PILOT PRIORS: SAFELY GATING WARM STARTS FOR CONTEXTUAL BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Warm-starting contextual bandits has become an increasingly attractive method for reducing early-stage regret. Warm start data can come from historical logs, transfer pipelines, or be generated synthetically via simulators or large language models. However, the quality of such priors is often unknown: priors aligned with target tasks can reduce downstream regret, while misaligned priors can increase regret relative to cold starts. This paper studies the pre-deployment setting in which a single candidate prior is available, but only a short amount of real target-task data can be collected before source trust is fixed for the next deployment phase. We propose the Pilot Priors framework, which uses a short target-task pilot to certify a candidate warm prior relative to a cold-start baseline. The framework produces accept, reject, or inconclusive decisions from certified bounds on prior error, and satisfies a $1 - \delta$ correctness guarantee for hard decisions about prior-error ordering. We further derive a sufficient condition for decisiveness through a margin-versus-width condition that links pilot informativeness to prior quality. Finally, for the inconclusive regime, we introduce a soft trust-calibration rule that interpolates between warm and cold starts rather than forcing a brittle all-or-nothing decision. The framework is evaluated in both a synthetic setting and two deployment scenarios based on real-world data. We show that the certificate becomes decisive as pilot information increases in synthetic experiments and detects harmful LLM-generated priors in one real-world setting. In another real-world setting, the certificate remains appropriately conservative while the soft calibration reduces harm from a reward-permuted null.

1 INTRODUCTION

Contextual multi-armed bandits provide a framework for sequential decision making under uncertainty, yet they frequently suffer from high initial regret during the cold start phase (Chu et al., 2011; Abbasi-Yadkori et al., 2011). A common mitigation is to warm start learning from auxiliary information before fine-tuning on target-task data. Such priors may come from historical logs (Saito et al., 2021), transfer pipelines (Deng et al., 2025), or be generated synthetically via simulators or large language models (Alamdari et al., 2024). When these priors are aligned with the downstream task, warm-starting can reduce uncertainty, shorten the cold-start phase, and improve early regret (Oetomo et al., 2021; Bayley et al., 2026). When the prior is misaligned, however, pretraining on such data can increase regret. The learner performs worse than if it had started cold. The core practical problem is not whether warm starts can help, but how to decide whether a candidate prior should be trusted given a target task.

Prior studies in warm-starting contextual bandits show that auxiliary structure can reduce early regret when transfer is favorable, largely framing the problem as how to exploit priors efficiently over a finite horizon (Oetomo et al., 2021). Robust warm start methods, such as ARROW-CB, instead study how to combine supervised and bandit feedback online when the two sources may be misaligned, using adaptive reweighting to mitigate harmful transfer (Zhang et al., 2019). Related work on prior misspecification and Bayesian sensitivity gives a quantitative language for how decision quality degrades when a prior is wrong, while contextual bandits formalize the idea that a learner should remain close to a trusted fallback when evidence for improvement is weak (Simchowitz et al., 2021; Kazerouni et al., 2017). What remains missing is a pre-deployment rule for the following setting: a candidate warm start has already been constructed, its task-specific quality is unknown, and only a

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.

short amount of real pilot data can be collected before source trust is fixed for the next deployment phase. In deployed systems, a bad prior can waste the early horizon, while rejecting a useful one leaves scarce feedback underused. A pilot-based rule should therefore make this trust decision auditable before deployment. Prior-error analyses make this problem mathematically legible, but primarily in a post hoc sense: they explain why a warm start helped or hurt after the fact, rather than turning prior quality into a task-specific pre-deployment rule (Bayley et al., 2026).

This study bridges the gap by turning prior-quality assessment into a pilot-based certification problem. Building on the prior-error perspective of Bayley et al. (2026), our framework uses a short real pilot to produce a certificate comparing the candidate prior to a cold-start baseline. The certificate either accepts the prior, rejects it in favor of a cold start, or leaves the decision inconclusive. For the inconclusive regime, we introduce a soft trust-calibration rule that blends warm and cold according to pilot evidence rather than forcing an all-or-nothing choice. Our contributions are as follows:

- We formulate prior-quality assessment as a pre-deployment decision problem for warm-started contextual bandits, using a short real pilot to compare a single candidate warm prior against a cold-start baseline on the target task.
- Our work proves that the resulting pilot-based certificates satisfy a $1-\delta$ correctness guarantee for hard decisions about cold-relative prior-error ordering, rather than finite-horizon regret, and we characterize when the certificate becomes decisive through a margin-versus-width condition that links pilot informativeness to prior quality.
- We introduce a certificate-derived soft trust-calibration rule for the inconclusive regime, yielding a deployment policy that interpolates between warm and cold when hard certification is not yet justified.
- We validate the framework in a controlled synthetic setting and two real-data deployment scenarios: LLM-to-human alignment determined through a conjoint survey, and historical policy shift on KuaiRand, a short-video recommendation dataset with both standard recommendation logs and randomized intervention logs. Results show the certificate produces clean accept/reject/inconclusive structure in synthetic data, identifies harmful LLM priors in the conjoint study, and remains appropriately conservative under the shifts in KuaiRand while still yielding policy-level improvements via soft calibration. In representative downstream comparisons, the soft rule preserves $\approx 90\%$ of the aligned COVID warm-start gain, removes the full regret cost of the corrupted COVID prior, and reduces the KuaiRand reward-shuffled SNIPS loss by 69% relative to ungated warm start.

2 Preliminaries

2.1 Problem setup

We consider a linear contextual bandit over rounds $t = 1, \dots, T$. At round t, the learner observes feature vectors $\{x_{t,a}\}_{a \in \mathcal{A}_t}$ for the available actions, selects $a_t \in \mathcal{A}_t$, and observes only the reward of the chosen action. We assume the standard linear reward model $r_t(a_t) = x_{t,a_t}^\top \theta^\star + \xi_t$, where $\theta^\star \in \mathbb{R}^d$ is the unknown target parameter. Let

$$a_t^\star \in \arg\max_{a \in \mathcal{A}_t} x_{t,a}^\top \theta^\star, \qquad r_t^\star := x_{t,a_t^\star}^\top \theta^\star.$$

The instantaneous regret is $r_t^\star - x_{t,a_t}^\top \theta^\star$, and cumulative regret is the sum of these terms over the deployment horizon.

We follow the prior-centered view of warm-started linear contextual bandits introduced in (Bayley et al., 2026). In the rest of this section, we provide a brief summary to make our paper self-contained.

2.2 Warm and cold initialization

A candidate warm prior m is represented by a precision matrix $A_0^{(m)} \succ 0$ and mean vector $\theta_0^{(m)} \in \mathbb{R}^d$. In the ridge-type warm starts considered throughout this paper, these are induced from auxiliary data,

$$A_0^{(m)} := X_m^\top X_m + \tau_{\mathrm{pre}} I, \qquad b_0^{(m)} := X_m^\top \tilde{y}_m, \qquad \theta_0^{(m)} := \left(A_0^{(m)}\right)^{-1} b_0^{(m)}, \tag{1}$$

where X_m and $\tilde{y}_m$ denote the auxiliary design and targets associated with candidate m, and $\tau_{\mathrm{pre}} > 0$ is a regularization parameter.

A warm-started LinUCB (Li et al., 2010) instance initialized from candidate m begins deployment with

$$V_0^{(m)} = A_0^{(m)}, \qquad \hat{\theta}_0^{(m)} = \theta_0^{(m)},$$

and then updates on real target-task feedback. As a baseline, we also consider the cold-start prior

$$(A_0^{\mathrm{cold}}, \theta_0^{\mathrm{cold}}) = (I, 0),$$

which corresponds to an uninformative initialization with unit regularization. Equivalently, it matches (1) with $\tau_{\mathrm{pre}} = 1$ and no auxiliary data.

2.3 Prior error as a scaffold

Following Bayley et al. (2026), the effect of warm-starting can be organized around the prior-error quantity

$$\mathcal{B}_0^{(m)} := \|\theta_0^{(m)} - \theta^\star\|_{A_0^{(m)}}. \tag{2}$$

This measures how far the candidate prior is from the target parameter in the geometry induced by its precision matrix. Smaller values correspond to better-aligned priors; larger values correspond to priors that may induce negative transfer.

The corresponding cold-start quantity is

$$\mathcal{B}_0^{\mathrm{cold}} := \|\theta_0^{\mathrm{cold}} - \theta^\star\|_{A_0^{\mathrm{cold}}} = \|\theta^\star\|_2. \tag{3}$$

A key consequence of the prior-centered analysis is that, for a warm-started run with candidate m,

$$V_t^{(m)} := A_0^{(m)} + \sum_{s \le t} x_{s,a_s} x_{s,a_s}^\top,$$

the estimation error can be bounded in the form

$$\|\hat{\theta}_t^{(m)} - \theta^\star\|_{V_t^{(m)}} \;\le\; \beta_t^{(m)}(\delta) + \mathcal{B}_0^{(m)}, \tag{4}$$

with high probability, where $\beta_t^{(m)}(\delta)$ denotes the self-normalized confidence radius associated with $V_t^{(m)}$. We use the notation $\beta_t^{(\cdot)}(\delta)$ generically throughout the paper; the specific pilot version used in our certification construction is defined in Section 3. Here $V_t^{(m)}$ is the design matrix of a warm-started deployment run and should not be confused with the shared pilot design matrix introduced later.

For the present paper, (4) is used only as a scaffold: it shows that pretraining enters through a prior-dependent confidence radius and the unobserved quantity $\mathcal{B}_0^{(m)}$. Ordering the initial parameter set alone is insufficient to determine deployment regret since both the confidence radius and design matrix depend on initialization and the choices made in subsequent steps.

2.4 From post hoc prior error to pre-deployment certification

The strength of the prior-error view is that it explains why the same prior can help on one task and hurt on another. Its limitation is that $\mathcal{B}_0^{(m)}$ is not observable at deployment time because it depends on the unknown target parameter $\theta^\star$. On its own, it therefore remains a post hoc explanatory quantity. We use the pilot to make this quantity operational: Section 3 constructs certified upper and lower bounds on $\mathcal{B}_0^{(m)}$ and $\mathcal{B}_0^{\mathrm{cold}}$, then converts these bounds into gating rules.

3 Pilot Priors Framework

We leverage real pilot data to construct certified bounds on the prior error of each candidate warm prior and of the cold-start baseline. These bounds induce a cold-relative certification rule that labels each candidate prior as accept, reject, or abstain before source trust is fixed for the next deployment phase. We then extend this hard rule with a soft trust-calibration mechanism that preserves the certified endpoints while interpolating in the inconclusive regime.

3.1 Pilot Confidence Region

Assume access to a real pilot of length T_{pilot} collected on the target task using a shared pilot policy π_{pilot}. Let $(x_t, r_t)_{t=1}^{T_{\text{pilot}}}$ denote the chosen pilot contexts and observed rewards, with

$$r_t = x_t^\top \theta^\star + \xi_t, \qquad t = 1, \dots, T_{\text{pilot}},$$

where $\theta^\star \in \mathbb{R}^d$ is the unknown target parameter and (ξ_t) is conditionally sub-Gaussian. We additionally assume a known bound $\|\theta^\star\|_2 \le S$. In applications, S can be chosen from domain constraints, feature and reward normalization, or a conservative validation estimate. Increasing S preserves validity but widens the certificate.

We form the pilot ridge estimator

$$V_t^{\text{pilot}} := \tau_{\text{pilot}} I + \sum_{s \le t} x_s x_s^\top, \qquad \hat{\theta}_t^{\text{pilot}} := \big(V_t^{\text{pilot}}\big)^{-1} \sum_{s \le t} r_s x_s, \tag{5}$$

where $\tau_{\text{pilot}} > 0$ is a pilot regularization parameter.

Theorem 1 (Pilot confidence ellipsoid)**.** [1] *Fix $\delta \in (0,1)$. With probability at least $1-\delta$, for all $t \in \{1, \dots, T_{\text{pilot}}\}$,*

$$\|\hat{\theta}_t^{\text{pilot}} - \theta^\star\|_{V_t^{\text{pilot}}} \le \beta_t^{\text{pilot}}(\delta) + \sqrt{\tau_{\text{pilot}}}\, S, \tag{6}$$

where

$$\beta_t^{\text{pilot}}(\delta) := \sigma \sqrt{2 \log \frac{\det(V_t^{\text{pilot}})^{1/2}}{\det(\tau_{\text{pilot}} I)^{1/2} \delta}}. \tag{7}$$

Equivalently, the set

$$\mathcal{C}_t^{\text{pilot}}(\delta) := \Big\{ \theta : \|\theta - \hat{\theta}_t^{\text{pilot}}\|_{V_t^{\text{pilot}}} \le \beta_t^{\text{pilot}}(\delta) + \sqrt{\tau_{\text{pilot}}}\, S \Big\} \tag{8}$$

contains $\theta^\star$ simultaneously for all $t \le T_{\text{pilot}}$ with probability at least $1-\delta$.

Theorem 1 provides a high-probability confidence region for the unknown target parameter based only on the pilot data. We now transfer the pilot uncertainty into the geometry of the candidate priors.

