LEARN THE USER, SERVE THEM BETTER: FROM USER MODELS TO PERSONALIZED LLM DECISIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

An online personalized LLM learns from user feedback. While pairwise questions reveal which response a user prefers, absolute acceptability questions ask whether a response is good enough. With a limited feedback budget, our proposed DUAL (Dual-feedback User-Adaptive Learning) learns each user's preferences and acceptance threshold to decide whether to serve either response or abstain. An established information-gain score selects the question type, while a calibrated cutoff and quota completion determine when to ask. Under independent uniform candidate features, a correctly specified logistic user with moderate true logits, and continuing answer quotas, we prove that DUAL's original recursive estimates converge almost surely to the true preference weights and acceptance threshold. We also characterize information coverage and bound the online estimate's distance to a joint fit. A controlled development simulation isolates question choice: at identical question times and 940 answers, information selection achieves 15.85% area under the serving-mistake learning curve (AUC), versus 16.45% for a separately calibrated fixed allocation. A common-calibration development study compares complete policies at 940 answers: DUAL achieves 14.735% AUC versus 15.206% for adaptive acceptability-only questioning. These simulations use PersonalLLM features and controlled acceptability feedback. Exploratory PRISM results leave the benefit of mixing feedback after population initialization unresolved. Community Alignment supports personal choice-probability prediction and information-based question selection over entropy. The complete policy's human serving benefit remains open.

1 INTRODUCTION

A user may prefer response A to B but find both unsuitable. Pairwise questions reveal their ordering, while absolute acceptability questions ask whether a response meets their standard. The latter can inform both preferences and acceptability, helping the system decide whether to serve either response or abstain.

Because frequent feedback requests can disrupt the user experience, the system must decide which question to ask and when to ask it. Does combining absolute acceptability and pairwise preference feedback improve serving over spending the same answer budget on acceptability questions? We study this question through *Dual-feedback User-Adaptive Learning* (DUAL), an online policy that uses a shared user belief to allocate feedback and decide whether to serve response A, serve response B, or abstain from serving either response. Figure 1 summarizes the interaction, with serving committed before feedback.

DUAL maintains a Gaussian belief over preference weights and an unknown acceptance threshold. Existing optimal-design methods study absolute scores and rankings of item lists (Mukherjee et al., 2024). Here, absolute feedback is a binary judgment about whether one response is good enough. DUAL chooses between that judgment and a preference comparison on the current pair, places questions within fixed answer quotas, and uses the same belief to serve or abstain. Its established local D-optimal score requires no serving target bank (Lindley, 1956; Chaloner & Verdinelli, 1995). The paper makes three contributions.

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.

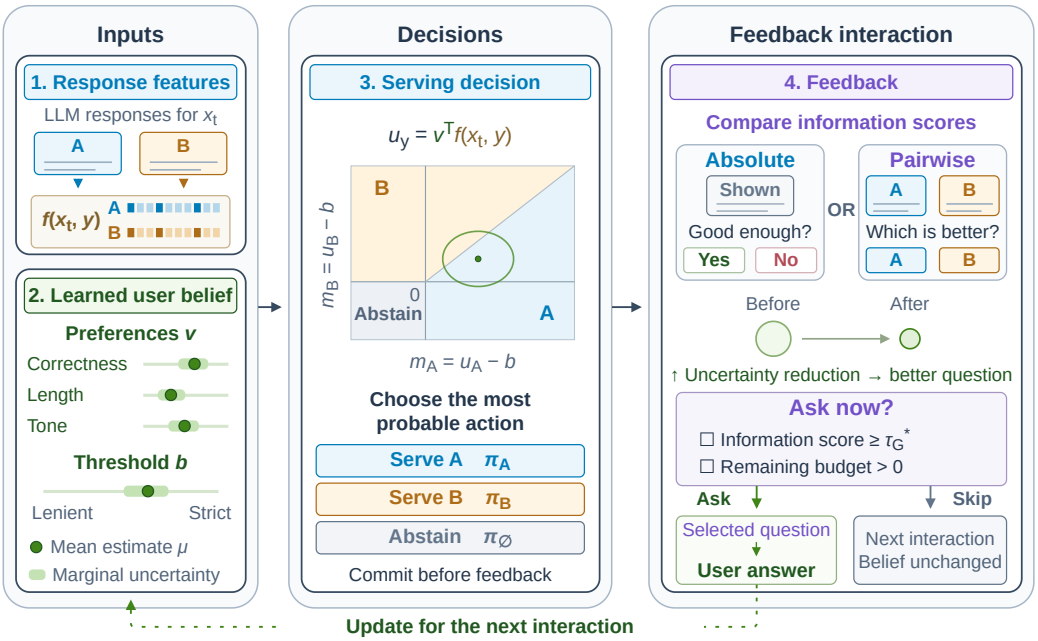


Figure 1: DUAL's online interaction loop. Steps 1–2 supply response features and a user belief. Step 3 integrates belief over the action regions and commits the most probable action. The dot marks the projected mean. Step 4 compares local approximate information scores before asking or skipping. Feedback can follow abstention. Answers update the belief, while skipping preserves it. *Quota completion can also trigger a question below the cutoff (Procedure 1). Feature labels and predicted uncertainty changes are schematic. Native features are reward-model scores.

1. **An acceptance-aware streaming policy.** We formulate binary acceptability and pairwise question selection on the current response pair, with an unknown acceptance threshold, fixed answer quotas and serving committed before feedback.
2. **Convergence of the original updates.** Under an explicit stochastic feature and feedback model with continuing answer quotas, we prove almost-sure convergence of the original recursion to the true user parameters (Theorem 3). The proof controls estimation error before establishing information growth. We also characterize information coverage and give a checkable bound on the online estimator's distance to a joint fit.
3. **Controlled and component evaluations.** A fixed-time simulation isolates adaptive question choice and supports it over calibrated fixed allocations. Complete-policy simulations assess timing, costs and representation. PRISM tests feedback after population initialization, while Community Alignment tests personal choice prediction. The human-data studies assess distinct components.

2 Related Work

Preference learning and personalization. Pairwise models underpin reward modeling and RLHF (Bradley & Terry, 1952; Christiano et al., 2017; Ouyang et al., 2022). Pluralistic and personalized alignment represents heterogeneous values (Kirk et al., 2024a; Sorensen et al., 2024). PRISM collects participant-specific feedback (Kirk et al., 2024b), while PersonalLLM constructs persistent reward-model-induced profiles (Zollo et al., 2025). Personalized RLHF and reward factorization learn compact user models (Li et al., 2024; Shenfeld et al., 2025). We keep response features fixed and study feedback for learning both preference and acceptability.

Bayesian design and active preference elicitation. Information acquisition follows classical Bayesian design (Lindley, 1956; Chaloner & Verdinelli, 1995). BALD and preference elicitation

connect queries to posterior uncertainty and recommendations (Houlsby et al., 2011; Viappiani & Boutilier, 2020). Mukherjee et al. (2024) design queries over item lists to learn rankings from absolute scores or ranking feedback. DUAL instead learns an unknown acceptance threshold from binary acceptability feedback alongside preferences, with two available question types and serving committed before feedback. Label-plus-comparison active learning also has established query-complexity theory (Kane et al., 2017). Our setting uses noisy logistic feedback, streaming query availability and acceptance-aware serving. EPIG targets information about predictions under an input distribution (Bickford Smith et al., 2023), whereas our decision-aware comparator targets expected serving loss.

Timing, feedback budgets and abstention. Active learning and multi-source annotation address label allocation and costs (Lewis & Gale, 1994; Settles, 2009; Settles et al., 2008; Donmez & Carbonell, 2008; Yan et al., 2012). Value-of-information formulations also optimize stopping (Monahan, 1980). We study fixed answer quotas and unequal response costs. Unlike dueling bandits (Yue et al., 2009), our objective includes acceptability and incorrect abstention, with a separate feedback allowance. Selective prediction studies abstention (Chow, 1970; El-Yaniv & Wiener, 2010). Under cost-sensitive decision principles (Elkan, 2001), our loss penalizes both incorrect abstention and incorrect serving.

Recursive learning and approximation. The Gaussian update is a static logistic Kalman recursion (de Vilmarest & Wintenberger, 2021). Their global logistic result bounds cumulative risk for a curvature-modified update with independent observations. Theorem 3 instead proves almost-sure convergence of the original update under adaptive questioning, for correctly specified users, uniform candidates, moderate true logits and continuing quotas. Table 13 compares the updates, assumptions and guarantees. Generalized self-concordant analysis supplies deterministic localization bounds for logistic objectives (Bach, 2010; Sun & Tran-Dinh, 2019; Marteau-Ferey et al., 2019). We apply these bounds to the selected-feedback residual to check distance to a joint fit. The resulting refit rule is tested only in a separate pilot.

3 Problem Formulation

Learner and user state. DUAL's learner estimates a user model from response features and collected answers, while the user's true preferences and acceptance standard remain hidden. Figure 1 shows the features in step 1, the user belief in step 2, the serving decision in step 3, and feedback in step 4. At interaction t, prompt x_t has two candidate responses A_t and B_t. Each response y has a fixed feature vector $f(x_t, y) \in \mathbb{R}^d$ available to the learner. Our simulation uses the external PersonalLLM benchmark (Zollo et al., 2025), with $d = 10$ reward-model scores as features.

We model both (i) which response user i prefers through utility $u_i(x_t, y)$, and (ii) whether it meets the user's standard through an acceptance margin $m_i(x_t, y)$. The preference weights $v_i \in \mathbb{R}^d$ specify how much each feature contributes to utility, and $b_i \in \mathbb{R}$ is the minimum acceptable utility. Together they form the user state η_i:

$$\eta_i = [v_i; b_i], \qquad u_i(x_t, y) = v_i^\top f(x_t, y), \qquad m_i(x_t, y) = u_i(x_t, y) - b_i. \tag{1}$$

Here $^\top$ denotes transpose. The dot product is the user's weighted sum of feature scores. A response is acceptable when its margin is nonnegative.

Reference action and mistakes. Write $u_A = u_i(x_t, A_t)$, $u_B = u_i(x_t, B_t)$, and $m_A = u_A - b_i$, $m_B = u_B - b_i$. The reference action $a_i^*(t)$ serves the preferred acceptable response, or abstains if neither is acceptable:

$$a_i^*(t) = \begin{cases} \textsc{Abstain}, & \max(u_A, u_B) < b_i, \\ \textsc{Serve-A}, & u_A \geq b_i \text{ and } u_A \geq u_B, \\ \textsc{Serve-B}, & u_B \geq b_i \text{ and } u_B > u_A. \end{cases} \tag{2}$$

Equality at the threshold is acceptable, and acceptable utility ties go to A. Our zero–one loss assigns zero to the reference action and one to any other action. A mistake can therefore mean serving an unacceptable response, abstaining when an acceptable response exists, or choosing the inferior acceptable response. Equal penalties measure agreement with the reference, not equal harm. We also separate these errors and test unequal penalties (Appendices D.2 and E.3). PRISM additionally tests accepting either response on an acceptable tie.

Available feedback. An absolute question asks whether a displayed response y is acceptable, whereas a pairwise question asks whether A is preferred to B. The subscripts a and p denote these two channels. The binary answer $r_a = 1$ means acceptance, while $r_p = 1$ means preference for A. A zero denotes the opposite answer. Each question has a design vector z, the input to its logistic answer model. With $\sigma(s) = 1/(1 + e^{-s})$,

$$z_a = [f(x_t, y); -1], \qquad \Pr(r_a = 1 \mid \eta_i, z_a) = \sigma(\eta_i^\top z_a), \tag{3}$$

$$z_p = [f(x_t, A_t) - f(x_t, B_t); 0], \qquad \Pr(r_p = 1 \mid \eta_i, z_p) = \sigma(\eta_i^\top z_p). \tag{4}$$

The final entry -1 makes $\eta_i^\top z_a = u_i(x_t, y) - b_i$, so absolute feedback depends on the acceptance threshold. The zero in z_p removes the threshold, leaving $\eta_i^\top z_p = u_A - u_B$. Both channels use unit logistic temperature, so the utility scores enter σ without rescaling. Answers are stochastic under these probabilities. The noise scale is fixed rather than estimated from each user's answers.

User belief and serving probabilities. After interaction t, DUAL represents uncertainty about the user state by a Gaussian distribution $q_{i,t}(\eta) = \mathcal{N}(\mu_{i,t}, \Sigma_{i,t})$. Its mean $\mu_{i,t} \in \mathbb{R}^{d+1}$ is the current parameter estimate. The full covariance matrix $\Sigma_{i,t} \in \mathbb{R}^{(d+1)\times(d+1)}$ records uncertainty and dependence among parameters. Its inverse $\Lambda_{i,t} = \Sigma_{i,t}^{-1}$ is the precision matrix. The simulation starts from $\mu_{i,0} = 0$ and $\Lambda_{i,0} = 2I$, where I is the identity matrix. PRISM uses a population-fitted starting belief (Appendix E.2). Each possible user state implies one reference action. The belief probability of each action region is

$$\begin{aligned} \pi_A &= \Pr_q(m_A \geq 0, u_A - u_B \geq 0), \\ \pi_B &= \Pr_q(m_B \geq 0, u_B - u_A > 0), \\ \pi_\varnothing &= 1 - \pi_A - \pi_B. \end{aligned} \tag{5}$$

Step 3 of Figure 1 depicts these regions in the acceptance-margin plane. Both negative margins imply abstention. Otherwise, the reference response has the larger margin, with boundary ties resolved by Equation (2). DUAL chooses the most probable reference action using fixed tie-breaking rules. This integrates uncertainty over user states, rather than using only the mean.

4 DUAL: Mixed Feedback with Adaptive Timing

At interaction t, DUAL uses features of A_t, B_t and the belief $q_{t-1} = \mathcal{N}(\mu_{t-1}, \Sigma_{t-1})$ to commit a serving action (Figure 1, steps 1–3). Step 4 selects and optionally asks a question, then updates to q_t after the answer. We suppress the user index i. Main experiments use only these recursive updates. Certificate-triggered refitting is tested in a separate pilot (Appendix C.9).

4.1 Choosing a Question

After committing the serving action, DUAL constructs the two questions in Equations (3)–(4). The absolute question concerns the served response. After abstention, DUAL instead displays the response with the larger acceptance prediction at the mean μ_{t-1}, with ties going to A. The pairwise question displays both responses. Thus feedback can still be requested after abstention.

For either design $z \in \{z_a, z_p\}$, $p = \sigma(\mu_{t-1}^\top z)$ predicts answer one at the current mean. The Bernoulli variance $h = p(1 - p)$ is largest at $p = 0.5$. It is the logistic loss's curvature evaluated at that mean, which makes the following information score a local approximation:

$$G(z; \mu, \Sigma) = \tfrac{1}{2} \log(1 + h z^\top \Sigma z) \geq 0. \tag{6}$$

Here $\mu = \mu_{t-1}$ and $\Sigma = \Sigma_{t-1}$. The score combines answer variability h with parameter uncertainty along the queried direction, $z^\top \Sigma z$. This Bayesian D-optimal criterion measures the increase in log-determinant precision (Lindley, 1956; Chaloner & Verdinelli, 1995), using only the belief q_{t-1} and current responses A_t, B_t. The decision-aware comparator instead scores expected risk reduction over a serving target bank (Appendix A.4).

At equal projected variance, the score favors the less predictable answer. An absolute question can therefore inform both preference and threshold without always having the larger score. Appendix A.6 gives a numerical example.

4.2 Deciding when to spend the answer budget

Procedure 1 uses the selected score from Equation (6), the calibrated threshold τ_G, and the remaining allowance. Each interaction permits one question at most, costing one answer in either channel. This count does not measure human effort.

Procedure 1 (DUAL's timing rule)**.** Let K_t be the unused answer allowance in the current block and L_t the number of eligible interactions remaining, including this one.

1. If this interaction is ineligible or $K_t = 0$, skip feedback.
2. Otherwise, select $c_t \in \{a, p\}$ with the larger score and ask that question if its score reaches τ_G or $L_t = K_t$.
3. If asked, record its design $z_t = z_{c_t}$ and answer r_t, consume one allowance, and apply Equations (7)–(8). Otherwise, keep $q_t = q_{t-1}$.

For the confirmatory experiment, we chose 94 shared 100-interaction blocks with ten answers each, matching all questioning policies at evaluation boundaries. The 10% allowance was selected on development users as a diminishing-returns trade-off (Appendix A.4). DUAL uses $\tau_G \approx 0.0255$, frozen on 200 development profiles. Appendix A.1 specifies the exact threshold and comparison schedules. Interactions 9,401–9,402 are ineligible, leaving 940 answers in total. The condition $L_t = K_t$ completes the quota without future feedback, so adaptation changes question positions, not their total count. Blocks can permit bursts across boundaries, assessed in Section 6.

4.3 Updating the belief after feedback

Procedure 1 supplies z_t, r_t only when feedback is requested. With $p_t = \sigma(\mu_{t-1}^\top z_t)$, a quadratic approximation of the new logistic loss gives the Gaussian update (de Vilmarest & Wintenberger, 2021, Algorithm 1):

$$\Lambda_t = \Lambda_{t-1} + p_t(1 - p_t) z_t z_t^\top, \tag{7}$$

$$\mu_t = \mu_{t-1} + \Lambda_t^{-1}(r_t - p_t) z_t. \tag{8}$$

The added loss curvature increases precision, while the prediction error $r_t - p_t$ moves the mean along the design direction, scaled by the updated covariance. The resulting belief is used at the next interaction. The derivation and implementation checks appear in Appendix A.6.

5 Structural Analysis

We analyze information coverage and approximation to the joint fit (Equation (12)), then establish true-user convergence under an explicit stochastic model.

Channel geometry and acquisition. For design z and curvature h from Section 4, the local Fisher information $\mathcal{I}(\mu; z)$ is the question's contribution to precision at the current mean:

$$\Lambda^+ = \Lambda + \mathcal{I}(\mu; z), \qquad \mathcal{I}(\mu; z) = h z z^\top. \tag{9}$$

Since $\eta = [v; b]$, the threshold occupies the last coordinate, so $\mathcal{I}_{bb} = \mathcal{I}_{d+1,d+1}$ is the last diagonal entry. It is zero for a pairwise design and equals $h > 0$ for an absolute design at a finite logit. Pairwise feedback thus provides no direct threshold precision increment, although parameter dependence can allow an indirect threshold update. For this rank-one update, the matrix determinant lemma is

$$\frac{\det(\Lambda + h z z^\top)}{\det \Lambda} = 1 + h z^\top \Lambda^{-1} z. \tag{10}$$

Taking half the logarithm gives the D-optimal score G in Equation (6), the local Gaussian entropy reduction.

Coverage of the available questions. For the recursion without full-history refits, index only answered steps $t = 1, \ldots, n$. Initialize $\Lambda_0 = \lambda I$, with $\lambda > 0$ and I the identity matrix for $k = d + 1$ user parameters. For channel $c \in \{a, p\}$, write $\widetilde{z}_t(c) = \sqrt{h_t(c)} z_t(c)$ for its design scaled by the

square root of logistic curvature. Its outer product is the precision increment. Substitution into Equation (6) gives $G = \frac{1}{2}\log(1 + \widetilde{z}_t(c)^\top \Lambda_{t-1}^{-1}\widetilde{z}_t(c))$. The increasing logarithm makes maximizing G equivalent to maximizing the quadratic term.

Proposition 1 (Coverage inherited from the available questions)**.** *Suppose $\|\widetilde{z}_t(c)\|_2 \le U$ for some $U > 0$. Along the realized adaptive history, let $\bar{z}_t$ be a feasible scaled design at each answered step with $\sum_{t=1}^{n} \bar{z}_t \bar{z}_t^\top \succeq \kappa n I$, where $\kappa > 0$. Then greedy information-based acquisition gives*

$$\lambda_{\min}(\Lambda_n) \ge \frac{\kappa n}{(1 + U^2/\lambda)\, k\, \log(1 + nU^2/(k\lambda))}. \tag{11}$$

Here U bounds the scaled design length, and the matrix condition requires the feasible questions to supply at least κn information in every unit direction. The smallest eigenvalue $\lambda_{\min}(\Lambda_n)$ measures precision in the least-constrained direction. Its lower bound grows on the order of $n/\log n$ under this condition. The prompt stream and the rule deciding when to ask must supply that coverage. This controls model precision, not true-user accuracy. The development audit checks finite prefixes (Appendix C).

A checkable approximation certificate. To check the online mean in Equation (8), define the joint maximum a posteriori (MAP) fit to the n selected answers:

$$F_n(\eta) = \tfrac{\lambda}{2}\|\eta\|_2^2 + \sum_{t=1}^{n}[\log(1 + e^{z_t^\top \eta}) - r_t z_t^\top \eta], \qquad \widehat{\eta}_n = \arg\min_{\eta} F_n(\eta). \tag{12}$$

Write $\|x\|_H = (x^\top H x)^{1/2}$. At μ_n, define

$$g_n = \nabla F_n(\mu_n), \qquad H_n = \nabla^2 F_n(\mu_n),$$
$$\nu_n = (g_n^\top H_n^{-1} g_n)^{1/2}, \qquad \rho_n = \max_{t \le n}(z_t^\top H_n^{-1} z_t)^{1/2}, \qquad \chi_n = \rho_n \nu_n.$$

The gradient g_n vanishes at the joint-fit optimum. The full-history Hessian H_n need not equal the recursive precision Λ_n.

Proposition 2 (Recursive-to-MAP certificate)**.** *If $\chi_n < 1$,*

$$\|\mu_n - \widehat{\eta}_n\|_{H_n} \le R_n := -\log(1 - \chi_n)/\rho_n, \tag{13}$$
$$(1 - \chi_n)H_n \preceq \nabla^2 F_n(\widehat{\eta}_n) \preceq (1 - \chi_n)^{-1} H_n. \tag{14}$$

For $\rho_n = 0$, use the continuous limit $R_n = \nu_n$.

For $\rho_n > 0$, the condition compares the gradient size ν_n, measured using H_n^{-1}, with the curvature-variation scale $1/\rho_n$. When $\chi_n \ge 1$, this certificate supplies no bound, leaving approximation accuracy unresolved. The computable certificate applies generalized self-concordant localization bounds (Sun & Tran-Dinh, 2019; Marteau-Ferey et al., 2019) (Appendix C.3). Without refits, the full-history gradient sums signed logistic-update remainders (Equation (34)), preserving cancellations lost by separate worst-case bounds. Appendix C reports the development audit of this bound.

We now study whether updates (7)–(8) recover the true preference weights and acceptance threshold $\eta^\star$ as the number n of answered questions increases.

Theorem 3 (Convergence to the true user parameters)**.** *Suppose independent candidate features are uniform on $[-1, 1]^d$, feedback follows the correctly specified logistic model with fixed parameter $\eta^\star$, and every true question logit has magnitude at most $\operatorname{arcosh}(3)$. Under the conditions in Appendix C.4, run exact DUAL without refits indefinitely, with a fixed positive information cutoff and Q answers per L-interaction block, $1 \le Q \le L$. Almost surely,*

$$\mu_n \to \eta^\star, \qquad \lambda_{\min}(\Lambda_n) = \Omega_\omega(n/\log n), \qquad \|\mu_n - \widehat{\eta}_n\|_2 \to 0.$$

Proposition 2's certificate is eventually valid and its Euclidean error bound $R_n/\sqrt{\lambda_{\min}(H_n)}$ vanishes. If fresh pairs have nonzero true decision margins almost surely, both their conditional mistake probability and the realized average mistake rate tend to zero.

Here Ω_ω allows path-dependent positive constants, and the largest modeled variance is $O_\omega(\log n/n)$. *Proof sketch.* We first bound squared estimation error weighted by Λ_n, without assuming bounded online estimates. Quota completion and greedy question selection make minimum precision grow faster than this error, forcing $\mu_n \to \eta^\star$. Gradient bounds and Proposition 2 give a vanishing Euclidean bound to the joint fit, while belief concentration yields vanishing serving error under the stated margins. The complete proof is in Appendix C.4.

Separate refitting pilot. The certificate triggers 181 full refits, compared with 990 after every answer and 174 under the Newton screen. Both screens meet the 0.10-logit tolerance at every checked decision (Appendix C.9).

Finite-budget serving guarantees require valid statistical confidence and margins (Appendix C.1). We do not instantiate the experimental learner's confidence radius or use it to stop questioning.

