\rho, \qquad 0 \le \rho \le \frac{21}{40}.$$

Case (b): $\{A, B\} \subseteq S_1$. Both A and B remain in the terminal menu regardless of the assistant's round-2 action. Since B is the global optimum of η_1 and A is the global optimum of η_2, the terminal items are respectively B and A under every continuation, regardless of the reports. Expected regret is therefore zero, while assistant value is $\frac{1}{2}\tilde{U}_a(B) + \frac{1}{2}\tilde{U}_a(A) = \frac{1}{2} \cdot \frac{3}{5} + \frac{1}{2} \cdot 0 = \frac{3}{10}$. Thus, this case contributes only the point $(\rho, V) = (0, \frac{3}{10})$.

Case (c): $S_1 = \{M, Z\}$. Here, the prescribed signalling rule separates the two types. Conditional on the round-1 report, the only unrevealed payoff-relevant items are A and B. At the final assistant move, arbitrary behavioral randomization therefore produces convex combinations of the deterministic continuation points:

S_2	η_1 branch $(r, \tilde{U}_a)$	η_2 branch $(r, \tilde{U}_a)$
$\varnothing$	$(\frac{1}{2}, 1)$	$(\frac{11}{20}, 1)$
$\{A\}$	$(\frac{2}{5}, 0)$	$(0, 0)$
$\{B\}$	$(0, \frac{3}{5})$	$(\frac{1}{2}, \frac{3}{5})$
$\{A, B\}$	$(0, \frac{3}{5})$	$(0, 0)$

The upper concave envelopes are $v_1(\rho_1) = \frac{3}{5} + \frac{4}{5}\rho_1$ on $[0, \frac{1}{2}]$ for the η_1 branch and $v_2(\rho_2) = \frac{20}{11}\rho_2$ on $[0, \frac{11}{20}]$ for the η_2 branch. Writing $\rho = \frac{1}{2}(\rho_1 + \rho_2)$, the best value in this case is bounded by $\frac{3}{10} + \frac{2}{5}\rho_1 + \frac{10}{11}\rho_2$. Since $\frac{10}{11} > \frac{2}{5}$, regret is allocated first to the η_2 branch. The resulting upper boundary is

$$f(\rho) = \frac{3}{10} + \frac{20}{11}\rho \quad \text{for } 0 \le \rho \le \frac{11}{40}, \qquad f(\rho) = \frac{4}{5} + \frac{4}{5}\Big(\rho - \frac{11}{40}\Big) \quad \text{for } \frac{11}{40} \le \rho \le \frac{21}{40},$$

and $f(\rho) = 1$ for $\rho \ge \frac{21}{40}$. The two pieces agree at $\rho = \frac{11}{40}$, where both equal $\frac{4}{5}$, and f is concave.

We now compare the three cases. On $[0, \frac{21}{40}]$, $f \ge g$: the two functions agree at $\rho = 0$ and $\rho = \frac{21}{40}$, with values $\frac{3}{10}$ and 1, respectively, while f has the larger initial slope, $\frac{20}{11} > \frac{4}{3}$. Since both are concave and piecewise linear, f dominates throughout. Case (b) contributes exactly $(0, \frac{3}{10}) = (0, f(0))$. For regret satisfying $\rho > \frac{21}{40}$, the trivial bound $\tilde{U}_a \le 1 = f(\rho)$ applies. Thus, conditional on any round-1 slate, every attainable regret-value pair lies on or below the graph of the concave

function f. Behavioral randomization over round-1 slates takes convex combinations of these pairs, so concavity of f gives the same bound for every behavioral Markov policy.

At $\epsilon = \frac{11}{80} < \frac{11}{40}$, this gives

$$V \le f\Big(\frac{11}{80}\Big) = \frac{3}{10} + \frac{20}{11} \cdot \frac{11}{80} = \frac{11}{20}.$$

The proposed policy attains equality: its η_1 branch has zero regret, while its η_2 branch mixes $\varnothing$ and $\{A\}$ with equal probability, giving $\rho_1 = 0$ and $\rho_2 = \frac{11}{40}$. Hence, π is an ex-ante welfare-constrained best response to the fixed signalling rule σ, establishing Definition 5.1**(ii)**.

Finally, welfare feasibility follows from $W = \epsilon$, user sequential optimality was verified above, and the on-path round-1 posteriors follow Bayes' rule; beliefs at all remaining information sets can be specified consistently as required. Thus, conditions **(i)**-**(iv)** of Definition 5.1 hold. The resulting strategic assessment has value $\frac{11}{20}$, whereas $\mathrm{OPT}(\frac{11}{80}) = \mathrm{OPT}_{\mathrm{nf}}(\frac{11}{80}) = \frac{29}{60}$. Therefore, $\frac{11}{20} > \frac{29}{60}$, so strategic feedback strictly outperforms truthful feedback on this instance. □

The construction highlights why strategic feedback can outperform truthful feedback in this instance. Under truthful reporting, the two types separate only when the menu contains both A and B. But any such menu gives each type its global optimum, leaving the assistant with low payoff. Strategic reporting allows separation without revealing these items: at the probe $\{M, Z\}$, where truthful choice would pool the types, type η_2 separates itself by reporting the dominated item Z.

This separation allows the assistant to use the ex-ante welfare budget more effectively. On the η_2 branch, it can obtain the assistant-preferred item M at a favorable payoff-per-regret rate and compensate η_2 with its preferred item A on the complementary realization. Thus, two features of the construction work together: strategic reporting separates the types at a probe where truthful reporting does not, and unresolved continuation randomization lets the assistant trade utility across realizations after that separation.

The second feature is essential to escaping the collapse. By Proposition F.4, if the assistant's randomization is instead resolved and publicly observed before the user reports, every equilibrium in the public-randomization game has value $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. The particular truthful-pooling pattern used in this construction, however, is only one way to generate a strategic-over-truthful gap; we do not claim that such pooling is necessary in general.

F.13 Proof of Theorem 5.4

We now prove the characterization in Theorem 5.4. The public-randomization collapse of Proposition F.4 already identifies every equilibrium value in this variant with $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. What remains is to connect equality with the truthful value $\mathrm{OPT}(\epsilon)$ to the incentive condition in part **(iii)** of the theorem. We first formulate that condition on the full public deviation-closed domain and show that it implies truthful user optimality. We then use these results to prove the three-way equivalence.

For the public-randomization game, let $\sigma^{\mathrm{tr,pub}}$ denote the seed-independent augmentation of the truthful Markov rule:

$$\sigma_t^{\mathrm{tr,pub}}(\cdot \mid A, q, B, c, \eta) := \sigma_t^{\mathrm{tr}}(\cdot \mid A, q, B, \eta), \qquad c \in \mathcal{C}.$$

Here, for each seed c, we write $\sigma_t^{\mathrm{tr},c}(\cdot \mid A, q, B, \eta) := \sigma_t^{\mathrm{tr,pub}}(\cdot \mid A, q, B, c, \eta)$ for the sectional truthful rule. These truthful reporting rules do not depend on the off-path selector.

Now fix an augmented off-path selector μ^{pub}. For each $c \in \mathcal{C}$, define its sectional selector by $\mu_t^c(\cdot \mid A, q, B, \mathbf{y}) := \mu_t^{\mathrm{pub}}(\cdot \mid A, q, B, c, \mathbf{y})$ and write

$$\begin{aligned}\bar{\tau}_t^{\mathrm{tr},\mu^{\mathrm{pub}},c}(q, A, B, \mathbf{y}) &:= \tau_t^{\sigma^{\mathrm{tr,pub}},\mu^{\mathrm{pub}},\mathrm{pub}}(q, A, B, c, \mathbf{y}) \\ &= \tau_t^{\sigma^{\mathrm{tr},c},\mu^c}(q, A, B, \mathbf{y}).\end{aligned}$$

Thus, fixing the public seed reduces truthful reporting and belief updating to the corresponding baseline sectional problem. We can therefore extend the deviation-complete DIC condition of Definition F.8 to the public-randomization class by imposing it in every seed section.

Definition F.12 (Deviation-complete DIC under public randomization)**.** *Let $\pi^{\mathrm{pub}} = (\nu, (\pi^c)_{c\in\mathcal{C}})$ be a public-randomization policy and fix an augmented off-path selector μ^{pub}. The policy π^{pub} is* deviation-complete dynamically incentive compatible *relative to μ^{pub} if, for every seed $c \in \mathcal{C}$, round t, every type η, every augmented information set $(A, q, B, c) \in \mathcal{I}_t^{\mathrm{pub}}(\sigma^{\mathrm{tr,pub}}, \mu^{\mathrm{pub}})$, and every myopic maximizer $\mathbf{y}^\star \in \arg\max_{\mathbf{x}\in B} U_\eta(\mathbf{x})$,*

$$\mathbf{y}^\star \in \arg\max_{\mathbf{y}\in B} U_{t+1}^{\pi^c, \bar{\tau}^{\mathrm{tr},\mu^{\mathrm{pub}},c}}\big(B, \bar{\tau}_t^{\mathrm{tr},\mu^{\mathrm{pub}},c}(q, A, B, \mathbf{y}), \eta\big).$$

The condition is imposed on the full augmented deviation-closed domain, not only at information sets reached with positive probability under truthful play. Accordingly, for the fixed selector μ^{pub}, each π^c must be specified on the truthful sectional domain $\mathcal{D}(\sigma^{\mathrm{tr},c}, \mu^c)$, equivalently on the c-section of $\mathcal{D}^{\mathrm{pub}}(\sigma^{\mathrm{tr,pub}}, \mu^{\mathrm{pub}})$. Because the seed is publicly observed and remains fixed within a section, it need not appear as an additional coordinate of the continuation value; the superscripted sectional update rule already identifies the relevant seed.

The baseline result of Lemma F.7 now applies separately to every seed section. This gives the link between the incentive condition just defined and the user-optimality requirement for a public strategic assessment.

Corollary F.5 (Deviation-complete DIC implies public user optimality)**.** *Assume (A1). If π^{pub} is deviation-complete dynamically incentive compatible relative to μ^{pub}, then truthful signalling satisfies Definition F.11**(iii)** at every information set in $\mathcal{I}_t^{\mathrm{pub}}(\sigma^{\mathrm{tr,pub}}, \mu^{\mathrm{pub}})$. Consequently, no behavioral Markov reporting deviation in the augmented game, including one that conditions on the revealed seed, is profitable.*

Proof. Fix $c \in \mathcal{C}$. The truthful c-section is exactly the baseline problem $(\pi^c, \sigma^{\mathrm{tr},c}, \mu^c)$ on $\mathcal{D}(\sigma^{\mathrm{tr},c}, \mu^c)$, with posterior update $\tau_t^{\sigma^{\mathrm{tr},c},\mu^c} = \bar{\tau}_t^{\mathrm{tr},\mu^{\mathrm{pub}},c}$. Under Assumption (A1), Lemma F.7 therefore implies that truthful reporting is supported on continuation-optimal reports at every information set in this section. Since c was arbitrary, the same condition holds at every augmented information set, which is precisely Definition F.11**(iii)**. The final claim then follows from Corollary F.4. □

The role of the off-path selector differs slightly between the two directions of the characterization. Deviation-complete DIC in the baseline model is defined relative to a baseline selector μ, while Definition F.12 uses an augmented selector μ^{pub} that also conditions on the public seed. In the **(ii)⇒(iii)** direction of Theorem 5.4 below, the publicly implemented non-adaptive policy will satisfy deviation-complete DIC for *every* augmented selector μ^{pub}. The reverse direction requires only one selector for which the condition holds, and that same selector is used to construct the public strategic assessment. This is why **(iii)** requires the existence of some μ^{pub}, rather than imposing the condition for every possible off-path selector.

We now have the ingredients needed to prove Theorem 5.4. The public-randomization collapse identifies every equilibrium value with $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$, while the preceding corollary connects deviation-complete DIC to truthful user optimality on the full augmented domain. It remains to show how these two facts yield the three equivalences in the theorem.