3.2 Certified Bounds on Prior Error

For candidate model m, let $\mathcal{B}_0^{(m)}$ be the prior error as defined in (2). Since $\theta^\star$ is unknown, our goal is to use the pilot ellipsoid to obtain computable upper and lower bounds on $\mathcal{B}_0^{(m)}$. For each candidate model m, define

$$d^{(m)} := \theta_0^{(m)} - \hat{\theta}_{T_{\text{pilot}}}^{\text{pilot}}, \qquad \gamma^{(m)} := \sqrt{\lambda_{\max}\Big(A_0^{(m)} \big(V_{T_{\text{pilot}}}^{\text{pilot}}\big)^{-1} \Big)}. \tag{9}$$

Also define the pilot radius

$$\rho_{\text{pilot}}(\delta) := \beta_{T_{\text{pilot}}}^{\text{pilot}}(\delta) + \sqrt{\tau_{\text{pilot}}}\, S. \tag{10}$$

Theorem 2 (Certified prior-error interval)**.** *Fix $\delta \in (0,1)$. On the event*

$$\theta^\star \in \mathcal{C}_{T_{\text{pilot}}}^{\text{pilot}}(\delta),$$

the prior error of each candidate model m satisfies

$$\underline{\mathcal{B}}^{(m)}(\delta) \le \mathcal{B}_0^{(m)} \le \overline{\mathcal{B}}^{(m)}(\delta), \tag{11}$$

where

$$\overline{\mathcal{B}}^{(m)}(\delta) := \|d^{(m)}\|_{A_0^{(m)}} + \gamma^{(m)} \rho_{\text{pilot}}(\delta), \tag{12}$$

and

$$\underline{\mathcal{B}}^{(m)}(\delta) := \max\Big\{ 0,\ \|d^{(m)}\|_{A_0^{(m)}} - \gamma^{(m)} \rho_{\text{pilot}}(\delta) \Big\}. \tag{13}$$

The interval construction is simultaneous for any fixed finite collection of candidate priors because all candidate bounds are deterministic consequences of the same pilot confidence event.

[1] This applies the self-normalized concentration bound of Abbasi-Yadkori et al. (2011) with conditional sub-Gaussian noise σ and pilot context x_t measurable before ξ_t; see Appendix A.

The certified interval construction applies to the cold-start baseline $(A_0^{\mathrm{cold}}, \theta_0^{\mathrm{cold}}) = (I, 0)$, yielding

$$\mathcal{I}^{\mathrm{cold}} = \big[\underline{\mathcal{B}}^{\mathrm{cold}}(\delta), \overline{\mathcal{B}}^{\mathrm{cold}}(\delta)\big].$$

Concretely,

$$d^{\mathrm{cold}} = -\hat{\theta}^{\mathrm{pilot}}_{T_{\mathrm{pilot}}}, \qquad \gamma^{\mathrm{cold}} = \sqrt{\lambda_{\max}\Big(\big(V^{\mathrm{pilot}}_{T_{\mathrm{pilot}}}\big)^{-1}\Big)} = \frac{1}{\sqrt{\lambda_{\min}\big(V^{\mathrm{pilot}}_{T_{\mathrm{pilot}}}\big)}}.$$

Thus, the pilot yields comparable uncertainty-aware intervals for both the warm prior and cold-start baseline on the same target task. In the experiments below, we evaluate one candidate prior at a time.

3.3 Cold-Relative Certification Rule

We can now convert the certified prior-error intervals into a relative deployment rule. The central design choice is to compare each candidate warm prior directly against the cold-start baseline on the same pilot, rather than against a fixed external threshold. Let

$$\mathcal{I}^{(m)} = \big[\underline{\mathcal{B}}^{(m)}(\delta), \overline{\mathcal{B}}^{(m)}(\delta)\big]$$

denote the certified interval for candidate m, and let

$$\mathcal{I}^{\mathrm{cold}} = \big[\underline{\mathcal{B}}^{\mathrm{cold}}(\delta), \overline{\mathcal{B}}^{\mathrm{cold}}(\delta)\big]$$

denote the certified interval for cold-start.

Definition 3 (Cold-relative certification rule)**.** *For a candidate prior m, define:*

$$\begin{aligned} \textsc{Accept} \quad & \textit{if} \quad \overline{\mathcal{B}}^{(m)}(\delta) \le \underline{\mathcal{B}}^{\mathrm{cold}}(\delta), \\ \textsc{Reject} \quad & \textit{if} \quad \underline{\mathcal{B}}^{(m)}(\delta) > \overline{\mathcal{B}}^{\mathrm{cold}}(\delta), \textit{ and} \\ \textsc{Inconclusive} \quad & \textit{otherwise}. \end{aligned}$$

Theorem 4 (Correctness of hard cold-relative decisions)**.** *Fix $\delta \in (0, 1)$. With probability at least $1 - \delta$, the following statements hold for any candidate prior m:*

- *If candidate m is* Accept*, then $\mathcal{B}_0^{(m)} \le \mathcal{B}_0^{\mathrm{cold}}$;*
- *If candidate m is* Reject*, then $\mathcal{B}_0^{(m)} > \mathcal{B}_0^{\mathrm{cold}}$.*

Equivalently, on the pilot confidence event, the hard cold-relative rule makes no incorrect accept or reject decisions. The same guarantee holds for any fixed finite candidate set on the shared pilot confidence event.

Theorem 4 formalizes the main guarantee of the hard gate: on the pilot confidence event, each hard decision orders the original warm and cold prior errors. Overlapping intervals lead to abstention. We next introduce a margin-versus-width condition that links decisiveness to the relationship between the unobserved population gap $\mathcal{B}_0^{(m)} - \mathcal{B}_0^{\mathrm{cold}}$ and the observed certificate widths.

For each candidate, define the population margin

$$\mathcal{M}^{(m)} := \big|\mathcal{B}_0^{(m)} - \mathcal{B}_0^{\mathrm{cold}}\big|, \tag{14}$$

the individual certificate widths

$$\mathcal{W}^{(m)}(\delta) := \overline{\mathcal{B}}^{(m)}(\delta) - \underline{\mathcal{B}}^{(m)}(\delta), \qquad \mathcal{W}^{\mathrm{cold}}(\delta) := \overline{\mathcal{B}}^{\mathrm{cold}}(\delta) - \underline{\mathcal{B}}^{\mathrm{cold}}(\delta),$$

and the combined certificate width

$$\mathcal{W}_+^{(m)}(\delta) := \mathcal{W}^{(m)}(\delta) + \mathcal{W}^{\mathrm{cold}}(\delta). \tag{15}$$

Corollary 5 (Margin-versus-width condition)**.** *Fix $\delta \in (0, 1)$. On the pilot confidence event of Theorem 1, which occurs with probability at least $1 - \delta$:*

1. *If $\mathcal{B}_0^{(m)} \le \mathcal{B}_0^{\mathrm{cold}}$ and $\mathcal{M}^{(m)} > \mathcal{W}_+^{(m)}(\delta)$, then candidate m is certified as* Accept.

2. *If* $\mathcal{B}_0^{(m)} > \mathcal{B}_0^{\mathrm{cold}}$ *and* $\mathcal{M}^{(m)} > \mathcal{W}_+^{(m)}(\delta)$*, then candidate* m *is certified as* REJECT.

Moreover, the combined width admits the pilot-level bound

$$\mathcal{W}_+^{(m)}(\delta) \le 2\big(\gamma^{(m)} + \gamma^{\mathrm{cold}}\big)\rho_{\mathrm{pilot}}(\delta), \tag{16}$$

so a sufficient pilot-level condition for decisiveness is

$$\mathcal{M}^{(m)} > 2\big(\gamma^{(m)} + \gamma^{\mathrm{cold}}\big)\rho_{\mathrm{pilot}}(\delta). \tag{17}$$

Corollary 5 characterizes when a short pilot suffices. Once the population gap between a candidate prior and the cold-start baseline exceeds the combined certificate width, the pilot is informative enough to escape the inconclusive regime. The right-hand side of (17) depends on the pilot through the radius $\rho_{\mathrm{pilot}}(\delta)$ and the conversion factors $\gamma^{(m)}, \gamma^{\mathrm{cold}}$. As $V_{T_{\mathrm{pilot}}}^{\mathrm{pilot}}$ grows in the positive semi-definite order, the conversion factors decrease, scaling on the order of $T_{\mathrm{pilot}}^{-1/2}$ under an isotropic design, while $\rho_{\mathrm{pilot}}(\delta)$ grows only logarithmically. The combined width contracts with pilot size, enabling decisive certification of progressively smaller gaps. The condition is sufficient but conservative: the actual intervals may separate even when the bounds do not prove separation.

3.4 SOFT TRUST CALIBRATION

In finite-pilot regimes and noisier environments, the hard rule above may yield inconclusive outcomes. We therefore introduce an endpoint-clamped soft trust-calibration mechanism, preserving the certified endpoints while interpolating between warm start and cold start when the pilot is inconclusive.

For any interval $\mathcal{I} = [L, U]$, define its midpoint and width by

$$c(\mathcal{I}) := \frac{L+U}{2}, \qquad w(\mathcal{I}) := U - L. \tag{18}$$

Write

$$c^{(m)} := c(\mathcal{I}^{(m)}), \qquad w^{(m)} := w(\mathcal{I}^{(m)}), \qquad c^{\mathrm{cold}} := c(\mathcal{I}^{\mathrm{cold}}), \qquad w^{\mathrm{cold}} := w(\mathcal{I}^{\mathrm{cold}}).$$

We define the normalized relative advantage score

$$\mathcal{S}^{(m)} := \frac{c^{\mathrm{cold}} - c^{(m)}}{\frac{1}{2}\big(w^{(m)} + w^{\mathrm{cold}}\big)}. \tag{19}$$

This score is positive when the midpoint of the warm interval is smaller than the midpoint of the cold interval, and is scaled by the average interval width so that it heuristically balances relative advantage against uncertainty.

Definition 6 (Soft trust weight)**.** *Fix a temperature parameter* $\kappa > 0$*. The soft trust weight of candidate* m *is*

$$\omega^{(m)} := \begin{cases} 1, & \text{if candidate } m \text{ is ACCEPT}, \\ 0, & \text{if candidate } m \text{ is REJECT}, \\ \sigma\big(\kappa \mathcal{S}^{(m)}\big), & \text{if candidate } m \text{ is INCONCLUSIVE}, \end{cases} \tag{20}$$

where $\sigma(z) = 1/(1+e^{-z})$ *is the logistic sigmoid.*

The resulting deployment prior is obtained by convex interpolation of the cold and warm sufficient statistics. This construction recovers the warm prior on ACCEPT and the cold start prior on REJECT and is well defined for all $\omega^{(m)} \in [0,1]$. Appendix A.6 gives the explicit construction and proves the endpoint and continuity properties. We treat the soft rule as a certificate-derived extension for the ambiguous middle regime rather than as an equivalent certified decision rule. The next section empirically evaluates these hard and soft rules across several warm start regimes.

4 EXPERIMENTAL SETUP

We evaluate the pilot-certification framework in three settings: a controlled synthetic environment where $\theta^\star$ is known, and two real-data deployment scenarios with distinct prior sources—LLM-generated preferences (COVID conjoint) and historical production logs (KuaiRand).

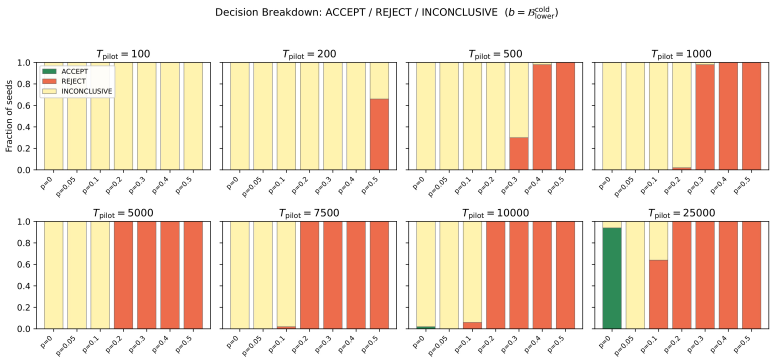

Figure 1: Fraction of seeds assigned to ACCEPT, REJECT, and INCONCLUSIVE across preference-flip rates and pilot sizes in the Synthetic setting

4.1 SHARED PROTOCOL

For each candidate prior m, we construct $(A_0^{(m)}, \theta_0^{(m)})$ from auxiliary data, collect a shared pilot of length T_{pilot} from the target environment, and compute certified intervals for both the candidate and the cold-start baseline. These intervals induce the cold-relative labels ACCEPT, REJECT, and INCONCLUSIVE. In the primary evaluation, the pilot is used only for certification, and deployment is evaluated on a held-out suffix. Appendix B.9.1 evaluates pilot-inclusive ablations.

We compare five policies on the deployment suffix: COLD (always cold), WARM (always warm), CONSERVATIVE (warm iff ACCEPT), PERMISSIVE (warm unless REJECT), and SOFT, which blends warm and cold in the inconclusive regime using the trust weight of Section 3 and recovers the hard endpoints otherwise. Results are reported as two complementary summaries: a *decision-group audit*, which asks whether labels predict downstream behavior, and a *policy comparison*, which asks whether obeying the gate improves deployment. Full dataset construction, shared hyperparameters, candidate definitions, feature maps, and additional evaluations are deferred to Appendix B.

4.2 SETTINGS

Synthetic. Linear contextual bandit with known $\theta^\star$ and $\mathcal{B}_0^{(m)}$. Candidates are corrupted by preference-flip rates and target-shift magnitudes, enabling direct oracle diagnostics: certified-interval coverage, decision correctness, and the pilot-size-versus-width relation of Corollary 5.

COVID conjoint (Kriner et al., 2020). Candidates are GPT-4o-generated priors fit via the CBLI construction (Alamdari et al., 2024), with preference-flipping (Bayley et al., 2026) for controlled misalignment. The target is real human responses from a conjoint study of 1,970 respondents. The prior and target come from *epistemically separate generative processes*, so the certificate's task is to detect geometric misalignment from pilot evidence alone.

KuaiRand (Gao et al., 2022). Candidates are warm priors fit from historical adaptive logs on KuaiRand-Pure. The target is the random-intervention stream from the same platform. We cluster target items into $K = 7$ arms via balanced-mass k-means on item features, with no target rewards used in arm construction. Deployment is evaluated by SNIPS replay reward lift versus cold start on the randomized target stream. Section 5.3 focuses on the cleaner `old_full` versus `shuffled_old` contrast. Appendix B.7.2 reports results for the additional `recent` prior.

5 RESULTS AND DISCUSSION

We evaluate the framework at two levels. First, we ask whether the cold-relative certificate behaves as predicted by theory. Second, we ask whether acting on the certificate improves downstream deployment relative to always warming or always starting cold. The three settings play complementary roles: synthetic validates the certification mechanism with known ground truth, COVID tests LLM-to-human alignment, and KuaiRand tests historical policy shift. Additional results are in Appendix B.