6 Experiments

We test question choice, timing, costs and representation, followed by component studies on recorded human feedback. Table 4 separates their evidence roles.

6.1 Design and comparisons

We use PersonalLLM's 9,402 online prompts, separate 1,000-prompt bank, and ten reward-model features, standardized using training data. Each user–prompt pair receives a deterministic A/B selection from eight responses, frozen independently of outcomes and order. Policies share this pair map, prompt order, and compatible feedback randomness. We construct acceptance thresholds from training-utility quantiles at outcome-blind assigned acceptance rates $0.30, 0.50, 0.70$. A frozen random stream generates only the selected binary answer. These thresholds and labels are synthetic. Reference actions use latent utility even when answers are noisy (Appendix A). For bank error $e_i(t)$ over horizon T, average error during learning is

$$\mathrm{AUC}_i = T^{-1} \int_0^T e_i(t)\, dt, \tag{15}$$

where $e_i(t)$ linearly interpolates errors at interaction zero, every 100 interactions and the final horizon on a frozen 1,000-pair bank (Appendix A.3). The bank never updates the learner. Only the decision-aware development comparator sees its features, and no bank labels, utilities or metrics reach acquisition. Users have equal aggregate weight and form the bootstrap units. The 971-pair duplicate-excluded view is sensitivity only. We assess actual online mistakes separately.

The original study selects the representation, objective, timing and budget on 200 simulated development profiles, then evaluates the frozen 10% design once on 1,000 official held-out simulated profiles. Four planned contrasts use 20,000 paired-user bootstrap draws. A separate native-feature comparison uses 100 common calibration and 200 fresh simulated assessment profiles, averaging three realizations within each assessment user. Tunable policies share the calibration objective and five-candidate search budget. Thresholds and a three-phase schedule are frozen before assessment. Its six planned answer/exposure contrasts retain their joint multiplicity correction. Development choices and reproducibility details appear in Appendix A, with the native protocol in Appendix D.2.

6.2 Does mixing feedback improve serving?

Can a fixed mixture replace adaptive question choice? In a controlled development simulation, every policy receives 940 answers at identical positions over 9,402 interactions. AUC on 100 calibration profiles selects the fixed pairwise fraction from ten values, including both single-channel policies. On 200 further profiles, information selection achieves 15.85% AUC against 16.45% for the best fixed allocation, which uses only absolute questions (Table 1). Thus adapting the question type helps at fixed question times. Appendix D.5 gives the protocol.

Question choice	AUC (%)	Difference [95% CI]
Calibrated fixed allocation	16.45	—
Information-selected mixed	15.85	-0.60 $[-0.82, -0.38]$

Table 1: Development comparison at identical question times on 200 fresh simulated profiles, each with three realizations and 940 answers. The paired difference is information selection minus the calibrated fixed allocation, in AUC percentage points.

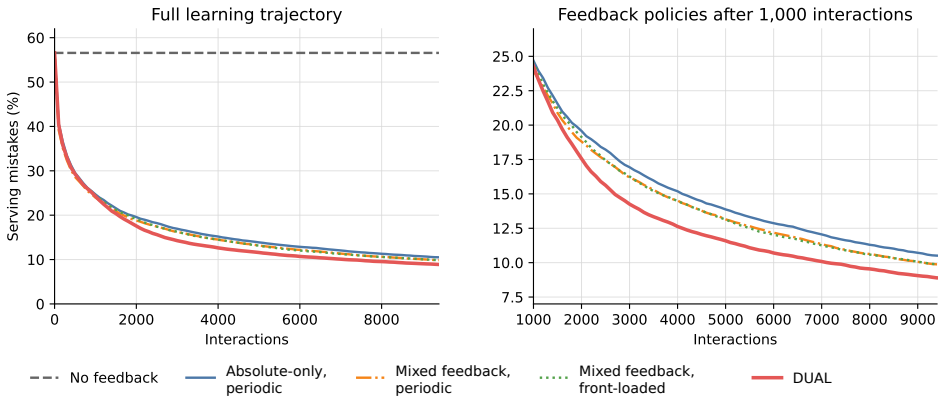


Figure 2: Evaluation-bank mistakes over 9,402 interactions with held-out simulated profiles. The right panel enlarges results after 1,000 interactions. Answer counts match at every checkpoint.

On 1,000 held-out simulated profiles, DUAL reduces AUC from 16.50% for periodic absolute feedback to 14.62% (Table 2). This complete-policy comparison combines channel choice and timing. Holding periodic positions fixed, information-based channel selection reaches 15.78%. All four planned contrasts survive multiplicity correction (Appendix B.6).

Policy	**Questions**	**AUC**	**Final**
No feedback	0	56.57%	56.57%
Absolute-only, periodic	940	16.50%	10.51%
Mixed feedback, periodic	940	15.78%	9.88%
Mixed feedback, front-loaded	940	15.85%	9.88%
DUAL	940	**14.62%**	**8.89%**

Table 2: Confirmatory results on 1,000 held-out simulated profiles with controlled acceptability feedback at the frozen 10% question budget. Front-loading occurs within each 100-interaction block.

Comparison with adaptive absolute-only questioning. In the separate common-calibration development study, both policies adapt question timing and receive 940 answers. DUAL lowers AUC from 15.206% to 14.735%, a difference of -0.471 percentage points with adjusted 95% interval $[-0.756, -0.194]$ (Table 18). This uses the same calibration objective and search budget for both policies. Exploratory online-mistake analysis also favors DUAL (-0.513 points, $[-0.798, -0.236]$), including fewer wrong-response choices when at least one response is acceptable (Appendix D.2).

6.3 What does adaptive timing contribute?

With information-based channel selection fixed, DUAL improves on periodic and within-block front-loaded placement in the held-out study (Table 2, Figure 2). DUAL also improves on the separately calibrated three-phase schedule, but its full-horizon comparison with confidence timing is unresolved (Table 18).

During interactions 6,001–9,400 of the held-out study, 98.58% of questions are quota-forced (Table 8). The comparisons therefore assess information timing together with quota completion. A strict rolling-cap intervention and an exploratory smaller-budget study assess this scheduling choice further (Appendix B.5).

In the native comparison, confidence timing has lower mean bank error through 100 answers, while DUAL has lower means at 250, 500 and 940. These are prefixes of 940-answer policies, not tests designed for short histories. Against confidence timing, DUAL makes fewer false abstentions but more serves when neither response is acceptable. Their adjusted online-total comparison is unresolved (Appendix D.2).

6.4 How do costs and model mismatch change the result?

With one exposure per absolute question and two per pairwise question, DUAL's tested controller has 1.326 points higher AUC than absolute-only adaptive (Appendix D.2). A new 120-profile study separates cost-normalized scoring from quota completion. Information per exposure reduces the mixed controller's AUC from 16.33% to 15.38%, with adjusted difference -0.94 points $[-1.41, -0.44]$. Absolute-only has the lowest mean at 14.92% (Appendix D.3). These assumed costs do not measure human effort.

A hidden quadratic preference component reverses the frozen timing advantage. Recalibrating the threshold gives partial mean recovery, but does not establish restoration of the matched-model benefit under the broader multiplicity sensitivity (Appendices D.1 and D.4).

6.5 Which findings extend to recorded human feedback?

These retrospective component studies use separate learners and targets, without measuring live outcomes or feedback effort.

PRISM: decisions from a population-initialized user model. We fit a shared starting belief on 64 PRISM participants and select features and regularization on 89 development participants (Kirk et al., 2024b). All policies share 24 surface and embedding features. We evaluate 90 previously reported participants, each with ten history pairs and four test pairs. Updating policies request five answers at interactions $2, 4, 6, 8, 10$, using rating cutoff 70.

Mixed feedback reaches 58.14% AUC, compared with 58.89% without personal updates and 57.86% for absolute-only feedback (Table 3). Both differences remain unresolved. This exploratory follow-up reuses the assessment cohort (Appendix E.2). The original cold-start studies remain in Appendix E.3.

Feedback	Answers	AUC (%)	Mixed minus row [95% CI]
Population only	0	58.89	-0.75 $[-3.32, +1.81]$
Absolute-only	5	57.86	$+0.28$ $[-0.50, +1.06]$
Pairwise-only	5	60.69	-2.56 $[-5.47, +0.24]$
Mixed, information-selected	5	58.14	—

Table 3: Exploratory PRISM results. All policies share the starting belief and features. Differences are mixed minus row in AUC percentage points, with negative values favoring mixed feedback.

Community Alignment: preference learning and question choice. Ten personal answers improve choice log loss on 180 participants over no history and others' answers to identical questions (Zhang et al., 2025). On 160 further participants, information selection reduces five-answer log loss from 1.4004 under entropy selection to 1.3617. The difference survives multiplicity adjustment and shared-prompt resampling, while the random-selection comparison and correct-choice accuracy gains are unresolved. These findings support person-specific probability prediction and informative question selection in the categorical MAP extension (Appendix E.5).

7 Conclusion

DUAL learns preferences and acceptability to guide feedback and serving. Under the explicit stochastic conditions of Theorem 3, its original updates converge to the true user parameters. The finite-history certificate separately checks approximation to a joint fit. At fixed question times, adaptive question choice improves on a calibrated fixed allocation. Complete-policy simulations also favor DUAL over periodic and adaptive absolute-only questioning at equal answer counts. The advantage depends on feedback costs and representation.

With a population prior, PRISM leaves the additional five-answer gain and the benefit of mixing unresolved. Community Alignment supports personal histories and informative questions for choice prediction. The complete policy's human serving benefit remains open.

USE OF AI

We used ChatGPT and Codex to support the research and preparation of this manuscript. ChatGPT assisted with discussing the feasibility of research ideas, exploring proof strategies, identifying related literature, and suggesting improvements to writing and manuscript structure. The authors wrote the mathematical proofs and used ChatGPT for an additional mathematical check.

Codex assisted with code planning, refinement, edge-case handling, code review, and running and checking experiments. This included assistance with code for constructing dataset acceptance thresholds and generating synthetic users for small tests. Experimental design combined author-proposed choices with AI suggestions, which the authors evaluated before making the final decisions. Codex also helped refine manuscript passages initially drafted by the authors.

The authors reviewed code logic, inspected small experimental runs in detail before conducting larger experiments, and read the sources suggested by AI. AI assistance with experimental results consisted of reporting outputs and providing additional detail on request. The authors developed the interpretations and scientific explanations of those results. The authors take responsibility for the final manuscript, mathematical claims, implementation, and reported results.

REFERENCES

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010. doi: 10.1214/09-EJS521. URL https://doi.org/10.1214/09-EJS521.

Freddie Bickford Smith, Andreas Kirsch, Sebastian Farquhar, Yarin Gal, Adam Foster, and Tom Rainforth. Prediction-oriented Bayesian active learning. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7331–7348. PMLR, 2023. URL https://proceedings.mlr.press/v206/bickfordsmith23a.html.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995. doi: 10.1214/ss/1177009939.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

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

Joseph de Vilmarest and Olivier Wintenberger. Stochastic online optimization using Kalman recursion. *Journal of Machine Learning Research*, 22(223):1–55, 2021. URL https://jmlr.org/papers/v22/20-618.html.

Pinar Donmez and Jaime G. Carbonell. Proactive learning: Cost-sensitive active learning with multiple imperfect oracles. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pp. 619–628, 2008. doi: 10.1145/1458082.1458165.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Charles Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence*, 2001.

Louis Faury, Marc Abeille, Clement Calauzenes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3052–3060. PMLR, 2020. URL https://proceedings.mlr.press/v119/faury20a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning, 2011. URL https://arxiv.org/abs/1112.5745.

Daniel M. Kane, Shachar Lovett, Shay Moran, and Jiapeng Zhang. Active classification with comparison queries, 2017. URL https://arxiv.org/abs/1704.03564.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6:383–392, 2024a. doi: 10.1038/s42256-024-00820-y.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale.

The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 105236–105344, 2024b. doi: 10.52202/079017-3342. URL https://arxiv.org/abs/2404.16019.

David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3–12, 1994.

Xinyu Li, Ruiyang Zhou, Zachary C. Lipton, and Liu Leqi. Personalized language modeling from personalized human feedback, 2024. URL https://arxiv.org/abs/2402.05133.

Dennis V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956. doi: 10.1214/aoms/1177728069.

Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Globally convergent Newton methods for ill-conditioned generalized self-concordant losses. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1907.01771.

George E. Monahan. Optimal stopping in a partially observable Markov process with costly information. *Operations Research*, 28(6):1319–1334, 1980. doi: 10.1287/opre.28.6.1319.

Subhojyoti Mukherjee, Anusha Lalitha, Kousha Kalantari, Aniket Deshmukh, Ge Liu, Yifei Ma, and Branislav Kveton. Optimal design for human preference elicitation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 90132–90159, 2024. doi: 10.52202/079017-2861.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, 2009.

Burr Settles, Mark Craven, and Lewis Friedland. Active learning with real annotation costs. In *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*, pp. 1–10, 2008.

Idan Shenfeld, Felix Faltings, Pulkit Agrawal, and Aldo Pacchiano. Language model personalization via reward factorization. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=E7Tu5yjqXw.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46280–46302. PMLR, 2024. URL https://proceedings.mlr.press/v235/sorensen24a.html.

Tianxiao Sun and Quoc Tran-Dinh. Generalized self-concordant functions: A recipe for Newton-type methods. *Mathematical Programming*, 178:145–213, 2019. doi: 10.1007/s10107-018-1282-4. URL https://arxiv.org/abs/1703.04599.

Paolo Viappiani and Craig Boutilier. On the equivalence of optimal recommendation sets and myopically optimal query sets. *Artificial Intelligence*, 286:103328, 2020. doi: 10.1016/j.artint.2020.103328.

Yan Yan, Romer Rosales, Glenn Fung, Faisal Farooq, Bharat Rao, and Jennifer Dy. Active learning from multiple knowledge sources. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pp. 1350–1357. PMLR, 2012. URL https://proceedings.mlr.press/v22/yan12.html.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The K-armed dueling bandits problem. In *Proceedings of the Conference on Learning Theory*, 2009.

Lily Hong Zhang, Smitha Milli, Karen Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, Bhaktipriya Radharapu, Jane Yu, Vidya Sarma, Kris Rose, and Maximilian Nickel. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. *arXiv preprint arXiv:2507.09650*, 2025. URL https://arxiv.org/abs/2507.09650.

Thomas P. Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. PersonalLLM: Tailoring LLMs to individual preferences. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2409.20296.

A Additional Experimental Details

Study	Evidence role	Supported decision or finding
Original simulation	Held-out confirmation	DUAL improves serving under the matched simulator.
Native controls	Fresh-profile development	DUAL improves over absolute-only adaptive at equal answers. A separate fixed-time study favors adaptive question choice over a calibrated allocation.
Controller follow-ups	Fresh-profile development	Cost-aware scoring reduces the exposure loss. Timing recalibration gives partial mean recovery under the omitted feature.
Certificate pilot	Generated numerical study	A bound controls when to refit the same history. Newton performs similarly.
PRISM	Recorded-data component studies and exploratory reuse	The population model supplies most observed performance. Mixed versus absolute-only remains unresolved. Original cold-start studies are retained.
Community Alignment	Frozen recorded-data assessments	Personal history and question choice improve probability prediction in separate preference tasks.

Table 4: Each study tests a distinct question, with its own population, learner or evaluation task.

A.1 Reproducibility and evidence provenance

The protocols specify populations, calibration, feedback availability and evaluation before reporting each comparison. Independent replay and numerical checks reconstruct the recorded updates, predictions and statistical reports. These checks establish agreement with the declared computation, not generalization to other users.

The supplementary package's `REPRODUCIBILITY.md` and `PAPER_TO_CODE.md` describe the available workflows, retained implementations and result snapshots, including participant-level derived metrics. The export `MANIFEST.json` records file hashes. The guide distinguishes generated validation from authentic-data reproduction and identifies historical workflows that are not standalone.

The original held-out simulation is distinct from the later native development cohorts, generated refit pilot and recorded-human assessments (Table 4). Earlier studies with unequal calibration, learned-noise mixtures, changed serving targets or statistical stopping remain in the research archive and do not support claims about the fixed learner evaluated here.

The confirmatory timing threshold is $\tau_G = 0.025515678376729425$, frozen on the original 200 development profiles. This is the value stored in the experiment configuration. Section 4 rounds it for readability. Exact acquisition-score ties select the absolute question. Ties among the highest serving probabilities favor A, then B, then abstention. All confirmatory questioning policies use 94 consecutive, nonoverlapping blocks of 100 interactions, with ten answers per block. Periodic questioning uses positions $10, 20, \ldots, 100$, while front-loading uses $1, \ldots, 10$ within each block. Interactions 9,401–9,402 are ineligible. These shared blocks match answer counts at each 100-interaction evaluation boundary. The 10% budget choice is reported in Appendix A.4.

A.2 PersonalLLM construction

The online split contains 9,402 prompts and 75,216 candidate responses. The evaluation split contains 1,000 prompts and 8,000 candidates. Reward-model features are standardized using online-split means and standard deviations before applying a user's ten-dimensional profile. Official prompt identifiers and source roles are disjoint. Because 29 evaluation prompts are linked to duplicate text, the complete 1,000-prompt view is primary and a 971-prompt duplicate-excluded view is reported only as sensitivity.

The 200 development profiles are drawn from the frozen symmetric Dirichlet generator. The 1,000 confirmatory profiles are the official simulated PersonalLLM evaluation weight vectors and are not used for design selection. In both populations, the acceptance stratum is assigned by ordering a stable hash of user identity and then cycling through target rates 0.30, 0.50, and 0.70. For target rate

ρ_i, the threshold is

$$b_i = Q_{1-\rho_i}\left(\{v_i^\top f(x,y) : (x,y) \text{ is in the online split}\}\right),$$

using linear quantile interpolation. This makes the threshold reproducible and balances easy-, medium-, and high-acceptance users without using policy outcomes. Absolute and pairwise labels follow Equations (3)–(4) with unit temperature.

A.3 Checkpoint integration

Let $e_{i,\ell}$ be user i's evaluation-bank mistake rate at interaction t_ℓ. The reported normalized learning-curve area is

$$\mathrm{AUC}_i = \frac{1}{T}\sum_{\ell=1}^{L} \frac{e_{i,\ell-1} + e_{i,\ell}}{2}(t_\ell - t_{\ell-1}),$$
$$(t_0, \ldots, t_L) = (0, 100, 200, \ldots, 9400, 9402), \qquad T = 9402.$$

Thus Equation (15) integrates the piecewise-linear interpolation of checkpoint errors. The final interval has width two. This is evaluation-bank error averaged over learning time. Actual online mistakes are instead scored on each encountered pair before its feedback.

A.4 Development question-objective comparison

All updating policies in Table 5 receive the same 1,880 periodic question opportunities. Pairwise-only feedback provides no direct threshold information, which helps explain its poor performance. The comparison between information-based and decision-aware selection remains unresolved. We select information gain for confirmation because it has the best point estimate, improves over random channel selection, does not require the target bank, and is much cheaper to score.

Policy	Mistake AUC	Final mistake
No feedback	0.5644	0.5644
Absolute-only, periodic	0.1324	0.0799
Pairwise-only	0.2148	0.1792
Random channel	0.1407	0.0843
Mixed feedback, periodic	**0.1261**	**0.0746**
Decision-aware	0.1277	0.0748

Table 5: Question-selection policies with the same native-RM10 learner and periodic 20% feedback schedule.

Information-selected minus decision-aware AUC is -0.00166 with a paired 95% interval [-0.00394, 0.00061]. Information-selected minus random-channel AUC is -0.01467 [-0.01809, -0.01120]. The information-based selector averages 0.463 seconds per user and the decision-aware selector 98.46 seconds per user under the same fixed-timing runner. This boundary measures selection work, not complete serving latency. The complete replay and scheduling measurements in Appendix B.7 use a different instrumentation boundary and do not establish the same speedup for the full system.

Development budget selection. Budgets of 5%, 10% and 20% yield the values in Table 6. We select 10% as a diminishing-returns trade-off on development users before the held-out comparison. This is a design choice for the controlled study, not an estimate of users' preferred interruption rate.

Budget	Questions/user	Mistake AUC	Final mistake
5%	470	18.84%	12.08%
10%	940	14.85%	9.18%
20%	1,880	12.14%	7.08%

Table 6: DUAL development performance under equal block quotas. Ten percent is the selected diminishing-returns trade-off, not an estimated global optimum.

Model assumptions and common-temperature sensitivity. The learner's ten-score basis matches the native simulator. The zero-mean prior with precision $2I$, unit logistic temperature and acceptance strata were fixed before held-out evaluation. Changing both true channel temperatures to 0.5 or 2 preserves DUAL's advantage over periodic mixed feedback in a development sensitivity (Appendix F.1). Such a common rescaling remains within the linear-logistic model family. The hidden-feature studies in Appendices D.1 and D.4 test structural mismatch separately. Unknown unequal channel scales and drifting preferences are outside this fixed-learner assessment.

A.5 Action-probability numerics

For two Gaussian margins with means a, b, standard deviations s_a, s_b, and correlation ρ, the implementation evaluates their joint nonnegative probability as a bivariate normal CDF. The batched path uses a Gauss–Legendre approximation to Plackett's one-dimensional integral. We check it against a separate adaptive quadrature implementation. Singular limits are reduced to deterministic or univariate probabilities. Tests reject negative variances, nonfinite values, and action probabilities that fail to partition one within tolerance. Values are clipped only within numerical tolerance, and the three action regions are renormalized.

Answer predictions and latent decision regions. For a query design z, the Gaussian belief induces $S = z^\top \eta \sim \mathcal{N}(m, s^2)$, where $m = z^\top \mu_{t-1}$ and $s^2 = z^\top \Sigma_{t-1} z$ before interaction t. Three probabilities have different roles:

$$p_{\text{plug}} = \sigma(m), \qquad p_{\text{pred}} = \int_{-\infty}^{\infty} \sigma(m + su)\phi(u)\, du, \qquad p_{\text{sign}} = \Phi(m/s),$$

where ϕ and Φ are the standard-normal density and CDF. The first predicts a noisy answer at the belief mean, while the second averages that prediction over parameter uncertainty. For an absolute design, the third is the probability that the latent utility meets the acceptance threshold. At $s = 0$, use $p_{\text{pred}} = \sigma(m)$ and $p_{\text{sign}} = \mathbf{1}\{m \geq 0\}$.

The local update and information score use p_{plug}. After abstention, the absolute query targets the response with the larger p_{plug}, with canonical A breaking ties. The decision-aware comparator weights its hypothetical answer branches in Equation (17) by p_{pred}, evaluated with 41-point Gauss–Hermite quadrature. Serving instead integrates the joint latent action regions in Equation (5). These quantities are not interchangeable. For example, at $m = 2, s = 2$, they are approximately $0.881, 0.775, 0.841$, respectively.

A.6 Interaction and comparison definitions

The likelihoods and full-covariance belief are defined in Section 3. Oracle utilities generate only selected feedback and evaluation targets, never acquisition inputs. The serving target uses latent utility even when feedback is noisy. For example, utilities $(0.8, 0.6)$ imply serving A at threshold 0.5 and abstaining at 0.9, despite the same pairwise margin. Strict boundaries have zero mass under a continuous nonsingular Gaussian, while canonical ties cover degenerate cases. Point-estimate serving is a supporting sensitivity from the same state. Appendix A.5 specifies Gaussian integration and distinguishes answer predictions from serving probabilities.

Why question choice changes. An absolute question can inform both preference and threshold, but need not have the larger score. Suppose the designs have equal projected variance, $z_a^\top \Sigma z_a = z_p^\top \Sigma z_p = 0.5$. If predictions at μ_{t-1} give $p_a = 0.50$ for accepting the displayed response and $p_p = 0.95$ for preferring A to B, the scores in Equation (6) are $G_a = 0.0589$ and $G_p = 0.0117$ nats. DUAL selects the absolute question because the pairwise answer is already relatively predictable. Reversing these probabilities reverses the scores and selects the pairwise question.