Theorem 5.4 (When truthful feedback is optimal)**.** *Assume (A1) and (A2), and fix ϵ for which a welfare-feasible non-adaptive control exists. In the public-randomization variant, the following are equivalent:* **(i)** *every welfare-feasible equilibrium has value $\mathrm{OPT}(\epsilon)$;* **(ii)** $\mathrm{OPT}(\epsilon) = \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$*;* **(iii)** *some truthful-optimal policy and off-path beliefs make truthful reporting optimal after every possible report, including zero-probability ones.*

Proof. **(i)⇔*(ii)***. By Proposition F.4, every equilibrium in the public-randomization game has value $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. Therefore, every equilibrium has the truthful-feedback value $\mathrm{OPT}(\epsilon)$ if and only if $\mathrm{OPT}_{\mathrm{nf}}(\epsilon) = \mathrm{OPT}(\epsilon)$.

(ii)⇒*(iii)*. Let D attain $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ and implement this menu lottery publicly using the construction in the proof of Proposition F.4: the terminal-menu plan is drawn before round 1, its seed is revealed, and the corresponding deterministic non-adaptive control is then followed. Every seed label, including those assigned zero probability, is associated with a deterministic non-adaptive control.

Fix a seed $c \in \mathcal{C}$ and an arbitrary augmented off-path selector μ^{pub}, and write

$$\bar{\tau}_t^c(q, A, B, \mathbf{y}) := \tau_t^{\sigma^{\mathrm{tr,pub}}, \mu^{\mathrm{pub}}, \mathrm{pub}}(q, A, B, c, \mathbf{y}) = \tau_t^{\sigma^{\mathrm{tr},c}, \mu^c}(q, A, B, \mathbf{y}).$$

Conditional on c, the assistant uses a slate kernel $\kappa_t^c(\cdot \mid A)$ that depends only on the accumulated menu and not on the posterior generated by the user's reports. Reports may change the posterior, so $\bar{\tau}_t^c$ need not be constant in $\mathbf{y}$, but the non-adaptive control never uses that posterior.

We show by backward induction that $U_t^{\pi^c, \bar{\tau}^c}(A, q, \eta)$ is independent of q. At $T+1$, $U_{T+1}^{\pi^c, \bar{\tau}^c}(A, q, \eta) = \max_{\mathbf{x} \in A} U_\eta(\mathbf{x})$, which does not depend on q. If the continuation value at $t+1$ is independent of the posterior, then the posterior induced by the current report cannot affect that continuation value; because the non-adaptive slate kernel also ignores the posterior, the same is true at t. Hence, at every augmented information set, including zero-probability ones, the map $\mathbf{y} \mapsto U_{t+1}^{\pi^c, \bar{\tau}^c}(B, \bar{\tau}_t^c(q, A, B, \mathbf{y}), \eta)$ is constant. Every report, and therefore every myopic maximizer, is continuation-optimal.

Moreover, because a non-adaptive control has no posterior argument, its lift is well defined on the entire truthful sectional domain for every augmented selector μ^{pub}, as required by Definition F.12. Since the argument did not use any property of μ^{pub}, the resulting public policy is deviation-complete dynamically incentive compatible relative to every augmented off-path selector. Under **(ii)**, its value is $\mathrm{OPT}_{\mathrm{nf}}(\epsilon) = \mathrm{OPT}(\epsilon)$, so it is truthful-optimal. This establishes **(iii)**.

(iii)⇒(i). Let $\pi^\star = (\nu^\star, (\pi^{\star,c})_{c \in \mathcal{C}})$ be a truthful-optimal public policy that is deviation-complete dynamically incentive compatible relative to some augmented off-path selector μ^{pub}. For each seed c, let $(\pi^{\star,c}, \sigma^{\mathrm{tr},c}, \mu^c)$ be the corresponding sectional triple. By Corollary F.5, truthful signalling satisfies Definition F.11**(iii)** at every augmented information set. By Corollary F.4, no complete behavioral Markov reporting deviation, including one that conditions on the revealed seed, is profitable.

It remains to verify assistant optimality and welfare feasibility. With truthful signalling fixed, condition Definition F.11**(ii)** is the welfare-constrained truthful-feedback optimization restricted to the public policy class. Since $\pi^\star$ belongs to this class and attains the unrestricted truthful value $\mathrm{OPT}(\epsilon)$ by truthful optimality, it also attains the restricted optimum. Welfare feasibility follows from the same truthful optimality hypothesis, while belief consistency follows from the truthful Bayes updates of Definition F.10, completed off path by the same selector μ^{pub}. Thus, $(\pi^\star, \sigma^{\mathrm{tr,pub}}, \mu^{\mathrm{pub}})$ is a welfare-feasible public strategic assessment with value $\mathrm{OPT}(\epsilon)$. Hence, $\mathrm{OPT}^{\max}_{\mathrm{strat,pub}}(\epsilon) \geq \mathrm{OPT}(\epsilon)$. Conversely, Proposition F.4 gives $\mathrm{OPT}^{\max}_{\mathrm{strat,pub}}(\epsilon) = \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$, and Theorem 5.2(i) gives $\mathrm{OPT}_{\mathrm{nf}}(\epsilon) \leq \mathrm{OPT}(\epsilon)$. Therefore, $\mathrm{OPT}^{\max}_{\mathrm{strat,pub}}(\epsilon) = \mathrm{OPT}(\epsilon)$, proving **(i)**. □

F.14 Characterization of the Strict Gap

The preceding results characterize not only when truthful feedback and the best public-randomization equilibrium have the same value, but also when the comparison is strict. Because every public-randomization equilibrium attains $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$, truthful feedback strictly outperforms the best such equilibrium exactly when adaptation to truthful feedback strictly improves on the best non-adaptive control. The characterization in Theorem 5.4 gives an equivalent incentive condition.

Proposition F.7 (Strict gap under public randomization)**.** *Assume (A1) and (A2), and fix ϵ for which a welfare-feasible non-adaptive control exists. In the public-randomization game,*

$$\mathrm{OPT}^{\max}_{\mathrm{strat,pub}}(\epsilon) < \mathrm{OPT}(\epsilon) \quad \Longleftrightarrow \quad \mathrm{OPT}_{\mathrm{nf}}(\epsilon) < \mathrm{OPT}(\epsilon).$$

Moreover, these equivalent inequalities hold if and only if there is no truthful-optimal policy $\pi^\star$ in the public-randomization class and augmented off-path selector μ^{pub} for which $\pi^\star$ is deviation-complete dynamically incentive compatible relative to μ^{pub} in the sense of Definition F.12.

Proof. By Proposition F.4, every public-randomization equilibrium has value $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$, so $\mathrm{OPT}^{\max}_{\mathrm{strat,pub}}(\epsilon) = \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. The first equivalence follows.

By Theorem 5.4, $\mathrm{OPT}_{\mathrm{nf}}(\epsilon) = \mathrm{OPT}(\epsilon)$ if and only if there exist a truthful-optimal public-randomization policy $\pi^\star$ and an augmented off-path selector μ^{pub} such that $\pi^\star$ is deviation-complete

dynamically incentive compatible relative to μ^{pub}. Negating this equivalence gives the final characterization. □

The restriction to public randomization is essential. In the unrestricted baseline game, unresolved continuation randomization can raise the best strategic equilibrium value above the non-adaptive benchmark. In Proposition F.5, $\mathrm{OPT}^{\max}_{\mathrm{strat}}(\frac{1}{5}) = \mathrm{OPT}(\frac{1}{5}) = \frac{1}{4} > \frac{2}{9} = \mathrm{OPT}_{\mathrm{nf}}(\frac{1}{5})$. Thus, $\mathrm{OPT}_{\mathrm{nf}}(\epsilon) < \mathrm{OPT}(\epsilon)$ does not generally imply $\mathrm{OPT}^{\max}_{\mathrm{strat}}(\epsilon) < \mathrm{OPT}(\epsilon)$ in the baseline game. Importantly, the policy in Proposition F.5 is itself truthful-optimal and deviation-complete dynamically incentive compatible in the sense of Definition F.8. There is no contradiction with Proposition F.7: its continuation randomization remains unresolved when the user reports and therefore falls outside the public-randomization class.

The strict case in Proposition F.7 can indeed occur. Consider the instance of Proposition F.5 with $k = 5 = |X|$. Preferences are strict and Assumption (A2) holds, while the calculations in that example give $\mathrm{OPT}_{\mathrm{nf}}(\frac{1}{5}) = \frac{2}{9} < \frac{1}{4} = \mathrm{OPT}(\frac{1}{5})$. Hence, by Proposition F.4, $\mathrm{OPT}^{\max}_{\mathrm{strat,pub}}(\frac{1}{5}) = \frac{2}{9} < \frac{1}{4} = \mathrm{OPT}(\frac{1}{5})$. We do not use the decoy-anchor family as a strict-preference witness for the public-collapse result. Although that family remains the witness for the unbounded Price of Truthful Feedback in the baseline maximin comparison, type η_2 is indifferent between g and o, so Assumption (A1) fails.

G Synthetic Instances

We use two synthetic constructions for distinct experimental purposes. The learning experiments use a single four-item, two-type instance constructed to study the mechanisms in Theorems E.1 and E.2. The instance is designed to admit informative belief-consistent strategic reporting. Opposing preferences over continuation outcomes would be self-defeating, as an assistant accounting for strategic behavior could invert the resulting signal. We therefore give both types the same ranking over the continuation outcomes but different preference intensities, while an initial pair of items distinguishes the types. The remaining benchmark, dissociation, regime, and scalability experiments use a separate family of randomly generated permutation-utility instances described in Section G.6.

G.1 Items, Types, Utilities, and Prior

The item set is $X = \{a, b, g_z, g_0\}$ and the type space is $\mathcal{E} = \{\eta_0, \zeta\}$ with uniform prior $p = (\frac{1}{2}, \frac{1}{2})$; the true type is η_0. User utilities are:

	a	b	g_z	g_0
η_0	0.35	0.15	1.00	0.20
ζ	0.15	0.35	0.70	0.50

Table 2: User utilities in the learning instance.

The pair $\{a, b\}$ is a probe: a is preferred by η_0 and b by ζ. The continuation pair $\{g_z, g_0\}$ gives both types a preference for g_z, but with different intensities: the utility gaps are 0.65 for η_0 and 0.20 for ζ.

G.2 Assistant Policy and Horizon

We use a horizon of $T = 3$ and an adaptive assistant policy that makes future recommendations depend smoothly on the posterior. The assistant reveals $\{a, b\}$ in the first round and, in rounds 2 and 3, reveals $\{g_z\}$ with probability $\varsigma(\beta_a(q(\eta_0) - q(\zeta)))$ and $\{g_0\}$ otherwise, where ς is the logistic function. Thus, reports can affect future recommendations by shifting the posterior toward one type or the other. The logistic policy keeps this dependence continuous, which is needed for the fixed-point construction of the belief-consistent reporting rule. A deterministic threshold policy would instead make continuation values discontinuous in beliefs.

G.3 Likelihoods and Belief State

We evaluate two interpretations of the same strategic feedback, as in Definition E.1. The truthful-likelihood posterior $\widetilde{q}$ treats each report as generated by the finite-temperature multinomial-logit model of Section B.5, while the strategic-likelihood posterior $q^\star$ uses the reporting probabilities of the strategic data-generating process. The process itself carries the belief state $q^\dagger$ generated by its reporting rule. As all three belief sequences start from the prior, $q^\star = q^\dagger$ along any realized history.

G.4 Strategic Data-Generating Process and Parameter Values

At each information state, let $V_t^\eta(\mathbf{y})$ denote the continuation value to type η from reporting $\mathbf{y}$. We define the reporting rule by

$$\sigma_t(\mathbf{y} \mid A, q, B, \eta) = \frac{\exp\{\lambda V_t^\eta(\mathbf{y})\}}{\sum_{\mathbf{z}\in B} \exp\{\lambda V_t^\eta(\mathbf{z})\}}, \qquad \mathbf{y} \in B,$$

with $\lambda = 4$. Thus, larger λ places more probability on reports with higher continuation value. These continuation values use the assistant policy and the beliefs generated by the same reporting rule, so the rule is defined by a fixed-point condition over the interaction tree. It is therefore belief-consistent and, for finite λ, full-support by construction. We do not interpret this process as an equilibrium; these are the properties required by Theorems E.1 and E.2.

For the two posterior interpretations, we set the inverse temperature of the truthful multinomial-logit likelihood to $\beta = 4$, while the assistant policy uses $\beta_a = 6$ to control how sharply future recommendations respond to the posterior difference $q(\eta_0) - q(\zeta)$. We simulate the $T = 3$ rounds specified above and evaluate both learning results after the final round, $t = 3$, using failure probability $\rho = 0.1$ for their finite-sample bounds.