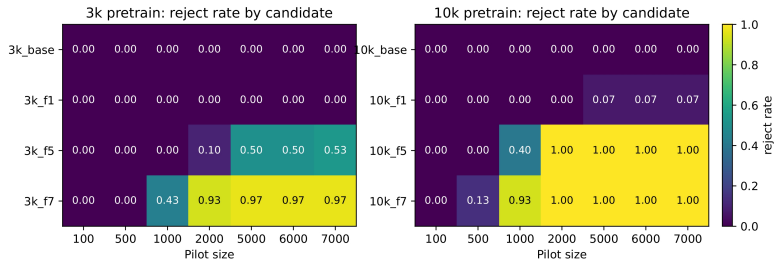

Figure 2: COVID dataset reject rates by candidate prior and pilot size for (left) $n_s = 3$k and (right) $n_s = 10$k synthetic pretraining samples. The base prior is unflipped; `f1`, `f5`, and `f7` reverse 10%, 50%, and 70% of resolved synthetic preference labels before fitting.

Table 1: KuaiRand source-size sweep using SNIPS replay reward lift versus cold start at $T_{\text{pilot}} = 20{,}000$. Positive values indicate improvement over cold. Confidence intervals are paired bootstrap 95% intervals over the fixed-deploy windows; p-values are Wilcoxon signed-rank tests. Percentages report $100 \times \Delta_{\text{SNIPS}}/V_{\text{cold}}$, with $V_{\text{cold}} = 0.1014$. Full replay details are in Appendix B.6.

Source size	Candidate / contrast	Δ_{SNIPS}	95% CI	p
10,000	`old_full`	$+0.0145$ $(+14.3\%)$	$[+0.0108, +0.0183]$	0.002
10,000	`shuffled_old`	-0.0058 (-5.7%)	$[-0.0116, +0.0006]$	0.160
100,000	`old_full`	$+0.0118$ $(+11.7\%)$	$[+0.0074, +0.0170]$	0.002
100,000	`shuffled_old`	-0.0189 (-18.7%)	$[-0.0270, -0.0105]$	0.006
10,000	`old_full` − `shuffled_old`	$+0.0203$	$[+0.0153, +0.0252]$	0.002
100,000	`old_full` − `shuffled_old`	$+0.0308$	$[+0.0234, +0.0379]$	0.002

5.1 Synthetic Results

The synthetic setting evaluates the certification mechanism itself. Figure 1 shows that the certificate becomes decisive in the correct direction as pilot information increases. With small pilots, most cases remain Inconclusive. As the pilot grows, Reject decisions appear first for misaligned priors, while Accept decisions emerge only for the cleanest and best-aligned priors. This is the margin-versus-width pattern predicted by the theory: when the pilot is weak, the confidence region is too wide to separate Warm from Cold, and as the pilot grows, the width contracts and the rule becomes decisive.

5.2 COVID-19 Results

The COVID experiment asks whether the certificate can detect misalignment when the prior is produced by a language model and the target is real human behavior. Figure 2 shows that Reject rates increase with candidate corruption and with pilot size. The same ordering appears at both pretraining scales, but the separation is sharper at $10k$, where the prior is more informative and therefore more capable of harming deployment when misaligned. Inconclusive cases remain where the pilot is too small, or the candidate is only mildly misaligned. In this setting, the pilot must discriminate between a candidate prior and the cold baseline using only a limited amount of human data, while the downstream learner is still a finite-horizon contextual bandit. When the candidate is only mildly misaligned, or when the pilot is not yet large enough, the certified interval for the candidate overlaps with the cold interval, so neither Accept nor Reject is justified. The certificate is therefore responding to target compatibility, not merely to how the synthetic prior was constructed.

5.3 KuaiRand Results

KuaiRand is the hardest setting because the prior and target come from the same platform but different logging policies. We compare `old_full`, the earlier standard-policy source, with `shuffled_old`, a reward-permuted null built from that same source. Here, the hard certificate remains Inconclusive, which is the expected behavior when pilot information is insufficient to separate warm from cold under policy shift. The setting is also intrinsically noisier than the synthetic and COVID experiments; however, Table 1 shows that this conservative certificate does not imply the absence of downstream signal. Under SNIPS replay reward lift, `old_full` improves over cold at $s = 10{,}000$ and $s = 100{,}000$,

Table 2: Representative downstream policy comparison across settings. Entries report deployment effect versus cold start. Positive values indicate improvement, negative values indicate harm. For synthetic and COVID, the effect is $-\Delta_{\text{regret}}$. For KuaiRand, the effect is SNIPS replay reward lift.

Setting	Candidate / regime	Warm	Conservative	Soft
Synthetic	`flip_p=0.00`, $T_{\text{pilot}} = 25000$	$+3.7867$	$+3.5319$	$+3.7889$
Synthetic	`misalign_||Δ||=2.0`, $T_{\text{pilot}} = 5000$	-169.5610	0.0000	0.0000
COVID	`10k_base`, $T_{\text{pilot}} = 2000$	$+80.267$	0.0000	$+72.600$
COVID	`10k_f7`, $T_{\text{pilot}} = 2000$	-338.6003	0.0000	0.0000
KuaiRand	`old_full`, $s = 100{,}000$, $T_{\text{pilot}} = 20{,}000$	$+0.0118$	0.00000	$+0.0103$
KuaiRand	`shuffled_old`, $s = 100{,}000$, $T_{\text{pilot}} = 20{,}000$	-0.0189	0.00000	-0.0059

while `shuffled_old` is non-beneficial at $s = 10{,}000$ and harmful at $s = 100{,}000$. The small non-monotonicity between `old_full` sizes is consistent with policy shift: a larger historical prior is more concentrated but not necessarily better matched to the target replay windows. The pairwise contrast is significant at both source sizes, indicating that the replay evaluation distinguishes transferable historical signal from the reward-permuted null even when the hard certificate abstains.

5.4 Cross-setting downstream effects

Table 2 summarizes downstream policy behavior. Conservative uses the certified prior-error comparison but often falls back to cold, losing value when a prior is helpful. Soft tracks Warm on aligned priors and pulls back toward Cold on harmful ones: it preserves most of the COVID base-prior gain, matches Warm on the synthetic aligned prior, preserves most of the KuaiRand `old_full` SNIPS gain, and reduces the KuaiRand reward-shuffled SNIPS loss from -0.0189 to -0.0059. Scale differences across settings reflect reward units rather than methodological inconsistency: synthetic uses unbounded linear rewards, COVID uses binary pairwise choices, and KuaiRand uses a binary engagement signal with a base rate of approximately 8.5%. Within each setting, the directional pattern is consistent: Soft retains much of the benefit when the prior is helpful and reduces harm when the prior is misaligned. Appendix B.8.3 compares fixed source weights, Appendix B.9.1 evaluates pilot reuse, and Appendix B.10 considers alternative pilot-based rules.

6 Limitations

Our guarantee certifies prior-error ordering, not finite-horizon regret directly. On the pilot confidence event, the hard rule makes no incorrect Accept or Reject decisions relative to cold start in terms of $\mathcal{B}_0$, but the realized regret still depends on the deployment horizon, exploration rule, action availability, and feature geometry. The hard rule is conservative by design. The margin-versus-width condition is sufficient, not necessary, so small population gaps, weak pilot coverage, or high-dimensional feature maps can leave the certificate inconclusive even when a useful prior exists. This behavior is visible in KuaiRand, where policy-level signal remains despite hard-rule abstention. Regarding the soft trust-calibration rule, it is certificate-derived but not itself certified. It shares the hard rule's endpoints and is well defined by convex interpolation, but intermediate trust weights do not inherit the same $1-\delta$ correctness guarantee. Finally, our real-data evidence is non-live and linear. COVID uses sequential evaluation on a fixed held-out set of human responses, while KuaiRand uses offline clustered-arm replay rather than live deployment. SNIPS corrects for unequal induced arm propensities in the replay reward estimate, but does not remove limitations from the arm abstraction, finite overlap, or offline evaluation of adaptive policies. Extending the framework to nonlinear policies, pilot policies for source discrimination, multiple simultaneous priors, and live online evaluation remains future work.

7 Conclusion

We introduced a pilot-based certification procedure that turns warm start prior-quality assessment into a pre-deployment decision. A short real pilot produces certified cold-relative intervals on $\mathcal{B}_0^{(m)}$, yielding accept, reject, and inconclusive labels with a $1-\delta$ correctness guarantee for the hard prior-error decisions. We provide a theoretical backing that characterizes when the hard rule becomes decisive, and a soft trust-calibration rule interpolates between warm and cold in the inconclusive regime. Across three experiments, we demonstrate that the framework works when the ground truth is known, can detect LLM-human misalignment, and remains conservative under historical policy shift while still reducing harm through the soft rule.

AI USE STATEMENT

In this work, we used generative AI tools (LLMs). They are part of the experimental setup because the COVID-19 priors are generated from GPT-4o through the CBLI construction. This is described in Section 4 and Appendix B.4. The tools were additionally used for grammar, and spelling checks. The authors reviewed all AI-assisted work and take responsibility for its contents.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

Parand A. Alamdari, Yanshuai Cao, and Kevin H. Wilson. Jump Starting Bandits with LLM-Generated Prior Knowledge. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19821–19833, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1107. URL https://aclanthology.org/2024.emnlp-main.1107.

Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(59):1823–1840, 2008. URL http://jmlr.org/papers/v9/bartlett08a.html.

Adam Bayley, Xiaodan Zhu, Raquel Aoki, Yanshuai Cao, and Kevin H. Wilson. Jump Start or False Start? A Theoretical and Empirical Evaluation of LLM-initialized Bandits, 2026. URL https://arxiv.org/abs/2604.02527.

Changxiao Cai, T. Tony Cai, and Hongzhe Li. Transfer learning for contextual multi-armed bandits, 2024. URL https://arxiv.org/abs/2211.12612.

Brian M Cho, Ana-Roxana Pop, Kyra Gan, Sam Corbett-Davies, Israel Nir, Ariel Evnine, and Nathan Kallus. Cspi-mt: Calibrated safe policy improvement with multiple testing for threshold policies, 2024. URL https://arxiv.org/abs/2408.12004.

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

Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, JMLR Workshop and Conference Proceedings, pp. 208–214, 2011.

Mingwei Deng, Ville Kyrki, and Dominik Baumann. Transfer learning in latent contextual bandits with covariate shift through causal transportability, 2025. URL https://arxiv.org/abs/2502.20153.

Chongming Gao, Shijun Li, Yuan Zhang, Jiawei Chen, Biao Li, Wenqiang Lei, Peng Jiang, and Xiangnan He. KuaiRand: An Unbiased Sequential Recommendation Dataset with Randomly Exposed Videos. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, CIKM '22, pp. 3953–3957, 2022. doi: 10.1145/3511808.3557624. URL https://doi.org/10.1145/3511808.3557624.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence, 2016. URL https://arxiv.org/abs/1602.04589.

Joey Hong, Branislav Kveton, Manzil Zaheer, and Mohammad Ghavamzadeh. Hierarchical bayesian bandits, 2022. URL https://arxiv.org/abs/2111.06929.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models, 2016. URL https://arxiv.org/abs/1407.4443.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi-Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits, 2017. URL https://arxiv.org/abs/1611.06426.

Sarah Kreps, Sandip Prasad, John S. Brownstein, Yulin Hswen, Brian T. Garibaldi, Baobao Zhang, and Douglas L. Kriner. Factors Associated With US Adults' Likelihood of Accepting COVID-19 Vaccination. *JAMA Network Open*, 3(10):e2025594–e2025594, 10 2020. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2020.25594. URL https://doi.org/10.1001/jamanetworkopen.2020.25594.

Douglas Kriner, Sarah Kreps, John S Brownstein, Yulin Hswen, Baobao Zhang, and Sandip Prasad. Replication Data for: Factors Associated With US Adults' Likelihood of Accepting COVID-19 Vaccination: Evidence From a Survey and Choice-Based Conjoint Analysis, 2020. URL https://doi.org/10.7910/DVN/6BSJYP.

Romain Laroche, Paul Trichelair, and Rémi Tachet des Combes. Safe policy improvement with baseline bootstrapping, 2019. URL https://arxiv.org/abs/1712.06924.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670. ACM, 2010.

Bastian Oetomo, R. Malinga Perera, Renata Borovica-Gajic, and Benjamin I. P. Rubinstein. Cutting to the chase with warm-start contextual bandits. In *2021 IEEE International Conference on Data Mining (ICDM)*, pp. 459–468, 2021. doi: 10.1109/ICDM51629.2021.00057.

OpenAI. GPT-4 Technical Report, 2024. URL https://arxiv.org/abs/2303.08774.

Marek Petrik, Yinlam Chow, and Mohammad Ghavamzadeh. Safe policy improvement by minimizing robust baseline regret, 2016. URL https://arxiv.org/abs/1607.03842.

Yufei Ruan, Jiaqi Yang, and Yuan Zhou. Linear bandits with limited adaptivity and learning distributional optimal design, 2021. URL https://arxiv.org/abs/2007.01980.

Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation, 2021. URL https://arxiv.org/abs/2008.07146.

Max Simchowitz, Christopher Tosh, Akshay Krishnamurthy, Daniel Hsu, Thodoris Lykouris, Miroslav Dudík, and Robert E. Schapire. Bayesian decision-making under misspecified priors with applications to meta-learning, 2021. URL https://arxiv.org/abs/2107.01509.

Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/39027dfad5138c9ca0c474d71db915c3-Paper.pdf.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2380–2388, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/thomas15.html.

Andrea Zanette, Kefan Dong, Jonathan Lee, and Emma Brunskill. Design of experiments for stochastic contextual linear bandits, 2021. URL https://arxiv.org/abs/2107.09912.

Chicheng Zhang, Alekh Agarwal, Hal Daumé III, John Langford, and Sahand N Negahban. Warm-starting contextual bandits: Robustly combining supervised and bandit feedback, 2019. URL https://arxiv.org/abs/1901.00301.