Derivation of the online update. For the selected design and answer z_t, r_t, the new negative log-likelihood is $\ell_t(\eta) = \log(1 + e^{z_t^\top \eta}) - r_t z_t^\top \eta$. At the previous mean μ_{t-1}, its gradient is $(p_t - r_t)z_t$ and its Hessian is $h_t z_t z_t^\top$, where $p_t = \sigma(z_t^\top \mu_{t-1})$ and $h_t = p_t(1 - p_t)$. Adding its second-order Taylor approximation to the current Gaussian negative log density gives, up to constants, the quadratic in the mean increment δ:

$$Q_t(\delta) = \tfrac{1}{2}\delta^\top \Lambda_{t-1}\delta + (p_t - r_t)z_t^\top \delta + \tfrac{1}{2}h_t(z_t^\top \delta)^2.$$

Setting its gradient to zero gives $(\Lambda_{t-1} + h_t z_t z_t^\top)\delta = (r_t - p_t)z_t$, hence the mean and precision updates in Equations (7)–(8). This is one Newton step from the previous mean for the Gaussian prior plus the new logistic loss. It is also the logistic form of the static Kalman recursion in de Vilmarest & Wintenberger (2021, Algorithm 1). This correspondence identifies the update, without importing that work's stochastic-design guarantees for our adaptively selected observations.

Decision-aware comparator. For a bank of M target response pairs and a Gaussian belief q, let $\pi_m^q(\alpha)$ be the belief probability that action $\alpha \in \{\text{SERVE-A}, \text{SERVE-B}, \text{ABSTAIN}\}$ is the reference action on pair m, computed as in Equation (5). The target-bank risk is

$$\mathcal{R}(q) = \frac{1}{M} \sum_{m=1}^{M} \left(1 - \max_{\alpha} \pi_m^q(\alpha)\right). \tag{16}$$

Before feedback at interaction t, the current belief is q_{t-1}. For channel $c \in \{a, p\}$ and hypothetical answer $r \in \{0, 1\}$, let $q_t^{c,r}$ be the belief obtained by applying Equations (7)–(8) to q_{t-1} with design z_c and answer r. The answer probabilities integrate the logistic response model over q_{t-1}. The question's value is the expected one-step risk reduction:

$$V_t(c) = \mathcal{R}(q_{t-1}) - \sum_{r \in \{0,1\}} \Pr(r \mid c, \mu_{t-1}, \Sigma_{t-1}) \mathcal{R}(q_t^{c,r}). \tag{17}$$

The comparator evaluates both hypothetical answers on every target using only target features and the current belief. DUAL's information score avoids this target-bank evaluation. Appendix A.4 compares their fixed-timing performance and selector computation.

B Stability, Burden, and Development Sensitivities

We computed these post hoc diagnostics from saved query traces, learner predictions and outcome summaries for the 5,000 completed held-out runs. The original policies and trajectories remained unchanged. These analyses extend the four prespecified comparisons. The strict-cap and free-stopping experiment uses a separate 200-user development cohort.

B.1 Low-answer learning curve

Table 7 reads exact held-out checkpoints at which every questioning policy has received the stated number of answers. Information-based channel selection helps from the first ten answers. DUAL is initially tied with front-loading because both ask early within a block, then separates by 250 answers.

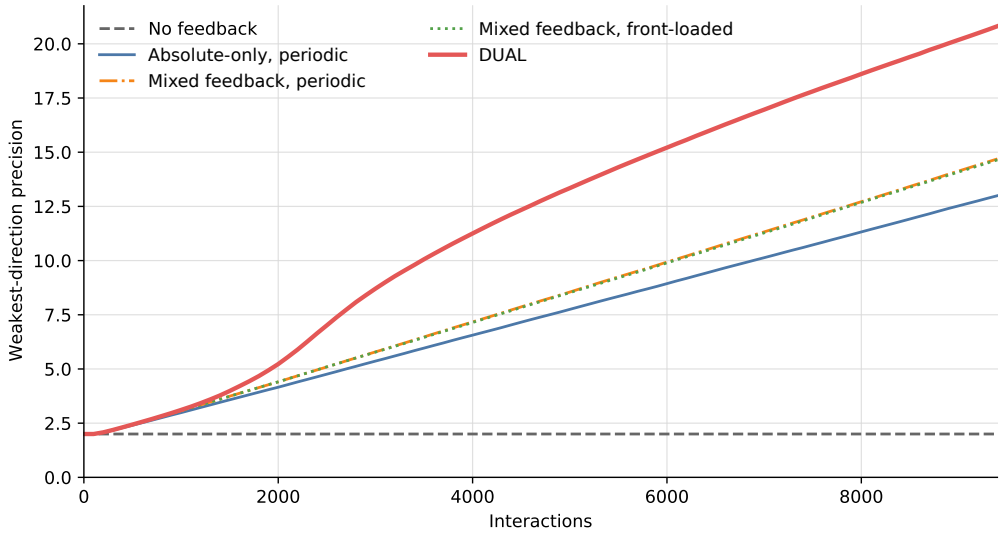
Answers	No feedback	Absolute-only, periodic	Mixed feedback, periodic	Mixed feedback, front-loaded	DUAL
10	56.57	40.67	**39.20**	40.38	40.38
20	56.57	36.59	**35.11**	35.81	35.81
50	56.57	29.54	**28.61**	29.28	29.28
100	56.57	24.68	**23.96**	24.48	24.25
250	56.57	18.22	17.44	17.43	**15.63**
500	56.57	13.87	13.15	13.14	**11.59**

Table 7: Held-out mistake rate (%) at exact low-answer checkpoints. Bold marks the best questioning policy in each row.

The full-trajectory effects are heterogeneous. Periodic mixed feedback helps 615 and harms 385 users relative to absolute-only periodic feedback. DUAL helps/harms 690/310 versus front-loading and 621/379 versus periodic mixed feedback. There are no ties. By contrast, absolute-only periodic feedback helps all 1,000 users relative to no feedback.

B.2 When DUAL's adaptive timing is active

Table 8 separates questions chosen because their information score exceeded the frozen threshold from questions forced near an aligned block boundary to satisfy the common ten-question quota.



Policy	Weakest precision	Condition number
Absolute-only, periodic	12.99	81.38
Mixed feedback, periodic	14.69	76.43
Mixed feedback, front-loaded	14.64	76.62
DUAL	**20.81**	**59.58**

Figure 3: Mean weakest-direction posterior precision across 1,000 held-out users. All questioning policies receive the same cumulative question count at every 100-interaction checkpoint. The embedded table gives final weakest precision and condition number. Higher precision and lower condition number indicate a better-constrained, more even fitted belief, not verified accuracy about the true user.

The DUAL information threshold triggers all questions in the first 1,000 interactions and most questions through interaction 3,000. Later, the fitted state is sufficiently concentrated that the quota guard supplies most questions. The pairwise share remains comparatively stable, so this transition primarily concerns *when* to ask rather than a collapse to one feedback channel.

Interactions	Natural trigger	Quota-forced	Pairwise share
1–1,000	100.00%	0.00%	23.55%
1,001–3,000	85.07%	14.93%	24.80%
3,001–6,000	10.43%	89.57%	22.25%
6,001–9,400	1.42%	98.58%	21.94%

Table 8: Composition of DUAL's 940 questions per held-out user across four chronological segments. "Natural" means that the frozen information threshold triggered the question. "Quota-forced" means that a causal late fill was required to reach ten questions in the current aligned block.

B.3 Information is balanced across user-state directions

For an eleven-dimensional Gaussian state, the smallest eigenvalue of the posterior precision matrix describes the least-constrained linear combination of the ten preference weights and acceptance threshold. Larger values mean that even this weakest direction is more tightly constrained. This is an internal model diagnostic, not direct evidence that the latent user vector is ground-truth accurate.

Figure 3 shows that DUAL gains more weakest-direction precision after the early stage. Its lower final condition number also indicates more even precision across parameter directions.

B.4 LOCAL QUESTION BURSTS

Policy	Absolute	Pairwise	Responses shown	Max rolling 100
Absolute-only, periodic	940.0	0.0	940.0	10
Mixed feedback, periodic	713.0	227.0	1,167.0	10
Mixed feedback, front-loaded	712.5	227.5	1,167.5	10
DUAL	725.5	214.5	1,154.5	20

Table 9: Mean held-out question mix and response exposure per user. Every row uses 940 binary answers. Rolling maxima describe local question frequency, not the aligned quota.

Aligned blocks do not impose a strict rolling cap. Every held-out user has more than ten DUAL questions in at least one overlapping 100-interaction window, and the largest count is twenty (Table 9). The following intervention directly tests a strict cap, rather than inferring its effect from burst correlations.

B.5 STRICT ROLLING CAP AND FREE STOPPING

We test DUAL variants using its frozen score and threshold on all 200 development users. The strict policies permit at most ten questions in every overlapping 100-interaction window. "Forced" spends all 940 answers using a causal feasibility guard, whereas "free" asks only above the threshold and may leave budget unused. Table 10 includes DUAL's aligned controller and periodic mixed feedback reference.

Development policy	Questions	AUC	Final mistake
Mixed feedback, periodic	940.0	16.46%	10.21%
DUAL (aligned, forced)	940.0	**14.85%**	**9.18%**
Strict rolling, forced	940.0	15.44%	9.95%
Aligned adaptive, free	344.2	16.33%	12.64%
Strict rolling, free	344.3	16.30%	12.52%

Table 10: Development-only burden intervention. "Strict rolling" caps every overlapping 100-interaction window at ten questions.

Strict-forced minus DUAL AUC is +0.59 points $[0.31, 0.86]$, while strict-forced minus periodic is -1.02 $[-1.55, -0.49]$. Strict-free saves 595.7 answers (63.4%) relative to 940 and has an unresolved AUC difference from periodic (-0.16 points $[-0.77, 0.45]$). This interval does not establish equivalent quality. The free policy demonstrates a burden–accuracy trade-off. A claim of preserved quality at lower burden would require a prespecified noninferiority or equivalence margin and an appropriate paired analysis. All results in this subsection remain development evidence.

Placement at a smaller answer count. An exploratory follow-up gives first-reference and periodic timing exactly 344 answers on the same development cohort. First-reference asks at the first score exceeding each block's initial score, with endpoint fallback. It improves AUC by 2.8219 points, adjusted interval $[-3.6682, -1.9481]$, while asking substantially earlier within blocks. Appendix F.2 reports the protocol and both planned contrasts. The result does not isolate a pure information-score effect.

B.6 ACCEPTANCE STRATA AND MULTIPLICITY SENSITIVITY

The held-out population contains deterministic target-acceptance strata of 30%, 50%, and 70%, assigned from stable user identifiers before policy evaluation. Table 11 shows that every primary comparison has the same favorable direction in all three strata. These are descriptive subgroup means without subgroup-adjusted inference.

Planned intervals and multiplicity sensitivity. All four planned paired 95% intervals exclude zero. In AUC points, absolute-only periodic feedback minus no feedback is -40.07

Target acceptance	Absolute–none	Mixed–absolute	DUAL–front	DUAL–periodic
30% ($n = 334$)	-33.92	-0.57	-1.07	-0.97
50% ($n = 333$)	-38.79	-0.73	-1.13	-1.42
70% ($n = 333$)	-47.53	-0.85	-1.50	-1.10

Table 11: Mean left-minus-right mistake-AUC differences in percentage points within each target-acceptance stratum. Absolute and mixed denote periodic absolute-only and mixed feedback. Front denotes front-loaded mixed feedback. Negative values favor the left method.

$[-40.51, -39.65]$, and periodic mixed feedback minus absolute-only periodic feedback is -0.71 $[-0.89, -0.54]$. DUAL minus front-loaded and periodic is -1.23 $[-1.39, -1.08]$ and -1.16 $[-1.38, -0.94]$. These prespecified intervals make no familywise-error claim. As a conservative post-result sensitivity, we also use 98.75% percentile intervals for the four comparisons, applying a Bonferroni adjustment with a nominal 95% familywise coverage target. All four remain below zero (Table 12). This does not change the primary analysis. It shows that the conclusion is not sensitive to this simple multiplicity correction.

Left minus right	Δ AUC	Bonferroni 98.75% interval
Absolute-only – no feedback	-40.07	[-40.64, -39.52]
Mixed feedback – absolute-only	-0.71	[-0.93, -0.50]
DUAL – mixed, front-loaded	-1.23	[-1.44, -1.04]
DUAL – mixed, periodic	-1.16	[-1.44, -0.88]

Table 12: Post-result multiplicity sensitivity for the four planned paired effects, in AUC percentage points. Updating policies in the first two rows use periodic placement. Each interval uses 20,000 paired user resamples.

B.7 Recorded computational overhead

Using the same measurement procedure for complete trajectory replays, absolute-only periodic feedback, periodic mixed feedback, and front-loaded mixed feedback take 49.50, 49.72, and 49.44 seconds per user, respectively. DUAL takes 49.66 seconds for the replay plus 5.56 seconds for causal scheduling, or 55.22 seconds combined. Its selector itself accounts for 0.46 seconds of the scheduling pass. Thus the recorded DUAL overhead is about 11% relative to periodic mixed feedback for these operations. These per-user timings measure computation, not the wall-clock duration of the parallel held-out job.

C Proofs and Development Certificate Audit

This section proves the results in Section 5 and reports a separate diagnostic of saved histories for the 200 development users. The propositions are deterministic statements about a realized feedback sequence. The statistical corollary becomes probabilistic only when supplied with a valid confidence result for the exact regularized logistic estimator.

For a positive-definite matrix A, write $\|x\|_A = (x^\top A x)^{1/2}$. The state dimension is $k = d + 1$ (eleven in our experiment), and the initial precision is $\Lambda_0 = \lambda I$ with $\lambda = 2$.

C.1 Conditional user confidence and serving guarantees

Conditional transfer to user confidence. Closeness to the joint fit does not by itself establish accuracy about the true user. The next result requires a separate confidence guarantee for the exact MAP estimate (Faury et al., 2020).

Theorem 4 (Confidence transfer)**.** *Suppose an exact-MAP result gives, with probability at least* $1 - \delta_n$, $\|\widehat{\eta}_n - \eta_i^*\|_{\nabla^2 F_n(\widehat{\eta}_n)} \le \beta_n$, *and* $\Lambda_n \preceq b_n H_n$. *When* $\chi_n < 1$, *the same event implies*

$$\|\mu_n - \eta_i^*\|_{\Lambda_n} \le B_n^{\mathrm{rec}} := \sqrt{b_n}\left(R_n + \frac{\beta_n}{\sqrt{1-\chi_n}}\right). \tag{18}$$

The smallest b_n is the observable largest generalized eigenvalue of (Λ_n, H_n). We do not instantiate a valid β_n for the experimental learner, so this theorem supplies a conditional relation rather than an operational user-confidence guarantee.

From uncertainty to serving mistakes. For a pair, write $a = [f_A; -1]$, $b = [f_B; -1]$, $d = a - b$. Its action regions use oriented constraints

$$\mathcal{Z}_A = \{a, d\}, \qquad \mathcal{Z}_B = \{b, -d\}, \qquad \mathcal{Z}_\varnothing = \{-a, -b\}. \tag{19}$$

For example, utilities $(0.8, 0.6)$ require serving A at threshold 0.5 and abstaining at 0.9, despite the same pairwise margin 0.2. Absolute feedback distinguishes these thresholds. Serving A requires both acceptability and preference, while abstention requires neither response to be acceptable. The next result requires each true margin to dominate estimation error and posterior spread in its direction. Small parameter error alone is insufficient near a decision boundary.

Corollary 5 (Three-action mistake and AUC control)**.** *Let $C_n = \{\chi_n < 1\}$ and $\gamma > \Phi^{-1}(3/4)$, where Φ is the standard-normal CDF. On C_n and the confidence event in Theorem 4, posterior action-probability argmax serving under uniform zero–one loss is correct if both nonzero $z \in \mathcal{Z}_{a_i^*}$ satisfy*

$$z^\top \eta_i^* > (B_n^{\mathrm{rec}} + \gamma)\sqrt{z^\top \Lambda_n^{-1} z}. \tag{20}$$

Let $M_i(\epsilon)$ be the bank fraction whose smaller oriented true margin per $\|z\|_2$ is at most ϵ, and $m_n = \lambda_{\min}(\Lambda_n)$. Set $Q_i(n) = M_i((B_n^{\mathrm{rec}} + \gamma)/\sqrt{m_n})$ on C_n, and $Q_i(n) = 1$ otherwise, without evaluating an uncertified radius. For fixed n,

$$\begin{aligned} \Pr\{e_i(n) \le Q_i(n)\} &\ge 1 - \delta_n, \\ \mathbb{E}[e_i(n)] &\le \mathbb{E}[\mathbf{1}_{C_n} Q_i(n)] + \Pr(C_n^c) + \delta_n. \end{aligned} \tag{21}$$

If $M_i(\epsilon) \le C\epsilon^\alpha$, the certificate remains active, and $B_n^{\mathrm{rec}} = \widetilde{O}(1)$, Proposition 1 gives endpoint rate $\widetilde{O}(n^{-\alpha/2})$ on the confidence event and normalized AUC rate through N answers $\widetilde{O}(N^{-\min\{\alpha/2,1\}})$, up to averaged confidence- and certificate-failure terms. With $n(t) = \Theta(t)$ answers by interaction t, the same rates apply in interaction time.

The true margins support analysis, not an observable skipping rule. Repeated stopping requires valid time-uniform user confidence or a summable failure budget as well as certificate validity. Fixed-n control is insufficient. We use the certificate to control approximation error, with no operational stopping claim. Appendix C gives the proofs and rate details. Convergence beyond the explicit model in Appendix C.4 is not asserted (de Vilmarest & Wintenberger, 2021).

C.2 Proof of greedy information coverage

At answered step t, let $c \in \{a, p\}$ index the available feedback channels. Scale each candidate design by the square root of its logistic curvature:

$$\widetilde{z}_t(c) = \sqrt{h_t(c)} z_t(c), \qquad \ell_t(c) = \widetilde{z}_t(c)^\top \Lambda_{t-1}^{-1} \widetilde{z}_t(c).$$

Because $x \mapsto \log(1 + x)$ is increasing, the information-gain rule selects a candidate $\widetilde{z}_t$ that maximizes $\ell_t(c)$. Let $\ell_t = \widetilde{z}_t^\top \Lambda_{t-1}^{-1} \widetilde{z}_t$. The matrix determinant lemma gives

$$\log \frac{\det \Lambda_n}{\det \Lambda_0} = \sum_{t=1}^{n} \log(1 + \ell_t). \tag{22}$$

Since $\Lambda_{t-1} \succeq \lambda I$ and every candidate has norm at most U, $0 \le \ell_t \le U^2/\lambda$. For $C_U = 1 + U^2/\lambda$, the elementary inequality $\log(1+x) \ge x/(1+x)$ therefore yields

$$\log \frac{\det \Lambda_n}{\det \Lambda_0} \ge \frac{1}{C_U} \sum_{t=1}^{n} \ell_t \tag{23}$$

$$\ge \frac{1}{C_U} \sum_{t=1}^{n} \bar{z}_t^\top \Lambda_{t-1}^{-1} \bar{z}_t \tag{24}$$

$$\ge \frac{1}{C_U} \sum_{t=1}^{n} \bar{z}_t^\top \Lambda_n^{-1} \bar{z}_t \tag{25}$$

$$= \frac{1}{C_U} \operatorname{tr}\left(\Lambda_n^{-1} \sum_{t=1}^{n} \bar{z}_t \bar{z}_t^\top \right) \tag{26}$$

$$\ge \frac{\kappa n}{C_U} \operatorname{tr}(\Lambda_n^{-1}) \ge \frac{\kappa n}{C_U \lambda_{\min}(\Lambda_n)}. \tag{27}$$

The second line uses greedy maximality, the third uses $\Lambda_{t-1}^{-1} \succeq \Lambda_n^{-1}$, and the penultimate line uses the comparator coverage assumption. In the other direction, $\operatorname{tr}(\Lambda_n) \le k\lambda + nU^2$. Applying the arithmetic–geometric mean inequality to the eigenvalues gives

$$\log \frac{\det \Lambda_n}{\det \Lambda_0} \le k \log\left(1 + \frac{nU^2}{k\lambda}\right). \tag{28}$$

Combining Equations (27) and (28) proves Proposition 1. The proof is pathwise: the candidate sets and comparator may depend on all previous selected queries and answers, but the comparator must be available at the step where it is invoked.

C.3 Proof of the recursive-to-MAP certificate

The ridge term makes F_n strongly convex, so its minimizer $\widehat{\eta}_n$ is unique. Let $h(s) = \sigma(s)(1-\sigma(s))$. Logistic curvature satisfies

$$\left| \frac{\mathrm{d}}{\mathrm{d}s} \log h(s) \right| = |1 - 2\sigma(s)| \le 1. \tag{29}$$

This is the generalized self-concordant property used in analyses of logistic regression (Bach, 2010). For any displacement v, it implies

$$e^{-c(v)} H_n \preceq \nabla^2 F_n(\mu_n + v) \preceq e^{c(v)} H_n, \qquad c(v) = \max_{t \le n} |z_t^\top v|. \tag{30}$$

The fixed ridge term also obeys these inequalities because the exponential factors straddle one.

Set $\Delta = \mu_n - \widehat{\eta}_n$ and $D = \|\Delta\|_{H_n}$. By Cauchy–Schwarz in the H_n-geometry,

$$|z_t^\top \Delta| \le (z_t^\top H_n^{-1} z_t)^{1/2} D \le \rho_n D. \tag{31}$$

Since $\nabla F_n(\widehat{\eta}_n) = 0$, integration along the segment from μ_n to $\widehat{\eta}_n$, followed by Equation (30), gives

$$\Delta^\top g_n = \int_0^1 \Delta^\top \nabla^2 F_n(\mu_n - s\Delta)\Delta \, \mathrm{d}s \tag{32}$$

$$\ge D^2 \int_0^1 e^{-s\rho_n D} \, \mathrm{d}s = D \frac{1 - e^{-\rho_n D}}{\rho_n}. \tag{33}$$

The left side is at most $D\|g_n\|_{H_n^{-1}} = D\nu_n$. The case $D = 0$ is immediate. Otherwise, cancellation of D gives

$$\frac{1 - e^{-\rho_n D}}{\rho_n} \le \nu_n.$$

When $\chi_n = \rho_n \nu_n < 1$, inversion yields $D \le -\log(1-\chi_n)/\rho_n = R_n$. If $\rho_n = 0$, the continuous limit is $D \le \nu_n$. Finally, applying Equation (30) at $v = -\Delta$ and using $e^{-\rho_n R_n} = 1 - \chi_n$ proves Equation (14).

Correspondence with prior localization bounds. Proposition 2 is a trace-computable specialization of generalized self-concordant theory. To compare in the same coordinates, write $\eta = \mu_n + H_n^{-1/2} v$. The Hessian at $v = 0$ is the identity and the maximum transformed design norm is ρ_n. The transformed ridge is positive semidefinite with zero third derivative. Proposition 2 of Bach (2010), equation (7), then gives radius $4\nu_n$ when $\chi_n \le 1/2$. The logarithmic radius is tighter on that domain and also applies for $1/2 < \chi_n < 1$.

This comparison does not establish a new localization result. Proposition 9 of Sun & Tran-Dinh (2019), at order two, supplies the gradient-map inequality underlying Equation (33). More explicitly, set $G = \{z_1, \dots, z_n\}$ in Marteau-Ferey et al. (2019, Appendix A.1). Their local radius is $r_\lambda(\mu_n) = 1/\rho_n$, their Newton decrement is ν_n, and $t(v) = \max_t |z_t^\top v|$. Lemma 5 gives $t(\Delta) \le T = -\log(1 - \chi_n)$. Corollary 1 gives $\phi(t(\Delta))\|\Delta\|_{H_n} \le \nu_n$, where $\phi(t) = (1 - e^{-t})/t$ is decreasing. Consequently,

$$\|\Delta\|_{H_n} \le \frac{\nu_n}{\phi(T)} = \frac{-\log(1 - \chi_n)}{\rho_n} = R_n.$$

Their exponential Hessian comparison gives the same Hessian factors. The zero case uses continuous limits. Thus both the radius and its validity condition coincide with existing theory. We connect them to the recursive residual below and a declared current-query refit contract.