G.5 Population of Strategic Reporting Rules

The fixed-point condition above can admit multiple belief-consistent reporting rules for the same instance. Holding all instance primitives and parameter values fixed, we numerically search for distinct solutions and find 90 reporting rules, each satisfying the fixed-point condition by residual of at most 10^{-9}. Relabelling reports generates 39 additional solutions by symmetry, giving 129 reporting rules in 12 symmetry classes. We evaluate all 129 rules. This set contains all relabellings of the solutions found by the search, but the search is not guaranteed to find every solution of the fixed-point condition.

G.6 Randomly Drawn Permutation-Utility Instances

The remaining synthetic studies draw from a common family of random instances. An instance is specified by the catalogue size $|X|$, number of user types K, horizon T, slate size k, and welfare budget ϵ. For each type, utilities over the catalogue are an independent uniformly random permutation of the evenly spaced values $0, \frac{1}{|X|-1}, \ldots, 1$; the assistant's utilities are generated independently in the same way, and the prior is uniform over types. Thus, each instance consists of $K + 1$ independent strict rankings on a common utility scale. The construction has no item features or latent utility structure. Distinct utilities ensure Assumption (A1), which we also verify directly on every instance, while the common scale makes additive value gaps comparable across configurations. Seeds are determined by the experiment configuration and instance index, making instance generation independent of execution order and parallelization.

Experimental configurations.

- Babbling verification (Section M.3): $|X| \in 6, 8, 10$, $K \in 3, 5$, 100 instances for each pair, with $T = 2$, $k = |X|$, and $\epsilon = 0.1$.
- Informativeness–value dissociation (Section M.4): three instances with $|X| = 10$, $K = 3$, $T = 2$, $k = |X|$, and $\epsilon = 0.1$.
- Descriptive regime study (Section M.5): 200 instances with $|X| = 6$, $K = 3$, $T = 2$, and $k = |X|$ at each of six welfare budgets.

- Scalability study (Section M.9): starting from $|X| = 10$, $K = 3$, $T = 2$, and $k = 3$, we vary one dimension at a time and generate three instances per configuration.

The first three studies set $k = |X|$, ensuring Assumption (A2) whenever it is required by the corresponding theoretical result. The scalability study varies catalogue and slate size separately and therefore does not impose this assumption; it is purely computational and invokes no theorem.

H Real-data benchmark instantiations

The experiments evaluating the learning results of Section 4 use controlled instances, which allow the hypothesis constants to be computed exactly. We separately evaluate the value results using preferences estimated from Coat (Schnabel et al., 2016) and SUSHI (Kamishima et al., 2010). Both datasets are converted into model instances using the common procedure below; dataset-specific details are given in Sections H.2 and H.3, followed by a comparison in Section M.8.

H.1 Shared Conversion Pipeline

From ratings to types. We encode each catalogue item using one-hot representations of the dataset's item attributes. For each respondent with at least two usable observations, we fit a ridge regression of rescaled ratings on these features. We then cluster the fitted coefficient vectors by k-means into $K = 5$ types. Cluster proportions define the type prior, and each type's item utilities are obtained from its cluster centroid and min-max normalized to $[0, 1]$. We use the same estimation, clustering, and normalization procedure for both datasets.

Assistant utility. The datasets contain no assistant objective, so we construct the assistant's item utility from two item-feature blocks and normalize it to the same scale. This utility is not estimated from user behavior and is chosen to differ from the fitted user objectives.

Candidate pools. Each instance contains a subset of the catalogue selected to preserve preference heterogeneity. We first include each type's favorite item, then add the assistant's highest-utility items until the pool is full; a random remainder is used only if slots remain. At the configurations studied here, the first two steps always fill the pool, so each pool is determined entirely by the fitted types and assistant utility. Pools therefore vary across instances only when the fitted type favorites vary.

Strategic interaction. Neither dataset contains observations of users strategically manipulating feedback to influence later recommendations. The data provide item structure and estimated preference heterogeneity. We simulate the strategic interaction on these fitted primitives. The resulting evidence therefore concerns the model's behavior with preferences estimated from real data, not observed strategic behavior in either dataset.

Assumption checks. We check Assumptions (A1) and (A2) on every fitted instance without modifying utilities or breaking ties to make them hold. We compute the optimal value under truthful feedback $\mathrm{OPT}_{\mathrm{truthful}}(\epsilon)$, the best randomized non-adaptive value $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$, and its deterministic counterpart $\mathrm{OPT}_{\mathrm{nf}}^{\mathrm{det}}(\epsilon)$. We also verify the uninformative equilibrium of Theorem 5.1, whose value equals $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. None of these comparisons requires Assumptions (A1) or (A2), so we do not apply the collapse or public-randomization results to these instances.

H.2 Coat

Coat contains ratings over a 300-item coat catalogue (Schnabel et al., 2016). We use the self-selected training fold and obtain ridge estimates for 290 respondents, which we cluster into $K = 5$ types. The assistant utility is constructed from the jacket-type and colour feature blocks.

We generate 40 instances with $|X| = 12$, $T = 2$, $k = 3$, and $\epsilon = 0.1$. The candidate pools are all distinct and collectively contain 50 of the 300 catalogue items. The assistant's seven highest-utility items appear in every pool, while the remaining five items are the fitted type favorites and vary across instances. Thus, the Coat instances vary in both their candidate pools and fitted preference types.

Assumption (A1) fails on all 40 instances because some fitted types have utility ties, and Assumption (A2) fails because $k = 3 < 12 = |X|$. For each instance, we compute the truthful-feedback value $\mathrm{OPT}_{\mathrm{truthful}}(\epsilon)$, the randomized and deterministic non-adaptive values $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$

and $\mathrm{OPT}_{\mathrm{nf}}^{\mathrm{det}}(\epsilon)$, and the certified candidate assessments, including their assistant values, welfare feasibility, and first-round informativeness. Table 8 and section M.6 report the results.

H.3 SUSHI

SUSHI contains stated preferences from $5{,}000$ respondents over a 100-item catalogue (Kamishima et al., 2010). Each respondent rated ten items on a five-point scale. Items have three categorical attributes (style, major group, and minor group) and three continuous attributes, which we discretize into empirical quartiles to obtain a one-hot encoding. Ridge estimation succeeds for all $5{,}000$ respondents, which we cluster into $K = 5$ types. The assistant utility is constructed from the style and major-group feature blocks.

We generate 20 instances with $|X| = 10$, $T = 2$, $k = 3$, and $\epsilon = 0.1$. All 20 share the same ten-item candidate pool. Under the selection rule of Section H.1, the five fitted type favorites and the assistant's five highest-utility items fill the pool, and the fitted type favorites are identical across seeds. The fitted type structure still varies: cluster shares differ by up to 0.016, centroid utilities change, and the resulting truthful-feedback and non-adaptive values are distinct across all 20 instances. Variation across SUSHI instances therefore reflects fitted preference structure, not candidate-pool variation.

Assumption (A1) holds on all 20 instances, while Assumption (A2) fails because $k = 3 < 10 = |X|$. For each instance, we compute the truthful-feedback value $\mathrm{OPT}_{\mathrm{truthful}}(\epsilon)$, the randomized and deterministic non-adaptive values $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ and $\mathrm{OPT}_{\mathrm{nf}}^{\mathrm{det}}(\epsilon)$, and the certified candidate assessments, including their assistant values, welfare feasibility, and first-round informativeness. Exact optimization over the public-randomization equilibrium class is not computed at this configuration because it is computationally prohibitive. We therefore make no comparison between its optimum and the truthful-feedback value on SUSHI.

I Computational Procedures

This section describes the computational procedures used in the learning experiments.

Finding belief-consistent reporting rules. We solve the belief-consistency condition as a fixed-point problem over the entire interaction tree. Starting from each of 300 initializations, we iterate the fixed-point map to convergence and retain a reporting rule only if its residual in probability space is at most 10^{-9}. The search is completed before any trajectories are simulated, so experimental outcomes cannot affect which reporting rules are included.

Completing the set under report symmetry. The labels attached to reports do not themselves affect the accumulated menu. Consequently, consistently relabelling reports in a solution produces another solution of the same fixed-point problem. After the numerical search, we therefore generate all reporting rules obtained from each discovered solution by these exact relabellings and verify each one using the same 10^{-9} residual criterion. This gives the discovered set of 129 reporting rules described in Section G.5. We group rules related by such relabellings into symmetry classes.

Representative reporting rules. For displays requiring representative examples, we classify each reporting rule by whether it satisfies the cumulative misspecification condition of Theorem E.1 and the cumulative separation condition of Theorem E.2, and select the lowest-indexed rule in each resulting class. The ordering is fixed before simulation, so selection does not depend on simulated outcomes or on how often the numerical search finds a rule. Local stability is recorded only as a diagnostic and does not affect selection.

Trajectory simulation and validation. For each reporting rule, we simulate independent feedback streams while evolving the accumulated menu and the process, truthful-likelihood, and strategic-likelihood belief sequences together. We store the realized slates, menus, reports, and belief paths. As a validation check, both posterior sequences are reconstructed from the stored feedback and verified against their simulated values.

Deterministic seed derivation. The seed for each trajectory is obtained by hashing the tuple $(\text{base seed}, \text{continuation-value convention}, \text{reporting-rule index}, \text{trajectory index})$ and reducing the digest modulo 2^{63}. Seeds therefore depend only on predeclared identifiers, never on observed out-

comes, the number of trajectories requested, or execution order. This makes the simulation reproducible, parallelisable, and resumable, while keeping the canonical and sensitivity streams disjoint.

J Evaluation Metrics

Let $\widetilde{q}_T$ and $q^\star_T$ denote the terminal truthful-likelihood and strategic-likelihood posteriors, respectively, as defined in Definition E.1, and let η_0 denote the true type. We report the following quantities in Appendix L.

False-type dominance. For either posterior $\pi \in \{\widetilde{q}, q^\star\}$, we measure $\mathbb{P}[\max_{\zeta \neq \eta_0} \pi_T(\zeta) > \pi_T(\eta_0)]$, the probability that some false type ends with higher posterior probability than the true type. We report this proportion with a 95% Wilson score interval, which remains well behaved when the estimated probability is near zero or one.

Mean terminal posterior on the true type. For either posterior $\pi \in \{\widetilde{q}, q^\star\}$, we report $\mathbb{E}[\pi_T(\eta_0)]$, the mean terminal probability assigned to the true type, together with its standard error.

Paired posterior difference. We report $\mathbb{E}[q^\star_T(\eta_0) - \widetilde{q}_T(\eta_0)]$ with a paired standard error. The difference is computed within each realized interaction because both posteriors are updated from the same feedback stream.

Theorem-condition classes. Each reporting rule is classified by whether it satisfies the cumulative-drift condition of Theorem 4.1, the cumulative-separation condition of Theorem 4.2, both, or neither. For each rule, we compute exhaustively over all positive-probability histories the minimum likelihood probability α_{dom}, the cumulative drift γ toward a false type under the truthful-likelihood interpretation, and the cumulative separation $\gamma_\star$ between the true and false types under the strategic-likelihood interpretation. These quantities are computed exactly rather than estimated from simulated trajectories, with values below 10^{-9} treated as zero.

Secondary diagnostics. We also record whether the cumulative drift toward a false type is large enough to exceed the finite-sample deviation term in Theorem 4.1. For each reporting rule, we record whether it is locally stable under the fixed-point iteration and which report-relabelling symmetry class it belongs to; these diagnostics are discussed in Appendix L.5, and the symmetry classes are constructed in Appendix I. Finally, we record the smallest report probability encountered along each realized path, as a diagnostic for the full-support condition of Theorems 4.1 and 4.2, and whether the off-path posterior selector of Definition E.1 is ever invoked.

Aggregation. Trajectories are first summarized within each reporting rule, and the resulting rule-level quantities are then averaged with equal weight. Because every reporting rule in the main experiment has exactly 100 trajectories, this average equals the trajectory-pooled estimate. We also report equal weighting across symmetry classes as a robustness check in Section L.1. The 129 reporting rules form the discovered symmetry-closed set described in Section G.5, not a random sample of strategic behaviors. Counts and averages therefore describe this set and are not population estimates.