A Proofs of Theorems in Main Body

A.1 Proof of Theorem 1

Proof. From the definition of the pilot ridge estimator,

$$\hat{\theta}_t^{\text{pilot}} = \big(V_t^{\text{pilot}}\big)^{-1} \sum_{s \le t} r_s x_s, \qquad V_t^{\text{pilot}} = \tau_{\text{pilot}} I + \sum_{s \le t} x_s x_s^\top,$$

and the reward model $r_s = x_s^\top \theta^\star + \xi_s$, we obtain

$$\sum_{s \le t} r_s x_s = \sum_{s \le t} x_s x_s^\top \theta^\star + \sum_{s \le t} \xi_s x_s.$$

Therefore,

$$V_t^{\text{pilot}} \hat{\theta}_t^{\text{pilot}} = \big(V_t^{\text{pilot}} - \tau_{\text{pilot}} I\big)\theta^\star + \sum_{s \le t} \xi_s x_s,$$

and hence

$$V_t^{\text{pilot}} \big(\hat{\theta}_t^{\text{pilot}} - \theta^\star\big) = \sum_{s \le t} \xi_s x_s - \tau_{\text{pilot}} \theta^\star.$$

Multiplying by $\big(V_t^{\text{pilot}}\big)^{-1/2}$ and taking Euclidean norms gives

$$\|\hat{\theta}_t^{\text{pilot}} - \theta^\star\|_{V_t^{\text{pilot}}} \le \left\| \sum_{s \le t} \xi_s x_s \right\|_{\left(V_t^{\text{pilot}}\right)^{-1}} + \tau_{\text{pilot}} \|\theta^\star\|_{\left(V_t^{\text{pilot}}\right)^{-1}}.$$

For the first term, apply Theorem 1 of Abbasi-Yadkori et al. (2011) with initial matrix $V = \tau_{\text{pilot}} I$, noise scale $R = \sigma$, and the pilot feature sequence. With respect to the pilot filtration, x_s is measurable before observing ξ_s, and the conditional sub-Gaussian assumption is

$$\mathbb{E}\big[e^{\lambda \xi_s} \mid \mathcal{F}_{s-1}\big] \le e^{\lambda^2 \sigma^2 / 2}, \qquad \lambda \in \mathbb{R}.$$

The self-normalized inequality gives, with probability at least $1 - \delta$, simultaneously over pilot prefixes,

$$\left\| \sum_{s \le t} \xi_s x_s \right\|^2_{(V_t^{\text{pilot}})^{-1}} \le 2\sigma^2 \log\left(\frac{\det(V_t^{\text{pilot}})^{1/2}}{\det(\tau_{\text{pilot}} I)^{1/2} \delta} \right).$$

Taking square roots and using the definition in (7) yields, for all $t \le T_{\text{pilot}}$,

$$\left\| \sum_{s \le t} \xi_s x_s \right\|_{\left(V_t^{\text{pilot}}\right)^{-1}} \le \beta_t^{\text{pilot}}(\delta).$$

For the second term, since $V_t^{\text{pilot}} \succeq \tau_{\text{pilot}} I$, we have

$$\big(V_t^{\text{pilot}}\big)^{-1} \preceq \tau_{\text{pilot}}^{-1} I,$$

and therefore

$$\tau_{\text{pilot}} \|\theta^\star\|_{\left(V_t^{\text{pilot}}\right)^{-1}} = \sqrt{\tau_{\text{pilot}}} \cdot \sqrt{\tau_{\text{pilot}} (\theta^\star)^\top \big(V_t^{\text{pilot}}\big)^{-1} \theta^\star} \le \sqrt{\tau_{\text{pilot}}}\, \|\theta^\star\|_2 \le \sqrt{\tau_{\text{pilot}}}\, S.$$

Combining the two bounds yields

$$\|\hat{\theta}_t^{\text{pilot}} - \theta^\star\|_{V_t^{\text{pilot}}} \le \beta_t^{\text{pilot}}(\delta) + \sqrt{\tau_{\text{pilot}}}\, S$$

simultaneously for all $t \le T_{\text{pilot}}$ with probability at least $1 - \delta$. The equivalent set-based statement follows immediately from the definition of $\mathcal{C}_t^{\text{pilot}}(\delta)$. □

A.2 AUXILIARY NORM CONVERSION LEMMA

Lemma 7 (Norm conversion). *Let $V \succ 0$ and $A \succeq 0$. Then for any $v \in \mathbb{R}^d$,*

$$\|v\|_A \le \sqrt{\lambda_{\max}(AV^{-1})}\,\|v\|_V.$$

Proof. For any $v \in \mathbb{R}^d$,

$$\|v\|_A^2 = v^\top A v = v^\top V^{1/2}\big(V^{-1/2}AV^{-1/2}\big)V^{1/2}v.$$

Since $V^{-1/2}AV^{-1/2}$ is positive semidefinite, its Rayleigh quotient is bounded by its largest eigenvalue:

$$v^\top A v \le \lambda_{\max}\big(V^{-1/2}AV^{-1/2}\big)\, v^\top V v.$$

Using the similarity relation

$$\lambda_{\max}\big(V^{-1/2}AV^{-1/2}\big) = \lambda_{\max}(AV^{-1}),$$

we obtain

$$\|v\|_A^2 \le \lambda_{\max}(AV^{-1})\|v\|_V^2.$$

Taking square roots yields the claim. □

A.3 PROOF OF THEOREM 2

Proof. Fix a candidate model m and suppose that

$$\theta^\star \in \mathcal{C}^{\text{pilot}}_{T_{\text{pilot}}}(\delta).$$

By definition of the pilot ellipsoid,

$$\|\theta^\star - \hat\theta^{\text{pilot}}_{T_{\text{pilot}}}\|_{V^{\text{pilot}}_{T_{\text{pilot}}}} \le \rho_{\text{pilot}}(\delta).$$

Using the decomposition

$$\theta_0^{(m)} - \theta^\star = \big(\theta_0^{(m)} - \hat\theta^{\text{pilot}}_{T_{\text{pilot}}}\big) + \big(\hat\theta^{\text{pilot}}_{T_{\text{pilot}}} - \theta^\star\big),$$

and the triangle inequality in the $A_0^{(m)}$-norm, we obtain

$$\mathcal{B}_0^{(m)} = \|\theta_0^{(m)} - \theta^\star\|_{A_0^{(m)}} \le \|d^{(m)}\|_{A_0^{(m)}} + \|\hat\theta^{\text{pilot}}_{T_{\text{pilot}}} - \theta^\star\|_{A_0^{(m)}}.$$

Applying Lemma 7 with

$$A = A_0^{(m)}, \qquad V = V^{\text{pilot}}_{T_{\text{pilot}}},$$

gives

$$\|\hat\theta^{\text{pilot}}_{T_{\text{pilot}}} - \theta^\star\|_{A_0^{(m)}} \le \gamma^{(m)}\|\hat\theta^{\text{pilot}}_{T_{\text{pilot}}} - \theta^\star\|_{V^{\text{pilot}}_{T_{\text{pilot}}}} \le \gamma^{(m)}\rho_{\text{pilot}}(\delta).$$

Hence

$$\mathcal{B}_0^{(m)} \le \|d^{(m)}\|_{A_0^{(m)}} + \gamma^{(m)}\rho_{\text{pilot}}(\delta) = \overline{\mathcal{B}}^{(m)}(\delta).$$

For the lower bound, the reverse triangle inequality in the $A_0^{(m)}$-norm gives

$$\mathcal{B}_0^{(m)} = \|\theta_0^{(m)} - \theta^\star\|_{A_0^{(m)}} \ge \Big|\|d^{(m)}\|_{A_0^{(m)}} - \|\hat\theta^{\text{pilot}}_{T_{\text{pilot}}} - \theta^\star\|_{A_0^{(m)}}\Big|.$$

Using the same norm-conversion bound as above,

$$\mathcal{B}_0^{(m)} \ge \|d^{(m)}\|_{A_0^{(m)}} - \gamma^{(m)}\rho_{\text{pilot}}(\delta).$$

Since $\mathcal{B}_0^{(m)} \ge 0$, we conclude

$$\mathcal{B}_0^{(m)} \ge \max\Big\{0,\ \|d^{(m)}\|_{A_0^{(m)}} - \gamma^{(m)}\rho_{\text{pilot}}(\delta)\Big\} = \underline{\mathcal{B}}^{(m)}(\delta).$$

This proves (11). Simultaneous validity over a finite set $\mathcal{M}$ follows because the only random event used is the pilot confidence event, which holds uniformly with probability at least $1-\delta$. □

A.4 Proof of Theorem 4

Proof. Work on the event that all certified intervals are valid simultaneously, which by Theorem 2 occurs with probability at least $1 - \delta$.

If candidate m is ACCEPT, then by definition

$$\overline{\mathcal{B}}^{(m)}(\delta) \le \underline{\mathcal{B}}^{\mathrm{cold}}(\delta).$$

Since the certified intervals are valid on the event under consideration,

$$\mathcal{B}_0^{(m)} \le \overline{\mathcal{B}}^{(m)}(\delta) \le \underline{\mathcal{B}}^{\mathrm{cold}}(\delta) \le \mathcal{B}_0^{\mathrm{cold}}.$$

Thus $\mathcal{B}_0^{(m)} \le \mathcal{B}_0^{\mathrm{cold}}$.

If candidate m is REJECT, then by definition

$$\underline{\mathcal{B}}^{(m)}(\delta) > \overline{\mathcal{B}}^{\mathrm{cold}}(\delta).$$

Again using interval validity,

$$\mathcal{B}_0^{(m)} \ge \underline{\mathcal{B}}^{(m)}(\delta) > \overline{\mathcal{B}}^{\mathrm{cold}}(\delta) \ge \mathcal{B}_0^{\mathrm{cold}}.$$

Thus $\mathcal{B}_0^{(m)} > \mathcal{B}_0^{\mathrm{cold}}$. This establishes correctness of the decisive outcomes. □

A.5 Proof of Corollary 5

Proof. Work on the event that the candidate and cold-start certified intervals are simultaneously valid, which by Theorem 2 occurs with probability at least $1 - \delta$.

From interval validity,

$$\underline{\mathcal{B}}^{(m)}(\delta) \le \mathcal{B}_0^{(m)} \le \overline{\mathcal{B}}^{(m)}(\delta), \qquad \underline{\mathcal{B}}^{\mathrm{cold}}(\delta) \le \mathcal{B}_0^{\mathrm{cold}} \le \overline{\mathcal{B}}^{\mathrm{cold}}(\delta). \tag{21}$$

Hence

$$\overline{\mathcal{B}}^{(m)}(\delta) - \mathcal{B}_0^{(m)} \le \overline{\mathcal{B}}^{(m)}(\delta) - \underline{\mathcal{B}}^{(m)}(\delta) = \mathcal{W}^{(m)}(\delta), \tag{22}$$

and

$$\mathcal{B}_0^{\mathrm{cold}} - \underline{\mathcal{B}}^{\mathrm{cold}}(\delta) \le \overline{\mathcal{B}}^{\mathrm{cold}}(\delta) - \underline{\mathcal{B}}^{\mathrm{cold}}(\delta) = \mathcal{W}^{\mathrm{cold}}(\delta). \tag{23}$$

Proof of item (1) of Corollary 5. Suppose $\mathcal{B}_0^{(m)} \le \mathcal{B}_0^{\mathrm{cold}}$ and

$$\mathcal{M}^{(m)} = \mathcal{B}_0^{\mathrm{cold}} - \mathcal{B}_0^{(m)} > \mathcal{W}_+^{(m)}(\delta).$$

Then

$$\begin{aligned}
\underline{\mathcal{B}}^{\mathrm{cold}}(\delta) - \overline{\mathcal{B}}^{(m)}(\delta) &= \big[\mathcal{B}_0^{\mathrm{cold}} - \big(\mathcal{B}_0^{\mathrm{cold}} - \underline{\mathcal{B}}^{\mathrm{cold}}(\delta)\big)\big] - \big[\mathcal{B}_0^{(m)} + \big(\overline{\mathcal{B}}^{(m)}(\delta) - \mathcal{B}_0^{(m)}\big)\big] \\
&\ge \big[\mathcal{B}_0^{\mathrm{cold}} - \mathcal{W}^{\mathrm{cold}}(\delta)\big] - \big[\mathcal{B}_0^{(m)} + \mathcal{W}^{(m)}(\delta)\big] \\
&= \mathcal{M}^{(m)} - \mathcal{W}_+^{(m)}(\delta) > 0.
\end{aligned}$$

Therefore $\overline{\mathcal{B}}^{(m)}(\delta) < \underline{\mathcal{B}}^{\mathrm{cold}}(\delta)$, so by Definition 3 candidate m is certified as ACCEPT.

Proof of item (2) of Corollary 5. Suppose $\mathcal{B}_0^{(m)} > \mathcal{B}_0^{\mathrm{cold}}$ and

$$\mathcal{M}^{(m)} = \mathcal{B}_0^{(m)} - \mathcal{B}_0^{\mathrm{cold}} > \mathcal{W}_+^{(m)}(\delta).$$

By the symmetric slack bounds

$$\mathcal{B}_0^{(m)} - \underline{\mathcal{B}}^{(m)}(\delta) \le \mathcal{W}^{(m)}(\delta), \qquad \overline{\mathcal{B}}^{\mathrm{cold}}(\delta) - \mathcal{B}_0^{\mathrm{cold}} \le \mathcal{W}^{\mathrm{cold}}(\delta),$$

we obtain

$$\begin{aligned}
\underline{\mathcal{B}}^{(m)}(\delta) - \overline{\mathcal{B}}^{\mathrm{cold}}(\delta) &\ge \big[\mathcal{B}_0^{(m)} - \mathcal{W}^{(m)}(\delta)\big] - \big[\mathcal{B}_0^{\mathrm{cold}} + \mathcal{W}^{\mathrm{cold}}(\delta)\big] \\
&= \mathcal{M}^{(m)} - \mathcal{W}_+^{(m)}(\delta) > 0,
\end{aligned}$$

so $\underline{\mathcal{B}}^{(m)}(\delta) > \overline{\mathcal{B}}^{\mathrm{cold}}(\delta)$, so candidate m is certified as REJECT.