Why the certificate is needed. Consider the recursion initialized at mean zero and precision λI, without replacing the state by full-history fits. Re-index answered steps so that μ_{t-1} is the pre-update mean. Let $p_t = \sigma(z_t^\top \mu_{t-1})$ and $h_t = p_t(1 - p_t)$. The recursion in Equation (8) satisfies the exact identity

$$\nabla F_n(\mu_n) = \sum_{t=1}^{n} z_t \left[\sigma(z_t^\top \mu_n) - \sigma(z_t^\top \mu_{t-1}) - h_t z_t^\top (\mu_n - \mu_{t-1})\right]. \tag{34}$$

To see this, multiply each mean update by Λ_t, sum over t, and exchange the order of the curvature sums. The resulting identity is

$$\lambda \mu_n + \sum_{t=1}^{n} h_t z_t z_t^\top (\mu_n - \mu_{t-1}) = \sum_{t=1}^{n} (r_t - p_t) z_t;$$

adding the likelihood-score terms produces Equation (34).

Thus the exact-MAP residual is a sum of local logistic Taylor remainders. A numerical reconstruction of the saved histories verified this identity to maximum absolute error 1.01×10^{-12}. However, bounding every remainder separately with $\sup_s |\sigma''(s)| = 1/(6\sqrt{3})$ was too loose to certify any of the 200 development users. Cancellation among signed residuals matters, which is why Proposition 2 uses the realized total gradient rather than claiming unconditional convergence of the raw recursion. Related global analyses of static logistic Kalman recursions introduce a curvature floor and impose stochastic-design conditions (de Vilmarest & Wintenberger, 2021). The finite-history certificate does not require those modifications or assumptions. The convergence result below uses an explicit stochastic candidate model and retains the original recursion.

C.4 Convergence of the online belief

We prove Theorem 3 for the original recursion without projection, a curvature floor, or full-history refits. The proof connects the two preceding propositions to a statistical convergence result. Proposition 1 controls the belief's spread when available questions cover every parameter direction. Proposition 2 compares its center with a joint fit at a fixed answer count. The new argument proves that the online mean approaches the true user parameters and that the Euclidean bound to the joint fit tends to zero as answers accumulate.

The key step first bounds growth of the precision-weighted error without assuming coverage or bounded estimates. We then establish coverage along the policy's actual history and show that precision grows faster than this error. Dividing the weighted error by the least-direction precision proves convergence of the mean. Afterward, the largest covariance eigenvalue vanishes because it equals the reciprocal of the smallest precision eigenvalue. The user is fixed and correctly specified throughout this argument, so the joint-fit optimum and the true parameters remain distinct objects until their limits are established.

Assumptions and information order. Fix one user with $k = d + 1$ parameters $\eta^\star = (v^\star, b^\star)$, where $v^\star$ contains preference weights and $b^\star$ is the acceptance threshold. At interaction t, the candidate features X_t^A, X_t^B are independent uniform draws from $[-1, 1]^d$. Pairs are independent across interactions and independent of the past. Initialize $\mu_0 = 0$ and $\Lambda_0 = \lambda I$, with $\lambda > 0$. The belief after n answers is $\mathcal{N}(\mu_n, \Lambda_n^{-1})$. Thus μ_n is the current parameter estimate and Λ_n is its precision matrix. Use exact arithmetic and exact Gaussian action probabilities. The serving and absolute-target rules are those of Section 4, with fixed tie conventions. Thus the feasible designs are

$$z_t^a = (X_t, -1), \qquad z_t^p = (X_t^A - X_t^B, 0), \qquad X_t \in \{X_t^A, X_t^B\}.$$

Both satisfy $\|z_t^c\|_2 \le Z := \sqrt{\max\{d+1, 4d\}}$. The coverage argument below permits any history-dependent choice of X_t from the current pair.

Repeat blocks of L interactions with exactly Q answers, where $1 \le Q \le L$. All interactions are eligible. Procedure 1 uses a fixed cutoff $\tau_G > 0$, greedy information selection, and its quota-completion rule. Count an answer as above-cutoff if its score reaches τ_G, including when the quota trigger also applies. Let T_n be the interaction of answer n, and let $m(t)$ count answers strictly before interaction t. Then $T_n \le L\lceil n/Q \rceil$.

Let $\mathcal{F}_{n-1}$ contain the history through answer $n-1$, $\mathcal{Q}_n$ the information immediately before answer n, and $\mathcal{F}_n$ the history after it. Thus $\mathcal{F}_{n-1} \subseteq \mathcal{Q}_n \subseteq \mathcal{F}_n$. The chosen design z_n, serving action, and ask decision belong to $\mathcal{Q}_n$. The binary answer satisfies

$$\mathbb{E}[r_n \mid \mathcal{Q}_n] = \sigma(z_n^\top \eta^\star) =: p_n^\star. \tag{35}$$

This conditional law allows adaptively selected questions. We do not assume the selected observations are independent.

Index Equations (7)–(8) by answers in this proof:

$$p_n = \sigma(z_n^\top \mu_{n-1}), \quad h_n = p_n(1-p_n), \quad \Lambda_n = \Lambda_{n-1} + h_n z_n z_n^\top, \quad \mu_n = \mu_{n-1} + \Lambda_n^{-1} z_n (r_n - p_n).$$

The quantities p_n, h_n, Λ_n are $\mathcal{Q}_n$-measurable. Only the new mean uses answer r_n. The joint maximum a posteriori (MAP) estimate $\widehat{\eta}_n$ minimizes F_n from Equation (12) on the same selected history.

The true-logit assumption is

$$\max\{\|v^\star\|_1 + |b^\star|,\, 2\|v^\star\|_1\} \le B_0 := \operatorname{arcosh}(3). \tag{36}$$

It bounds both channels' true logits and keeps their answer probabilities between approximately 0.146 and 0.854. This restricts the generating user, not the learner's iterates. We write O_ω and Ω_ω for eventual upper and lower bounds with finite constants that may depend on the sample path. In particular, the precision claim means that almost every path has $c > 0$ and $n_0 < \infty$ such that $\lambda_{\min}(\Lambda_n) \ge c\, n/\log n$ for every $n \ge n_0$. This is a lower growth bound, not an equality or a claim that the parameter error decreases after every answer. Norms without a subscript are Euclidean.

The bound B_0 satisfies $\sigma'(B_0) = 1/8$, which ensures the nonnegative restoring term in Step 1. The logit restriction enters through this step.

1. Bound error growth before using coverage. Put $\mathbf{e}_n = \mu_n - \eta^\star$, $a_n = z_n^\top \mathbf{e}_{n-1}$, and define

$$V_n = \mathbf{e}_n^\top \Lambda_n \mathbf{e}_n, \qquad s_n = z_n^\top \Lambda_n^{-1} z_n, \qquad \ell_n = h_n z_n^\top \Lambda_{n-1}^{-1} z_n.$$

Here V_n measures squared error using accumulated precision. The leverage s_n uses the updated precision before the answer. For fixed $\eta^\star$, a_n, s_n, ℓ_n are $\mathcal{Q}_n$-measurable, whereas $\mathbf{e}_n, V_n$ are $\mathcal{F}_n$-measurable. Truth-dependent quantities need not be observable to the learner.

Expanding the mean update and applying the matrix determinant lemma gives

$$V_n - V_{n-1} = h_n a_n^2 + 2a_n(r_n - p_n) + s_n(r_n - p_n)^2, \tag{37}$$

$$h_n s_n = \frac{\ell_n}{1+\ell_n} \le \log(1+\ell_n),$$

$$\sum_{j \le n} h_j s_j \le \log \frac{\det \Lambda_n}{\det \Lambda_0} \le J_n := k \log\left(1 + \frac{nZ^2}{4k\lambda}\right). \tag{38}$$

	Static logistic Kalman analysis	DUAL, Theorem 3
Update	Global result uses curvature $\max\{n^{-\beta}, h_n\}$, $0 < \beta < 1/2$; it eventually equals the original update.	Uses $h_n = p_n(1-p_n)$ throughout, without projection or refits.
Observations	Independent, identically distributed bounded observations; $\mathbb{E}[XX^\top]$ is invertible.	Independent uniform candidate pairs; the selected question and answer times depend on past feedback, with continuing block quotas.
User model	A finite population-risk minimizer; correct logistic specification is not required.	Fixed correctly specified logistic user, with true logits bounded by $\operatorname{arcosh}(3)$.
Guarantee	High-probability entry into a local region and cumulative excess-risk bounds for the modified recursion.	Almost-sure parameter convergence, growing minimum precision, a vanishing joint-fit error bound, and serving consistency under nonzero margins.

Table 13: Scope of the closest recursive logistic analysis (de Vilmarest & Wintenberger, 2021, Algorithm 3, Proposition 9 and Theorem 10) and our result. The basic update equations coincide before that work's curvature modification. The assumptions and guarantees differ, so neither column is a uniform strengthening of the other.

The leverage identity follows from Sherman–Morrison. The last inequality uses $h_j \le 1/4$ and the arithmetic–geometric mean bound on the eigenvalues of Λ_n.

For a true logit u and prediction error a, define

$$D(a,u) = 2a\{\sigma(u+a) - \sigma(u)\} - \sigma'(u+a)a^2.$$

When $|u| \le B_0$, this quantity is nonnegative for every a. Indeed, with $\bar{h} = \int_0^1 \sigma'(u+va)\,dv$, we have $D = (2\bar{h} - \sigma'(u+a))a^2$. If $|u+a| \le B_0$, then $\bar{h} \ge 1/8$ and $\sigma'(u+a) \le 1/4$. Otherwise every point on the segment has absolute value at most $|u+a|$, giving $\bar{h} \ge \sigma'(u+a)$.

We also need a uniform tail bound. For any finite B, set $\delta_B = \sigma(-B) > 0$. If $|u| \le B$ and $r = |a|$ is sufficiently large, the logistic probability difference has the sign of a and magnitude at least $\delta_B/2$, while $\sigma'(u+a) \le e^{B-r}$. Therefore

$$D(a,u) \ge \delta_B r - e^{B-r}r^2 \ge \delta_B r/2$$

eventually, uniformly over $|u| \le B$.

Taking the conditional expectation in Equation (37), with $u_n = z_n^\top \eta^\star$, gives

$$\mathbb{E}[V_n - V_{n-1} \mid \mathcal{Q}_n] = -D(a_n, u_n) + s_n \mathbb{E}[(r_n - p_n)^2 \mid \mathcal{Q}_n].$$

Since $s_n \le K := Z^2/\lambda$, choose a finite A such that $D(a,u) \ge K$ for $|a| \ge A$ and $|u| \le B_0$. The expected increment is then nonpositive outside this interval. Inside it, $D \ge 0$ and $h_n \ge \sigma'(B_0 + A) > 0$. For a finite deterministic constant C,

$$\mathbb{E}[V_n \mid \mathcal{Q}_n] \le V_{n-1} + Ch_n s_n. \tag{39}$$

For any fixed $\gamma > 0$, put $d_n = \log^{1+\gamma}(e+n)$ and $w_n = h_n s_n$. Equation (38) and Abel summation imply

$$\sum_{n \ge 1} \frac{w_n}{d_n} \le K_\gamma < \infty$$

for a deterministic K_γ. Specifically, the partial sums of w_n are $O(\log(e+n))$, so the summation-by-parts tail is bounded by a constant times $\int_1^\infty [(e+x)\log^{1+\gamma}(e+x)]^{-1}\,dx < \infty$. Consequently,

$$M_n = \frac{V_n}{d_n} + C\left(K_\gamma - \sum_{j \le n} \frac{w_j}{d_j}\right)$$

is a nonnegative supermartingale in $(\mathcal{F}_n)$. To see this, first condition on $\mathcal{Q}_n$ in Equation (39), use $d_n \ge d_{n-1}$, and then apply the tower property. Finite-time integrability follows from $\|\mu_n -$

$\mu_{n-1}\| \le Z/\lambda$ and $\lambda_{\max}(\Lambda_n) \le \lambda + nZ^2/4$. The nonnegative supermartingale convergence theorem yields

$$V_n = O_\omega(\log^{1+\gamma}(e+n)). \tag{40}$$

Since $\Lambda_n \succeq \lambda I$, choosing $0 < \gamma < 1$ also gives

$$\max_{j \le n} \|\mu_j\| = O_\omega(\log^{(1+\gamma)/2}(e+n)) = o_\omega(\log(e+n)).$$

2. Show that available questions cover all parameters. Let N_n count above-cutoff answers among the first n answers. Each adds at least $2\tau_G$ to log-determinant precision, so $N_n \le J_n/(2\tau_G)$. After b blocks, let S_b be the actual answered interaction set and J_b^{tail} the deterministic grid of the last Q positions per block. A block with no above-cutoff answer follows this grid exactly. Every other block contributes at most $2Q$ positions to the symmetric difference. Hence

$$|S_b \triangle J_b^{\text{tail}}| \le 2QN_{Qb} = O(\log(e+b)). \tag{41}$$

Define the feasible absolute design $\xi_t = (X_t, -1)$ at every interaction along the actual history, including skipped interactions. Let $\mathcal{H}_{t-1}$ be the history before the current pair. For a unit vector $\theta = (\theta_v, \theta_b)$, write $Y_A = \theta_v^\top X_t^A - \theta_b$, and similarly Y_B. Conditional on the history,

$$s := \mathbb{E}[Y_A^2] = \|\theta_v\|^2/3 + \theta_b^2 \ge 1/3, \qquad \mathbb{E}[Y_A^4] \le 3s^2.$$

For the fourth-moment bound, if $S = \theta_v^\top X_t^A$ and $w = \mathbb{E}S^2$, then $\mathbb{E}S^4 = 3w^2 - (2/15)\sum_i \theta_{v,i}^4$. Expanding $(S - \theta_b)^4$ gives the bound. Cauchy–Schwarz applied to $\mathbb{E}[Y_A^2 \mathbf{1}_{\{Y_A^2 \ge s/2\}}] \ge s/2$ yields $\mathbb{P}(Y_A^2 \ge s/2 \mid \mathcal{H}_{t-1}) \ge 1/12$. Both candidates meet this bound with probability at least $1/144$. Whichever candidate defines ξ_t, $(\theta^\top \xi_t)^2 \ge \min\{Y_A^2, Y_B^2\}$, so

$$\mathbb{E}[\xi_t \xi_t^\top \mid \mathcal{H}_{t-1}] \succeq \alpha I, \qquad \alpha = 1/864.$$

On the deterministic grid J_b^{tail}, the difference between the sum of these matrices and their conditional means is a matrix of bounded martingale sums. Entrywise Azuma bounds and Borel–Cantelli in fixed dimension make its operator norm $o(b)$ almost surely. Thus the grid sum is at least $(\alpha/2)QbI$ eventually. Equation (41) removes only $O(\log(e+b))$ terms when passing to S_b, each with norm at most $d+1$. Added terms are positive semidefinite, so the actual sum is eventually at least $(\alpha/4)QbI$. Retaining the previous completed block for incomplete prefixes gives, almost surely for all sufficiently large n,

$$\sum_{j \le n} \xi_{T_j} \xi_{T_j}^\top \succeq c_0 nI, \qquad c_0 = \alpha/8 = 1/6912. \tag{42}$$

These are available absolute questions, regardless of which channel was actually selected.

3. Turn coverage into consistency. The proof of Proposition 1 also applies to a prefix-dependent comparator matrix $\mathsf{A}_n = \sum_{j \le n} \bar{z}_j \bar{z}_j^\top$, where $\bar{z}_j$ is any feasible design scaled by the square root of its current curvature. Greediness gives

$$\sum_{j \le n} \ell_j \ge \operatorname{tr}(\Lambda_n^{-1} \mathsf{A}_n) \ge \frac{\lambda_{\min}(\mathsf{A}_n)}{\lambda_{\min}(\Lambda_n)}.$$

Since $\ell_j \le Z^2/(4\lambda)$, $\sum_j \ell_j \le (1 + Z^2/(4\lambda))J_n$. Therefore

$$\lambda_{\min}(\Lambda_n) \ge \frac{\lambda_{\min}(\mathsf{A}_n)}{(1 + Z^2/(4\lambda))J_n}. \tag{43}$$

Choose $\gamma = 1/2$ in Equation (40). Every feasible curvature in the first n answers is at least

$$c_n^h = \tfrac{1}{4} \exp\left(-Z \max_{j \le n} \|\mu_j\|\right).$$

The maximum grows as $O_\omega(\log^{3/4}(e+n))$, so $c_n^h \ge n^{-\epsilon}$ eventually for every fixed $\epsilon > 0$. Using the absolute comparators in Equation (42) with $\epsilon = 1/2$, Equation (43) yields $\lambda_{\min}(\Lambda_n) = \Omega_\omega(\sqrt{n}/\log(e+n))$. It follows that

$$\|\mathbf{e}_n\|^2 \le \frac{V_n}{\lambda_{\min}(\Lambda_n)} = O_\omega\left(\frac{\log^{5/2}(e+n)}{\sqrt{n}}\right) \to 0.$$

Consistency makes the entire sequence μ_n bounded almost surely. Only now can we use the positive constant $c_\omega^h = \frac{1}{4}\exp(-Z\sup_j\|\mu_j\|)$. Repeating the transfer gives $\lambda_{\min}(\Lambda_n) = \Omega_\omega(n/\log(e+n))$. Equation (40), with $\gamma = 2\epsilon$, then gives

$$\|\mu_n - \eta^\star\| = O_\omega\left(\frac{\log^{1+\epsilon}(e+n)}{\sqrt{n}}\right) \quad \text{for every fixed } \epsilon > 0. \tag{44}$$

Countably many positive rational exponents give a common probability-one event if needed. In particular, $\|\Lambda_n^{-1}\|_{\mathrm{op}} = O_\omega(\log(e+n)/n)$.

4. Relate the online mean to the joint fit and its certificate. The localization proof of Proposition 2 applies at any reference point a. At that point, define

$$g = \nabla F_n(a), \qquad H = \nabla^2 F_n(a), \qquad \nu = \|g\|_{H^{-1}}, \qquad \rho = \max_{j\le n}\|z_j\|_{H^{-1}}.$$

For $\epsilon = \|g\|/\lambda_{\min}(H)$, we have $\rho\nu \le Z\epsilon$. Thus if $\rho\nu \le 1/2$,

$$\|a - \widehat{\eta}_n\| \le \frac{-\log(1-\rho\nu)}{\rho\sqrt{\lambda_{\min}(H)}} \le 2\epsilon. \tag{45}$$

The last step uses $-\log(1-x) \le 2x$ on $[0, 1/2]$, with the continuous limit when $\rho = 0$.

Because $h_j \le 1/4$, $\sum_{j\le n} z_j z_j^\top \succeq 4(\Lambda_n - \lambda I)$. Bounded true logits therefore give $\lambda_{\min}(H_n^\star) = \Omega_\omega(n/\log(e+n))$, where $H_n^\star = \nabla^2 F_n(\eta^\star)$. The gradient

$$g_n^\star = \lambda\eta^\star + \sum_{j\le n} z_j(p_j^\star - r_j)$$

is a fixed vector plus a sum of bounded martingale differences by Equation (35). Azuma and Borel–Cantelli give $\|g_n^\star\| = O_\omega(\sqrt{n\log(e+n)})$. Applying Equation (45) at $\eta^\star$ shows

$$\|\widehat{\eta}_n - \eta^\star\| = O_\omega(\log^{3/2}(e+n)/\sqrt{n}).$$

Together with Equation (44), this also bounds $\|\mu_n - \widehat{\eta}_n\|$ at the same order.

At the online mean, logistic curvature comparison yields $H_n \succeq e^{-Z\|\mu_n - \eta^\star\|}H_n^\star$. The global upper bound $\lambda_{\max}(\nabla^2 F_n(\eta)) \le \lambda + nZ^2/4$ and $\nabla F_n(\widehat{\eta}_n) = 0$ imply

$$\epsilon_n := \frac{\|g_n\|}{\lambda_{\min}(H_n)} \le \frac{(\lambda + nZ^2/4)\|\mu_n - \widehat{\eta}_n\|}{\lambda_{\min}(H_n)} = O_\omega(\log^{5/2}(e+n)/\sqrt{n}) \to 0.$$

Consequently $\chi_n \le Z\epsilon_n \to 0$, and eventually $R_n/\sqrt{\lambda_{\min}(H_n)} \le 2\epsilon_n \to 0$. This establishes a vanishing Euclidean bound. It does not assert that the Hessian-norm radius R_n itself vanishes.

5. Translate the result to timing and serving. At interaction t, every feasible score satisfies

$$G_t(c) \le \tfrac{1}{2}\log\left(1 + \frac{Z^2}{4\lambda_{\min}(\Lambda_{m(t)})}\right) \to 0.$$

Thus only finitely many answers reach the fixed positive cutoff. Every sufficiently late block asks at its last Q positions, and $m(t)/t \to Q/L$. Continued learning is supplied by the repeated quota.

Define $\mathcal{E}_n = \|\mu_n - \eta^\star\|^2 + \mathrm{tr}(\Lambda_n^{-1}) \to 0$. This is the mean squared distance of a draw from the represented Gaussian to $\eta^\star$. Markov's inequality therefore makes its probability outside any fixed ball about $\eta^\star$ tend to zero.

For a fresh pair, let Δ be the smallest absolute value of

$$(v^\star)^\top(X^A - X^B), \qquad (v^\star)^\top X^A - b^\star, \qquad (v^\star)^\top X^B - b^\star.$$

Assume $\mathbb{P}(\Delta = 0) = 0$, which holds in the uniform model when $v^\star \neq 0$. A perturbation smaller than $\Delta/(2Z)$ preserves all three signs and the correct action. Gaussian mass outside this ball is at most $4Z^2\mathcal{E}_n/\Delta^2$. If $\Delta^2 > 8Z^2\mathcal{E}_n$, the true action has probability greater than $1/2$ and must be selected. Writing F_Δ for the fresh-pair margin distribution gives

$$\mathbb{P}(\text{mistake at } t \mid \mathcal{H}_{t-1}) \le F_\Delta\left(\sqrt{8Z^2\mathcal{E}_{m(t)}}\right) \to 0.$$

The difference between the realized mistake indicator and this conditional risk is a bounded martingale difference. Azuma and Borel–Cantelli make its time average vanish. The conditional risks' Cesaro mean also vanishes, proving the realized average-error claim. For each fixed pair with nonzero margins the action is eventually correct, although fresh pairs can keep approaching a boundary.

Extension beyond moderate true logits. Drop Equation (36) while retaining a fixed finite $\eta^\star$ and all other assumptions. Define

$$E_D = \{D(a_n, z_n^\top \eta^\star) \geq 0 \text{ for all sufficiently large } n\}.$$

Let $\mathcal{C}$ collect the parameter, precision, MAP, certificate, timing, and Gaussian-concentration conclusions above. For this fixed model and parameter,

$$\mathbb{P}(E_D \cap \mathcal{C}^c) = 0. \tag{46}$$

The serving conclusions also hold there under the same boundary condition. This implication does not establish $\mathbb{P}(E_D) = 1$ or a positive lower bound on that probability.

To prove it without conditioning the feedback law on a future event, fix a deterministic N. All true logits have magnitude at most $B = Z\|\eta^\star\| < \infty$. The tail bound for D still holds, so Equation (39) holds at every step where $D \geq 0$, for a finite constant depending on B. Let

$$\zeta_n^{(N)} = \mathbf{1}\{D(a_j, z_j^\top \eta^\star) \geq 0 \text{ for all } N \leq j \leq n\}, \qquad n \geq N.$$

This indicator is $\mathcal{Q}_n$-measurable. Freeze the energy before the first failure:

$$W_{N-1}^{(N)} = V_{N-1}, \qquad W_n^{(N)} = W_{n-1}^{(N)} + \zeta_n^{(N)}(V_n - V_{n-1}), \qquad \widetilde{w}_n^{(N)} = \zeta_n^{(N)} h_n s_n.$$

The process $W^{(N)}$ is nonnegative. On active steps it agrees with V, and otherwise it remains constant. Hence $\mathbb{E}[W_n^{(N)} \mid \mathcal{Q}_n] \leq W_{n-1}^{(N)} + C\widetilde{w}_n^{(N)}$. For the same d_n and K_γ as above,

$$\frac{W_n^{(N)}}{d_n} + C\left(K_\gamma - \sum_{j=N}^{n} \frac{\widetilde{w}_j^{(N)}}{d_j}\right)$$

is a nonnegative, integrable supermartingale after taking the tower expectation. It yields $W_n^{(N)} = O_\omega(d_n)$. On paths with no failure after N, $W_n^{(N)} = V_n$. Taking the countable union over N establishes the same weak growth bound on E_D outside a null set. Coverage holds outside its own null set. Steps 3–5 then prove Equation (46).