Monte Carlo uncertainty and theoretical bounds. Wilson intervals and standard errors quantify Monte Carlo uncertainty and shrink as the number of simulated interactions increases. By contrast, $c_t(\rho)$ in Theorem 4.1 and $c^\star_t(\rho)$ in Theorem 4.2 are finite-sample deviation terms determined by the interaction horizon and full-support constant. They do not shrink with additional simulated interactions and are not combined with the Monte Carlo intervals.

K Reproducibility

The main learning experiment uses base seed 20260821, 300 initializations for the fixed-point search, and 100 trajectories for each reporting rule. The remaining parameters and their roles are specified with the learning instance in Section G. Reporting rules are retained only when their fixed-point residual is at most 10^{-9}.

Trajectory seeds are deterministic functions of the base seed, continuation-value convention, reporting-rule index, and trajectory index, as described in Section I. They are therefore independent

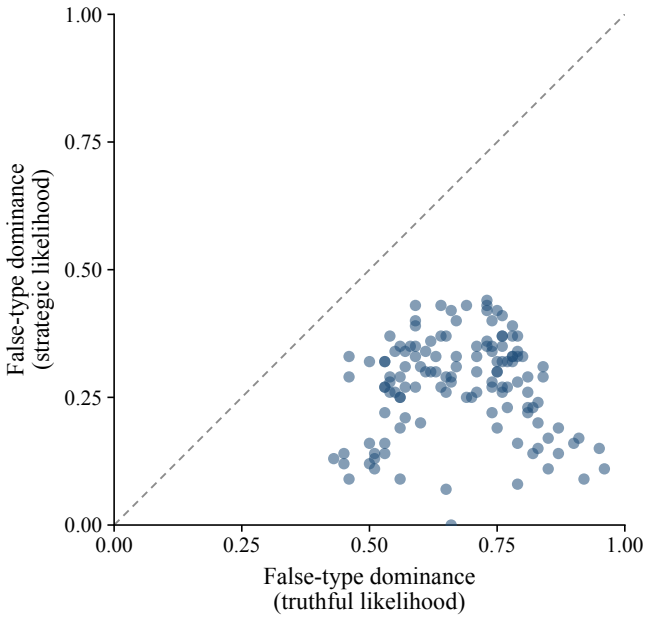

Figure 1: False-type dominance under two interpretations of the same strategic feedback. Each point represents one of the 129 reporting rules and gives the empirical probability, over 100 interactions, that a false type ends with higher posterior probability than the true type under the truthful-likelihood (horizontal axis) and strategic-likelihood (vertical axis) interpretations. Both posteriors are computed from the same realized feedback stream. Every point lies below the 45-degree line.

of execution order and parallelization, and the main and sensitivity analyses use disjoint random streams.

As validation checks, we reconstruct both posterior sequences from the stored slates and reports and verify that they match the posteriors obtained during simulation. In the main experiment, the strategic-likelihood posterior also coincides with the belief state of the data-generating process throughout all 12,900 interactions, and the off-path posterior selector is never invoked. All reported aggregates are recomputed from the trajectory-level results.

L Empirical Results

This section reports the empirical results. All numbers are computed from the frozen artifacts of Section K, and the reported quantities are defined in Section J.

L.1 Paired Truthful- versus Strategic-Likelihood Inference

Configuration. We evaluate all 129 belief-consistent reporting rules described in Section G.5, with 100 simulated interactions per rule and 12,900 interactions in total. In every interaction, the truthful-likelihood and strategic-likelihood posteriors are updated from the same realized feedback stream, so they differ only in how that feedback is interpreted.

Results. A false type ends with higher posterior probability than the true type in 67.8% of interactions under the truthful-likelihood interpretation, compared with 27.4% under the strategic-likelihood interpretation. The corresponding 95% Wilson intervals are $[0.670, 0.687]$ and $[0.267, 0.282]$. The mean terminal posterior on the true type increases from 0.368 to 0.524 when strategic behavior is accounted for. This improvement holds for every one of the 129 reporting rules, with a mean rule-level increase of 0.156 and a range from 0.021 to 0.367. Moreover, no reporting rule has higher false-type dominance under the strategic interpretation. Figure 1 shows the rule-level comparison.

Because both posteriors are computed from the same realized feedback, these differences arise from the likelihood used to interpret strategic reports rather than from different observed data. The empirical results illustrate the misspecification result of Theorem 4.1: treating strategic reports as truthful can systematically shift belief toward a false type. They should not, however, be interpreted as a numerical demonstration of the concentration guarantee in Theorem 4.2. At the three-round horizon studied here, the mean strategic-likelihood posterior on the true type is 0.524, close to its prior value of 0.5. The empirical finding is that accounting for strategic behavior substantially reduces incorrect inference, not that the posterior has already concentrated on the true type.

Some of the 129 reporting rules are related by exact relabellings of reports, as described in Section I. As a robustness check, giving equal weight to the 12 report-relabelling classes changes strategic-likelihood false-type dominance from 0.274 to 0.227 and changes the other reported quantities by at most 0.002, so the comparison is not driven by unequal class sizes. The 129 reporting rules all come from the single controlled instance of Section G, and 39 are obtained by exact report relabelling of numerically discovered rules (Section I). They are therefore not independent draws or a random sample of strategic behaviors, so the reported frequencies describe this set rather than a broader population.

L.2 Learning Conditions and Empirical Behavior

Configuration. We use the same $12{,}900$ interactions as in Section L.1 and group the 129 reporting rules according to whether they satisfy the cumulative-drift condition of Theorem 4.1, the cumulative-separation condition of Theorem 4.2, both, or neither. These conditions and their computation are described in Section J.

Results. Of the 129 reporting rules, 24 satisfy both conditions, 5 only the cumulative-drift condition, 86 only the cumulative-separation condition, and 14 neither. False-type dominance under the truthful-likelihood interpretation is 0.836, 0.738, 0.635, and 0.654 across these four groups, respectively, compared with 0.225, 0.368, 0.282, and 0.281 under the strategic-likelihood interpretation. Consistent with Theorem 4.1, the misspecification effect is strongest when the cumulative-drift condition holds. In particular, the 29 rules with positive cumulative drift have truthful-likelihood false-type dominance of 0.819, compared with 0.638 for the remaining 100 rules. However, 91 of those 100 rules still have false-type dominance above 0.5, illustrating that the theorem's condition is sufficient rather than necessary.

The cumulative-separation condition should likewise not be interpreted as implying concentration at the three-round horizon. For three reporting rules satisfying the condition of Theorem 4.2, the mean terminal $q^\star(\eta_0)$ is slightly below its prior value. Their cumulative separation is only about 0.012, for which the finite-sample bound is not informative. Because all 129 reporting rules come from the single controlled instance of Section G, their distribution across the four theorem-condition classes does not indicate how frequently these conditions hold among strategic reporting rules more generally.

L.3 Continuation-Value Sensitivity

Configuration. As a sensitivity check, we replace the main convention, which evaluates continuation values under optimal subsequent reporting, with an alternative that averages future continuation utility under the reporting rule itself. All other parameters are held fixed. Because this convention enters the fixed-point condition defining the reporting rule, the alternative produces a different set of 93 reporting rules, each evaluated with 100 interactions. The two sets are therefore compared qualitatively rather than as matched reporting rules.

Results. Under the alternative convention, false-type dominance is 0.692 under the truthful-likelihood interpretation and 0.216 under the strategic-likelihood interpretation. The mean rule-level increase in the terminal posterior on the true type is 0.167 and is positive for all 93 reporting rules. Thus, the direction and magnitude of the difference between the two likelihood interpretations are similar to those in Section L.1, despite the alternative construction producing a different set of reporting rules. Table 3 summarizes the comparison.

Table 3: Sensitivity to the continuation-value convention. Because the two conventions define different fixed-point problems, they produce different sets of reporting rules and are compared qualitatively rather than as matched populations.

Convention	Rules	Dominance ($\widetilde{q}$)	Dominance ($q^\star$)	Mean $q^\star - \widetilde{q}$
Main	129	0.678	0.274	+0.1557
Alternative	93	0.692	0.216	+0.1667

L.4 Horizon Scaling of the Learning Results

Experimental design. Both learning results in Theorems 4.1 and 4.2 concern evidence accumulated over rounds, whereas Section L.1 evaluates the paired likelihood comparison at a single horizon. We therefore repeat the experiment at $T = 2, 3, 4$. At each horizon, we hold fixed the item set, utility functions, prior, assistant-policy form, softmax inverse temperature, truthful-choice inverse temperature, assistant-policy sharpness, residual tolerance, 100 simulated interactions per reporting rule, and the deterministic seed construction of Section K. Changing T, however, changes the belief-consistency problem itself. We therefore solve separately for the reporting rules at each horizon and do not match rules across horizons; all cross-horizon comparisons are between separately constructed populations. As in the main experiment, local stability is recorded only as a diagnostic and is never used to select reporting rules.

Reporting-rule populations. At $T = 2$ and $T = 3$, we use the same procedure as in Section I: multi-start numerical search followed by exact closure under report relabellings. A report relabelling permutes report labels while preserving the fixed-point problem, so applying such a relabelling to a belief-consistent reporting rule produces another belief-consistent rule. We group rules connected by these relabellings into *report-relabelling classes*; membership in the same class records this symmetry relation, regardless of whether a rule was found directly by the numerical search or added during closure. This gives 3 reporting rules in 2 report-relabelling classes at $T = 2$ and 129 rules in 12 classes at $T = 3$.

At $T = 4$, exact relabelling closure becomes computationally infeasible: applying it to only 18 discovered rules already generated more than 18,600 rules without terminating. We therefore use the 127 rules found directly by the numerical search. The $T = 4$ population is thus discovery-only: it is not report-relabelling-closed, has no associated relabelling-class partition, and is not claimed to contain every belief-consistent reporting rule.

The $T = 3$ population allows us to assess the effect of omitting relabelling closure. Comparing the 90 directly discovered rules with the full 129-rule closed set changes the central empirical quantities by at most 0.00988. The largest difference among all reported quantities is 0.030, attained by an extreme order statistic, while the certificate quantities γ, $\gamma_\star$, the accumulated evidence, the deviation terms, and the full-support constant agree to numerical precision. We therefore use the $T = 4$ discovery-only population as a descriptive extension of the horizon comparison, not as an equivalent substitute for a report-relabelling-closed population.

Empirical learning across horizons. At each horizon, we compute false-type dominance under the truthful- and strategic-likelihood interpretations, the mean terminal posterior assigned to the true type under each interpretation, and the paired difference $q^\star_T(\eta_0) - \widetilde{q}_T(\eta_0)$ on the same realized interaction. The aggregation follows Section J. Table 4 and fig. 2 report these results.

False-type dominance under the truthful-likelihood interpretation rises from 0.560 at $T = 2$ to 0.678 at $T = 3$ and 0.786 at $T = 4$, while the mean terminal posterior on the true type falls from 0.423 to 0.368 and then 0.291. The strategic-likelihood posterior behaves differently. Its mean posterior on the true type remains close to the prior of 0.5, taking values 0.525, 0.524, and 0.521, although its false-type dominance also rises from 0.087 to 0.274 and 0.315. Consequently, the paired posterior difference grows from 0.103 to 0.156 to 0.231 and remains positive for every reporting rule at every horizon. Thus, the widening advantage of the strategic interpretation is driven mainly by increasingly misleading inference under the misspecified truthful likelihood, rather than by increasingly accurate inference under the strategic likelihood.

Table 4: Empirical learning by horizon under the main continuation-value convention. Reporting-rule populations are constructed separately at each horizon and are not matched across horizons. The $T = 4$ population is discovery-only and is not report-relabelling-closed. Dominance is the empirical probability that a false type ends with higher posterior probability than the true type.