Proof of (16)**.** From (12)–(13),

$$\begin{aligned}\mathcal{W}^{(m)}(\delta) &= \overline{\mathcal{B}}^{(m)}(\delta) - \underline{\mathcal{B}}^{(m)}(\delta) \\ &\le \big[\|d^{(m)}\|_{A_0^{(m)}} + \gamma^{(m)}\rho_{\text{pilot}}(\delta)\big] - \big[\|d^{(m)}\|_{A_0^{(m)}} - \gamma^{(m)}\rho_{\text{pilot}}(\delta)\big] \\ &= 2\gamma^{(m)}\rho_{\text{pilot}}(\delta),\end{aligned}$$

with equality whenever the lower bound is not floored at zero. The same argument for cold-start gives

$$\mathcal{W}^{\text{cold}}(\delta) \le 2\gamma^{\text{cold}}\rho_{\text{pilot}}(\delta).$$

Adding the two inequalities yields

$$\mathcal{W}_+^{(m)}(\delta) \le 2\big(\gamma^{(m)} + \gamma^{\text{cold}}\big)\rho_{\text{pilot}}(\delta),$$

which is (16). The sufficient condition (17) follows immediately by substituting this upper bound into the decisiveness condition $\mathcal{M}^{(m)} > \mathcal{W}_+^{(m)}(\delta)$. □

A.6 Hybrid Prior Construction

Given $\omega^{(m)} \in [0,1]$, we define the hybrid prior by convex interpolation of the cold and warm sufficient statistics:

$$A_0^{\text{hybrid},(m)} := (1-\omega^{(m)})A_0^{\text{cold}} + \omega^{(m)}A_0^{(m)}, \tag{24}$$

$$b_0^{\text{hybrid},(m)} := (1-\omega^{(m)})b_0^{\text{cold}} + \omega^{(m)}b_0^{(m)}, \tag{25}$$

and

$$\theta_0^{\text{hybrid},(m)} := \big(A_0^{\text{hybrid},(m)}\big)^{-1} b_0^{\text{hybrid},(m)}. \tag{26}$$

Proposition 8 (Basic properties of the hybrid prior)**.** *For every candidate m:*

1. $A_0^{\text{hybrid},(m)} \succ 0$ *for all* $\omega^{(m)} \in [0,1]$.
2. *If candidate m is* ACCEPT*, then*
$$(A_0^{\text{hybrid},(m)}, \theta_0^{\text{hybrid},(m)}) = (A_0^{(m)}, \theta_0^{(m)}).$$
3. *If candidate m is* REJECT*, then*
$$(A_0^{\text{hybrid},(m)}, \theta_0^{\text{hybrid},(m)}) = (A_0^{\text{cold}}, \theta_0^{\text{cold}}).$$
4. $A_0^{\text{hybrid},(m)}$*,* $b_0^{\text{hybrid},(m)}$*, and* $\theta_0^{\text{hybrid},(m)}$ *vary continuously as functions of the scalar weight* $\omega^{(m)}$.

Proof. Fix a candidate m and write $\omega := \omega^{(m)}$ for brevity.

Proof of item (1). Both A_0^{cold} and $A_0^{(m)}$ are positive definite. Hence for any nonzero $v \in \mathbb{R}^d$,

$$v^\top A_0^{\text{hybrid},(m)} v = (1-\omega)v^\top A_0^{\text{cold}} v + \omega v^\top A_0^{(m)} v > 0,$$

since $\omega \in [0,1]$. Therefore $A_0^{\text{hybrid},(m)} \succ 0$.

Proof of item (2). If candidate m is ACCEPT, then $\omega = 1$ by Definition 6. Substituting into (24)–(26) yields

$$A_0^{\text{hybrid},(m)} = A_0^{(m)}, \qquad b_0^{\text{hybrid},(m)} = b_0^{(m)}, \qquad \theta_0^{\text{hybrid},(m)} = \theta_0^{(m)}.$$

Proof of item (3). If candidate m is REJECT, then $\omega = 0$, so

$$A_0^{\text{hybrid},(m)} = A_0^{\text{cold}}, \qquad b_0^{\text{hybrid},(m)} = b_0^{\text{cold}}, \qquad \theta_0^{\text{hybrid},(m)} = \theta_0^{\text{cold}}.$$

Proof of item (4). $A_0^{\mathrm{hybrid},(m)}$ and $b_0^{\mathrm{hybrid},(m)}$ are affine functions of ω, hence continuous on $[0,1]$. Since the map $A \mapsto A^{-1}$ is continuous on the cone of positive definite matrices and part (1) shows that $A_0^{\mathrm{hybrid},(m)} \succ 0$ for all $\omega \in [0,1]$, it follows that

$$\theta_0^{\mathrm{hybrid},(m)} = \big(A_0^{\mathrm{hybrid},(m)}\big)^{-1} b_0^{\mathrm{hybrid},(m)}$$

is also continuous in ω. □

The soft rule is continuous as a function of the scalar trust weight $\omega^{(m)}$, but the implemented score-to-policy map is endpoint-clamped. For INCONCLUSIVE outcomes we set $\omega^{(m)} = \sigma(\kappa \mathcal{S}^{(m)})$, while for certified ACCEPT and REJECT outcomes we set $\omega^{(m)} = 1$ and $\omega^{(m)} = 0$, respectively. The map from the certificate score $\mathcal{S}^{(m)}$ to the final deployment prior is piecewise and need not be continuous at the hard decision boundaries. We note this as a design choice; the soft rule interpolates only in the ambiguous middle regime and preserves the certified hard endpoints exactly.

B SUPPLEMENTARY MATERIAL

This Appendix consists of a concise review of related works, information regarding the datasets used, and information needed to replicate the results we present in the paper. Additional sections describe pilot retention and decision timing, κ interpretations, and comparisons with other source-trust rules.

B.1 RELATED WORKS

As discussed in Section 1, warm priors can arise from several sources, including historical logs, transfer pipelines, and LLM-generated synthetic data. In this supplementary discussion, we focus on the literature most directly tied to our methodological contribution: warm-started contextual bandits, source mismatch, safe fallback behavior, and adjacent decision-theoretic notions of abstention and confidence-based identification.

B.1.1 WARM-STARTED CONTEXTUAL BANDITS AND TRANSFER

A substantial literature studies how auxiliary data can reduce the cold-start cost of contextual bandits. Prior work shows that warm-starts can improve finite-horizon performance by shrinking early uncertainty and accelerating learning when source and target tasks are sufficiently aligned (Oetomo et al., 2021; Saito et al., 2021). Related work on transfer learning for contextual bandits similarly characterizes when shared structure across tasks reduces regret and when transfer becomes harmful (Cai et al., 2024; Hong et al., 2022; Deng et al., 2025). This line largely asks how to *exploit* auxiliary information efficiently once it is available. Our setting is different: we assume a single candidate prior has already been constructed, its target-task quality is unknown, and only a short real pilot can be collected before the remaining horizon must be committed.

B.1.2 ROBUST WARM-STARTING UNDER SOURCE MISMATCH

The closest algorithmic neighbor is work on robustly combining supervised warm-start data with online contextual-bandit feedback under mismatch. In particular, ARRoW-CB adaptively reweights offline and online signals to mitigate harmful transfer when the two sources diverge (Zhang et al., 2019). This is an important methodological precedent, but it addresses a different operational problem: trust is learned *during* deployment through online adaptation. Our objective is instead pre-deployment certification. We use a short target-task pilot to decide whether to warm-start, revert to cold start, or interpolate between the two before the remaining deployment horizon is spent.

B.1.3 NON-REACTIVE AND LIMITED-ADAPTIVITY DEPLOYMENT

Non-reactive and limited-adaptivity deployment studies sequential decision making when operational constraints limit how frequently a policy can adapt to newly collected data. Non-reactive experimental designs consider data collection under fixed policies while limited-adaptivity methods study settings in which policy updates come in batches or through a small number of switches. Zanette et al. (2021) study experimental design for stochastic contextual linear bandits under non-reactive data collection,

motivated in part by the operational cost of frequent policy changes. Ruan et al. (2021) analyze linear bandits under limited adaptivity, including batch-learning and rare-policy-switch settings. These works provide precedent for separating information collection from a subsequent policy phase rather than assuming unrestricted online learning.

B.1.4 Misspecified priors and conservative fallback behavior

Our theoretical perspective also connects to work on decision making under misspecified priors and on conservative contextual bandits. Prior-misspecification analyses provide a language for how decision quality degrades when a prior is wrong, while conservative contextual-bandit methods formalize the principle that a learner should remain close to a trusted fallback when evidence for improvement is weak (Simchowitz et al., 2021; Kazerouni et al., 2017). We build on both ideas, but our contribution is different in form: we derive a pilot-based, cold-relative certification rule for a single candidate warm prior, together with a soft trust-calibration mechanism for the inconclusive regime.

B.1.5 High-confidence and safe policy improvement

High-confidence policy improvement literature considers when evidence justifies departing from a baseline. Some methods use batch reinforcement-learning approaches to provide probabilistic guarantees on proposed-policy quality (Thomas et al., 2015), while others constrain and formulate policy learning and improvement based on uncertainty (Laroche et al., 2019; Petrik et al., 2016). More recently, Cho et al. (2024) study baseline-relative improvement for threshold policies with multiple candidate tests. Our work shares the principle of evidence-dependent choice between a candidate and fallback. These methods target policy performance under their respective assumptions, whereas our hard rule certifies the cold-relative ordering of $\mathcal{B}_0$ for prior initializations. Accordingly, our certificate does not imply their policy-value guarantees. Downstream bandit regret also depends on the deployment horizon, action availability, exploration rule, and feature geometry.

B.1.6 Confidence-based identification and abstention

Our decisiveness analysis is also conceptually related to two broader areas of work. Fixed-confidence best-arm identification studies adaptive sampling and stopping rules that identify the best arm with high probability while minimizing sample complexity (Kaufmann et al., 2016; Garivier & Kaufmann, 2016). Their works are similar in that the ability to commit to a recommendation requires sufficient separation between alternatives relative to the statistical evidence available. Committing to a decision based on a population signal overcoming a statistical uncertainty width can be seen as analogous to our inconclusive regime and soft trust calibration.

Selective classification, or classification with a reject option, studies when a learner should abstain rather than commit, typically through error–reject or risk–coverage trade-offs (Chow, 1970; Bartlett & Wegkamp, 2008). Chow characterizes the Bayes-optimal error–reject trade-off, while Bartlett and Wegkamp give a statistical learning treatment with convex surrogate losses and fast rates under margin conditions. Similarly to the previous paragraph, these works provide a useful conceptual background for our inconclusive regime, where insufficient evidence leads to abstention rather than commitment. Our specific target is a pilot-based comparison of a candidate warm prior's error with the cold baseline, rather than selective prediction or best-arm identification.

B.2 Synthetic Validation Pipeline

B.2.1 Data-Generating Process

The synthetic environment is a linear contextual bandit with ambient dimension $d = 10$. At each round t, a sleeping-bandit action set is formed by independently including each of $K = 10$ arms with probability 0.7; the learner observes contexts only for available arms and must choose among them. Contexts are drawn i.i.d. from $\mathcal{N}(0, I_d)$ and rewards follow the standard linear model

$$r_t(a) = x_{t,a}^\top \theta^\star + \xi_t, \qquad \xi_t \sim \mathcal{N}(0, \sigma^2),$$

with reward noise $\sigma = 0.1$. The target parameter $\theta^\star \in \mathbb{R}^d$ is drawn uniformly from the unit sphere at the start of each seed. The pilot stream is drawn from the same distribution: $T_{\max} = 25{,}000$ pilot

contexts are sampled i.i.d. from $\mathcal{N}(0, I_d)$ and rewards are computed under the same $\theta^\star$ and noise level. Pilot prefixes of length T_{pilot} are nested within this single stream, so all pilot sizes are evaluated on the same underlying pilot data for a given seed.

B.2.2 Warm Prior Construction

For each seed, a synthetic design matrix $X_{\text{syn}} \in \mathbb{R}^{n_s \times d}$ is drawn i.i.d. from $\mathcal{N}(0, I_d)$ with $n_s = 500$. Two families of warm priors are constructed.

Preference-flip Priors. For each flip rate $p \in \{0.00, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50\}$, synthetic regression labels are generated as

$$\tilde{y}_i = x_i^\top \theta^\star + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma_s^2),$$

with label noise $\sigma_s = 0.1$. A fraction p of labels is then selected uniformly at random and negated ($\tilde{y}_i \mapsto -\tilde{y}_i$), introducing controlled misalignment while keeping the feature geometry fixed. The warm prior is fit from the perturbed labels via ridge regression,

$$A_0^{(m)} = X_{\text{syn}}^\top X_{\text{syn}} + \tau_{\text{pre}} I, \qquad \theta_0^{(m)} = \left(A_0^{(m)}\right)^{-1} X_{\text{syn}}^\top \tilde{y},$$

with regularization $\tau_{\text{pre}} = 1.0$.

Target-shift Priors. For each shift magnitude $\|\Delta\| \in \{0.5, 1.0, 2.0\}$, a displacement vector Δ of the prescribed norm is drawn in a uniformly random direction. The synthetic labels are generated under the displaced parameter $\theta_{\text{syn}} = \theta^\star + \Delta$ with no preference flipping ($p = 0$). The warm prior is then fit from these labels via the same ridge construction. This produces a prior whose θ_0 is systematically offset from $\theta^\star$ by a known, geometrically controlled amount.

The cold-start baseline is always $(A_0^{\text{cold}}, \theta_0^{\text{cold}}) = (I, 0)$.

B.2.3 Deployment Protocol

Each seed runs $T_{\text{deploy}} = 2000$ rounds. The deployment environment is freshly sampled per seed: contexts are drawn i.i.d. from $\mathcal{N}(0, I_d)$ and rewards are computed under the same $\theta^\star$ used to construct the pilot. A warm-started LinUCB instance is initialized from the candidate prior $(A_0^{(m)}, \theta_0^{(m)})$ and run with exploration parameter $\alpha = 1.0$. Cumulative pseudo-regret is computed as the sum of instantaneous regret $x_{t,a_t^\star}^\top \theta^\star - x_{t,a_t}^\top \theta^\star$ over the deployment horizon, where $a_t^\star$ is the best available arm at round t under $\theta^\star$.