C.5 Proof of confidence transfer

On the exact-MAP confidence event and $C_n = \{\chi_n < 1\}$, Equation (14) implies

$$\|\widehat{\eta}_n - \eta_i^*\|_{H_n} \leq \frac{\beta_n}{\sqrt{1 - \chi_n}}. \tag{47}$$

The triangle inequality and Proposition 2 then give

$$\|\mu_n - \eta_i^*\|_{H_n} \leq R_n + \frac{\beta_n}{\sqrt{1 - \chi_n}}.$$

If $\Lambda_n \preceq b_n H_n$, then $\|x\|_{\Lambda_n} \leq \sqrt{b_n}\|x\|_{H_n}$ for every x, proving Theorem 4. The theorem accepts any valid exact-estimator radius β_n. Self-normalized martingale analyses for logistic models provide one route under predictable bounded designs, bounded parameters, regularization, and a well-specified feedback model (Faury et al., 2020). It does not assert that those premises hold under arbitrary user behavior.

C.6 Proof of the three-action risk bound

Let $\widetilde{\eta} \sim \mathcal{N}(\mu_n, \Lambda_n^{-1})$ denote a posterior draw, and let E_n be the exact-MAP confidence event, with $\Pr(E_n^c) \leq \delta_n$. On $E_n \cap C_n$, the event in Equation (18) holds and every oriented constraint z satisfies

$$z^\top \mu_n \geq z^\top \eta_i^* - B_n^{\mathrm{rec}} \sqrt{z^\top \Lambda_n^{-1} z}. \tag{48}$$

Under Equation (20), the standardized posterior mean of each of the true action's two constraints is therefore larger than γ. A union bound gives

$$\Pr(\widetilde{\eta} \text{ violates the true action region}) \leq 2\Phi(-\gamma) < \frac{1}{2}. \tag{49}$$

The three action regions partition the posterior mass, so the true region has strictly more than one half and is the unique posterior-MAP action. The strict margin premise excludes deterministic boundary ties, which the canonical-A rule handles operationally.

For evaluation pair j, define

$$D_{ij} = \min_{z \in \mathcal{Z}_{a_i^*(j)},\, z \neq 0} \frac{z^\top \eta_i^*}{\|z\|_2}, \qquad M_i(\epsilon) = \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}\{D_{ij} \leq \epsilon\}. \tag{50}$$

Because $z^\top \Lambda_n^{-1} z \leq \|z\|_2^2 / m_n$, every pair with $D_{ij} > (B_n^{\mathrm{rec}} + \gamma)/\sqrt{m_n}$ is correct on $E_n \cap C_n$. Hence

$$e_i(n) \leq M_i\left(\frac{B_n^{\mathrm{rec}} + \gamma}{\sqrt{m_n}}\right) \quad \text{on } E_n \cap C_n.$$

On C_n^c, use the trivial bound $e_i(n) \leq 1 = Q_i(n)$, because the recursive confidence radius is not defined there. Therefore $e_i(n) \leq Q_i(n)$ throughout E_n, giving the first line of Equation (21). Pointwise, $e_i(n) \leq Q_i(n) + \mathbf{1}_{E_n^c}$, so

$$\mathbb{E}[e_i(n)] \leq \mathbb{E}[Q_i(n)] + \delta_n = \mathbb{E}[\mathbf{1}_{C_n} Q_i(n)] + \Pr(C_n^c) + \delta_n.$$

This proves the expectation bound without assuming independence between certificate validity and statistical confidence. In particular, conditioning on C_n alone would not preserve a $1-\delta_n$ confidence probability.

If $M_i(\epsilon) \leq C\epsilon^\alpha$, Proposition 1 gives $m_n = \Omega(n/\log n)$, and the transferred radius is $\widetilde{O}(1)$ uniformly on the certified histories, then the expected endpoint error is $\widetilde{O}(n^{-\alpha/2}) + \Pr(C_n^c) + \delta_n$. Summing this sequence and dividing by the horizon gives normalized AUC rate $\widetilde{O}(N^{-\min\{\alpha/2,1\}})$, assuming both the confidence- and certificate-failure terms decay compatibly. The same decomposition applies to the nonnegative trapezoidal weights used for the recorded AUC. Converting to interaction time requires the cumulative answer count $n(t) = \Theta(t)$, rather than only a positive final budget fraction. The confirmatory block schedule satisfies $n(100j) = 10j$ for $j = 0, \ldots, 94$, with $n(t) = 0.1t + O(1)$ between block boundaries. This changes constants but not the exponent and establishes the stated interaction-time conversion in Corollary 5.

C.7 Development-only certificate audit

We evaluated the observable certificate on the 200 development-user information-selected mixed-feedback trajectories, each containing 1,880 selected feedback updates. At every checkpoint we independently optimized Equation (12) on the realized selected trace. No held-out profile or held-out outcome was used. Table 14 shows that the condition $\chi_n < 1$ already held for most users at 25 answers and for all users from 500 answers onward.

Answers	Certified users	Median χ_n	Median R_n	Median actual MAP gap
25	180/200	0.505	0.376	0.306
50	188/200	0.490	0.442	0.339
100	193/200	0.423	0.443	0.361
200	196/200	0.349	0.407	0.341
500	200/200	0.247	0.379	0.334
1,000	200/200	0.183	0.341	0.309
1,500	200/200	0.150	0.318	0.289
1,880	200/200	0.135	0.302	0.283

Table 14: Recursive-to-MAP certificate on development trajectories. Both the certificate radius R_n and actual recursive-to-MAP gap use the local H_n-norm. These are diagnostics, not held-out confirmation of a theorem premise.

At 1,880 answers, the largest χ_n was 0.349. The median ratio of the certificate radius to the actual gap was 1.068 and the largest was 1.206, so the bound was valid without being orders of magnitude loose. Across users, the generalized eigenvalues comparing recursive precision Λ_n with H_n

lay within 0.899 and 1.169 in the worst case. The recursive and exact MAP states differed by a median 1.86% in relative Euclidean norm (95th percentile 3.28%, maximum 6.18%). On the development reference bank, their predicted probabilities differed by a median RMSE of 0.271 percentage points, and their point-serving actions agreed for 99.66% of pairs on average (minimum 98.7% across users).

The same audit also checked the premise behind Proposition 1. The sequence of available curvature-weighted absolute questions reached rank eleven for every user. For all users and all audited prefixes from 100 answers onward, $\lambda_{\min}(\sum_{t\le n} \bar{z}_t \bar{z}_t^\top)/n$ was at least 0.00512. The median user's worst-prefix value was 0.00911. The median fitted growth exponents were 1.061 for this absolute comparator and 1.070 for the queries selected by information-based acquisition. By contrast, pairwise-only designs have a zero threshold coordinate and cannot be full rank in the eleven- dimensional state. The selected trajectories used a median 76.62% absolute questions, consistent with the need to cover the threshold as well as relative preference directions.

On these development histories, the audit supports agreement with the joint fit and coverage of the user-parameter directions. The assumptions in Propositions 1 and 2 still apply. The audit supplies neither a statistical radius β_n nor population-wide confidence or margin guarantees, so it does not establish the complete serving guarantee or held-out performance.

C.8 A SHADOW REFIT SCREEN ON FRESH DUAL HISTORIES

We give the approximation certificate a concrete diagnostic use on the DUAL Q940 histories from Appendix D.2. The 200 fresh simulated assessment profiles and three realizations give 600 histories. We inspect each at $n = 5, 10, 20, 50, 100, 250, 500, 940$ answers, for 4,800 checkpoints. The screen is evaluated after the answer, on the same visible response pair. Serving at that interaction was already committed before feedback. It neither changes that action nor installs a refit into the learner.

Let $\mathcal{Z}_n^{\mathrm{vis}}$ contain the absolute-A, absolute-B and pairwise designs constructible from that visible pair. This three-design diagnostic does not expand the acquisition policy's two available questions. From Proposition 2, when $\chi_n < 1$,

$$B_n = R_n \max_{z \in \mathcal{Z}_n^{\mathrm{vis}}} \sqrt{z^\top H_n^{-1} z} \quad \text{bounds} \quad \max_{z \in \mathcal{Z}_n^{\mathrm{vis}}} |z^\top (\mu_n - \widehat{\eta}_n)|.$$

We freeze a tolerance of 0.10 logit units before assessment and recommend KEEP exactly when the valid bound is at most this tolerance. Otherwise, we request REFIT. A failed certificate condition is not proof of an inaccurate recursive estimate. This bound concerns the selected history's joint MAP fit on the current designs, not true-user probabilities or unseen queries. The logistic function's $1/4$-Lipschitz property gives at most 0.025 difference in the corresponding mode-logistic probabilities, not a guarantee about Gaussian three-action serving.

The tolerance is a declared numerical budget, not a deployment-derived risk requirement or an optimized accuracy–computation trade-off.

On the same histories and tolerance we compare the certificate with the Newton heuristic $\nu_n \max_z \sqrt{z^\top H_n^{-1} z}$ and the strong-convexity gradient bound $\max_z \|z\|_2 \, \|g_n\|_2/2$. Both maxima use $z \in \mathcal{Z}_n^{\mathrm{vis}}$, and g_n is the full regularized objective gradient at μ_n. MAP reference fits pass independent derivative, residual and Hessian checks. After accounting for the bounded numerical reference error, all 4,800 cases could be classified relative to the tolerance.

Screen	Keep	Refit requests	Unnecessary requests	Missed violations
Certificate	824	3976	2205	0
Newton heuristic	1017	3783	2012	0
Gradient bound	28	4772	3001	0

Table 15: Shadow refit recommendations at 4,800 checkpoints. A missed violation retains a state exceeding the declared logit tolerance. An unnecessary request asks to refit a state already within tolerance. These are diagnostic counts on correlated histories, not executed model refits or population error-rate guarantees.

There are 3,029 states actually within tolerance and 1,771 exceeding it. The certificate retains 17.2% overall, while the Newton heuristic retains 21.2% and the gradient bound 0.6%, all without observed missed violations. The certificate is particularly conservative between 20 and 250 answers, despite becoming more permissive at the last checkpoint (Table 35). Its 72.0% retention at 940 answers is therefore not representative of the full learning trajectory. The certificate supplies a deterministic bound under its stated assumptions. The observed absence of Newton misses does not give that heuristic the same mathematical justification. This experiment nevertheless establishes no empirical screening advantage over Newton.

No refit is installed, so this diagnostic leaves subsequent questions and serving states unchanged. Instrumented gate and checked-reference timings do not measure production savings. The separate installed-refit pilot below tests actual changes to the learner.

Comparison with published rigorous bounds. On the same saved quantities, the whitened Bach radius $4\nu_n$ is valid in 3,539 cases and keeps only six, compared with 4,543 valid logarithmic bounds and 824 keeps. Neither misses a tolerance violation. The equivalent Marteau-Ferey et al. (2019) expression gives exactly the same 824 keeps and 3,976 refit requests as the logarithmic expression. This illustrates the looseness of the particular 2010 radius on these histories, not an improvement over the later localization result.

C.9 An operational refit pilot and fixed stress cases

We separately test whether the approximation certificate can control installed corrections. The pilot has ten generated users and five refit rules, each with 1,000 interactions and 100 scheduled answers: 50 trajectories and 5,000 answers. It uses DUAL's eleven-dimensional learner, prior precision $2I$, unit feedback scales and the original feedback-after-abstention rule. Generator seeds are fixed across arms. An answer arrives every ten interactions. The refit rule does not change this allowance. Before each later decision, the controller checks the current visible pair's absolute-A, absolute-B and pairwise designs. As in the shadow study, these are three diagnostic designs, not three available feedback questions.

Generated users, streams and seeds. Each user has an independent fixed vector $\theta = (w, \tau) \sim \mathcal{N}(0, 0.7^2 I_{11})$. Each coordinate of each response's ten-dimensional feature vector is independently $\mathcal{N}(0, 0.7^2)$, for both the 1,000 online pairs and a separate, fixed bank of 64 evaluation pairs per user. Answers are Bernoulli with probability $\sigma(z^\top\theta)$, using the three designs $[a, -1]$, $[b, -1]$ and $[a-b, 0]$. Evaluation occurs at interactions 0, 250, 500, 750 and 1,000. Its serving target is the latent noiseless A/B/abstain action from (w, τ), and its bank is inaccessible to question selection. All arms share a user's features, parameters and keyed answer randomness, although selected questions and histories can differ.

Appendix F.4 specifies the exact seeds and answer-key construction.

The certificate rule keeps the recursive state exactly when its valid bound B_n from Appendix C.8 is at most 0.10 logit. Otherwise, it requests a refit. Newton replaces R_n in that bound by ν_n, at the same tolerance. The remaining rules refit after every answer, after every ten answers, or never. A requested refit constructs the checked full-history logistic MAP, validates its residual and Hessian, and atomically installs its mean and full Hessian precision. The prior enters once, and subsequent answers continue the ordinary recursion. Certificate evaluations use the full selected-history gradient and Hessian at the current state, recomputed whenever the state or history changes. Thus Proposition 2 remains applicable after a refit. Equation (34) and Proposition 1 concern the unrefitted recursion and are not asserted for this hybrid trajectory. There is no extra refit after the terminal answer, so every-answer and every-ten-answer rules make 99 and nine refits per user, respectively. The operational requirement is fidelity to a specified expensive reference fit: a 0.10 mode-logit gap implies at most 0.025 difference in the corresponding mode-logistic answer probability, since $\sigma' \le 1/4$. This is a declared numerical tolerance, not a deployment-derived risk budget or a bound on integrated Gaussian serving probabilities.

Relative to every-answer refitting, the certificate avoids 809/990 refits (81.72%) and reduces observed whole-controller CPU by 21.41%. Newton's observed reduction is 22.69%. Both screens have zero false KEEP decisions, zero post-gate tolerance violations and no unresolved ordinary-pilot

Refit rule	Refits	CPU (s)	Violations	Max. logit gap	Final error
Never	0	11.679	6864	1.471343	17.9688%
Every answer	990	17.288	0	$< 10^{-8}$	17.6563%
Every ten answers	90	12.080	347	0.469160	16.5625%
Certificate	181	13.587	0	0.073593	18.4375%
Newton heuristic	174	13.366	0	0.075456	18.4375%

Table 16: Generated operational pilot. Refits and controller CPU are totals over ten users. Violations count post-gate current-query mode-logit gaps above 0.10 relative to a checked MAP fit to the same history, out of 10,000 decisions per rule. Final serving error is a descriptive equal-user mean on the generated evaluation bank.

boundaries. Newton uses fewer refits and slightly less CPU, so this pilot establishes no empirical advantage of the certificate over Newton. At a fixed history, approximate point and query set, the certificate radius is at least the Newton norm, hence certificate KEEP implies Newton KEEP. However, their evolving trajectories need not coincide.

The certificate is conservative. Before all 181 triggered refits, the actual mode-logit approximation error on the current questions was already within 0.10. The same holds for Newton's 174. This does not test whether omitting those refits would preserve future states. Moreover, the certificate trajectory has 22 Gaussian serving-action disagreements with its same-history MAP reference among 10,000 predecision checks. The bound controls mode approximation on the current designs, not recursive covariance, Gaussian action probabilities or confidence about the true user. No serving-quality gain is established by Table 16.

CPU measurements come from one serial run in a fixed arm order, without timing repetitions. Whole-controller CPU includes gate and cache work, checked refits and installation, serving and selection, the answer provider, recursive updates and bookkeeping. It excludes evaluation, event serialization and independent replay. Nested component timers overlap and must not be added. The measured reduction is descriptive for this generated pilot, not a claim about production latency, total pipeline cost or a general speedup.

Constructed stress cases. A separate fixed panel contains 36 native visible-pair cases and three arbitrary single-query cases. Certificate and Newton agree on all native cases, with no post-gate violation. Both refit four numerically unresolved boundary cases. Two arbitrary-query cases expose a Newton false keep that the certificate avoids. These constructions distinguish a sufficient bound from a heuristic without establishing an operational advantage on ordinary histories. Appendix F.5 gives every construction and outcome.

D NATIVE-MODEL BASELINES AND STRUCTURAL SENSITIVITY

These studies ask whether the original learner benefits from choosing the feedback channel, how it responds to an omitted preference feature, and how it compares with simpler timing policies under common calibration. Each study uses its own development cohort and protocol.

D.1 FIXED-MIXTURE BASELINE AND NONLINEAR PREFERENCE STRESS

Table 17 retains two checks of the original learner. The mixture matches the calibration channel frequency under the same periodic question schedule. It was not optimized for serving quality. The nonlinear study adds a hidden quadratic preference component while keeping DUAL's linear learner and information timing rule. These are different studies, not a common policy ranking.

Exact nonlinear perturbation. For the standardized ten-score feature vector f, define

$$q(f) = \frac{1}{\sqrt{5}} \sum_{j=1}^{5} f_{2j-1} f_{2j}, \qquad \widetilde{u}_i(f) = w_i^\top f + 0.5 s_i \, \frac{q(f) - m_q}{s_q}.$$

Here m_q, s_q are the mean and population standard deviation of q over all training prompts and their eight responses. The quantity s_i is the corresponding standard deviation of $w_i^\top f$. These training statistics are reused at evaluation. The added component therefore has training standard deviation $0.5s_i$. Because it can covary with the linear component, this scale does not specify its fraction of total variance. For acceptance stratum $a_i \in \{0.3, 0.5, 0.7\}$, the threshold is recomputed as the linearly interpolated $(1 - a_i)$-quantile of the transformed training utilities $\widetilde{u}_i$. Both feedback temperatures remain one. The learner retains its ten linear features and prior precision $2I$. This historical stress keeps the information threshold fixed. Appendix D.4 separately tests recalibration.

Development comparison (first minus second)	Users	AUC difference [95% CI]
Periodic mixed feedback versus fixed 225-pairwise/715-absolute periodic mixture	100	-1.164 $[-1.554, -0.773]$
DUAL versus periodic mixed feedback, hidden quadratic preference	100	$+1.447$ $[0.606, 2.337]$

Table 17: Development checks in mistake-AUC percentage points. Negative favors the first method. The mixture row averages three realizations within each user. Intervals are within-study paired-user 95% intervals, not a joint familywise claim across these separate studies.

The information threshold was selected on the original 200 development profiles. These checks reuse the historical development design. Stored metrics, quotas and paired-user intervals were independently reconstructed without rerunning every historical learner. The nonlinear result identifies a limit of DUAL under this omitted-feature condition.

The common-calibration study below compares adaptive absolute-only, confidence timing and a learned phase schedule. A serving-optimized fixed mixture remains a different baseline from the frequency-matched mixture above.

D.2 Equally calibrated native-feature policy assessment

Population and fixed learner. The study in Section 6.2 draws 100 calibration and 200 fresh assessment profiles separately from the original $\mathrm{Dirichlet}(0.05, \ldots, 0.05)$ generator. The official 1,000 held-out profiles are excluded. Acceptance strata remain $0.30, 0.50, 0.70$, with thresholds constructed from each profile's training utilities. The fixed native ten-reward-model representation, eleven-parameter Gaussian recursion, zero prior mean, precision $2I$, unit feedback temperature, displayed-response absolute target and serving rule are unchanged. It reuses the known prompt distribution and evaluator-only bank, so it is prospective simulated development evidence, not new task or human validation.

Each history contains 9,402 interactions, of which the first 9,400 are feedback eligible. Calibration uses one realization per user. Assessment uses three realizations of response pairs, prompt order and feedback randomness, shared across policies and averaged within each user. The six policies and two budgets give 7,200 assessment histories, with 200 independent user clusters for inference.

Calibration and budget contracts. DUAL, absolute-only information timing and low-confidence timing use the same 100 calibration profiles, selection objective and five-candidate search budget, separately for each budget. Candidate thresholds are the pooled $0.50, 0.75, 0.90, 0.95, 0.99$ quantiles of the appropriate pre-feedback score stream from periodic calibration histories. We evaluate each candidate with a fresh learner and select the one with the lowest mean bank-mistake AUC, with final error, online error, threshold and candidate identifier as deterministic tie breakers. No assessment profile selects a setting.

Confidence timing uses the Gaussian learner's pre-feedback score $1 - \max(\pi_A, \pi_B, \pi_\varnothing)$, requesting feedback when this score meets its calibrated threshold, subject to the same quota and feasibility rules. Question type still follows information gain. These action probabilities are not an independently validated confidence set.

The phase schedule fits ten within-block positions from DUAL's selected calibration question counts, separately for interactions 1–3,000, 3,001–6,000 and 6,001–9,400. Ties favor earlier positions. The schedule uses calibration question positions, not assessment outcomes. All thresholds

and schedules are frozen before assessment. Q940 supplies exactly ten answers per aligned 100-interaction block, counting either feedback channel as one answer. The main text emphasizes this answer-budget comparison. E940 is reported as a supplementary cost-assumption sensitivity and supplies exactly ten response exposures. Absolute feedback is charged one exposure unit and pairwise feedback two. This counts displayed responses. It does not measure reading time, cognitive effort or a two-to-one human-cost ratio. Under E940, periodic and phase positions specify exposure tokens consumed by the same deterministic scheduling mechanism. A causal feasibility guard ensures the exact allowance can be spent using the remaining slots. Mixed acquisition ranks feasible questions by raw information, not information divided by exposure cost. All serving decisions precede same-interaction feedback. The separate calibrations and exact-spending guard make E940 a distinct controller, not a cost rescaling of a saved Q940 trajectory. We label the mixed information-adaptive policy DUAL under each stated contract. E940 is its exposure-controller variant.

Exact exposure-feasibility rule. Within each 100-interaction block, let $B = 10 -$ exposures used and $s = 101 - j$ at position j, including the current slot. Set $m = 2$ for mixed policies and $m = 1$ for absolute-only. When $B > 0$, an adaptive policy admits only question costs

$$\mathcal{C}(B, s) = \{c \in \{1, \ldots, m\} : c \le B,\ B - c \le m(s - 1)\}.$$

It selects the highest raw-information question among those costs, and asks if its timing score reaches the calibrated threshold or $B > m(s - 1)$, which forces a question to finish the block. Confidence timing uses $1 - \max_a \pi_a$ as its timing score. After asking, B decreases by the chosen cost. Once $B = 0$, no further feedback is allowed in that block. A type override means the unrestricted best question was not feasible. For periodic and phase policies, the next unconsumed token position triggers a question. A pairwise answer consumes that token and the following token, and a final single token permits only absolute feedback. Every block ends at $B = 0$, without borrowing or carrying exposure units. Thus an exposure-limited result combines the question values, their assumed costs, and this exact-spending controller.

Endpoints and planned inference. The primary endpoint is normalized evaluation-bank mistake AUC over the 9,402-interaction horizon. The learner selects A, B or abstention from its Gaussian action probabilities. Mistakes are disagreements with the reference action in Equation (2), defined by the true user's utilities and acceptance threshold. This endpoint counts all three error types, rather than only rejection of a served response. Final bank error and actual online pre-feedback error are separate outcomes. The fixed 1,000-prompt bank is primary, while the 971-prompt duplicate-excluded view is sensitivity only. The study prespecified six primary contrasts comparing DUAL with absolute-only adaptive, confidence timing and the phase schedule under each budget. Multiplicity adjustment includes all six comparisons across both budgets. We average the three realizations within each user before 20,000 paired-user bootstrap resamples (seed 134990203). Table 18 uses linear-percentile endpoints with tail probabilities $0.025/6$ and $1 - 0.025/6$, a Bonferroni adjustment over the six planned contrasts. Periodic controls are secondary. Intervals containing zero establish neither superiority nor equivalence.

Comparator	940 answers	940 exposures
Absolute-only adaptive	-0.471 $[-0.756, -0.194]$	$+1.326$ $[+1.000, +1.651]$
Confidence timing	-0.248 $[-0.582, +0.086]$	-0.026 $[-0.348, +0.306]$
Phase schedule	-1.150 $[-1.399, -0.912]$	-0.597 $[-0.964, -0.232]$

Table 18: Complete six-contrast native-feature development comparison: DUAL minus comparator within each budget, in mistake-AUC percentage points. Negative favors DUAL. All point estimates and intervals retain the original 20,000 paired-user bootstrap draws and six-comparison Bonferroni correction for a 95% familywise target. Confidence and phase policies retain information-based channel selection. The main text reports the absolute-only answer-budget contrast and discusses the unresolved confidence-timing comparison. E940 assumes a one-to-two exposure cost for absolute versus pairwise feedback. It does not establish their relative human effort.