T	Population	Rules	$\widetilde{q}$ dom.	$q^\star$ dom.	$\widetilde{q}_T(\eta_0)$	$q^\star_T(\eta_0)$	$q^\star - \widetilde{q}$
2	Closed	3	0.560	0.087	0.423	0.525	+0.103
3	Closed	129	0.678	0.274	0.368	0.524	+0.156
4	Discovery-only	127	0.786	0.315	0.291	0.521	+0.231

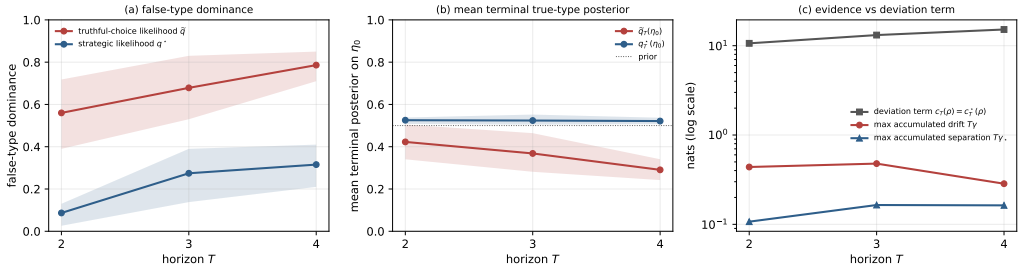

Figure 2: Learning and theorem quantities as the horizon grows under the main continuation-value convention. **(a)** False-type dominance under the truthful- and strategic-likelihood interpretations. **(b)** Mean terminal posterior on the true type; the dotted line marks the prior. **(c)** Largest accumulated evidence relative to the finite-sample deviation term, shown on a logarithmic scale. Lines report means across reporting rules and shaded bands the 10th–90th percentile range. Reporting rules are solved for separately at each horizon and are not matched across horizons. The $T = 4$ population contains only the numerically discovered rules and is not report-relabelling-closed.

Theorem hypotheses and finite-sample guarantees. We next ask how the horizon results relate to Theorems 4.1 and 4.2. Recall that γ is the uniform cumulative drift toward a false type under the truthful-likelihood interpretation, while $\gamma_\star$ is the uniform cumulative separation between the true and false types' strategic reporting distributions. Both are computed exhaustively over positive-probability histories, as described in Section J.

The two results impose different evidence conditions. For Theorem 4.1, $\gamma > 0$ means that, uniformly over positive-probability histories, the true type's strategic reports are cumulatively better explained by the truthful-choice likelihood of a false type than by that of the true type. For Theorem 4.2, $\gamma_\star > 0$ means that, uniformly over such histories, the strategic reporting distributions of the true and false types remain cumulatively separated. Thus, positive γ supports the misspecification mechanism of the first result, whereas positive $\gamma_\star$ supports learning under the correctly specified strategic likelihood. The formal definitions and their computation are given in Section J.

These hypotheses alone do not imply that the corresponding finite-sample conclusion is numerically informative at a given horizon. The accumulated evidence must also be large enough to overcome the finite-sample deviation term and the initial posterior odds. For Theorem 4.1, this requires $\log \frac{\widetilde{q}_1(\zeta)}{\widetilde{q}_1(\eta_0)} + T\gamma - c_T(\rho) > 0$, which guarantees that the truthful-likelihood posterior favors the false type ζ. For Theorem 4.2, the corresponding condition is $\log \frac{1-q^\star_1(\eta_0)}{q^\star_1(\eta_0)} - T\gamma_\star + c^\star_T(\rho) < 0$, which makes the theorem's upper bound on posterior mass assigned to false types informative. The prior is uniform in this experiment, so both prior-odds terms vanish exactly, and the complete conditions reduce here to comparing the accumulated evidence with the corresponding deviation term. We nevertheless evaluate the complete inequalities directly.

Table 5 shows that the evidence hypotheses frequently hold, but the finite-sample guarantees are not numerically informative at these horizons. Specifically, $\gamma > 0$ for 1, 29, and 8 reporting rules at $T = 2, 3, 4$, while $\gamma_\star > 0$ for 2, 110, and 118, respectively. However, no reporting rule satisfies either complete finite-sample inequality.

Table 5: Theorem quantities by horizon under the main continuation-value convention. Positive γ and $\gamma_\star$ correspond to the evidence hypotheses of Theorems 4.1 and 4.2, respectively. The final two columns report the number of reporting rules satisfying the corresponding complete finite-sample condition; none does so at a tested horizon.

T	$\max\gamma$	$\max\gamma_\star$	$\max T\gamma$	$\max T\gamma_\star$	$c_T(\rho)$	Thm. 4.1	Thm. 4.2
2	0.219	0.054	0.438	0.107	10.63	$0/3$	$0/3$
3	0.160	0.055	0.479	0.165	13.15	$0/129$	$0/129$
4	0.071	0.041	0.285	0.163	15.19	$0/127$	$0/127$

This does not contradict the theoretical learning results. The inequalities above are sufficient conditions for the high-probability conclusions of Theorems 4.1 and 4.2, not necessary conditions for the corresponding posterior behavior. Their failure therefore means only that the finite-sample bounds do not verify the observed behavior at these horizons. In particular, a false type may still be favored under the truthful-likelihood posterior when the bound in Theorem 4.1 is inconclusive, and the strategic-likelihood posterior may still be less prone to false-type dominance when the concentration bound in Theorem 4.2 is inconclusive.

The bounds are inconclusive because the accumulated evidence remains small relative to the deviation terms. The largest γ across reporting rules is 0.219, 0.160, and 0.071 at $T = 2, 3, 4$, giving maximum accumulated drift $T\gamma$ of 0.438, 0.479, and 0.285. Similarly, the largest $\gamma_\star$ is 0.054, 0.055, and 0.041, giving maximum accumulated separation $T\gamma_\star$ of 0.107, 0.165, and 0.163. By contrast, the deviation term grows from 10.63 to 13.15 and 15.19. Thus, even the strongest certified drift or separation at each horizon remains far below the threshold at which the corresponding finite-sample guarantee becomes informative. Accordingly, the horizon experiment illustrates the likelihood mechanisms underlying the two results, but does not numerically demonstrate their finite-sample high-probability conclusions.

Typical versus worst-case accumulated evidence. The fact that γ and $\gamma_\star$ do not increase with the horizon does not mean that additional rounds fail to generate evidence. The difference comes from how the theorem hypotheses are defined. Both require their evidence conditions to hold uniformly over positive-probability histories, so $T\gamma$ and $T\gamma_\star$ are determined by the *least favorable* positive-probability path. As the horizon grows, the number of such terminal paths increases from 12 at $T = 2$ to 84 at $T = 3$ and 636 at $T = 4$. Longer interactions therefore provide more opportunities to accumulate evidence, but also enlarge the set of paths over which the theorem requires a uniform guarantee.

Figure 3 illustrates this distinction. For the reporting rule with the largest γ at each horizon, the probability-weighted mean cumulative drift increases from 0.854 at $T = 2$ to 2.185 at $T = 4$, showing that a typical realized path accumulates substantially more evidence. Yet, its minimum cumulative drift over positive-probability paths, which equals $T\gamma$, changes from 0.4385 to 0.2851. The minimum cumulative strategic separation likewise remains nearly flat, from 0.1071 to 0.1619. Thus, additional rounds increase average evidence without strengthening the least favorable path on which the theorem's uniform condition is based.

This is not specific to the reporting rule that maximizes γ. Across the full reporting-rule population, the median probability-weighted mean cumulative drift rises from 0.252 at $T = 2$ to 1.440 at $T = 4$, and the median fraction of paths with positive cumulative drift rises from 0.50 to 0.92. At the same time, the median minimum cumulative drift falls from -0.093 to -0.982. Hence, evidence becomes stronger on average and along a larger share of realized paths, while increasingly adverse low-probability paths prevent the worst-case quantities required by the theorems from improving.

Sensitivity to the continuation-value convention. We repeat the horizon study using the alternative continuation-value convention of Section L.3. This convention defines a different belief-consistency problem, producing populations of 3, 93, and 185 reporting rules at $T = 2, 3, 4$. We therefore compare the two conventions only through their cross-horizon trends, not rule by rule.

The qualitative pattern is unchanged. Under the alternative convention, truthful-likelihood false-type dominance rises from 0.513 to 0.789, while the mean strategic-likelihood posterior on the true type remains close to the prior. The paired posterior difference grows from 0.073 to 0.242, and the

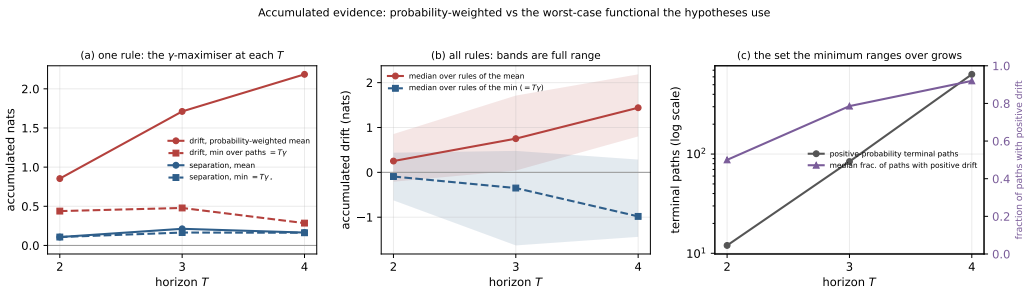

Figure 3: Typical and worst-case accumulated evidence as the horizon grows. **(a)** For the reporting rule with the largest γ at each horizon, the probability-weighted mean cumulative drift grows while the minimum over positive-probability paths, equal to $T\gamma$, does not. This panel concerns only the γ-maximizing rule at each horizon. **(b)** Across the full reporting-rule population, the median probability-weighted mean and minimum cumulative drift, with bands spanning the range across rules. **(c)** The number of positive-probability terminal paths over which the minimum is taken, together with the population median fraction of paths having positive cumulative drift.

largest γ falls from 0.219 to 0.094. As under the main convention, no reporting rule satisfies either complete finite-sample condition at any tested horizon. The horizon pattern is therefore not specific to the continuation-value convention used in the main experiment.

Computational limit at longer horizons. We stop the study at $T = 4$ because the fixed-point search becomes computationally prohibitive at $T = 5$. To assess feasibility, we run a $T = 5$ pilot without the local-stability diagnostic, which is not used for selection. A single random initialization takes approximately 352 seconds, implying about 29.3 hours for the 300-initialization search alone, before verification or trajectory simulation. Five pilot initializations locate one distinct reporting rule. This is only a diagnostic of the search at $T = 5$ and does not imply that the fixed-point problem has a unique solution. We therefore do not report $T = 5$ as an experimental horizon.

Summary of the horizon comparison. Across the tested horizons, interpreting strategic feedback through the truthful-choice likelihood becomes increasingly misleading, while the strategic-likelihood interpretation retains a growing paired advantage. That advantage is driven mainly by deterioration of the misspecified interpretation, not by increasing concentration of the strategic-likelihood posterior. Meanwhile, probability-weighted evidence accumulates even though the worst-case quantities entering the theorem hypotheses do not keep pace, and neither theorem's complete finite-sample guarantee becomes numerically informative.

These findings concern the separately constructed reporting-rule populations at $T = 2, 3, 4$. They do not establish behavior beyond $T = 4$, do not imply that the discovered reporting rules exhaust all belief-consistent solutions, and do not constitute a numerical validation of the finite-sample guarantees. The $T = 4$ results additionally describe a discovery-only population rather than a report-relabelling-closed one.

L.5 Additional Results

Representative reporting rules. Table 6 presents four belief-consistent reporting rules illustrating qualitatively different behaviors, selected according to Section I. The indices 64, 0, 2, and 56 identify the corresponding fixed-point solutions and are retained for reproducibility; the descriptions below characterize the rules independently of these indices.

All four rules have the functional form specified in Section G. At a reachable information state (A_s, q_s, B_s), a type-η user reports $\mathbf{y} \in B_s$ according to

$$\sigma_s(\mathbf{y} \mid A_s, q_s, B_s, \eta) = \frac{\exp(\lambda V_s^\eta(\mathbf{y} \mid A_s, q_s, B_s))}{\sum_{\mathbf{y}' \in B_s} \exp(\lambda V_s^\eta(\mathbf{y}' \mid A_s, q_s, B_s))}, \qquad \lambda = 4,$$

where $V_s^\eta(\mathbf{y} \mid A_s, q_s, B_s)$ is the type-η continuation value from reporting $\mathbf{y}$ under the assistant policy of Section G. These continuation values are evaluated using the beliefs generated by the

same reporting rule, so σ is defined by a fixed-point condition over the interaction tree. Because a report does not change the accumulated menu, its effect on continuation value operates through the posterior it induces. For this instance, the fixed-point condition determines the reporting rule at five information states: the first-round state with menu $\{a, b\}$ and four second-round states, with menus $\{a, b, g_z\}$ or $\{a, b, g_0\}$ depending on the first-round report and the assistant's second-round slate. In the final round, reporting is uniform by construction. The four representatives below are different solutions of this common fixed-point problem.