B.2.4 Certificate and Decision Protocol

After observing the pilot prefix of length T_{pilot}, the pilot ridge estimator is fit with regularization $\tau_{\text{pilot}} = 1.0$ and the pilot confidence radius $\rho_{\text{pilot}}(\delta)$ is computed per Theorem 1 with norm bound $S = 2.0$ and confidence parameter $\delta = 0.05$. Certified intervals $[\underline{\mathcal{B}}^{(m)}(\delta), \overline{\mathcal{B}}^{(m)}(\delta)]$ are computed for each warm prior and for the cold baseline. The cold-relative decision rule of Definition 3 is then applied to assign Accept, Reject, or Inconclusive. Because $\theta^\star$ is known in the synthetic setting, the oracle prior error $\mathcal{B}_0^{(m)}$ is computed directly and used to verify certified-interval coverage, decision correctness, and the pilot-size-versus-width scaling predicted by Corollary 5.

B.2.5 Soft Deployment Rule

For Inconclusive cases, the soft trust weight $\omega^{(m)}$ is computed from the normalized relative-interval advantage score and passed through a sigmoid with temperature $\kappa = 1.0$, as defined in Section 3. The hybrid prior $(A_0^{\text{hybrid}}, \theta_0^{\text{hybrid}})$ is the interpolation of warm and cold statistics, and a LinUCB instance is initialized from this hybrid prior and run on the same deployment environment.

B.2.6 Default Synthetic Configuration Summary

We present the default configurations we tested in Table 3 for reproduction purposes.

Table 3: Synthetic pipeline configuration.

Parameter	Value
Ambient dimension d	10
Synthetic pretraining size n_s	500
Number of arms K	10
Arm availability probability	0.7
Deployment horizon T_{deploy}	2,000
Pilot sizes T_{pilot}	$\{100, 200, 500, 1000, 5000, 7500, 10000, 15000, 25000\}$
Number of seeds	100
Reward noise σ	0.1
Label noise σ_s	0.1
Warm regularization τ_{pre}	1.0
Pilot regularization τ_{pilot}	1.0
Norm bound S	2.0
Confidence parameter δ	0.05
LinUCB exploration α	1.0
Soft temperature κ	1.0
Flip rates p	$\{0.00, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50\}$
Target-shift magnitudes $\|\Delta\|$	$\{0.5, 1.0, 2.0\}$

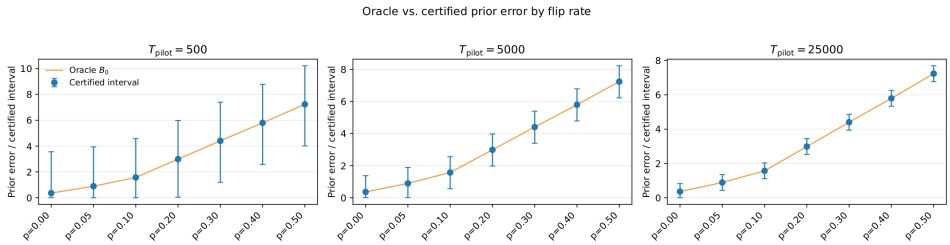

Figure 3: Oracle versus certified prior error for the synthetic flip family at representative pilot sizes. The oracle prior error lies within the certified interval across flip rates and pilot sizes, and the intervals tighten as pilot information increases.

B.3 Additional Synthetic Results

In this subsection, we provide additional synthetic diagnostics to help support the results and claims of the main body. We first show how the certified interval for prior error contains the oracle quantity in the controlled synthetic setting before showing how the uncertainty term contracts with pilot size. This follows the empirical mechanism underlying the behavior we describe in Corollary 5.

B.3.1 Oracle Versus Certified Prior Error

Figures 3 and 4 compare the oracle prior error $\mathcal{B}_0^{(m)}$ with its certified interval across pilot sizes for the flip and target-shift families. In both families, the oracle value lies within the certified interval across the displayed pilot sizes, providing a controlled synthetic check of interval validity consistent with Theorem 2. As the pilot grows, the intervals tighten substantially while preserving coverage.

B.3.2 Width Contraction

Figures 5 and 6 plot the certified width against pilot size for the two synthetic prior families. The widths contract monotonically as the pilot grows, which is consistent with the margin-versus-width mechanism underlying hard-rule decisiveness. In the flip family, different corruption levels have similar contraction profiles. In the target-shift family, the curves nearly overlap because the source covariance geometry is shared across shift magnitudes.

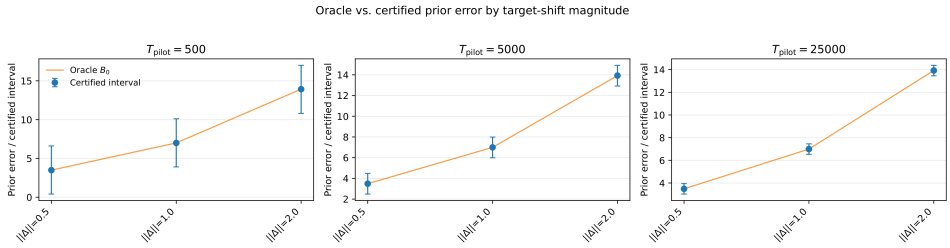

Figure 4: Oracle versus certified prior error for the synthetic target-shift family at representative pilot sizes. The oracle prior error is contained within the certified interval across shift magnitudes, with visibly narrower intervals at larger pilot sizes.

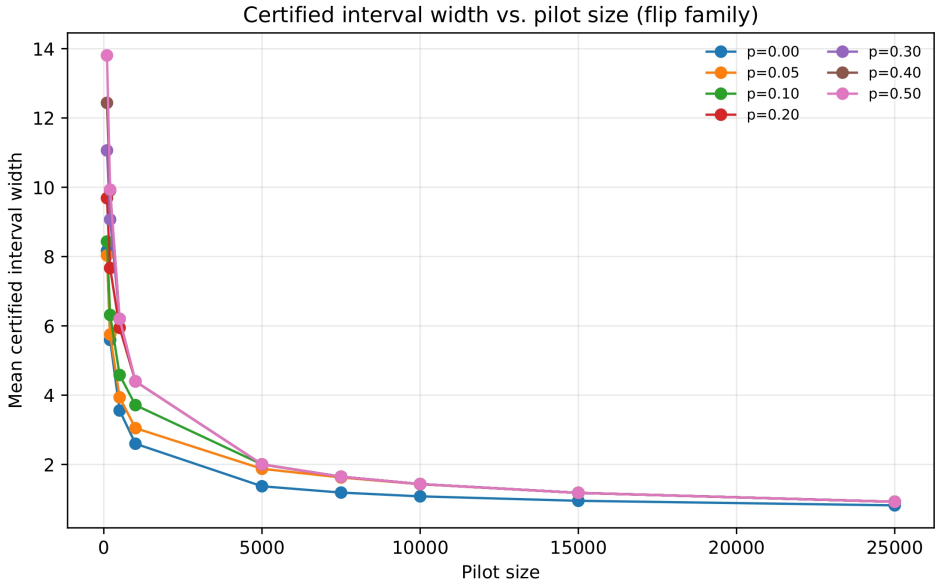

Figure 5: Certified width versus pilot size for the synthetic flip family. Interval width contracts steadily with pilot size across all flip rates, consistent with the pilot-driven tightening underlying Corollary 5.

B.3.3 Dimension Scaling Check

We include a controlled dimension-scaling check for the hard certificate. The goal is not to evaluate downstream soft-gating performance, but to test the certificate geometry predicted by Corollary 5. We fix the source-sample ratio at $n_s/d = 50$, use the same intermediate corruption level $p = 0.20$, and vary the ambient dimension $d \in \{10, 20, 50\}$. For each configuration, we report the combined certified width $W_+^{(m)} = W^{(m)} + W^{\text{cold}}$ and the hard-rule reject rate. The $p = 0.20$ prior is an oracle-worse-than-cold case in this diagnostic, so REJECT is the decisive outcome.

Table 4 shows that the combined width grows with dimension at fixed pilot size and contracts with pilot size at fixed dimension. The width ratios at $T_{\text{pilot}} = 1000$ are 1.49 for $d = 20$ relative to $d = 10$ and 2.31 for $d = 50$ relative to $d = 10$, close to the $\sqrt{d}$ trend expected under the isotropic synthetic design. The reject transition correspondingly shifts to larger pilot budgets: the certificate is fully decisive by $T_{\text{pilot}} = 5000$ for $d = 10$, by 7500 for $d = 20$, and by 25000 for $d = 50$. This supports the margin-versus-width interpretation: higher-dimensional pilots require more samples before the certified intervals separate enough to justify a hard decision.

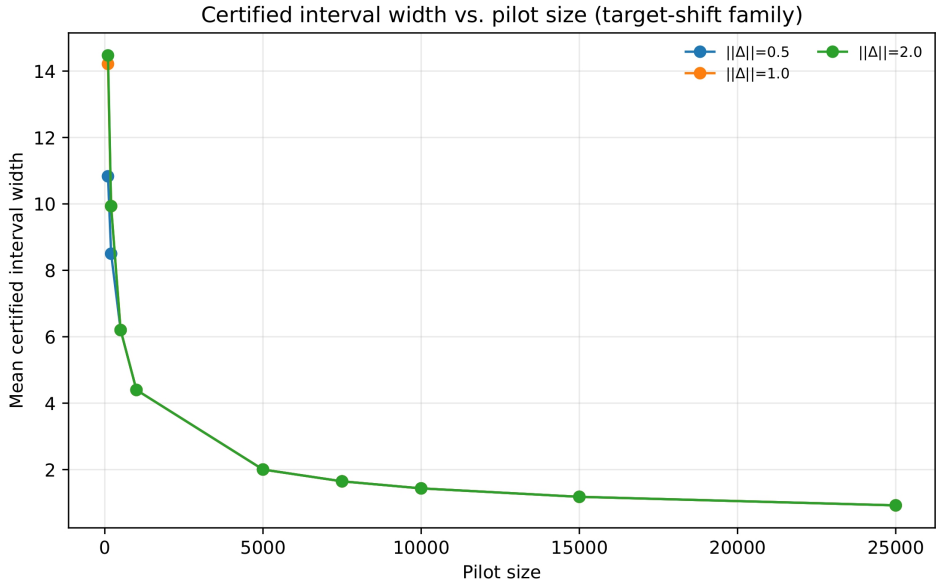

Figure 6: Certified width versus pilot size for the synthetic target-shift family. Width contracts with pilot size and is nearly identical across shift magnitudes because the source covariance geometry is shared across these priors.

Table 4: Synthetic dimension-scaling check with fixed source-sample ratio $n_s/d = 50$ and flip rate $p = 0.20$. Width columns report the mean combined certified width $W_+^{(m)} = W^{(m)} + W^{\text{cold}}$. Reject columns report the fraction of seeds assigned REJECT. Larger dimensions have wider certified intervals at the same pilot size and require larger pilots to become decisive.

d	n_s	$W_+^{(m)}$				Reject rate		
		1,000	5,000	10,000	25,000	5,000	10,000	25,000
10	500	4.97	2.26	1.62	1.04	1.00	1.00	1.00
20	1,000	7.40	3.55	2.55	1.65	0.95	1.00	1.00
50	2,500	11.48	6.64	4.77	3.09	0.00	0.32	1.00

B.3.4 DOWNSTREAM POLICY COMPARISON: FLIP FAMILY

Table 5 reports mean regret deltas versus cold start for the flip family across pilot sizes. For $p = 0.00$, the prior is perfectly aligned and WARM captures a consistent improvement of -3.787 at every pilot size. CONSERVATIVE does not fire ACCEPT until $T_{\text{pilot}} = 25{,}000$, reflecting the fact that the hard rule requires the margin to exceed the combined certificate width, which only occurs at the largest pilot budget. SOFT tracks WARM closely throughout, with the gap narrowing as the pilot grows. For $p = 0.10$, WARM remains helpful (-3.534), but CONSERVATIVE stays at cold throughout, while PERMISSIVE and SOFT partially preserve the benefit. For $p = 0.20$, WARM yields modest benefit (-1.949), but CONSERVATIVE stays at zero throughout and SOFT reverts to cold at $T_{\text{pilot}} \geq 5{,}000$. At $p = 0.20$, $T_{\text{pilot}} = 1{,}000$, SOFT slightly outperforms WARM (-2.931 versus -1.949); this reflects a regularization benefit from the convex blend rather than a systematic property of the rule. For $p \geq 0.30$, WARM is harmful. At $p = 0.30$, PERMISSIVE and SOFT slightly outperform cold at $T_{\text{pilot}} = 1{,}000$, and by $T_{\text{pilot}} \geq 5{,}000$ all gated policies have reverted to cold. For $p = 0.40$ and $p = 0.50$, all gated policies return cold at every pilot size.

B.3.5 DOWNSTREAM POLICY COMPARISON: TARGET-SHIFT FAMILY

Table 6 reports mean regret deltas for the target-shift family. All three shift magnitudes produce harmful warm priors, with WARM incurring increasing regret damage as $\|\Delta\|$ grows ($+15.724$,

Table 5: Synthetic flip family: mean regret delta versus cold start across pilot sizes. Negative values indicate improvement over cold start. COLD is omitted since its delta is zero by definition.