Exploratory online-error inference. A later analysis of the saved per-user outcomes adds paired intervals for the online means above, after those means and the primary results were known. Each

Policy	AUC	Final	Online	Answers	Exposures
Q940: equal answers					
DUAL	14.735	8.951	14.395	940.0	1158.0
Absolute-only adaptive	15.206	9.283	14.908	940.0	940.0
Mixed confidence timing	14.983	9.273	14.692	940.0	1141.1
Mixed phase schedule	15.885	10.017	15.534	940.0	1167.7
Mixed feedback, periodic	15.651	9.675	15.468	940.0	1168.2
Absolute-only, periodic	16.353	10.220	16.135	940.0	940.0
E940: equal response exposures					
DUAL (E940 controller)	16.532	11.535	16.206	601.9	940.0
Absolute-only adaptive	15.206	9.283	14.908	940.0	940.0
Mixed confidence timing	16.558	10.391	16.213	788.0	940.0
Mixed phase schedule	17.129	10.969	16.819	768.9	940.0
Mixed feedback, periodic	16.768	10.623	16.557	768.5	940.0
Absolute-only, periodic	16.353	10.220	16.135	940.0	940.0

Table 19: Complete native-feature development means. AUC, final bank error and online pre-feedback error are percentages, with lower values indicating better performance. Answer and exposure counts are means across users after within-user realization averaging. Identical absolute-only budget rules give identical Q940/E940 results. The two budgets have separate calibrations and are not a single policy evaluated at two points on one trajectory.

outcome is mistakes divided by 9,402 actual interactions, with decisions committed before feedback and the three realizations averaged within person. We use 100,000 paired-user bootstrap draws over the same 200 profiles (PCG64 seed 134917701), ordinary 95% percentile intervals, and a Bonferroni adjustment over the six new online contrasts. No new learner histories are generated. The original six bank-AUC comparisons retain their separate 20,000-draw intervals. Table 20 supports the same qualitative pattern on actual online mistakes: DUAL improves on absolute-only adaptive at equal answers, loses at equal exposure, and improves on the tested phase schedule under both contracts. Both adjusted confidence-timing comparisons remain unresolved, including Q940 where the ordinary 95% interval is favorable. This is supporting exploratory development evidence, not an additional held-out confirmation.

Budget	Comparator	Difference	95% interval	Adjusted interval
Q940	Absolute-only adaptive	-0.513	$[-0.724, -0.305]$	$[-0.798, -0.236]$
Q940	Confidence timing	-0.297	$[-0.546, -0.051]$	$[-0.632, +0.035]$
Q940	Phase schedule	-1.139	$[-1.319, -0.962]$	$[-1.382, -0.903]$
E940	Absolute-only adaptive	$+1.298$	$[+1.055, +1.537]$	$[+0.973, +1.620]$
E940	Confidence timing	-0.007	$[-0.242, +0.228]$	$[-0.325, +0.311]$
E940	Phase schedule	-0.613	$[-0.879, -0.348]$	$[-0.972, -0.255]$

Table 20: Exploratory native online-mistake contrasts, in percentage points: DUAL minus comparator within each budget. Each participant contributes the mean of three pre-feedback error rates over 9,402 interactions. The six contrasts share a Bonferroni correction (100,000 paired-user draws with 99.1667% marginal intervals). The original bank-AUC family is unchanged.

Exposure sensitivity. At equal answers, DUAL displays 1,158 responses versus 940 for absolute-only adaptive, a 23.2% difference in exposure rather than measured effort. At equal exposures it receives only 601.9 answers, and the AUC ordering reverses (Tables 18–19). Its AUC advantage over the phase schedule does not extend to final bank error, which is lower for phase and confidence timing. All six primary point-estimate directions persist in the duplicate-excluded view. Across realizations, directions agree except for the unresolved E940 confidence comparison.

Which online mistakes change? We decompose the saved Q940 pre-feedback action counts into three disjoint errors: serving when neither response is acceptable, abstaining when at least one is acceptable, and serving the other response when the oracle would serve. The third category includes both an unacceptable alternative and a worse acceptable alternative, which the saved action-

confusion counts cannot separate. Each rate divides its count by all 9,402 interactions, rather than by the number of serves. The three rates sum to each user's recorded online mistake rate.

This exploratory analysis uses the same 200 assessment profiles, averaging three realizations within each person. We retain all six component contrasts against absolute-only adaptive and confidence timing. Intervals use 100,000 paired-user bootstrap draws with PCG64 seed 134919001, linear percentiles and Bonferroni tails $0.025/6$ and $1 - 0.025/6$. They form a separate exploratory family and do not replace the original primary or online-total comparisons. A separate implementation reconstructs the counts, component sums and every resampling mean.

Online error	DUAL minus absolute-only adaptive	DUAL minus confidence
Serve when neither is acceptable	-0.002 $[-0.168, +0.169]$	$+0.345$ $[+0.116, +0.580]$
Abstain when one is acceptable	-0.189 $[-0.419, +0.048]$	-0.459 $[-0.789, -0.127]$
Serve the other response	-0.322 $[-0.456, -0.195]$	-0.184 $[-0.335, -0.037]$

Table 21: Exploratory differences in whole-history online error components, in percentage points per interaction. Negative favors DUAL. All six intervals use the stated Bonferroni adjustment with a 95% familywise target.

Relative to absolute-only adaptive, the clearest component improvement is fewer wrong-response choices when an acceptable response exists. Against confidence timing, DUAL reduces incorrect abstention and wrong-response choices but serves more often when neither response is acceptable. This trade-off is compatible with their unresolved adjusted online-total comparison. These whole-history counts do not isolate the effects of channel choice, timing or a particular learning phase.

Early learning and quota behavior. Table 22 reports error when each policy reaches the stated number of answers. Interaction positions and exposure counts can differ. These checkpoints show progress through policies designed for 940 answers. They do not establish the best policy when only a few answers are available. At the reported Q940 checkpoints through 100 answers, confidence timing has lower mean error than DUAL. DUAL has lower means at 250, 500 and 940 answers, and lower means than absolute adaptive at all eight reported checkpoints. This change in the descriptive ordering does not identify a switching point or validate a hybrid policy. E940 mixed-policy checkpoints at 940 answers are unavailable and are not interpolated.

Answers	DUAL	Absolute-only adaptive	Confidence timing
5	42.505	43.906	42.466
10	39.338	40.144	39.141
20	36.333	36.940	34.996
50	29.214	29.887	27.571
100	24.322	24.995	22.415
250	16.265	16.798	16.546
500	11.493	11.885	12.439
940	8.951	9.283	9.273

Table 22: Descriptive received-answer bank-error percentages under Q940. All 200 assessment users reach these checkpoints. The primary comparison remains the full-horizon AUC in Table 18.

DUAL uses 722.0 absolute and 218.0 pairwise answers under Q940. Under E940 the corresponding counts are 263.9 and 338.1. This documents the allocation difference without isolating its effect on AUC. Table 23 shows how the exact exposure contract interacts with question choice. A quota-forced ask occurs when the timing threshold would not request feedback but the remaining allowance requires it. A type override occurs when the unrestricted highest-information question is infeasible under the remaining exposure allowance and slots. In the final phase, 160.8 of 177.4 asks are quota-forced and 132.8 have type overrides. There are 213.6 type overrides over the full E940 history on average. Under Q940, 330.3 of the final phase's 340 asks are quota-forced, with no type overrides.

The E940 loss is specific to this assumed exposure cost and controller. Fewer received answers, raw rather than cost-normalized information ranking, and forced question-type changes are possible

E940 phase	Absolute	Pairwise	All asks	Quota-forced	Type overrides
1–3,000	190.0	55.0	245.0	1.0	4.2
3,001–6,000	59.2	120.4	179.6	104.1	76.6
6,001–9,400	14.7	162.7	177.4	160.8	132.8

Table 23: Observed phase diagnostics for DUAL's E940 controller. Entries are mean counts. Quota-forced asks and type overrides can overlap. The implemented policy combines raw information ranking with exact budget feasibility. These observations are not a causal decomposition of its loss against absolute-only questioning.

contributors. The study does not isolate their causal effects. It therefore establishes neither an intrinsic inefficiency of mixed feedback nor a measured human-effort ranking.

The study compares these native policies under common calibration, with uncertainty assessed over users. Its conclusions depend on the simulator, budget contract and frozen learner. Verification details are collected in Appendix A.1.

D.3 Separating exposure-aware scoring from quota completion

The earlier E940 controller ranks raw information and may force pairwise questions to spend its remaining quota. We separately change the scoring rule and the completion rule to test whether either explains that loss. This new study draws 24 calibration and 120 assessment profiles from the same Dirichlet(0.05) native-feature model, with PCG64 seeds 134920701 and 134920702. It retains 9,402 interactions, the original prior and recursive learner, unit feedback temperatures, Gaussian serving, the original balanced pair construction and the 1,000-pair evaluator. Each policy receives exactly ten response-exposure units per 100-interaction block, totaling 940. Absolute questions cost one unit and pairwise questions cost two. No original held-out profile is reused. Two realizations, 134930101 and 134930102, are averaged within each assessment user.

Two controller changes. The score is either the local information score from Equation (6) or that score divided by question cost. The legacy completion rule can defer questions until spending the remaining units requires cost-two answers. The alternative starts forcing questions when remaining exposure units equal remaining interaction slots. It can then complete using cost-one answers and never has to force a pairwise question. Both variants exclude unaffordable candidates. An absolute-only adaptive policy supplies the simpler comparison. Dividing by cost changes the greedy rule, so Proposition 1 does not directly certify the cost-normalized variants.

For each score type, periodic calibration histories supply five threshold candidates at pooled score quantiles 0.50, 0.75, 0.90, 0.95 and 0.99. Every policy evaluates all five on the same 24 users, choosing minimum mean mistake AUC, then final and online error, then candidate order. Thresholds are frozen before assessment profiles are drawn. Absolute-only selects q75, both raw-information variants and cost-normalized legacy completion select q50, and the combined cost-normalized/earlier-completion variant selects q99. Thus the comparisons include each policy's calibrated timing, not just an intervention on one decision within a common history.

Information per exposure with legacy completion reduces AUC from 16.326% to 15.384%, a 0.942-point improvement with adjusted interval $[-1.406, -0.444]$. Earlier completion alone gives a smaller 0.047-point improvement within the four-comparison family. Under the broader sensitivity in Appendix D.4, the information-per-exposure improvement remains favorable, while the earlier-completion contrast becomes unresolved. Combining the changes does not improve on absolute-only, which has the lowest observed mean at 14.918%. Of the combined policy's 902.2 answers, 880.8 are collected under the completion condition. Cost-aware scoring therefore repairs part of this mixed controller's loss, but these results do not establish mixed feedback as the preferable policy at this exposure budget. The calibration cohort is small, and inference conditions on its frozen choices and the known native prompt distribution.

Policy	AUC (%)	Answers	Pairwise	Completion	Overrides
Absolute-only adaptive	14.918	940.0	0.0	564.1	0.0
Raw information, legacy completion	16.326	712.1	227.9	88.3	67.4
Information/cost, legacy completion	15.384	823.2	116.8	104.9	87.7
Raw information, earlier completion	16.278	764.5	175.5	148.7	12.5
Information/cost, earlier completion	15.783	902.2	37.8	880.8	4.2

First policy minus second	AUC difference	Adjusted interval
Both changes − absolute-only	$+0.866$	$[+0.315, +1.440]$
Both changes − original mixed controller	-0.542	$[-1.140, +0.049]$
Information/cost − raw, legacy completion	-0.942	$[-1.406, -0.444]$
Earlier − legacy completion, raw score	-0.047	$[-0.091, -0.004]$

Table 24: New E940 controller assessment on 120 simulated profiles. Upper panel: means after averaging two realizations per user. All policies spend 940 exposures. Completion counts include questions that also meet the threshold, and overrides include exclusion of unaffordable pairwise questions. Lower panel: all four planned AUC contrasts, in points, using 20,000 paired-user draws, seed 134920703, and four-comparison Bonferroni adjustment. This cohort and calibration differ from Table 18.

D.4 Timing recalibration with an omitted preference feature

The historical nonlinear stress leaves two possibilities open: the timing threshold may need recalibration, or the missing feature may still limit information-based timing after recalibration. This study compares those policies on the same new users. It preserves the hidden quadratic construction in Appendix D.1 and keeps the learner's ten linear features, threshold coordinate, prior, feedback temperatures and serving rule. Every policy receives 940 answers, ten per 100-interaction block, over 9,402 interactions. Periodic mixed feedback asks every tenth interaction, while adaptive mixed feedback applies its information threshold and quota guard. The last two interactions receive no feedback.

We draw 24 calibration and 120 assessment profiles independently from Dirichlet(0.05), using PCG64 seeds 134920801 and 134920802 and the original balanced acceptance strata. Assessment averages realizations 134930101 and 134930102 within each user. No original held-out profile is reused. Separate periodic calibration histories under matched and nonlinear preferences provide candidate thresholds at score quantiles 0.50, 0.75, 0.90, 0.95 and 0.99. Each candidate is evaluated on all 24 calibration users. Selection minimizes mean bank-mistake AUC, then final error, online error and candidate index. Both choices are fixed before assessment profiles are drawn.

Matched calibration selects q75, a threshold of 0.0212016285. Nonlinear calibration selects q99, a threshold of 0.199570665. The frozen policy uses the matched choice under both user models. The recalibrated policy uses the nonlinear choice under nonlinear preferences. Thus only the timing threshold changes between the two nonlinear adaptive policies. Calibration still conditions inference, and the five-candidate search does not exhaust all possible timing rules.

Adaptive timing improves on periodic timing under matched preferences and loses under the omitted feature. Recalibration reduces mean nonlinear AUC by 0.744 points, but periodic timing remains 0.633 points lower within this study. The recalibrated policy asks 918.4 of its 940 questions under the completion condition, so its mean gain does not demonstrate stronger example-specific information timing. Under the broader sensitivity below, neither recalibration contrast excludes zero. These results support partial mean recovery after threshold adjustment, without establishing restoration of the matched-model advantage or diagnosing every consequence of the missing feature.

Sensitivity across the four new studies. The PRISM replication, controlled corruption, exposure-controller study and timing recalibration have 1, 16, 4 and 4 planned contrasts, respectively. Their tables retain those declared comparison families. As an additional conservative sensitivity, we apply Bonferroni adjustment across all 25 contrasts. The information-per-exposure improvement with legacy completion remains favorable, with interval $[-1.522, -0.328]$ points. The combined

User model	Mixed-feedback timing	AUC (%)	Online (%)	Pairwise	Completion
Matched	Periodic	15.969	15.764	227.2	0.0
Matched	Calibrated adaptive	14.931	14.562	219.0	572.4
Nonlinear	Periodic	21.714	21.914	225.3	0.0
Nonlinear	Frozen adaptive	23.091	23.123	214.8	562.5
Nonlinear	Recalibrated adaptive	22.347	22.537	221.4	918.4

First policy minus second	AUC difference	Adjusted interval
Matched adaptive − periodic	-1.037	$[-1.563, -0.501]$
Nonlinear frozen − periodic	$+1.377$	$[+0.768, +2.007]$
Nonlinear recalibrated − periodic	$+0.633$	$[+0.011, +1.267]$
Nonlinear recalibrated − frozen	-0.744	$[-1.442, -0.054]$

Table 25: Timing recalibration on 120 new simulated profiles, each with two realizations and 940 answers per policy. Upper panel: mean outcomes and completion counts, including asks that also meet the threshold. Online error is descriptive. Lower panel: all four planned AUC contrasts in points, using 20,000 paired-user draws, seed 134920803, and four-comparison Bonferroni adjustment. The broader sensitivity below qualifies the two recalibration contrasts.

exposure-controller changes remain worse than absolute-only, $[+0.182, +1.573]$. Earlier completion alone becomes unresolved, $[-0.103, +0.007]$.

For timing, matched adaptive minus periodic remains favorable, $[-1.713, -0.376]$, and nonlinear frozen minus periodic remains adverse, $[+0.613, +2.164]$. The nonlinear recalibrated-minus-frozen interval widens to $[-1.577, +0.103]$, and recalibrated-minus-periodic to $[-0.145, +1.411]$. In PRISM, the replication and all comparisons with absolute-only or alternation remain unresolved. Information selection remains favorable to pairwise-only under 10% and 20% pairwise corruption, with intervals $[-11.424, -0.143]$ and $[-11.115, -0.579]$. The broader sensitivity changes no other zero-exclusion conclusion. It does not replace the original study families or correct for every historical exploratory decision.

D.5 Fixed allocations and adaptive question choice

Question and shared protocol. This development study asks whether a serving-calibrated fixed allocation can replace information-based question selection. It uses the native ten reward-model features, the original Gaussian recursion, a zero initial mean and precision $2I$. All policies receive 940 answers at interactions $10, 20, \ldots, 9{,}400$ of the 9,402-interaction stream. The current belief determines the serving action before the question is asked. Only the selected answer is revealed. Absolute questions target the served response, or the response with higher predicted acceptability after abstention. The two policies differ in their choice between that absolute question and the current pairwise comparison.

Profiles and calibration. We draw 100 calibration profiles and 200 separate assessment profiles, with preference weights from $\mathrm{Dirichlet}(0.05\,\mathbf{1}_{10})$. Outcome-independent assignments select training-utility thresholds for acceptance rates $0.30, 0.50, 0.70$. Policies share the response-pair map, prompt order and compatible feedback randomness within a profile and realization. The fixed policy assigns a prescribed number of pairwise questions to a random permutation of the 940 question positions, determined before feedback. Calibration initially considers pairwise fractions 0, 0.25, 0.50, 0.75 and 1. After the calibration minimum occurs at zero, a refinement adds 0.025, 0.05, 0.10, 0.15 and 0.20, before assessment outcomes are inspected. The pairwise count is the nearest integer to 940 times the fraction. Mean calibration AUC selects the fraction, with final error and then the smaller fraction breaking ties. Zero remains the selected fraction.

Assessment and uncertainty. We assess three realizations per profile. Their seeds are 134930101, 134930102 and 134930103. Calibration uses the first realization seed. The independent profile-generation seeds are 134925251 for calibration and 134925252 for assessment. We evaluate the same 1,000-pair bank at interaction zero, every 100 interactions and the final horizon. The normal-

ized trapezoidal AUC is computed as in the main simulation. The bank does not inform question selection.

We average realizations within each assessment profile before resampling profiles. With 20,000 paired bootstrap draws and seed 134925253, information selection minus the selected fixed allocation is -0.599 percentage points, with a 95% interval $[-0.822, -0.382]$. The two planned comparisons are with the selected fixed allocation and with absolute-only feedback. Their Bonferroni-adjusted interval is $[-0.853, -0.351]$. These contrasts coincide because calibration selects absolute-only, so they provide one comparison rather than two independent findings. The initial five-fraction search and its refinement have the same selected comparator and assessment result.

Interpretation. The comparison isolates adaptive question choice at fixed question times. It does not test adaptive timing or optimize all possible fixed curricula. The profiles are newly generated development profiles, distinct from the original held-out simulated profiles.

E Retrospective studies with recorded human feedback

PRISM's graded ratings support acceptance-aware decisions, while Community Alignment's four-way choices support a categorical preference extension. These retrospective assessments use recorded human feedback, not new judgments or conversations produced by our policies. They do not measure interruption or reading effort. Dataset-specific provenance and outcome isolation are given below.

E.1 Surface representation for the original studies

The original PRISM studies and Community Alignment use twelve fixed measurements computed from prompt and response text: log word count, log mean sentence length, log paragraph count, the fraction of nonempty lines formatted as lists, repeated-bigram fraction, unique-word fraction, prompt-word coverage, response-word overlap with the prompt, a code-fence indicator, questions per sentence, hedge-word fraction, and second-person-word fraction. Count transforms are $\log(1 + x)$, and empty denominators are bounded below by one. Tokenization uses lowercase English word matches. The hedge lexicon is {may, might, perhaps, possibly, probably, likely, generally, usually, sometimes, could}. The second-person lexicon is {you, your, yours, yourself, yourselves}. These inexpensive surface measurements use no LLM scoring or human rating as a feature. They do not measure factual correctness or establish a complete representation of taste. Normalization is frozen separately for each study using earlier population data, with no current assessment outcomes used to choose features or scales.

E.2 PRISM with a population-initialized user model

Cohorts and information available to the learner. This exploratory follow-up uses the recorded PRISM pairs and rating-defined targets of the original studies. The 64 population, 89 development and 90 assessment participants are disjoint. Each development or assessment participant supplies ten history pairs and four separate evaluation pairs. The assessment participants were used previously, so the follow-up is not an untouched confirmation study. Recorded conversation order, fixed within-interaction pairs, rating cutoff 70, canonical acceptable-tie handling and the five-question schedule are retained.

Fixed representation. Each response has 24 features. Twelve are the original surface measurements. The remaining twelve comprise eleven principal components of a frozen sentence embedding and the cosine similarity between the prompt and response embeddings. We use Sentence Transformers (Reimers & Gurevych, 2019) with the `all-MiniLM-L6-v2` checkpoint.[1] Prompt and response texts are encoded separately, using normalized 384-dimensional embeddings and the checkpoint's 256-word-piece limit. The encoder is not fitted to PRISM. Principal components and feature normalization use only the 640 canonical-A responses from the population participants. Assessment features are transformed with these fixed quantities.

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. The checkpoint revision is `1110a243fdf4706b3f48f1d95db1a4f5529b4d41`.

Population initialization. A shared parameter vector supplies the same starting mean for every assessment participant. For each of the 640 population pairs, we construct two binary acceptability labels from the recorded ratings and one pairwise label when the ratings differ. This gives 1,280 absolute labels and 607 strict pairwise labels. Let $\mathcal{D}_{\text{pop}}$ denote these 1,887 design–label pairs, with designs defined by the feedback model. We obtain the 25-dimensional starting mean by minimizing

$$\mu_{\text{pop}} = \arg\min_{\eta \in \mathbb{R}^{25}} \left\{ \sum_{(z,r) \in \mathcal{D}_{\text{pop}}} \left[\log(1 + \exp(z^\top \eta)) - r z^\top \eta \right] + \|\eta\|_2^2 \right\}.$$

The objective uses prior precision $2I$ around zero. The personal recursion then starts from $\mu_{i,0} = \mu_{\text{pop}}$ and $\Lambda_{i,0} = 8I$. Thus the transferred quantity is the fitted mean, while the personal initial precision is separately selected. The population-fit Hessian is not imported as the personal precision. The acceptance threshold remains a boundary in feature space, distinct from the numeric rating cutoff.

Configuration selection and controls. The development criterion is the mean five-answer serving AUC averaged over information-selected mixed and absolute-only feedback. It does not maximize the difference between those policies. The initial 39 configurations vary the representation, population prior, initial precision, update rule and observation scale. The selected configuration combines the surface and embedding features with the population mean and initial precision $8I$. A later 20-configuration family of externally supervised quality features was explored after initial assessment results were seen. It does not change the selected configuration under the development criterion. The broader search is retained as exploratory provenance and does not supply a new confirmation claim.

All feedback policies receive the same representation, initial belief, pairs and question opportunities. Updating policies use the original Gaussian recursion and unit logistic scales. They receive five answers over ten history interactions, while the population-only reference receives none. Pairwise ties consume a question without a binary update. Acquisition sees only the current pair. We store predictions before calculating assessment scores.

Metrics and intervals. At each of the eleven history checkpoints, we evaluate the four-pair bank and integrate its mistake rate over the ten interactions. Participants have equal weight. This AUC measures disagreement with rating-defined serving actions, not rejection among served responses. Table 3 gives ordinary paired 95% bootstrap intervals using 200,000 participant resamples and seed 134925301. These intervals describe exploratory comparisons on a reused cohort. A broader 200-comparison sensitivity, covering the report's 161 contrasts, gives mixed-minus-control intervals of $[-5.56, 4.03]$ for population-only, $[-1.22, 1.81]$ for absolute-only and $[-8.33, 2.60]$ for pairwise-only feedback, in percentage points. This sensitivity was not a preregistered comparison family.