- **Type-independent uniform reporting (rule 64).** At every reachable information state, both types use the same reporting distribution:

  $$\sigma_s(\cdot \mid A_s, q_s, B_s, \eta_0) = \sigma_s(\cdot \mid A_s, q_s, B_s, \zeta).$$

  The strategic likelihood is therefore identical across types. In the Bayes update of Definition E.1, the likelihood term is common to both types and cancels, so the process belief remains at its prior, $q^\dagger_{s+1} = q^\dagger_s = p = (\frac{1}{2}, \frac{1}{2})$. Consequently, in rounds 2 and 3 the assistant reveals $\{g_z\}$ with probability

  $$\varsigma(\beta_a(q(\eta_0) - q(\zeta))) = \varsigma(0) = \tfrac{1}{2}.$$

  The fixed-point solution is uniform at each of the five information states: each of a and b is reported with probability $\frac{1}{2}$ in the first round, and each item is reported with probability $\frac{1}{3}$ at every second-round state. This also follows from the softmax specification: since reports affect continuation values only through the posterior they induce and no report changes that posterior under this rule, all reports in a given menu have the same continuation value. Hence

  $$\sigma_s(\mathbf{y} \mid A_s, p, B_s, \eta) = \frac{1}{|B_s|} \qquad \text{for every } \mathbf{y} \in B_s \text{ and both types}$$

  in rounds 1 and 2; final-round reporting is uniform by construction.

  This behavior parallels the babbling construction of Theorem 5.1, in which all types use the same reporting rule and the posterior remains at the prior. Here, however, the rule arises from the belief-consistent fixed-point construction used for the learning experiment and is not claimed to be an equilibrium. Because the two types have identical strategic reporting likelihoods, every strategic separation increment vanishes, $\gamma_\star = 0$, and $q^\star_s = q^\dagger_s = p$ throughout the interaction. Hence its strategic-likelihood false-type dominance is exactly zero.

- **Type-dependent reporting with positive drift and separation (rule 0).** For this solution, the two types use different reporting distributions on a positive-probability set of information states, so reports can change the posterior and the continuation values entering the softmax can differ across types. In the first round, with coordinates ordered as (a, b),

  $$\sigma_1(\cdot \mid \varnothing, p, \{a, b\}, \eta_0) = (0.084609, 0.915391),$$
  $$\sigma_1(\cdot \mid \varnothing, p, \{a, b\}, \zeta) = (0.324600, 0.675400).$$

  Thus, both types favor report b, but type η_0 does so substantially more strongly. At three of the four second-round states, both types report uniformly. At the state reached after first reporting b and then receiving slate $\{g_0\}$, the menu is (a, b, g_0) and

  $$\sigma_2(\cdot \mid b, \{g_0\}, \eta_0) = (0.117059, 0.117059, 0.765882),$$
  $$\sigma_2(\cdot \mid b, \{g_0\}, \zeta) = (0.264383, 0.264383, 0.471234).$$

  The rule therefore remains type-dependent at this continuation state.

  For this reporting rule, the cumulative truthful-likelihood drift toward the false type, γ, and the cumulative separation between the true and false types under their strategic reporting likelihoods, $\gamma_\star$, are defined as in Theorems 4.1 and 4.2 and computed as described in Section J. Their values are $\gamma = 0.160$ and $\gamma_\star = 0.055$. In particular, γ measures the accumulated advantage with which the truthful-choice likelihood explains the true type's strategic reports as coming from the false type, whereas $\gamma_\star$ measures the accumulated separation between the two types' strategic reporting distributions. Both exceed the 10^{-9} numerical gate, so this rule satisfies the cumulative-drift condition of Theorem 4.1 and the cumulative-separation condition of Theorem 4.2. Its false-type dominance is 0.91 under the truthful-likelihood interpretation and 0.17 under the strategic-likelihood interpretation.

- **A report relabelling of rule** 0 **(rule** 2**).** Rules 0 and 2 were both found directly by the numerical search and belong to the same report-relabelling class. They coincide at the first-round state and at three of the four second-round states. In particular, rule 2 has the same first-round reporting probabilities as rule 0:

$$\sigma_1(\cdot \mid \varnothing, p, \{a, b\}, \eta_0) = (0.084609, 0.915391),$$
$$\sigma_1(\cdot \mid \varnothing, p, \{a, b\}, \zeta) = (0.324600, 0.675400).$$

  The two rules differ only at the second-round state reached after first reporting b and then receiving slate $\{g_0\}$. At this state, whose menu is (a, b, g_0), rule 2 has

$$\sigma_2^{(2)}(\cdot \mid b, \{g_0\}, \eta_0) = (0.765882, 0.117059, 0.117059),$$
$$\sigma_2^{(2)}(\cdot \mid b, \{g_0\}, \zeta) = (0.471234, 0.264383, 0.264383).$$

  These probabilities are obtained from rule 0 by exchanging the reports a and g_0 for both types at this state and leaving every other state unchanged. This is a minimal report relabelling mapping rule 0 to rule 2; because several other states have uniform reporting distributions, additional permutations at those states leave the rules unchanged. The exact report-relabelling symmetry and its action on the interaction tree are described in Section I.

  This pair illustrates an important difference between the two learning quantities. Relabelling reports in the same way for both types permutes their strategic likelihoods together, leaving their cumulative separation unchanged:

$$\gamma_\star^{(2)} = \gamma_\star^{(0)} = 0.055.$$

  By contrast, the truthful-choice likelihood remains anchored to the utilities of the original items and is not relabelled together with the strategic reports. The truthful-likelihood drift therefore changes from $\gamma = 0.160$ for rule 0 to $\gamma = -0.100$ for rule 2. Thus, two reporting rules can have the same strategic separation while inducing different misspecification drift. Rule 2 has false-type dominance 0.83 under the truthful-likelihood interpretation and 0.15 under the strategic-likelihood interpretation.

- **Type-dependent reporting with negative drift and vanishing uniform separation (rule** 56**).** Rule 56 is another type-dependent solution of the same fixed-point problem. Its first-round reporting distribution is uniform for both types. Three of its four second-round states are also uniform. At the state reached after first reporting a and then receiving slate $\{g_0\}$, with coordinates ordered as (a, b, g_0), the reporting probabilities are

$$\sigma_2(\cdot \mid a, \{g_0\}, \eta_0) = (0.114077, 0.442961, 0.442961),$$
$$\sigma_2(\cdot \mid a, \{g_0\}, \zeta) = (0.247766, 0.376117, 0.376117).$$

  At the corresponding state following first-round report b, they are

$$\sigma_2(\cdot \mid b, \{g_0\}, \eta_0) = (0.442961, 0.114077, 0.442961),$$
$$\sigma_2(\cdot \mid b, \{g_0\}, \zeta) = (0.376117, 0.247766, 0.376117).$$

  Thus, unlike rule 64, rule 56 is type-dependent even though its first-round reports are uninformative.

  Its cumulative drift is $\gamma = -0.106$, so the positive accumulated advantage toward a false type required by Theorem 4.1 is absent. Its strategic separation constant is $\gamma_\star = 0$ at the 10^{-9} numerical gate, so the uniform cumulative-separation condition of Theorem 4.2 also fails. Importantly, $\gamma_\star = 0$ does not imply that the reporting rule is type-independent. Rather, it means that the positive uniform separation required by the theorem cannot be certified over all relevant positive-probability histories. The state-dependent probabilities above make the distinction explicit. Despite satisfying neither sufficient learning condition, rule 56 still exhibits the empirical likelihood contrast: its false-type dominance is 0.74 under the truthful-likelihood interpretation and 0.28 under the strategic-likelihood interpretation.

Reproducibility of the representative rules. The complete reporting profiles for all 129 fixed-point solutions are stored as log-odds vectors, from which the state-by-state report distributions

Table 6: Representative belief-consistent reporting rules illustrating different learning behaviors. Rule numbers identify the corresponding fixed-point solutions. False-type dominance is estimated from 100 interactions per rule.

Rule	Behavior	γ	$\gamma_\star$	Dominance ($\widetilde{q}$)	Dominance ($q^\star$)
64	Type-independent reporting	-0.062	0.000	0.66	0.00
0	Positive drift and separation	$+0.160$	0.055	0.91	0.17
2	Report relabelling of rule 0	-0.100	0.055	0.83	0.15
56	Neither learning condition	-0.106	0.000	0.74	0.28

$\sigma_s(\cdot \mid A_s, q_s, B_s, \eta)$ are recovered deterministically using the fixed interaction-tree ordering described in Section I. The rule indices are likewise assigned by a deterministic canonical ordering of the reporting profiles, so indices 0, 2, 56, and 64 identify the representatives above reproducibly. The stored profiles therefore suffice to reconstruct each representative reporting rule exactly.

Finite-sample bounds at the three-round horizon. With $\alpha_{\mathrm{dom}} = 0.0291$ and $\rho = 0.1$, the deviation terms are $c_t(\rho) = c_t^\star(\rho) = 13.15$ for every reporting rule because α_{dom} is constant across the set. The largest accumulated truthful-likelihood drift and strategic-likelihood separation are $t\gamma = 0.4786$ and $t\gamma_\star = 0.1646$, respectively. Thus, the corresponding deviation term exceeds the accumulated evidence for every one of the 129 reporting rules, and neither finite-sample bound is numerically informative at $T = 3$. This is a property of the bounds at this short horizon, not of Monte Carlo precision: the deviation terms depend on the horizon and α_{dom} and do not decrease with additional simulated interactions.

Report-relabelling and stability diagnostics. Within each report-relabelling class, the cumulative strategic-likelihood separation $\gamma_\star$ is invariant up to numerical tolerance, whereas the truthful-likelihood drift γ varies by as much as 0.70. This asymmetry is structural. The relabelling permutes both strategic likelihoods together when computing $\gamma_\star$, whereas γ compares the strategic likelihood with the utility-anchored truthful-choice likelihood, which is not permuted. The empirical effect is also not driven by local stability under the fixed-point iteration: truthful-likelihood false-type dominance is 0.681 across the 21 locally stable reporting rules and 0.678 across the remaining 108.

L.6 Illustrating the Two Interpretations

We compare the two interpretations of strategic feedback in Theorems 4.1 and 4.2 on a controlled four-item, two-type, three-round instance. We construct 129 belief-consistent strategic reporting rules and simulate 100 interactions under each. For every interaction, the truthful-likelihood posterior $\widetilde{q}$ and strategic-likelihood posterior $q^\star$ are updated from the same realized feedback stream, so the comparison isolates how that feedback is interpreted. The instance and reporting-rule construction are described in Section G, and the computational procedure in Section I.

Across the resulting 12,900 interactions, the truthful-likelihood posterior ends favoring a false type in 67.8% of interactions, compared with 27.4% under the strategic-likelihood posterior. Moreover, the mean terminal posterior on the true type is higher under the strategic interpretation for every one of the 129 reporting rules, with a mean increase of 0.156. Figure 1 shows the comparison across reporting rules. Thus, accounting for the strategic origin of the same feedback substantially reduces incorrect inference.

These results illustrate the misspecification mechanism of Theorem 4.1, including for reporting rules that do not satisfy its sufficient cumulative-drift condition. They should not be read as a numerical demonstration of the concentration guarantee in Theorem 4.2: at the three-round horizon, the finite-sample bounds are not numerically informative. The qualitative advantage of the strategic interpretation also persists under the alternative continuation-value convention, which produces a separate set of 93 reporting rules (Section L.3).

M Additional Empirical Details

M.1 Numerical Tolerance

We adjudicate equality between benchmark values at 10^{-7}. This tolerance applies only to benchmark comparisons; equilibrium verification, fixed-point residuals, and the theorem-condition gates use the tighter tolerances described in Sections I and J.