Label	T_{pilot}	Warm	Conservative	Permissive	Soft
`flip_p=0.00`	1,000	−3.787	0.000	−3.787	−3.751
	5,000	−3.787	0.000	−3.787	−3.767
	10,000	−3.787	−0.105	−3.787	−3.775
	25,000	−3.787	−3.532	−3.787	−3.789
`flip_p=0.05`	1,000	−3.743	0.000	−3.743	−3.723
	5,000	−3.743	0.000	−3.743	−3.742
	10,000	−3.743	0.000	−3.743	−3.744
	25,000	−3.743	0.000	−3.743	−3.746
`flip_p=0.10`	1,000	−3.534	0.000	−3.534	−3.605
	5,000	−3.534	0.000	−3.534	−3.596
	10,000	−3.534	0.000	−3.319	−3.382
	25,000	−3.534	0.000	−1.327	−1.334
`flip_p=0.20`	1,000	−1.949	0.000	−1.932	−2.931
	5,000	−1.949	0.000	0.000	0.000
	10,000	−1.949	0.000	0.000	0.000
	25,000	−1.949	0.000	0.000	0.000
`flip_p=0.30`	1,000	+1.923	0.000	−0.066	−0.077
	5,000	+1.923	0.000	0.000	0.000
	10,000	+1.923	0.000	0.000	0.000
	25,000	+1.923	0.000	0.000	0.000
`flip_p=0.40`	1,000	+14.899	0.000	0.000	0.000
	5,000	+14.899	0.000	0.000	0.000
	10,000	+14.899	0.000	0.000	0.000
	25,000	+14.899	0.000	0.000	0.000
`flip_p=0.50`	1,000	+51.601	0.000	0.000	0.000
	5,000	+51.601	0.000	0.000	0.000
	10,000	+51.601	0.000	0.000	0.000
	25,000	+51.601	0.000	0.000	0.000

Table 6: Synthetic target-shift family: mean regret delta versus cold start across pilot sizes. Negative values indicate improvement over cold start; positive values indicate harm. COLD is omitted since its delta is zero by definition.

Label	T_{pilot}	Warm	Conservative	Permissive	Soft
`misalign_‖Δ‖=0.5`	1,000	+15.724	0.000	+6.432	+1.213
	5,000	+15.724	0.000	0.000	0.000
	10,000	+15.724	0.000	0.000	0.000
	25,000	+15.724	0.000	0.000	0.000
`misalign_‖Δ‖=1.0`	1,000	+63.144	0.000	0.000	0.000
	5,000	+63.144	0.000	0.000	0.000
	10,000	+63.144	0.000	0.000	0.000
	25,000	+63.144	0.000	0.000	0.000
`misalign_‖Δ‖=2.0`	1,000	+169.561	0.000	0.000	0.000
	5,000	+169.561	0.000	0.000	0.000
	10,000	+169.561	0.000	0.000	0.000
	25,000	+169.561	0.000	0.000	0.000

+63.144, +169.561). For $\|\Delta\| \geq 1.0$, CONSERVATIVE and SOFT both return zero at every pilot size: the certificate correctly rejects these priors even at $T_{\text{pilot}} = 1{,}000$. For $\|\Delta\| = 0.5$, the prior is only mildly misaligned; at $T_{\text{pilot}} = 1{,}000$ the PERMISSIVE gate still carries residual harm (+6.432) because the pilot is not yet sufficient to trigger a full reject, while SOFT reduces this to +1.213. By $T_{\text{pilot}} = 5{,}000$, all gated policies have reverted to cold start for every shift magnitude.

B.4 COVID-19 Conjoint Deployment Pipeline

B.4.1 Dataset

The target environment is a choice-based conjoint survey administered in July 2020 to 1,970 U.S. respondents during the early COVID-19 pandemic (Kriner et al., 2020; Kreps et al., 2020). The replication data are released under the CC0 1.0 Universal public-domain dedication. In each task, respondents chose between two hypothetical vaccines defined by seven randomized attributes: efficacy, duration of protection, major side-effect risk, minor side-effect risk, FDA approval status, country of origin, and endorser. Each respondent completed multiple tasks, yielding approximately 9,800 real pairwise decisions. These real human responses serve as the target; the pilot is drawn from this pool, and the held-out deployment suffix is disjoint from the pilot.

B.4.2 CBLI Warm Start Construction

Candidate warm priors are constructed using the Contextual-Bandit with LLM Initialization (CBLI) framework introduced by Alamdari et al. (2024). The procedure uses a large language model to simulate synthetic conjoint responses, then fits a linear warm start prior from these synthetic labels.

Concretely, for each synthetic sample, the LLM is prompted to adopt the role of a user characterized by a randomly sampled demographic profile and to choose between two procedurally generated vaccine profiles defined by the same seven attributes as the real conjoint. The exact prompt template is given in Appendix B.4.6. For each pairwise task indexed by i, let $u_i \in \mathbb{R}^{d_u}$ denote the user embedding, and let $v_i^A, v_i^B \in \mathbb{R}^{d_v}$ denote the vaccine embeddings. The pairwise feature is constructed as the normalized outer product

$$x_i = \frac{\mathrm{vec}\big(u_i(v_i^A - v_i^B)^\top\big)}{\|\mathrm{vec}\big(u_i(v_i^A - v_i^B)^\top\big)\|_2}, \qquad x_i \in \mathbb{R}^{d_u d_v},$$

and the synthetic label y_i records the LLM's parsed preference. The warm prior for each arm is then fit via ridge regression on the synthetic (x_i, y_i) pairs, producing $(A_0^{(m)}, \theta_0^{(m)})$ in the block-lifted two-arm construction used for deployment.

Synthetic sample sizes control prior informativeness. We use pretraining sizes $n_s \in \{1\text{k}, 3\text{k}, 10\text{k}, 20\text{k}\}$ generated via GPT-4o (OpenAI, 2024). Under the CBLI assumption of well-aligned LLM responses, larger n_s should yield progressively better-aligned priors on the real human target.

B.4.3 Preference-Flipping for Controlled Misalignment

The CBLI pipeline produces priors of unknown task-specific quality: a prior fit on $n_s = 10{,}000$ synthetic GPT-4o responses may or may not align with real human preferences, and this alignment cannot be verified without access to real target labels. To evaluate the certification framework under varying degrees of misalignment, we use the preference-flipping procedure of Bayley et al. (2026), which introduces controlled label corruption to the synthetic dataset before the ridge fit.

For a base synthetic dataset of n_s labels y_i, the flip variant with rate $f \in [0, 1]$ is constructed by selecting a fraction f of the resolved labels ($y_i \neq 0$) uniformly at random and reversing their signs ($y_i \mapsto -y_i$). Zero or unresolved labels are left untouched. The perturbed label set is then fed through the same ridge fit as the unperturbed version, yielding a prior $(A_0^{(m,f)}, \theta_0^{(m,f)})$ whose feature geometry is unchanged but whose θ_0 is displaced progressively further from the real-human target as f increases. We generate flip variants at rates $f \in \{0.1, 0.5, 0.7\}$ for each pretraining size, denoted `base` (unperturbed), `f1`, `f5`, and `f7` respectively.

Preference flipping provides a controlled misalignment axis orthogonal to pretraining size. `base` priors vary only in prior strength and share nominal alignment quality subject to LLM noise; flip variants at fixed n_s hold the feature geometry fixed but push θ_0 away from the real-human target by a known amount. The certification framework should, in principle, detect high-flip variants as misaligned regardless of pretraining size, while leaving the `base` priors in the accept or inconclusive regime when the LLM and real human preferences are geometrically close. Whether this occurs empirically is addressed in Section 5.

B.4.4 Epistemic Asymmetry Between Prior and Target

A key feature of this scenario is that the synthetic prior and the real target come from *epistemically separate generative processes*. The LLM simulates fictitious users choosing between procedurally generated vaccines; the real conjoint records actual human responses from $1{,}970$ respondents to a fixed survey. These two processes share a feature space by construction, but no mechanism guarantees their reward-generating structures align. Consequently, a certificate that rejects a CBLI prior is not necessarily identifying a corrupted prior. It may be correctly detecting geometric divergence between the LLM's simulated preferences and real human preferences, even for an uncorrupted `base` prior. This distinguishes the COVID scenario from synthetic validation, where ground-truth alignment is controlled by construction, and from the KuaiRand scenario (Appendix B.6), where source and target come from the same platform and the gap between them reflects policy rather than process.

B.4.5 Default COVID Configuration Summary

We present the COVID experiment configuration in Table 7 for reproduction purposes.

Table 7: COVID conjoint pipeline configuration.

Parameter	Value
Synthetic pretraining sizes n_s	$\{1\text{k}, 3\text{k}, 10\text{k}, 20\text{k}\}$
Candidate variants	`base`, `f1`, `f5`, `f7`
Flip rates f	$\{0.0, 0.1, 0.5, 0.7\}$
Deployment horizon T_{deploy}	2,000
Pilot sizes T_{pilot}	$\{100, 500, 1000, 2000, 5000, 7000\}$
Number of seeds	30
Reward noise scale σ	1.0
Warm regularization τ_{pre}	1.0
Pilot regularization τ_{pilot}	1.0
Norm bound S	2.0
Confidence parameter δ	0.05
Soft temperature κ	1.0

B.4.6 LLM Prompt Template

The exact prompt used to elicit synthetic preferences from GPT-4o is reproduced from Alamdari et al. (2024):

> *Consider you are in the middle of the COVID pandemic, where vaccines are just being produced. Pretend to be the following user: [User]. Now you are given two vaccine choices for COVID. The description of each vaccine is as follows: [Vaccine A] Now the next one: [Vaccine B]. Which one do you take? A or B? Let's think step by step. Print the final answer as [Final Answer] at the end as well.*

Placeholders `[User]`, `[Vaccine A]`, and `[Vaccine B]` are replaced at generation time with the sampled demographic profile and the two randomly generated vaccine attribute sets.

B.5 Additional COVID Results

We provide additional experiments and regimes not present in the main text including additional heatmaps with 1k and 20k sizes as well as a downstream policy-figure. We also provide a fixed-pilot policy table.

B.5.1 Additional Reject-Rate Heatmaps

Figure 7 shows that at 1k pretraining the certificate does not reject any candidate across the pilot sizes considered. This is the fully INCONCLUSIVE regime where the synthetic prior is not yet strong enough for the pilot to cleanly separate WARM from COLD.

Figure 8 shows the opposite extreme. Even the unflipped 20k base prior accumulates nonzero reject rate as pilot size grows, reaching 0.13 at $T_{\text{pilot}} = 2000$ and 0.40 by 7000. This should not be read as

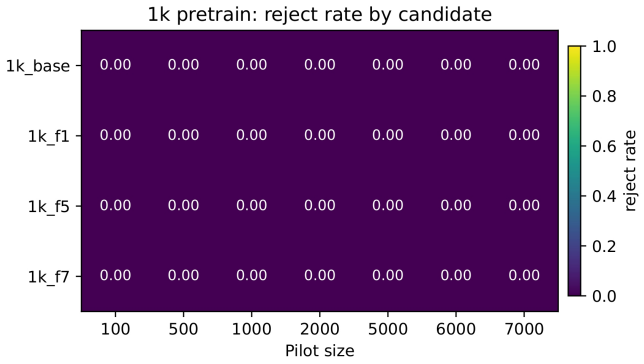

Figure 7: Reject-rate heatmap for the 1k pretraining regime. All candidates remain unrejected across pilot sizes, indicating that the certificate is not yet informative enough to separate warm from cold at this scale.

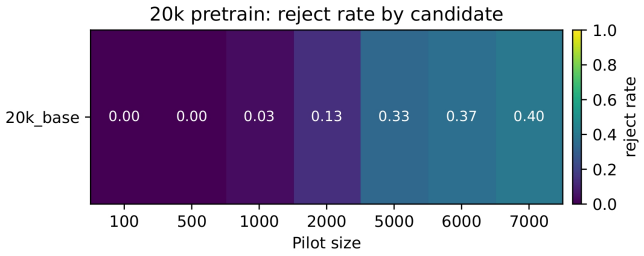

Figure 8: Reject-rate heatmap for the 20k pretraining regime. The unflipped base prior begins to accumulate nonzero reject rate as pilot size grows, reflecting persistent geometric mismatch between synthetic and human preferences despite larger synthetic pretraining.

a false reject. Rather, it reflects the fact that LLM-generated and human preferences remain distinct sources even when synthetic pretraining is large, as aforementioned in Section 5.

B.5.2 Downstream Policy Curves

Figure 9 reports regret delta versus cold across pilot sizes for the 1k, 3k, and 10k regimes. At 1k, the certificate remains effectively uninformative, and the downstream curves reflect this; aligned candidates are helpful, corrupted candidates are harmful, and the soft rule only partially mitigates the latter. At 3k and especially 10k, the same pilot-based signal becomes much more useful. The aligned candidates remain strongly beneficial, while the corrupted candidates separate sharply. Under Warm, the $f5$ and $f7$ priors incur large positive regret deltas, whereas Permissive and Soft pull those curves sharply downward as pilot size increases. This is the deployment-side counterpart to the heatmap evidence in Section 5 and Appendix B.5.

B.5.3 Representative Fixed-Pilot Comparison

Table 8 gives a representative cross-candidate comparison at $T_{\text{pilot}} = 2000$ for the 10k regime. For the aligned candidates (`10k_base`, `10k_f1`), Warm is beneficial and Permissive and Soft preserve most of that gain. For the corrupted candidates (`10k_f5`, `10k_f7`), naive Warm is harmful, while all gated variants revert to cold. In this setting, Conservative remains uniformly at zero because the hard rule does not certify acceptance, whereas Soft retains the advantage of adapting between the helpful and harmful cases across the broader pilot sweep.

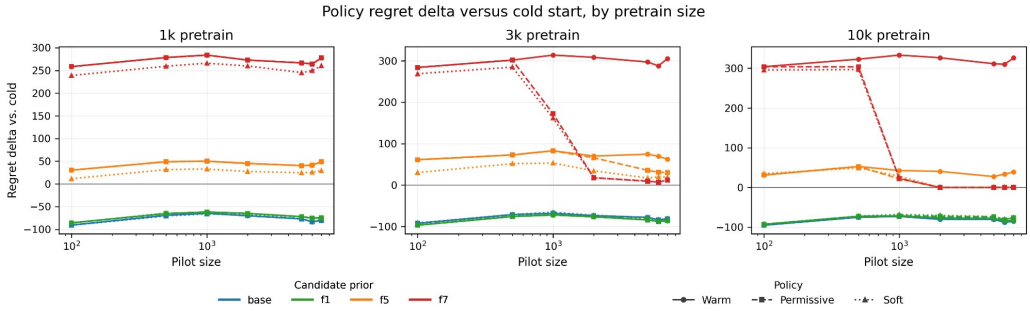

Figure 9: Mean regret delta versus cold start across pilot sizes for the 1k, 3k, and 10k pretraining regimes. Negative values indicate improvement over cold start. At 1k the certificate remains largely inconclusive, while at 3k and 10k the soft and permissive gates substantially reduce the cost of corrupted priors.