What the controls establish. Most of the selected model's performance is present before personal feedback: population-only AUC is 58.89%, compared with 58.14% for mixed feedback and 57.86% for absolute-only feedback. The intervals do not establish an additional five-answer gain or superiority from mixing. The original surface-only, zero-mean mixed learner has AUC 68.71% on these same participants. The selected model's difference from it is -10.57 points, with an ordinary 95% interval $[-16.51, -4.78]$ and broader sensitivity interval $[-22.03, 0.26]$. This change combines population training, representation and regularization, rather than isolating any one of them.

False-abstention AUC falls from 31.31% for the original mixed learner to 15.72% for the selected mixed learner. Unacceptable-serving AUC is 29.86% and 30.81%, respectively. The lower total AUC therefore should not be interpreted as fewer unacceptable responses served. The original cold-start experiment's advantage over pairwise-only feedback remains a result of that original configuration.

Sensitivity to budget, cutoff and allocation. At ten answers, selected-model AUC is 59.92% for mixed and 58.08% for absolute-only feedback, with difference 1.83 points and 95% interval $[-0.19, 3.93]$. At five answers, changing the cutoff to 60 gives 52.85% and 52.15%, while cutoff 80 gives 59.29% and 59.51%. The corresponding mixed-minus-absolute intervals are $[-0.50, 1.97]$

and $[-1.35, 0.83]$. A fixed-mixture comparison selects 25% pairwise feedback on development participants at the five-answer budget, then averages three predetermined allocation phases within each assessment participant. It obtains 58.30% AUC, compared with 58.14% for information selection. Their difference is -0.16 points, with interval $[-1.54, 1.17]$. These sensitivities leave the advantage of mixing unresolved.

E.3 Original PRISM: channel choice and acceptance-aware serving

Cohort, provenance and outcome isolation. A metadata-only selection uses all 89 development-role PRISM participants meeting the ten-interaction history and test-separation criteria after excluding 452 people in the documented recent-project scope (Kirk et al., 2024b). The exclusions comprise 308 earlier participants and 144 population/development participants in the preceding threshold pilot. The current protocol followed preference-transfer and threshold-learning pilots whose outcomes were known. Its cohort and protocol were fixed before its own outcomes, but it remains development evidence. Non-overlap is established with the documented pilots, not every prior project use.

Each person supplies ten earlier response pairs and four evaluator pairs, two from each of the last two recorded conversations. One pair per interaction is fixed without ratings. The 356 test prompts are distinct after case folding and whitespace normalization and disjoint from selected history and preceding pilot prompt groups. Completion timestamps order conversations retrospectively, without establishing live chronology. Later contexts come from the original platform. The original assessment-role cohort is excluded, with the separate replication in Appendix E.4.

Learner and interaction contract. We reuse only the twelve equations and normalization from 640 earlier responses belonging to 64 separate population-role participants. All channel arms start with mean zero and precision $2I$ over twelve preference weights and an acceptance threshold. The full-covariance sequential local Laplace update, unit feedback scales, Gaussian-region serving rule and information score follow DUAL at this adapted dimension. No population weights or learned reliability are imported. The primary schedule asks at interactions $2, 4, 6, 8, 10$, while a sensitivity schedule asks at all ten. All updating arms have the same pairs, order, initialization and question opportunities. We compare absolute-only, pairwise-only, fixed alternation and information selection, with a zero-question no-feedback reference. Alternation starts absolute for 45 people and pairwise for 44, fixed without outcomes.

Serving is committed before feedback. An absolute query targets the served response. After abstention, it targets the response with higher acceptance prediction at the belief mean, with canonical A breaking a tie. Questions after abstention are allowed. Acquisition sees only the current pair. An absolute query reveals its target's binary label. A pairwise query reveals only the relative rating, not either score or acceptability label. At most one query is asked per interaction, at unit cost in either channel. All trajectories and evaluator predictions are committed before evaluator ratings are opened. Thus this study tests *which channel* at fixed times, not adaptive timing.

Rating-defined targets and ties. The primary acceptability proxy is a recorded rating of at least 70 on the 0–100 scale. Cutoffs 60 and 80 are prespecified sensitivities. The learned threshold is a boundary in feature space, not the numeric rating cutoff. The correct serving action is abstention if both ratings are below the cutoff, and otherwise the higher-rated acceptable response. Canonical A resolves equally rated acceptable pairs for the primary outcome. A separate sensitivity counts either as correct. Tied pairwise answers consume one question but produce no binary update or replacement query. They do not imply that either response is acceptable. All pairs remain in serving evaluation. Only binary pairwise prediction metrics exclude rating ties.

Primary serving result. The primary endpoint is each person's four-pair evaluator-bank mistake rate, integrated over all eleven interaction checkpoints using trapezoidal AUC divided by ten, then averaged equally across participants. It differs from final accuracy and from mistakes on the actual ten learning interactions. Three planned information-minus-control comparisons use 20,000 paired-person bootstrap replicates (PCG64 seed 134917402) and Bonferroni-adjusted 98.333% percentile intervals. Table 26 gives every method and the full primary comparison family.

The supported improvement over pairwise-only is consistent with the need for acceptance information, but does not isolate its causal mechanism: channel content, selected examples and tied-answer

Method	Serving-mistake AUC ↓ Five questions	Ten questions	Final accuracy, five questions ↑ Acceptability	Pairwise
No feedback	72.753%	72.753%	52.95%	53.65%
Absolute-only	67.514%	63.933%	56.88%	52.25%
Pairwise-only	72.584%	69.649%	50.14%	51.40%
Fixed alternation	69.143%	68.581%	56.32%	53.37%
Information selection	66.685%	66.699%	59.55%	54.03%

Comparator	AUC difference (points)	Adjusted interval
Absolute-only	-0.829	$[-3.413, +1.451]$
Pairwise-only	-5.899	$[-10.604, -1.152]$
Fixed alternation	-2.458	$[-6.180, +1.124]$

Table 26: PRISM, 89 development participants and acceptability cutoff 70. Five questions is primary, while ten is a sensitivity. Final acceptability accuracy uses 712 responses and pairwise accuracy uses 337 strict comparisons, excluding 19 ties, with metrics averaged within person first. The lower panel gives all three primary information-minus-control AUC contrasts in points, retaining the three-comparison adjustment. Negative favors information selection. No feedback receives zero questions.

update counts differ. The selector requests 339 absolute and 106 pairwise questions across people, or 76.18% absolute. Its seven ties leave 438 binary updates, compared with 445 for absolute-only, 416 for pairwise-only (29 ties), and 430 for alternation (15 ties). Each updating method receives 445 total questions. The corresponding displayed-response counts are 551, 445, 890 and 667. These are descriptive exposure proxies, not measured effort or an exposure-matched comparison. The following exploratory analysis adds paired uncertainty and an error decomposition to the online means.

Online mistakes and error decomposition. We analyze the saved outcomes at the primary cutoff and five-question schedule without repeating trajectories or model fits. The new analysis was specified after the primary results and online means had been inspected. It uses 100,000 paired-person bootstrap replicates (PCG64 seed 134917702), resampling the 89 people rather than individual pairs. Ordinary 95% intervals accompany a conservative Bonferroni correction across all 27 new contrasts: three online-total comparisons and 24 component comparisons (four components, three controls, two evaluation contexts). The adjusted marginal level is 99.8148%. This exploratory family does not change the original three primary bank-AUC comparisons or the separate prediction analysis below.

Each mistake falls into one of four mutually exclusive categories. The system may serve an unacceptable response, abstain when an acceptable response exists, choose the strictly worse acceptable response, or choose B on an equally rated acceptable pair when the primary canonical action is A. The last is a tie-breaking penalty, not evidence that the user disliked the chosen response. Table 27 retains all four types, including the no-feedback reference. Bank components integrate the same eleven saved checkpoints as the primary AUC, whereas online components use the ten actual learning interactions before feedback. All component rates use all pairs in the respective context, not only the pairs that were served.

All three adjusted online-total comparisons remain unresolved (Table 28), although the pointwise 95% intervals favor information selection over pairwise-only and alternation. Relative to pairwise-only, the observed bank-AUC benefit is mainly fewer false abstentions: -7.472 points, adjusted interval $[-14.323, -0.969]$. The online false-abstention difference is -10.449 points $[-16.629, -4.494]$, but worse-acceptable choices increase by $+3.820$ points $[+0.674, +7.865]$. Reduced unacceptable serving is not established in either context. Tables 37 and 38 report every component contrast. This arithmetic decomposition does not isolate threshold learning causally: channel content, selected examples, and the tied-answer update counts above differ across policies. It supports a narrower interpretation of the primary benefit without changing the original endpoint.

Separate acceptability and preference prediction. For both binary tasks, a query design z has predicted answer probability $p = \sigma(m)$, where $m = z^\top \mu$ uses the fitted mean rather than integrating

Method	Total	Unacc.	False abs.	Worse acc.	Tie
Evaluation-bank component AUC					
No feedback	72.753	30.056	37.640	4.213	0.843
Absolute-only	67.514	31.334	27.219	8.118	0.843
Pairwise-only	72.584	31.306	33.778	6.882	0.618
Fixed alternation	69.143	31.952	28.764	7.570	0.857
Information selection	66.685	31.545	26.306	8.174	0.660
Actual pre-feedback online errors					
No feedback	74.157	23.146	46.067	4.607	0.337
Absolute-only	66.966	27.191	29.213	9.551	1.011
Pairwise-only	71.011	26.180	37.753	6.629	0.449
Fixed alternation	70.000	27.978	31.798	9.551	0.674
Information selection	66.854	28.202	27.303	10.449	0.899

Table 27: PRISM error decomposition, five questions and cutoff 70. All entries are percentages, equally weighted over 89 people. The four components sum to total error before rounding. Online rates use all ten learning interactions, while bank AUC integrates eleven checkpoints, not just the final state. Unacceptable-serve rates use all interactions or bank pairs as the denominator, not only served responses.

Comparator	Difference	95% interval	Adjusted interval
Absolute-only	-0.112	$[-1.573, +1.236]$	$[-2.360, +2.022]$
Pairwise-only	-4.157	$[-7.528, -0.674]$	$[-9.551, +1.282]$
Fixed alternation	-3.146	$[-5.618, -0.674]$	$[-7.079, +0.899]$

Table 28: Exploratory PRISM online-total-error contrasts: information selection minus comparator, in percentage points. All three adjusted intervals contain zero. They share the new 27-contrast correction with the component analyses. This does not alter the original three primary bank-AUC intervals.

over the Gaussian belief. For label $y \in \{0, 1\}$, log loss is $\log(1 + e^m) - ym$, evaluated with a stable log-sum-exp calculation without probability clipping. Brier score is $(p - y)^2$. Predictions are positive when $m \geq 0$. Scores are averaged within each participant and then equally across participants. These answer predictions differ from the joint Gaussian action-region probabilities used for serving.

All methods are evaluated on both tasks. At five questions, information selection has the highest observed accuracy, with ordinary 95% intervals $[55.20, 64.04]\%$ for acceptability and $[48.88, 59.08]\%$ for preference. The metrics were prespecified, while the interval analysis followed inspection of their means. Its 24 contrasts cover two tasks, three metrics and four comparators, including no feedback. The original 20,000 paired-person resampling indices give 99.7917% Bonferroni intervals.

Only acceptability accuracy versus pairwise-only remains favorable: $+9.410$ points $[+0.562, +18.258]$. Comparisons with absolute-only ($+2.669$, $[-2.949, +8.287]$), alternation ($+3.230$, $[-3.792, +10.112]$) and no feedback ($+6.601$, $[-2.669, +15.730]$) are unresolved. All preference-accuracy and proper-score contrasts are also unresolved. The selector's preference accuracy is only 0.375 points above no feedback. A larger resampling check preserves every conclusion (Appendix F.7).

The constant-0.5 reference has lower observed log loss and Brier score than the selector on both tasks (Table 29), but those paired differences are unresolved even under ordinary 95% intervals. The reference's non-50% accuracy reflects fixed prediction tie-breaking and label balance. Neither better nor worse probability quality is established. Surface features, cold start, sparse feedback and approximation error remain possible limitations, which this experiment cannot distinguish from inconsistent preferences.

Prespecified sensitivities. Table 30 reports every cutoff and budget. Absolute-only has lower observed serving AUC at ten questions for all three cutoffs and at five questions for cutoff 80. At

Method	Acceptability		Pairwise preference	
	Log loss	Brier	Log loss	Brier
No feedback	0.693147	0.250000	0.693147	0.250000
Absolute-only	0.723182	0.257623	0.735088	0.260145
Pairwise-only	0.748893	0.271386	0.723406	0.261522
Fixed alternation	0.723509	0.259734	0.725173	0.259623
Information selection	0.708907	0.252036	0.722987	0.257454

Table 29: Final PRISM probability-quality metrics at five questions and cutoff 70. Lower is better. All contrasts in these metrics remain unresolved in the exploratory adjusted family. Proper scores are not pure calibration measures (Guo et al., 2017).

ten questions and cutoff 70, it also has higher final acceptability accuracy than the selector (58.57% versus 57.72%), while pairwise-only has higher pairwise accuracy (54.21% versus 53.46%). These are descriptive sensitivities, not new superiority tests. The tie-equivalent primary-schedule AUCs are 71.910%, 66.671%, 71.966%, 68.287% and 66.025% in the method order of Table 26. None replaces the declared primary outcome.

Cutoff	Questions	No feedback	Absolute	Pairwise	Alternation	Information
60	5	75.562	68.553	74.677	71.053	67.612
60	10	75.562	64.508	72.781	69.522	67.865
70	5	72.753	67.514	72.584	69.143	66.685
70	10	72.753	63.933	69.649	68.581	66.699
80	5	68.258	63.301	69.284	65.941	64.228
80	10	68.258	60.520	67.177	65.323	62.767

Table 30: All prespecified PRISM serving-mistake AUC means, in percent. Cutoff 70 with five questions is primary. The remaining rows are sensitivity analyses without a new adjusted superiority claim.

Alternative error costs and numerical stability. Reweighting the four saved error components does not evaluate a new serving policy. The complete ten-setting, 60-contrast grid in Appendix F.7 leaves every comparison with absolute-only and alternation unresolved. When either acceptable response counts as sufficient, information selection improves over pairwise-only in both bank and online loss. Other cost choices can reverse online point estimates. Million-draw bootstrap checks preserve all zero-exclusion conclusions in this family and the separate 24-contrast prediction family (Appendix F.7).

With four evaluator pairs per person and rating-derived acceptability, the comparison supports acceptance information over pairwise-only feedback. The remaining error and the absolute-only comparison limit what can be recommended for serving. Appendix A.1 collects the numerical verification details.

E.4 PRISM: FIXED MIXED-VERSUS-ABSOLUTE REPLICATION

Fixed replication. We froze a single mixed-minus-absolute comparison on 90 further eligible participants before opening their outcomes. A metadata audit excluded participants appearing in six documented historical fitting or outcome files, as well as recent pilots. This establishes non-overlap within that recorded scope, not a guarantee of all prior project access. The two policies retain the original twelve features and normalization, zero-mean prior with precision $2I$, five questions at interactions $2, 4, 6, 8, 10$, ten history pairs and four later evaluator pairs. All 360 evaluator prompts are distinct and disjoint from the selected history and blocked earlier prompt groups. The cutoff remains 70. Predictions and independent replays were committed before opening evaluator ratings.

The earlier development difference was smaller than one AUC point, whereas this cohort was expected to have an ordinary 95% interval half-width near two points. We therefore fixed the replication despite its limited precision, without adding participants or changing the endpoint after assessment. Information selection yields 68.708% AUC and absolute-only 68.319%. The paired difference

is $+0.389$ points, with 95% percentile interval $[-1.361, +2.236]$ from 20,000 paired-participant bootstrap draws (PCG64 seed 134920601). This comparison remains unresolved. It establishes neither superiority nor equivalence. Final bank error is 68.889% versus 65.556%, and online pre-feedback error is 65.222% versus 65.778%, respectively. These secondary means are descriptive. Both policies receive five questions. Information selection averages 3.556 absolute questions, 1.444 pairwise questions and 4.922 binary updates, because pairwise ties consume a question without an update.

Controlled corruption on the development cohort. A separate sensitivity flips 10% or 20% of selected labels from one channel on the original 89 participants, leaving evaluation ratings unchanged. Five realizations are averaged within person. The learner retains fixed noise scales. No adjusted contrast establishes an advantage over absolute-only or alternation, or a change in the relative benefit under corruption. These synthetic perturbations do not test learned reliability. Appendix F.8 gives all sixteen contrasts.

E.5 Community Alignment: personal histories and informative questions

Recorded choice unit and representation. Community Alignment records a participant's preferred response among four LLM-generated options (Zhang et al., 2025). Our studies use English-assigned conversation openings with four complete, distinct replies from participants in India and the United States. Each original four-way answer is one feedback unit, not three independent pairwise labels. There are no per-response acceptability labels: an unchosen response is not necessarily unacceptable. Conversation identifiers do not establish chronology. Both studies are offline assessments under a fixed ordering or finite-pool selection rule, not replay of counterfactual live conversations. Normalized prompt groups are globally disjoint between learning and test within each study, but test prompts can be shared across participants.

The twelve text equations, normalization and model settings are frozen from earlier development before either assessment. A population fit uses 80 choices from eight separate people with zero-centered Gaussian precision 2. For a person's observed history $\mathcal{H}$, the categorical extension refits

$$\widehat{w}_{\mathcal{H}} = \arg\min_{w} \left\{ \sum_{i \in \mathcal{H}} \left[\log \sum_{j=1}^{4} \exp(w^\top x_{ij}) - w^\top x_{i,y_i} \right] + \frac{16}{2} \|w - w_{\text{pop}}\|^2 \right\}. \tag{51}$$

Temperature is one. Unlike the original PRISM zero-initialized binary recursion, this is a population-centered multinomial MAP fit. No assessment labels select its features, regularization or population weights. Predictions are committed before opening test choices. Log loss is the primary endpoint, with accuracy and multiclass Brier score secondary. Better probabilities need not change the highest-ranked response. For four candidate logits $a_j = w^\top x_j$, we compute $\ell_j = a_j - \text{logsumexp}_{k=1}^{4} a_k$ and $p_j = e^{\ell_j}$. For chosen class y, log loss is $-\ell_y$, without probability clipping, and Brier score is $\sum_{j=1}^{4}(p_j - \mathbf{1}\{j = y\})^2$, without division by four. Log losses use natural logarithms throughout.

Does personal history matter? The first frozen assessment uses 180 new participants (90 per country), ten learning choices and ten separate test choices each. A metadata-only donor search found too few fully matched single-person histories, so the main identity control pools answers from other people to the *exact same* ten questions, including identical prompt and response texts. Donors come from disjoint same-country reserves of 128 people per country and are fixed before outcomes. This holds question content constant but is not another single person's coherent history. An additional single-person history with different questions is exploratory and retains a content confound. Selection for donor-matched questions restricts the assessment cohort.

The full representation is also compared with its log-word-count coordinate alone, using the same population data, prior precisions and matched average prior logit-difference variance. This is a specified length control, not a separately optimized one-dimensional model. For the normalized full features x_{ij} of population choice i, let $\bar{x}_i = \frac{1}{4}\sum_{j=1}^{4} x_{ij}$, with coordinate one denoting log word

count. The length-only feature is

$$x_{ij}^{\mathrm{len}} = a x_{ij,1}, \qquad a^2 = \frac{\sum_{i,j} \|x_{ij} - \bar{x}_i\|_2^2}{\sum_{i,j} (x_{ij,1} - \bar{x}_{i,1})^2}.$$

Both sums use only the population choices. With the same isotropic prior precision, this scaling matches average prior variance of pairwise logit differences within a choice. Table 31 reports all six arms. The three primary contrasts use 20,000 paired participant bootstrap replicates stratified by country (seed 134916952), with 98.333% Bonferroni intervals. A prespecified crossed sensitivity independently resamples participants and 684 normalized test-prompt clusters (seed 134916953), using product multiplicities on the recorded choices.

Representation and history	Log loss ↓	Accuracy ↑	Brier ↓
Twelve features, no personal history	1.369257	39.33%	0.726624
Twelve features, own ten answers	1.305227	40.22%	0.695514
Twelve features, pooled others, identical questions	1.369027	37.94%	0.727338
Twelve features, one other person, different questions	1.394097	37.72%	0.737158
Length only, no personal history	1.388109	24.50%	0.750859
Length only, own ten answers	1.370653	29.94%	0.740106

Comparator	Difference	Participant interval	Crossed interval
No personal history	−0.06403	[−0.08784, −0.04039]	[−0.10547, −0.02508]
Pooled others, identical questions	−0.06380	[−0.09082, −0.03829]	[−0.10940, −0.02157]
Personalized length only	−0.06543	[−0.11160, −0.01779]	[−0.15068, +0.02070]

Table 31: Community Alignment personal-history assessment, 180 participants and 1,800 test choices. Upper panel: all six arms. No personal history uses the population fit, while pooled others match questions but mix donors. Lower panel: own-history twelve-feature log loss minus comparator, in nats. Both interval columns retain the three-comparison adjustment. Crossed resampling accounts for shared test prompts as well as participants.

Own-history probability prediction improves over both the population-only and identical-question pooled-other controls under both adjusted analyses (Table 31). This is evidence of useful person-specific information in the frozen features. Superiority over the personalized length control is unresolved under shared-prompt resampling. The observed improvements exceed the 0.05-nat planning target, but the intervals do not establish that minimum magnitude. Correct-choice accuracy increases by 0.89 points over no personal history, with crossed 95% interval $[-1.89, +3.68]$, so that accuracy gain remains unresolved. The analysis conditions on the fixed small population fit and donor bank. It does not prove stable preferences, complete feature coverage or superiority to LLM-based features.

Which recorded questions are informative? A separate assessment selects 160 further participants (80 per country), excluding previously used people and the entire reserved donor pool. Each person supplies ten available learning questions and ten test questions. Information, entropy and random policies receive the same candidate pool and answer budget. Candidate texts are visible, but the winner is revealed only after selecting its question. No learner sees evaluator choices during selection. The MAP fit in Equation 51 is recomputed after each revealed answer.

At the current MAP estimate let H be the regularized history Hessian, $p_j = \mathrm{softmax}(X\widehat{w})_j$, $\bar{x} = \sum_j p_j x_j$, and $F = \sum_j p_j (x_j - \bar{x})(x_j - \bar{x})^\top$. Information selection maximizes

$$G(X) = \tfrac{1}{2}\{\log\det(H + F) - \log\det H\}, \tag{52}$$

whereas entropy selection maximizes $-\sum_j p_j \log p_j$. The former is a local Gaussian D-optimal approximation, not exact expected information. For each participant, we average metrics over ten frozen random orders. We do not ensemble their probabilities or count them as additional independent people. The primary endpoint is five-answer mean test log loss. The two planned information-minus-control contrasts use 20,000 country-stratified paired bootstrap replicates (seed 134917012), with 97.5% Bonferroni intervals. Crossed resampling additionally uses 816 test-prompt clusters (seed 134917013), with the same multiplicity adjustment.

Answers	Information	Random	Entropy
0	1.407319	1.407319	1.407319
1	1.396749	1.398772	1.410823
2	1.387363	1.394489	1.408864
5	1.361693	1.376099	1.400443
10	1.359585	1.359585	1.359585

Comparator	Difference	Participant interval	Crossed interval
Random	-0.01441	$[-0.02476, -0.00463]$	$[-0.03005, +0.00029]$
Entropy	-0.03875	$[-0.06499, -0.01515]$	$[-0.07791, -0.00539]$

Table 32: Community Alignment question selection, 160 participants and 1,600 test choices. Upper panel: mean four-way test log loss. Five answers is primary, and all policies recover the same fit after all ten available answers. Lower panel: both five-answer information-minus-control contrasts, in nats. The original two-comparison adjustment is retained for participant and crossed intervals.

Information selection improves probability prediction over entropy under both adjusted analyses, but its improvement over random is unresolved under the crossed sensitivity (Table 32). Five-answer accuracies are 36.125%, 36.031% and 36.688% for information, random and entropy, respectively, without an established accuracy advantage. The five-answer information loss is descriptively close to the common ten-answer loss (gap 0.002108 nats), but no equivalence or noninferiority margin was specified. It is not proof that five answers replace ten.

The crossed intervals are approximate sensitivity analyses conditional on the fixed population fit, pools and random orders, not exact coverage guarantees. The findings concern short-history four-way probability prediction in the selected English, two-country cohorts. Numerical verification is described in Appendix A.1.