Recall that $\mathrm{OPT}(\epsilon)$ is the optimal welfare-feasible assistant value under truthful feedback, $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ is the optimal value when the assistant cannot condition future recommendations on user feedback, and $\mathrm{OPT}_{\mathrm{nf}}^{\mathrm{det}}(\epsilon)$ further restricts the non-adaptive assistant to a deterministic policy. To compute these benchmarks, we first enumerate the attainable pairs of expected user-welfare loss and assistant value. For $\mathrm{OPT}(\epsilon)$ and $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$, randomization between policies is allowed, so the benchmark is obtained from the upper concave envelope of these attainable pairs, evaluated at the welfare budget ϵ. For $\mathrm{OPT}_{\mathrm{nf}}^{\mathrm{det}}(\epsilon)$, which excludes such randomization, we instead take the best deterministic attainable point satisfying the welfare constraint. The value of a welfare-feasible strategic assessment is computed independently by the occupation-measure linear program of Section I. Thus, a comparison such as a certified assessment attaining $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ compares quantities obtained by independent computations.

We calibrate the 10^{-7} benchmark-equality tolerance by independently recomputing these quantities using exact rational arithmetic on a sample of instances. In this check, we enumerate the horizon-reachable terminal menus, compute their expected user-welfare losses and assistant values as rational numbers rather than floating-point approximations, construct the upper concave hull of these points exactly, and evaluate the resulting benchmark at ϵ using rational arithmetic. This provides an exact numerical reference against which the floating-point benchmark and the independently computed strategic-assessment value can be compared. On the audited instances, the strategic-assessment value from the occupation-measure linear program agrees with this reference to within 3.331×10^{-16}, while the floating-point computation of $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ differs from it by at most 1.083×10^{-9}. The latter discrepancy occurs only when the welfare constraint binds and the benchmark requires interpolation between two attainable welfare-loss/value points; it is exactly zero on the audited instances where the constraint does not bind.

As a complementary check on the full experiment, across all 600 instances of Section M.3 the largest observed difference between the independently computed strategic-assessment value and $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ is 1.167×10^{-9}, with median 2.1×10^{-10}. We therefore use 10^{-7} as the benchmark-equality tolerance, approximately 86 times the largest observed discrepancy and rounded to a power of ten. This choice is conservative rather than fitted to the observed comparisons: 10^{-8} would already classify all 600 differences as numerical equality, and 10^{-7} remains far below the $[0, 1]$ scale of assistant values. We use the same benchmark-equality tolerance for synthetic and fitted instances. Equilibrium verification, fixed-point residuals, and the theorem-condition gates instead use a tolerance of 10^{-9}.

M.2 Learning Experiment: Reporting-Rule Population and Sensitivity

The main learning experiment uses the 129 belief-consistent reporting rules in the discovered report-relabelling-closed set described in Appendices G.5 and I. Numerical search first identifies belief-consistent reporting rules, after which we apply the exact report relabellings that preserve the fixed-point problem. Rules connected by these relabellings form *report-relabelling classes*, regardless of whether a particular rule was found directly by the numerical search or added during closure. The resulting set contains 129 reporting rules in 12 report-relabelling classes. Closure is exact within the classes reached by the numerical search, but does not establish that every disconnected family of belief-consistent reporting rules has been found. The 129 rules therefore form a discovered set, not a random sample from a population of strategic behaviors, and we make no prevalence claim from their relative frequencies.

The unit of aggregation is the reporting rule: trajectories are first summarized within each rule and the resulting rule-level quantities are averaged with equal weight. Each rule contributes exactly 100 trajectories, so the equal-rule average coincides with the trajectory-pooled estimate. Equal weighting across the 12 report-relabelling classes is reported separately as a robustness check in Section L.1.

We also examine two changes to the construction of the reporting rules. First, under the alternative continuation-value convention of Section L.3, the belief-consistency problem yields a separate discovered population of 93 reporting rules. Because the convention changes the fixed-point problem, this population is analyzed separately and is never pooled with the main one. Second, we repeat the experiment at two additional signalling temperatures, again solving separately for the corresponding reporting rules. The qualitative ordering of the truthful- and strategic-likelihood posteriors persists under both changes.

Across the 129 reporting rules in the main experiment, the rule-level mean difference in terminal true-type posterior, $q_T^\star(\eta_0) - \widetilde{q}_T(\eta_0)$, has mean 0.1557 and median 0.1453. The difference is positive for every reporting rule; the main text reports the median.

M.3 Verification of the Babbling Equilibrium across Instances

Theorem 5.1 guarantees that whenever the non-adaptive problem is feasible, there exists a welfare-feasible strategic assessment with type-independent reports and assistant value $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. We test the construction directly on a large synthetic collection by constructing the corresponding babbling assessment and submitting it to the equilibrium certifier rather than assuming that the construction succeeds numerically.

Configuration. We consider $|X| \in \{6, 8, 10\}$ crossed with $K \in \{3, 5\}$, with $T = 2$, $k = |X|$, and $\epsilon = 0.1$. For each of the six configurations, we generate 100 independent instances from the permutation-utility family of Appendix G.6, for 600 instances in total. The permutation construction gives strict preferences by construction, which we additionally verify on every instance, while $k = |X|$ ensures Assumption (A2).

Result. The babbling assessment is certified on all 600 instances against the four conditions of Definition 5.1. Its first-round report has exactly zero mutual information with the user's type, the assistant's posterior remains at the prior, and reports do not affect subsequent slates. That is, the first report carries no information about the user's type. Its assistant value equals the best non-adaptive value $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ within the numerical tolerance of Appendix M.1. No instance times out or triggers a size guard. Figure 4(a) shows the value residuals, with one point per instance grouped by $(|X|, K)$.

We do not report the fraction of other candidate assessments that verify. Theorem 5.1 is an existence result whenever the non-adaptive problem is feasible, so failure of an unrelated heuristic candidate to verify would not provide evidence against that existence claim.

M.4 Dissociation between Informativeness and Equilibrium Value

The babbling result establishes that completely uninformative feedback can already attain the best non-adaptive value $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$, but it does not determine how informativeness and assistant value vary across equilibria of the *same* instance. We therefore look for instances admitting certified assessments with different levels of report informativeness but the same assistant value. This isolates whether informativeness and equilibrium value can vary independently without relying on comparisons across different instances.

We measure first-round informativeness by the normalized mutual information between the user's type η and first-round report Y_1 under the assessment:

$$\hat{I} := \frac{I(\eta; Y_1)}{H(\eta)} = \frac{1}{H(\eta)} \sum_{\eta, y} \Pr(\eta, Y_1 = y) \log \frac{\Pr(\eta, Y_1 = y)}{\Pr(\eta) \Pr(Y_1 = y)}. \tag{4}$$

Thus, $\hat{I} \in [0, 1]$: $\hat{I} = 0$ means that the first report carries no information about the user's type, while $\hat{I} = 1$ means that it identifies the type completely. The probabilities are those induced by the prior and the certified assessment.

Configuration. We use three instances with $|X| = 10$, $K = 3$, $T = 2$, $k = |X|$, and $\epsilon = 0.1$, drawn from the permutation-utility family of Appendix G.6. For each instance, we run an exact alternating best-response search from three initializations. The assistant step computes an exact welfare-constrained best response by an occupation-measure linear program, and the user step computes an exact sequential best response. Every initialization terminates at a period-1 fixed

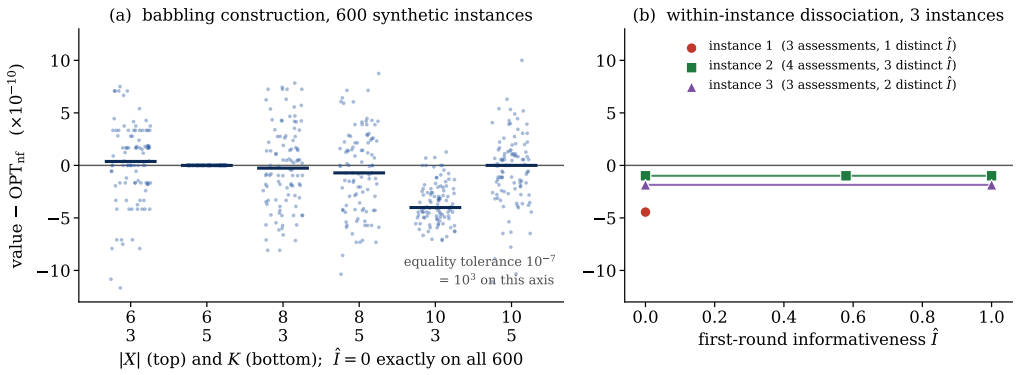

Figure 4: Informativeness does not track equilibrium value. **(a)** The babbling construction on 600 synthetic instances, one point per instance and grouped by $(|X|, K)$; $\hat{I} = 0$ exactly on all 600, and the assistant value equals $\mathrm{OPT}_{\mathrm{nf}}$ within 10^{-7}, which is 10^3 on the plotted axis. Horizontal bars show cell medians. **(b)** The separate three-instance dissociation study, showing the 10 certified assessments joined within each instance. Informativeness ranges from 0 to 1 while assistant value remains unchanged. Panel (b) is existential and supports no prevalence claim; none of its 10 assessments is outcome-determinate. The two panels represent different populations and are never pooled.

point, and none reaches the iteration limit. The procedure generates 16 assessments, of which 11 are distinct after exact deduplication and 10 satisfy the equilibrium verification conditions. The single rejected assessment fails user sequential optimality and no other condition. Informativeness is computed only after verification and is never used to generate or select assessments.

Result. Two of the three instances admit certified assessments with different informativeness but identical assistant value. On one instance, first-round informativeness $\hat{I}$ takes the values 0, 0.579, and 1 across assessments attaining the same value. Thus, the same instance admits completely uninformative, partially informative, and fully informative feedback without a corresponding difference in assistant value. Figure 4(b) shows all 10 certified assessments, joined within each instance.

Informativeness is therefore not by itself a proxy for equilibrium value: the two quantities can be separated within a fixed instance, rather than only across instances or through the worst-equilibrium construction of Theorem 5.1. This complements the value-collapse analysis of Section 5 rather than testing it. The two panels of Figure 4 should accordingly be read separately: panel (a) is a verification study over 600 independently generated instances, whereas panel (b) is an existential study over three instances. The two populations are never pooled.

The dissociation study above is existential: it demonstrates that informativeness and equilibrium value can differ independently, but does not estimate how frequently this occurs. All 10 certified assessments were also checked against Definition F.6, and none is outcome-determinate. In each case, the welfare-constrained assistant optimum randomizes at a reachable state, allowing a single type to reach two distinct terminal items. The study therefore provides no empirical instance of Theorem 5.3, whose outcome-determinacy hypothesis is not satisfied, and we make no such claim. The sample size of three was not predeclared: the three instances were fixed in advance as a validation stage, and the study ended after that stage once its existential objective had been met.

M.5 Exact Truthful-Strategic comparisons and a Descriptive Random Family

The exact constructions summarized in Table 7 illustrate three distinct relationships between truthful and strategic feedback: truthful feedback can outperform the worst strategic equilibrium, match the best strategic equilibrium, or be outperformed by a strategic equilibrium. These are exact examples rather than estimates from simulated interactions. Their utilities, welfare constraints, and resulting values are evaluated in rational arithmetic and verified against the corresponding closed-form constructions.

Table 7: Three distinct truthful–strategic comparisons on exact constructions. All values are evaluated in rational arithmetic. The first row uses Theorem 5.1, which identifies the worst-equilibrium value with the non-adaptive benchmark. The other two rows exhibit particular closed-form equilibria. The value $11/20$ in the final row is therefore the value of an exhibited equilibrium, not a computed maximum over all strategic assessments.

Construction	Comparison	Quantities	Established by
Decoy-anchor, 6 budgets	worst equilibrium $<$ truthful	$\mathrm{OPT}_{\mathrm{nf}}(\epsilon) < \mathrm{OPT}(\epsilon)$; $\mathrm{PoTF}(\epsilon) \in [1.56, 13.01]$	Theorem 5.1: $\mathrm{OPT}^{\min}_{\mathrm{strat}}(\epsilon) = \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$
Proposition F.5, $\epsilon = 1/5$	best equilibrium $=$ truthful	$\mathrm{OPT}^{\max}_{\mathrm{strat}}(\epsilon) = \mathrm{OPT}(\epsilon) = 1/4$; $\mathrm{OPT}_{\mathrm{nf}}(\epsilon) = 2/9$	exact closed-form equilibrium
Proposition F.6, $\epsilon = 11/80$	an equilibrium $>$ truthful	equilibrium value $= 11/20$; $\mathrm{OPT}(\epsilon) = \mathrm{OPT}_{\mathrm{nf}}(\epsilon) = 29/60$	exact closed-form equilibrium

Recall that $\mathrm{OPT}(\epsilon)$ is the optimal welfare-feasible assistant value under truthful feedback, while $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ is the optimal value when the assistant cannot condition future recommendations on user feedback. The price of truthful feedback, $\mathrm{PoTF}(\epsilon) = \mathrm{OPT}(\epsilon)/\,\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$, measures the multiplicative value of truthful adaptation relative to this non-adaptive benchmark. For the decoy-anchor construction, we evaluate the six welfare budgets $\epsilon \in \left\{\frac{1}{50}, \frac{1}{20}, \frac{1}{10}, \frac{3}{20}, \frac{1}{5}, \frac{1}{4}\right\}$, which give $\mathrm{PoTF}(\epsilon)$ values 13.005, 5.513, 3.025, 2.204, 1.800, and 1.563, respectively.

We complement these exact constructions with a descriptive random family from Appendix G.6. We draw 200 instances at each of the same six welfare budgets, with $|X| = 6$, $K = 3$, $T = 2$, and $k = |X|$, for 1,200 instances in total. At the benchmark-equality tolerance of Appendix M.1, 372 instances satisfy $\mathrm{OPT}(\epsilon) > \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$. Within this subset, the additive gap has median 0.022 and maximum 0.133, while $\mathrm{PoTF}(\epsilon)$ has median 1.034 and maximum 1.429.

This random family is descriptive and should not be read as estimating the frequency of the three strategic regimes in Table 7. In particular, it compares the truthful and non-adaptive benchmarks and does not establish that a strategic equilibrium exceeds truthful feedback. The three rows of Table 7 are instead established by the exact constructions and theoretical arguments stated there.

M.6 Additional Coat Results

The Coat instantiation of Appendix H.2 contains 40 fitted instances, split evenly between the two draw labels. Recall that $\mathrm{OPT}(\epsilon)$ is the optimal welfare-feasible value under truthful feedback and $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ the corresponding non-adaptive benchmark. Under the two draw labels separately, $\mathrm{OPT}(\epsilon) > \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ on 12 of 20 instances in one group and 17 of 20 in the other. The remaining instances satisfy equality at the tolerance of Appendix M.1, and none has $\mathrm{OPT}(\epsilon) < \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$.

The deterministic non-adaptive benchmark $\mathrm{OPT}^{\mathrm{det}}_{\mathrm{nf}}(\epsilon)$, which additionally excludes randomization over non-adaptive policies, lies below $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ on all 40 instances. Its median shortfall is 0.062, with range 0.006 to 0.177. We report this difference descriptively only: the theoretical argument that would turn it into an existence conclusion requires Assumption (A1), which fails on every Coat instance. Additional diagnostics for the public-randomization class did not yield a welfare-feasible comparison that could support a further result and are therefore not reported as empirical findings.

M.7 Additional SUSHI Results

The 20 SUSHI instances of Appendix H.3 share the same ten-item candidate pool and differ only in their fitted type structure. Consequently, the variation reported here reflects variation in the fitted type distributions and utilities at a fixed item set, not variation in the candidate pool.

The truthful benchmark $\mathrm{OPT}(\epsilon)$ exceeds the non-adaptive benchmark $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ on all 20 instances at the tolerance of Appendix M.1. The additive gap has median 0.0077 and range $[0.0060, 0.0084]$; the median values of $\mathrm{OPT}(\epsilon)$ and $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ are 0.719 and 0.712, respectively, and both benchmarks take distinct values across all 20 fitted instances.

Table 8: Benchmark results on the 40 Coat instances ($|X| = 12$, $K = 5$, $T = 2$, $k = 3$, $\epsilon = 0.1$). Values are medians with ranges across instances, and equality counts use the 10^{-7} benchmark-equality tolerance of Appendix M.1. The preference structure is estimated from Coat, but the strategic interaction is simulated; the respondents are not observed strategic users. Both (A1) and (A2) fail on every instance, so results requiring either assumption are not applied.

Quantity	Median [min, max] over 40 instances
Truthful constrained value $\mathrm{OPT}(\epsilon)$	$0.498\ [0.410,\ 0.621]$
Non-adaptive benchmark $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$	$0.493\ [0.394,\ 0.621]$
Additive gap $\mathrm{OPT}(\epsilon) - \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$	$0.010\ [0.000,\ 0.060]$
Instances with $\mathrm{OPT}(\epsilon) > \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$	$29/40$
Instances with $\mathrm{OPT}(\epsilon) = \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$	$11/40$
Instances with $\mathrm{OPT}(\epsilon) < \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$	$0/40$
Certified babbling assessment	$40/40$
Its first-round informativeness $\hat{I}$	0 on all 40
(A1) holds	$0/40$
(A2) holds	$0/40$

Exactly one assessment from the declared candidate family is certified on each instance. In every case, it is the babbling assessment of Theorem 5.1. Its value differs from $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ by at most 1.3×10^{-10}, its first-round informativeness is at most 1.5×10^{-16} (recall (4)), and it is welfare-feasible on every instance. The deterministic non-adaptive benchmark $\mathrm{OPT}_{\mathrm{nf}}^{\mathrm{det}}(\epsilon)$ lies below $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ on all 20 instances, with median shortfall 0.1038 and range $[0.1033, 0.1054]$. As for Coat, we report this relation descriptively and draw no existence conclusion from it.

Assumption (A1) holds on all 20 SUSHI instances, whereas Assumption (A2) holds on none. Exact optimization over the public-randomization class was attempted on every instance but was computationally unavailable at this configuration. We therefore attach no equality or strict comparison involving the public-class maximum. The 20 runs took 0.55 hours in total, with a median runtime of 98 seconds per instance.

M.8 Comparison across Real-Data Instances

The most consistent finding across the two real-data instantiations is the babbling result. On every Coat and SUSHI instance, we verify the type-independent assessment of Theorem 5.1. Its assistant value equals the non-adaptive benchmark $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ up to numerical precision, and its first-round informativeness is essentially zero. Thus, the existence result underlying the worst-equilibrium comparison is recovered on every fitted instance, despite the differences between the two datasets and their assumption profiles.

The value of truthful adaptation relative to the non-adaptive benchmark differs more across the fitted instances. Recall that $\mathrm{OPT}(\epsilon)$ is the optimal welfare-feasible assistant value under truthful feedback, whereas $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ is the optimal value when future recommendations cannot depend on user feedback. On SUSHI, $\mathrm{OPT}(\epsilon) > \mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ on all 20 instances, with a median additive gap of 0.008 and range $[0.006, 0.008]$. On Coat, the inequality holds on 29 of 40 instances, with a median gap of 0.017 among those 29 and a maximum of 0.060; the two benchmarks are equal on the remaining 11 instances. These spreads describe different sources of variation: the SUSHI instances share one candidate pool and vary in their fitted type structure, whereas the Coat instances also vary in their candidate pools.

Finally, the deterministic non-adaptive benchmark $\mathrm{OPT}_{\mathrm{nf}}^{\mathrm{det}}(\epsilon)$, which excludes randomization over non-adaptive policies, lies below $\mathrm{OPT}_{\mathrm{nf}}(\epsilon)$ on every instance of both datasets. The median shortfall is 0.104 on SUSHI and 0.062 on Coat. We report this relation descriptively only and draw no existence conclusion from it, because the assumptions needed for the corresponding theoretical implication are not satisfied uniformly across these instances.

M.9 Computational Scalability

This study measures how the computational cost of the empirical pipeline changes with the principal dimensions of an instance. It evaluates the implementation on a fixed grid of synthetic instances and makes no claim about asymptotic complexity or about the hypotheses of the theoretical results.

Instances and design. All instances are drawn from the permutation-utility family of Appendix G.6; no fitted or real-data instance is used. The baseline configuration is $|X| = 10$, $K = 3$, $T = 2$, $k = 3$, and $\epsilon = 0.1$. Starting from this configuration, we vary one dimension at a time: catalogue size over $\{6, 8, 10, 12, 14, 16, 18, 20\}$ with $k = 3$; the number of types over $\{2, 3, 4, 5, 6, 8\}$; the horizon over $\{2, 3, 4\}$; and slate size over $\{1, 2, 3, 5, 8, 10\}$ with $|X| = 10$. We additionally study their interaction by setting $|X| = k \in \{6, 8, 10, 12, 14\}$. This gives 28 configurations and three independently generated instances per configuration, for 84 runs in total. Every run is single-threaded, including the underlying numerical libraries.

Computational objects and censoring. We time five components separately: computation of the truthful and non-adaptive benchmarks; construction of the deviation-closed domain; enumeration of information sets; computation of the user's sequential best response; and verification of the resulting assessment. Each component has its own one-hour time limit. If a component reaches that limit, we record it as *censored* rather than treating one hour as its measured runtime. When we report a total time for a configuration containing a censored component, that component contributes its one-hour cap, so the resulting total is a lower bound.

Overall completion. All 84 runs execute without failure or size-guard termination. Of the 420 timed components, 414 complete and 6 are censored. All six censored measurements occur at $|X| = k = 14$: on each of the three instances, the user best response and assessment verification reach their one-hour limits. Every component completes at every other tested configuration.

Scaling in individual dimensions. At fixed slate size $k = 3$, increasing catalogue size from $|X| = 6$ to 20 raises the median total runtime from 0.7 seconds to 2,506.8 seconds. Intermediate values are 22.0 seconds at $|X| = 10$ and 222.1 seconds at $|X| = 14$. All components complete throughout this sweep. Over the same range, the number of states in the deviation-closed domain grows from 315 to 399,757.

The other one-dimensional sweeps separate the remaining effects. Increasing the number of types from $K = 2$ to 8 raises median runtime from 10.4 to 98.7 seconds, making it the mildest dimension tested. Increasing the horizon from $T = 2$ to 4 raises it from 19.8 to 320.1 seconds. Increasing slate size at $|X| = 10$ has a steeper initial effect: median runtime rises from 0.1 seconds at $k = 1$ to 114.0 seconds at $k = 5$, before leveling off at 139.7 seconds for $k = 8$ and 140.3 seconds for $k = 10$. This leveling occurs because, once k approaches the catalogue size, increasing the slate-size limit can no longer substantially enlarge the set of feasible slates.

Interaction between catalogue and slate size. The strongest scaling occurs when catalogue and slate size increase together. Along the $|X| = k$ sweep, median total runtime rises from 0.7 seconds at $|X| = k = 6$ to 140.3 seconds at 10 and 1,641.8 seconds at 12. At $|X| = k = 14$, the recorded total is at least 11,085.8 seconds because the user best response and assessment verification are both censored at one hour. This is the only tested configuration at which any component reaches its time limit.

The joint effect is substantially larger than increasing either dimension alone. A 14-item catalogue with $k = 3$ takes 222.1 seconds, while a 10-item catalogue with $k = 10$ takes 140.3 seconds; increasing both to $|X| = k = 14$ is the first configuration to encounter censoring. This comparison identifies an interaction between catalogue and slate size in the present implementation, not an impossibility result for the underlying computational problem.

Computational bottlenecks. The same components dominate across the sweeps. At the largest fully completed configuration in each of the four one-dimensional sweeps, assessment verification accounts for roughly half of total runtime, followed by the user best response at roughly one quarter to one third. These are also the only two components censored anywhere in the grid. Benchmark computation, construction of the deviation-closed domain, and information-set enumeration do not become the dominant component at the tested configurations.

Every configuration used elsewhere in the paper lies within the region in which all timed components complete: the largest catalogue used there contains 12 items at slate size 3, and the largest configuration with $k = |X|$ contains 10 items. The scaling study therefore identifies assessment verification and user best-response computation, particularly when catalogue and slate size grow together, as the practical computational bottlenecks for the present implementation.

These are measured runtimes on the stated grid and hardware, and we do not extrapolate them into an asymptotic complexity claim. A censored computation means only that it did not finish within its one-hour limit, not that it cannot be computed. Peak resident memory was also recorded as a process-level high-water mark, but we do not use it as a scaling measure.