F Supplementary Diagnostic Details

This appendix collects extended diagnostic protocols and complete comparison grids referenced in the preceding sections. The original endpoints and comparison families remain unchanged.

F.1 Common-temperature development sensitivity

We test how a shared change in feedback noise affects the selected policy. On a precommitted 100-user development subset, we keep the response representation, scalar acceptance threshold, prior, 10% question allowance, and all policy settings fixed. We change only the true temperature in both feedback channels to $T \in \{0.5, 1, 2\}$ while the learner continues to use the unit-temperature likelihood. Dividing the true absolute and pairwise logits by T leaves every hard serving action unchanged: $T = 0.5$ produces sharper feedback and $T = 2$ noisier feedback. Because the learner's preference weights and threshold are unconstrained, the same likelihood can be written with unit temperature and $\eta_i' = \eta_i / T$. This is a signal-strength and parameter-scaling sensitivity within the linear-logistic family, rather than structural misspecification. The fixed prior is not rescaled, so finite-feedback learning can still change. This experiment keeps the same scale in both channels.

True temperature	**Absolute-only, periodic**	**Mixed feedback, periodic**	**DUAL**
0.5 (sharper)	10.57%	9.90%	**8.79%**
1.0 (matched)	16.86%	16.22%	**14.86%**
2.0 (noisier)	27.49%	27.10%	**24.63%**

Table 33: Common-temperature sensitivity on 100 development profiles: mistake AUC (%). The matched row is the same-user reference. Lower is better.

At $T = 0.5$, periodic mixed feedback minus absolute-only periodic feedback is -0.66 percentage points with a paired 95% interval of [-0.97, -0.36], and DUAL minus periodic mixed feedback is

-1.11 [-1.53, -0.70]. At $T = 2$, periodic mixed feedback minus absolute-only periodic feedback is -0.39 [-1.26, 0.51] and therefore unresolved, whereas DUAL minus periodic mixed feedback is -2.47 [-3.87, -1.10]. Thus DUAL's advantage survives both tested changes in label consistency within this likelihood family. This is development sensitivity evidence, not a second held-out confirmation or an estimate of a user-specific temperature.

F.2 A SMALLER MATCHED-ANSWER TIMING COMPARISON

This exploratory follow-up reuses the 200-user development cohort in Appendix B.5, distinct from the fresh native assessment. Users, native RM10 features, pairs, order, answer realization, learner, question selector and evaluator are held fixed. Only periodic-344 and first-reference-344 are new. Strict-free and periodic-940 reuse saved controls without recalibration. Feedback after abstention follows the original contract, with serving committed before the answer and one question charged per answer in either channel.

The new policies divide the first 9,400 interactions into 344 consecutive blocks with endpoints $e_j = \lceil j\, 9400/344 \rceil$. Each block has 27 or 28 interactions and supplies exactly one answer. Periodic asks at the endpoint. First-reference records the maximum information score over the two candidates at the first interaction, then asks at the first strictly larger score or at the endpoint. The state is unchanged before the block's answer. Neither policy uses future labels or evaluator outcomes, and neither asks in the final two interactions. Their observed maximum asks per rolling 100 interactions are four and five, respectively.

Policy	Answers	AUC	Final error	Online error	Exposures
Periodic-344	344.000	22.7743%	15.7025%	22.5160%	426.360
First-reference-344	344.000	19.9524%	13.0580%	19.7399%	419.550
Strict-free (reused)	344.315	16.3004%	12.5165%	15.9785%	443.745
Periodic-940 (reused)	940.000	16.4604%	10.2090%	16.3458%	1164.860

Planned AUC contrast	Difference (points)	Adjusted interval
First-reference-344 minus periodic-344	-2.8219	$[-3.6682, -1.9481]$
Strict-free minus periodic-344	-6.4739	$[-7.2911, -5.6533]$

Table 34: Exploratory timing study on 200 historical development users. Upper panel: equal-user means, with AUC integrating evaluation-bank errors, final bank error and actual pre-feedback online error. Exposures count two responses for a pairwise question. Lower panel: both planned AUC contrasts in percentage points, using 20,000 paired-user draws and 97.5% marginal intervals (Bonferroni adjustment over two comparisons). Negative favors the first policy.

The two planned AUC contrasts use all 200 users, 20,000 paired whole-user bootstrap draws, shared indices and PCG64 seed 134994401. First-reference improves over periodic-344 for 146 users and worsens for 54, with no ties. Strict-free improves for 182 and worsens for 18. The 344 target was selected from this cohort's previously observed strict-free spending. Strict-free retains its original 940-answer allowance and threshold, receiving 312–372 answers per person (median 344). It matches average spending only, with 96 users below 344, five equal and 99 above.

First-reference asks at mean position 4.874 rather than 27.326. Of 68,800 answers, 66,310 exceed the reference and 2,490 use fallback. Earlier feedback, changed later states and question choices jointly determine the result. Through interaction 1,000, strict-free averages 100 answers, versus 36 for periodic-344 and 36.945 for first-reference. Its lower AUC than first-reference is descriptive, not a third primary contrast. No 344-capped adaptive policy is rerun. Mean questions after abstention are 138.140 and 167.035 for periodic and first-reference, respectively.

Secondary ordinary 95% intervals give first-reference minus periodic final error of -2.6445 points $[-3.3620, -1.9365]$. Strict-free minus periodic-940 has AUC difference -0.1600 $[-0.7676, +0.4603]$, but final error $+2.3075$ $[+1.7100, +2.9220]$. These newly recomputed intervals use the shared bootstrap and do not establish equivalence or noninferiority. Answer and exposure counts are not measured human effort.

F.3 Stage dependence of the native shadow refit screen

The shadow screen in Appendix C.8 evaluates the same 600 native histories at eight answer counts. These counts do not install refits. The complete prefix breakdown is retained below.

Answers	Actually within tolerance	Certificate keep	Newton keep	Gradient keep
5	270	102	116	25
10	252	43	52	3
20	263	10	17	0
50	312	11	14	0
100	376	8	16	0
250	428	15	28	0
500	545	203	280	0
940	583	432	494	0

Table 35: Counts out of 600 histories at each received-answer checkpoint. Certificate validity alone differs from certifying the fixed 0.10-logit tolerance. The latter can fail even when the recursive estimate is accurate.

F.4 Exact seeds for the generated operational refit pilot

The distributions, schedule and controls are specified in Appendix C.9. Let K be the UTF-8 encoding of strings joined by the byte `0x1f`. Its prefix is the namespace `run134/served-only-certificate-generated/20260914/v1` and realization string `1`. The historical namespace name does not restrict feedback eligibility. User IDs are `generated-cert-user-00` through `generated-cert-user-09`. Append `user`, `online-features` or `evaluation-features`, then the user ID, for three separate NumPy PCG64 generators. Their seed is the big-endian integer formed from the first 16 bytes of $\mathrm{SHA256}(K)$. Draw shapes are respectively 11, $(1000, 2, 10)$, and $(64, 2, 10)$. For an answer, append `selected-answer`, user ID, prompt ID `prompt-00001` through `prompt-01000`, and query ID `absolute-A-00`, `absolute-B-01` or `pair-00-01`. The leading 53 digest bits divided by 2^{53} give the uniform variate compared with $\sigma(z^\top\theta)$. Evaluator-only binary labels use the suffix `evaluation-answer`, user ID, zero-based pair index and channel index 0, 1 or 2 instead. These labels are distinct from the latent-action target used for serving error.

F.5 Prospectively fixed refit stress panel

The separate panel has 39 related constructed cases, not 39 users. Thirty-six use native visible-query geometry: four generated histories, three state recipes and three feature scales, always checking both absolute designs and the pairwise difference. Of these, 24 use artificial MAP offsets. The first two histories have 5 and 40 answers, with NumPy PCG64 seeds 20260914101 and 20260914201, respectively. In each stream draw $\theta \sim \mathcal{N}(0, 0.7^2 I_{11})$ first, then a fresh pair of ten-dimensional response features and one Bernoulli uniform per answer. Feature coordinate standard deviations are 0.15 and 0.7, respectively. Odd-numbered answers use $[a, -1]$, even-numbered answers use $[a - b, 0]$, with success probability $\sigma(z^\top\theta)$. The remaining histories repeat $z = 2e_1$. One has 40 answers, all one. The other has 100 answers, all one except at indices 20, 40, 60, 80 and 100. All start from zero mean and precision $2I_{11}$. For each history, the three tested means are the recursive mean, the checked joint MAP plus $0.025e_1$, and that MAP plus $0.5e_1$. All three retain the original recursive precision. Visible features are $a = se_1$, $b = -a$, for $s \in \{0.1, 1, 5\}$ in feature space, giving the $4 \times 3 \times 3 = 36$ native cases. The reference solver's parameter error bound is at most 10^{-9}. Its induced query-error margins determine whether a tolerance boundary is numerically resolved. Three additional algebraic examples instead use an arbitrary single query embedded in eleven dimensions. We report them separately because they do not instantiate the native three-query operational policy. Their exact construction uses prior precision $2I_{11}$ and forty identical observations $z_t = 2e_1$, $r_t = 1$, giving

$$F(\theta) = \|\theta\|_2^2 + 40\log(1 + e^{-2\theta_1}).$$

The three approximate points are $\theta = ae_1$ for $a \in \{1, 0.8, 0.75\}$, and the sole diagnostic query is $q = 0.2e_1$. They retain the same 0.10-logit tolerance. These points and the arbitrary query are constructed probes, not states or query sets sampled from the operational policy.

Geometry	Screen	Cases	Keep	Refit	False keep	Unresolved
Native visible pair	Certificate	36	11	25	0	4
Native visible pair	Newton	36	11	25	0	4
Arbitrary single query	Certificate	3	0	3	0	0
Arbitrary single query	Newton	3	2	1	2	0

Table 36: Fixed stress cases. "Unresolved" counts pre-gate numerical boundary cases at the nominal 0.10 tolerance. Both rules refit all four. False keep counts numerically resolved violations retained by a screen. Native cases have no post-gate violations, whereas the arbitrary single-query Newton cases have two.

Both screens agree on all 36 native cases, with no false KEEP or post-gate violation. Four pre-gate cases lie on the nominal tolerance boundary and remain numerically unresolved. Both rules refit them. Four other refits per rule are unnecessary for the immediate-query target. The three arbitrary-query examples show a possible Newton failure: two estimated gaps, 0.08018 and 0.09719, are below tolerance although the checked actual gaps are 0.11835 and 0.15835. The certificate bounds are 0.16183 and 0.35712, triggering refits. In the third example the certificate condition fails and Newton also requests a refit. These constructions distinguish a sufficient bound from a heuristic. They do not estimate failure prevalence or demonstrate an operational advantage on ordinary histories. The stress panel uses cold caches, whereas the pilot includes actual cache reuse, so their component timings are not directly comparable.

F.6 COMPLETE PRISM ERROR-COMPONENT CONTRASTS

These are the complete component intervals described in Appendix E.3, with the same 89 participants, five-question schedule, cutoff 70 and 27-comparison correction. The main appendix reports the component means and online-total contrasts.

Component	Comparator	Difference	95% interval	Adjusted interval
Unacceptable serve	Absolute-only	$+0.211$	$[-1.390, +1.798]$	$[-2.374, +2.725]$
Unacceptable serve	Pairwise-only	$+0.239$	$[-3.048, +3.469]$	$[-5.020, +5.525]$
Unacceptable serve	Fixed alternation	-0.407	$[-2.851, +2.051]$	$[-4.233, +3.539]$
False abstention	Absolute-only	-0.913	$[-2.654, +0.899]$	$[-3.652, +2.093]$
False abstention	Pairwise-only	-7.472	$[-11.699, -3.343]$	$[-14.323, -0.969]$
False abstention	Fixed alternation	-2.458	$[-5.463, +0.548]$	$[-7.267, +2.331]$
Worse acceptable	Absolute-only	$+0.056$	$[-0.941, +1.025]$	$[-1.593, +1.587]$
Worse acceptable	Pairwise-only	$+1.292$	$[-1.194, +3.820]$	$[-2.680, +5.323]$
Worse acceptable	Fixed alternation	$+0.604$	$[-1.124, +2.233]$	$[-2.261, +3.208]$
Acceptable tie	Absolute-only	-0.183	$[-0.548, +0.000]$	$[-0.913, +0.000]$
Acceptable tie	Pairwise-only	$+0.042$	$[-0.449, +0.646]$	$[-0.702, +1.124]$
Acceptable tie	Fixed alternation	-0.197	$[-0.520, +0.000]$	$[-0.772, +0.000]$

Table 37: All exploratory PRISM contrasts for evaluation-bank component AUC, in percentage points: information selection minus comparator. Negative means fewer errors of the stated type. These twelve contrasts are included in the new 27-comparison family (100,000 paired-person draws with 99.8148% marginal Bonferroni intervals). Acceptable-tie penalties retain the original canonical-action rule.

F.7 PRISM WEIGHTED-ERROR SENSITIVITY

Exploratory error-cost sensitivity. A further retrospective analysis holds all recorded actions fixed and assigns loss $L = c_U U + F + c_W W + c_T T$ to the four components above. False-abstention cost is one. We declare the complete grid $c_U \in \{1, 2, 5\}$, $c_W \in \{0, 1, 2\}$, $c_T = 0$, plus the

Component	Comparator	Difference	95% interval	Adjusted interval
Unacceptable serve	Absolute-only	$+1.011$	$[-0.225, +2.247]$	$[-0.899, +3.034]$
Unacceptable serve	Pairwise-only	$+2.022$	$[-0.449, +4.494]$	$[-1.910, +5.889]$
Unacceptable serve	Fixed alternation	$+0.225$	$[-1.798, +2.247]$	$[-3.034, +3.596]$
False abstention	Absolute-only	-1.910	$[-3.708, -0.225]$	$[-4.944, +0.787]$
False abstention	Pairwise-only	-10.449	$[-14.270, -6.742]$	$[-16.629, -4.494]$
False abstention	Fixed alternation	-4.494	$[-7.528, -1.461]$	$[-9.484, +0.225]$
Worse acceptable	Absolute-only	$+0.899$	$[-0.225, +2.135]$	$[-0.899, +2.809]$
Worse acceptable	Pairwise-only	$+3.820$	$[+1.798, +6.180]$	$[+0.674, +7.865]$
Worse acceptable	Fixed alternation	$+0.899$	$[-0.787, +2.697]$	$[-1.685, +3.754]$
Acceptable tie	Absolute-only	-0.112	$[-0.337, +0.000]$	$[-0.562, +0.000]$
Acceptable tie	Pairwise-only	$+0.449$	$[+0.112, +0.899]$	$[+0.000, +1.236]$
Acceptable tie	Fixed alternation	$+0.225$	$[-0.337, +0.787]$	$[-0.562, +1.124]$

Table 38: All exploratory PRISM contrasts for actual online-error components, in percentage points: information selection minus comparator. Negative means fewer errors of the stated type. These twelve contrasts are included in the new 27-comparison family (100,000 paired-person draws with 99.8148% marginal Bonferroni intervals). Acceptable-tie penalties retain the original canonical-action rule.

original unit-error reference $c_U = c_W = c_T = 1$, before computing the weighted comparisons. When $c_W = c_T = 0$, either acceptable response counts as sufficient, regardless of its ranking or a canonical tie break. These are assumed error costs, not measured human costs, a selected new endpoint, or a change to the serving policy. The original outcomes were known before this analysis.

All ten settings, two contexts and three comparators share 100,000 paired-person bootstrap draws (seed 134917703) and Bonferroni tails $0.025/60$ and $1 - 0.025/60$. The resulting 60-contrast family is separate from the primary three comparisons and 27-contrast decomposition. Tables 39–40 report every contrast in weighted loss per 100 pairs, not mistake percentages under nonunit costs. Component means in Table 27 determine the corresponding absolute weighted means.

(c_U, c_W, c_T)	vs. absolute-only	vs. pairwise-only	vs. alternation
$(1, 1, 1)$	-0.83 $[-4.63, +2.25]$	-5.90 $[-12.53, +0.75]$	-2.46 $[-7.47, +2.41]$
$(1, 0, 0)$	-0.70 $[-4.40, +2.88]$	-7.23 $[-14.46, -0.06]$	-2.87 $[-8.10, +2.36]$
$(1, 1, 0)$	-0.65 $[-4.51, +2.60]$	-5.94 $[-12.73, +0.83]$	-2.26 $[-7.34, +2.63]$
$(1, 2, 0)$	-0.59 $[-5.36, +3.26]$	-4.65 $[-13.41, +4.05]$	-1.66 $[-8.06, +4.37]$
$(2, 0, 0)$	-0.49 $[-6.49, +4.87]$	-6.99 $[-17.96, +3.50]$	-3.27 $[-11.21, +4.68]$
$(2, 1, 0)$	-0.44 $[-6.66, +4.67]$	-5.70 $[-16.70, +4.85]$	-2.67 $[-10.77, +5.49]$
$(2, 2, 0)$	-0.38 $[-7.16, +5.21]$	-4.41 $[-17.01, +7.81]$	-2.06 $[-11.38, +7.06]$
$(5, 0, 0)$	$+0.14$ $[-14.09, +13.07]$	-6.28 $[-32.72, +19.75]$	-4.49 $[-23.46, +15.13]$
$(5, 1, 0)$	$+0.20$ $[-14.22, +13.13]$	-4.99 $[-31.78, +21.21]$	-3.89 $[-23.56, +15.85]$
$(5, 2, 0)$	$+0.25$ $[-14.64, +13.56]$	-3.69 $[-31.58, +23.46]$	-3.29 $[-23.97, +17.48]$

Table 39: Evaluation-bank weighted-loss AUC: information selection minus each comparator, per 100 pairs, with intervals adjusted over all 60 exploratory contrasts. Negative favors information selection. The first row repeats the original unit-cost estimand under this new, wider interval family. The other rows assign zero penalty to acceptable ties.

At $(c_U, c_W, c_T) = (1, 0, 0)$, information selection improves over pairwise-only in both bank AUC (-7.233, adjusted interval $[-14.462, -0.056]$) and actual online loss (-8.427, $[-15.506, -1.685]$). The bank interval is close to zero. The other 58 adjusted contrasts contain zero, including every comparison with absolute-only and alternation. Giving additional weight to unacceptable serves or worse acceptable choices does not establish a favorable contrast in this grid. Some online point estimates reverse. This supports an acceptance-focused interpretation of the pairwise-only comparison, not uniformly better decisions under different costs. A cost-aware serving rule could change the actions themselves, but this fixed-action sensitivity does not evaluate one.

Bootstrap Monte Carlo sensitivity. Million-draw paired-person checks, with fixed seeds 134918642 and 134918643, preserve every zero-exclusion conclusion in the complete 60-contrast

(c_U, c_W, c_T)	vs. absolute-only	vs. pairwise-only	vs. alternation
$(1, 1, 1)$	-0.11 $[-2.47, +2.17]$	-4.16 $[-9.89, +1.91]$	-3.15 $[-7.42, +1.12]$
$(1, 0, 0)$	-0.90 $[-3.82, +1.91]$	-8.43 $[-15.51, -1.69]$	-4.27 $[-9.55, +0.67]$
$(1, 1, 0)$	$+0.00$ $[-2.47, +2.36]$	-4.61 $[-10.56, +1.57]$	-3.37 $[-7.98, +1.01]$
$(1, 2, 0)$	$+0.90$ $[-2.36, +4.27]$	-0.79 $[-7.79, +6.85]$	-2.47 $[-7.87, +3.26]$
$(2, 0, 0)$	$+0.11$ $[-3.93, +4.16]$	-6.40 $[-15.84, +2.70]$	-4.04 $[-11.35, +2.85]$
$(2, 1, 0)$	$+1.01$ $[-2.81, +4.83]$	-2.58 $[-11.39, +6.40]$	-3.15 $[-10.11, +3.48]$
$(2, 2, 0)$	$+1.91$ $[-2.70, +6.40]$	$+1.24$ $[-8.43, +11.16]$	-2.25 $[-10.00, +5.51]$
$(5, 0, 0)$	$+3.15$ $[-6.40, +13.07]$	-0.34 $[-21.05, +19.93]$	-3.37 $[-20.22, +12.81]$
$(5, 1, 0)$	$+4.04$ $[-5.73, +14.04]$	$+3.48$ $[-17.42, +23.86]$	-2.47 $[-19.48, +14.04]$
$(5, 2, 0)$	$+4.94$ $[-5.39, +15.51]$	$+7.30$ $[-14.19, +28.35]$	-1.57 $[-19.03, +15.84]$

Table 40: Actual pre-feedback online weighted loss: information selection minus each comparator, per 100 pairs, with intervals adjusted over all 60 exploratory contrasts. Negative favors information selection. The first row repeats the original unit-cost estimand under this new, wider interval family. The other rows assign zero penalty to acceptable ties.

cost family and 24-contrast prediction family. At costs $(1, 0, 0)$, the pairwise-only bank and online intervals become $[-14.565, -0.154]$ and $[-15.393, -1.685]$. The prediction-family acceptability-accuracy interval against pairwise-only becomes $[+0.702, +18.258]$ points. The original intervals, populations and comparison adjustments are retained.

F.8 PRISM CONTROLLED FEEDBACK CORRUPTION

On the original 89 development participants, we separately flip 10% or 20% of selected binary labels from one feedback channel. Evaluation ratings remain unchanged, and tied pairwise answers remain ties. A keyed random draw is shared across policies and nested across flip rates. Five fixed corruption realizations are averaged within each participant. They are not treated as additional people. The learner retains fixed channel temperatures and does not learn the corruption rate. This tests synthetic perturbations of recorded feedback, not naturally occurring inconsistency.

Corrupted channel	Absolute-only	Pairwise-only	Alternation	Information
None	67.514	72.584	69.143	66.685
Absolute, 10%	68.177	72.584	69.239	67.174
Absolute, 20%	69.514	72.584	69.930	68.076
Pairwise, 10%	67.514	72.404	69.031	66.610
Pairwise, 20%	67.514	72.643	69.303	66.809

Table 41: Controlled PRISM label corruption. Entries are mistake AUC in percent at five questions and cutoff 70. Clean cells are unchanged original controls. Each perturbed mean averages five realizations within each of 89 participants.

The declared comparison family contains twelve information-minus-control contrasts and four changes in the information-minus-absolute difference relative to clean feedback. We use 20,000 paired-person bootstrap draws (PCG64 seed 134920602) and Bonferroni intervals across all sixteen contrasts. Table 42 reports the complete family. Information selection remains better than pairwise-only under 10% and 20% pairwise corruption. No adjusted comparison establishes superiority over absolute-only or alternation. Nor do the four interaction contrasts establish that channel corruption changes the relative benefit over absolute-only. The clean study's original three-comparison inference is a separate family.

Corrupted channel	Contrast	Difference	Adjusted interval
Absolute, 10%	Information − absolute	-1.003	$[-3.725, +1.366]$
Absolute, 10%	Information − pairwise	-5.410	$[-11.130, +0.362]$
Absolute, 10%	Information − alternation	-2.065	$[-6.149, +2.003]$
Absolute, 10%	Change versus clean	-0.174	$[-1.084, +0.809]$
Absolute, 20%	Information − absolute	-1.438	$[-4.217, +0.817]$
Absolute, 20%	Information − pairwise	-4.508	$[-9.711, +1.015]$
Absolute, 20%	Information − alternation	-1.854	$[-5.286, +1.596]$
Absolute, 20%	Change versus clean	-0.610	$[-2.224, +0.974]$
Pairwise, 10%	Information − absolute	-0.904	$[-4.122, +1.685]$
Pairwise, 10%	Information − pairwise	-5.795	$[-11.242, -0.366]$
Pairwise, 10%	Information − alternation	-2.421	$[-6.578, +1.632]$
Pairwise, 10%	Change versus clean	-0.076	$[-0.475, +0.289]$
Pairwise, 20%	Information − absolute	-0.705	$[-3.837, +1.975]$
Pairwise, 20%	Information − pairwise	-5.834	$[-10.949, -0.789]$
Pairwise, 20%	Information − alternation	-2.494	$[-6.660, +1.513]$
Pairwise, 20%	Change versus clean	$+0.124$	$[-0.651, +1.215]$

Table 42: All sixteen planned corruption contrasts in AUC percentage points. Negative favors information selection. Change versus clean is the difference between the corrupted and clean information-minus-absolute contrasts. Intervals use a 95% familywise target.