Table 8: Representative COVID policy comparison at $T_{\text{pilot}} = 2000$ for the 10k pretraining regime. Entries are mean regret deltas versus cold start. Negative values indicate improvement over cold start; positive values indicate harm.

Candidate	Warm	Conservative	Permissive	Soft
`10k_base`	-80.267	0.000	-76.900	-72.600
`10k_f1`	-76.367	0.000	-72.567	-70.300
`10k_f5`	$+40.467$	0.000	0.000	0.000
`10k_f7`	$+338.600$	0.000	0.000	0.000

B.6 KuaiRand-Pure Historical Deployment Pipeline

B.6.1 Dataset

KuaiRand-Pure is a short-video recommendation dataset released by Kuaishou Technology (Gao et al., 2022). The dataset is released under the Creative Commons Attribution-ShareAlike 4.0 International license (CC BY-SA 4.0). It contains logged interactions from two distinct policy regimes on the same platform: a standard adaptive recommendation stream and a random-intervention stream in which items are served uniformly at random independent of user state. For each interaction, the dataset records the user, the logged item, several interaction signals (click, long-view, like, follow, share), and associated user- and item-side features. The random-intervention stream enables propensity-corrected replay evaluation of counterfactual policies on the same platform the adaptive prior was trained on. After clustering items into arms, we estimate clustered-arm logging propensities from target-stream arm mass fractions and use them for SNIPS evaluation.

B.6.2 Reward Definition and Preprocessing

We use the `long_view` signal as the binary reward, defined by the dataset as a view with a duration exceeding a platform-specific threshold. Among the available feedback signals, `long_view` is substantially denser than `click` (approximately 8.5% positive rate versus $< 1\%$), which provides sufficient signal for LinUCB to adapt within the deployment horizons we consider. The target bundle is the random-intervention stream containing $1{,}046{,}085$ arm-mapped rows after preprocessing.

Candidate source streams and the target stream are preprocessed independently. User-side features (`follow_user_num`, `fans_user_num`, `register_days`) and item-side features (`video_duration`, `like_cnt`, `share_cnt`) are extracted and, where heavy-tailed, transformed by $\log(1 + x)$ and standardized. Rows with missing user or item features are dropped. After preprocessing, the candidate source streams contain $1{,}096{,}253$ rows (`old_full` and `shuffled_old`) and $295{,}497$ rows (`recent`).

Table 9: KuaiRand arm construction on the `recent_main` target bundle. Mass fraction is the share of random-intervention target rows assigned to each arm and defines the clustered-arm propensity used for SNIPS replay evaluation. Mean reward $\bar{r}_a$ is the empirical `long_view` rate on the random-intervention stream; it is reported only as a descriptive diagnostic and was not used during arm construction.

Arm	Target mass (%)	Item count	$\bar{r}_a$
1	15.6	1,045	0.087
2	18.9	1,254	0.084
3	8.9	598	0.122
4	31.4	2,038	0.091
5	10.7	711	0.083
6	7.2	484	0.067
7	7.3	488	0.071

B.6.3 Arm Construction Via Balanced-Mass k-Means

Because KuaiRand items form a continuous feature space rather than a natural discrete action set, we cluster items into arms. A naive k-means on item features produces degenerate clusters in which a handful of items dominate; this yields replay acceptance rates below 0.1% on the random-intervention stream, making downstream evaluation impractical.

We therefore use a balanced-mass variant: we initialize $k_{\text{init}} = 30$ clusters via k-means on standardized item features, then iteratively merge the smallest cluster into its nearest neighbor until all remaining clusters carry at least a fraction $f_{\min} = 0.05$ of target-stream mass. On the `recent_main` target bundle, this procedure terminates at $K = 7$ arms with mass fractions $\{0.156, 0.189, 0.089, 0.314, 0.107, 0.072, 0.073\}$ and item counts ranging from 484 to $2{,}038$. Target rewards are not used at any stage of arm construction. Replay acceptance rates under this construction are approximately 10%, and the resulting arm mass fractions define the clustered-arm propensities used for SNIPS replay evaluation.

B.6.4 Arm Construction Diagnostics

Table 9 reports the final arm abstraction used in all KuaiRand experiments. The balanced-mass procedure uses item features and target-stream exposure counts, but not target rewards. The resulting arms are not exactly uniform, but every arm carries nontrivial target mass; this avoids the replay-starvation problem of naive clustering while preserving item-feature structure. The target mass fractions define the clustered-arm propensities used for SNIPS replay evaluation. The reward column is computed only after the arm map is fixed and is reported as a descriptive diagnostic.

B.6.5 Feature Map

For each interaction, the LinUCB feature vector is assembled as

$$x_{t,a} = \left[\, 1,\ \phi_{u_t},\ \psi_a,\ \phi_{u_t} \otimes_{\text{sub}} \psi_a \,\right] \in \mathbb{R}^{45},$$

where $\phi_u \in \mathbb{R}^{25}$ contains user-side covariates, $\psi_a \in \mathbb{R}^{10}$ contains arm-aggregated item features, and $\phi_u \otimes_{\text{sub}} \psi_a$ denotes a 9-dimensional interaction block constructed as pairwise products between three user features (`follow_user_num`, `fans_user_num`, $\log(1 + $ `register_days`$)$) and three item features (`video_duration`, `like_cnt`, `share_cnt`). The interaction block is included to avoid a nearly degenerate additive feature geometry. In diagnostics without the interaction block, the fitted θ_0 vectors for `old_full` and the reward-permuted null have cosine similarity above 0.99, meaning that the real and shuffled priors are almost collinear. With the interaction block, the same cosine drops to approximately 0.25 at source size 100, indicating that the enriched feature map substantially separates the real historical prior from the reward-permuted null while preserving their shared temporal and platform-level structure.

B.6.6 Replay Evaluation and SNIPS Reward

For each target row (u_t, a_t, r_t), LinUCB computes UCB scores over all K arms and selects $\hat{a}_t$. If $\hat{a}_t = a_t$, the row is accepted: the bandit observes reward r_t and updates its posterior. If $\hat{a}_t \neq a_t$,

the row is skipped. This is a replay evaluation on logged randomized feedback; because item-level randomization induces unequal mass over the clustered arms, we use self-normalized inverse-propensity weighting for the headline policy metric (Swaminathan & Joachims, 2015).

Let $p(a)$ denote the clustered-arm logging propensity, estimated by the target-stream mass fraction of arm a. For a policy π, we compute

$$\widehat{V}_{\mathrm{SNIPS}}(\pi) = \frac{\sum_t \mathbf{1}\{\pi(x_t) = a_t\} r_t / p(a_t)}{\sum_t \mathbf{1}\{\pi(x_t) = a_t\} / p(a_t)}.$$

We report SNIPS replay reward lift versus cold start,

$$\Delta_{\mathrm{SNIPS}}(\pi) = \widehat{V}_{\mathrm{SNIPS}}(\pi) - \widehat{V}_{\mathrm{SNIPS}}(\textsc{Cold}).$$

Positive values indicate improvement over cold.

B.6.7 Fixed-Deploy Window

We use a *fixed-deploy* windowing scheme to avoid two distinct sources of confounding: between-seed leakage across pilot and deployment, and within-seed confounding between pilot length and deployment epoch. The scheme is constructed so that no target row is ever used for both pilot and deployment within a seed, and so that each chosen T_{pilot} evaluates deployment on exactly the same target data.

Concretely, we partition the chronological target stream into $n_{\mathrm{seeds}} = 10$ non-overlapping windows, indexed by seed s with anchor $t^{(s)}_{\mathrm{anchor}}$ spaced uniformly across the stream. For each seed, the deployment window is fixed at

$$\big[\, t^{(s)}_{\mathrm{anchor}},\ t^{(s)}_{\mathrm{anchor}} + T_{\mathrm{deploy}} \,\big),$$

and the pilot is drawn from a strictly earlier interval

$$\big[\, t^{(s)}_{\mathrm{anchor}} - T_{\mathrm{pilot}},\ t^{(s)}_{\mathrm{anchor}} \,\big),$$

so that pilot and deployment are disjoint intervals of the stream within every seed. Varying T_{pilot} extends the pilot *backward* from $t^{(s)}_{\mathrm{anchor}}$ rather than shifting the deployment *forward*, so each pilot-size comparison evaluates deployment on the same target rows. Across seeds, the spacing between anchors is set so that deployment windows do not overlap each other either: for $T_{\mathrm{deploy}} = 10{,}000$ and the largest pilot length used in any experiment ($T^{\max}_{\mathrm{pilot}} = 20{,}000$), the inter-anchor stride satisfies $\Delta > T^{\max}_{\mathrm{pilot}} + T_{\mathrm{deploy}}$, ensuring that no target row appears in more than one seed's pilot-deployment pair. The full target stream thus provides 10 disjoint pilot-deployment pairs, and every reported result is an average over these 10 non-overlapping evaluations.

B.6.8 Default KuaiRand Configuration Summary

We present the KuaiRand experiment configuration in Table 10 for reproduction purposes.

B.7 Additional KuaiRand Source-Size Results

B.7.1 Source-Size Sweep Against Cold Start

Table 11 reports the full fixed-deploy source-size sweep at $T_{\mathrm{pilot}} = 20{,}000$ across all four source sizes. At $s = 100$, the source size is too small to support stable transfer. `old_full` is indistinguishable from cold (-0.0025, $p = 0.557$), while `shuffled_old` shows a small positive lift ($+0.0043$) with only marginal Wilcoxon evidence ($p = 0.064$). At $s = 1{,}000$, neither candidate provides useful transfer. The real historical prior remains statistically indistinguishable from cold (-0.0072, $p = 0.160$), while the reward-permuted null is significantly harmful (-0.0149, $p = 0.010$). At $s = 10{,}000$, the pattern reverses: `old_full` improves over cold by $+0.0145$ SNIPS reward lift (14.3% of $V_{\mathrm{cold}} = 0.1014$, $p = 0.002$), while `shuffled_old` remains non-beneficial (-0.0058, $p = 0.160$). At $s = 100{,}000$, `old_full` remains beneficial ($+0.0118$, 11.7% of cold, $p = 0.002$), while `shuffled_old` becomes clearly harmful (-0.0189, $p = 0.006$). The pairwise contrasts in the main table show that the real historical prior separates from the reward-permuted null once source size is sufficient ($+0.0203$ at $s = 10{,}000$ and $+0.0308$ at $s = 100{,}000$, both $p = 0.002$). Transferable historical signal appears at moderate-to-large source sizes, whereas matched source geometry without reward alignment does not.

Table 10: KuaiRand pipeline configuration.

Parameter	Value
Preprocessed bundles	`old_main`, `recent_main`
Target stream	`recent_main` random-intervention stream
Candidate priors	`old_full`, `recent`, `shuffled_old`
Historical source sizes s	$\{100, 1{,}000, 10{,}000, 100{,}000\}$
Number of arms K	7
Arm construction	Balanced-mass k-means on item features
Minimum arm mass	0.05
Cluster seed	20260415
Reward	`long_view`
Evaluation metric	SNIPS replay reward lift versus cold start
Clustered-arm propensities	Target-stream arm mass fractions
Pilot sizes T_{pilot}	$\{5{,}000, 20{,}000\}$
Deployment horizon T_{deploy}	10,000
Chronological windows per config	10
LinUCB exploration α	0.5
Warm regularization τ_{pre}	1.0
Pilot regularization τ_{pilot}	1.0
Reward noise scale σ	1.0
Norm bound S	2.0
Confidence parameter δ	0.05
Soft temperature κ	1.0

Table 11: Full KuaiRand source-size sweep using SNIPS replay reward lift versus cold start at $T_{\text{pilot}} = 20{,}000$. Positive values indicate improvement over cold. Confidence intervals are paired bootstrap 95% intervals over the 10 fixed-deploy windows, p-values are Wilcoxon signed-rank tests.

Source size	Candidate	Δ_{SNIPS}	95% CI	p
100	`old_full`	-0.0025	$[-0.0082, +0.0033]$	0.557
100	`shuffled_old`	$+0.0043$	$[+0.0011, +0.0082]$	0.064
1,000	`old_full`	-0.0072	$[-0.0153, +0.0009]$	0.160
1,000	`shuffled_old`	-0.0149	$[-0.0241, -0.0070]$	0.010
10,000	`old_full`	$+0.0145$	$[+0.0108, +0.0183]$	0.002
10,000	`shuffled_old`	-0.0058	$[-0.0116, +0.0006]$	0.160
100,000	`old_full`	$+0.0118$	$[+0.0074, +0.0170]$	0.002
100,000	`shuffled_old`	-0.0189	$[-0.0270, -0.0105]$	0.006

B.7.2 Temporal-Source Diagnostic

The main KuaiRand results focus on `old_full` versus `shuffled_old`, which isolates real historical reward signal from a reward-permuted null with matched source geometry. We also evaluate `recent`, a temporally closer but smaller historical source, as a diagnostic for temporal proximity. This comparison is secondary because `recent` changes both temporal distance and effective source coverage. At $s = 10{,}000$, `old_full` covers 99.4% of target items after arm mapping, whereas `recent` covers 87.3%.

Table 12 shows that temporal proximity alone does not determine prior quality. At $s = 1{,}000$, `recent` is close to cold start performance, while `old_full` is also not reliably beneficial and `shuffled_old` is harmful. Similar to the results in subsection B.7.1, transfer is weak at small source sizes. At $s = 10{,}000$, both real historical priors improve over cold, with `old_full` outperforming `recent` despite being older. The larger lift for `old_full` is consistent with the coverage trade-off: the older source is temporally farther from the target but covers more of the target item space at the same nominal source size. At $s = 100{,}000$, the two real priors converge to nearly identical performance, with `old_full` at $+0.0118$ and `recent` at $+0.0121$ (both $p = 0.002$), while `shuffled_old` is clearly harmful (-0.0189, $p = 0.006$). Real historical reward signal can transfer under policy shift, but temporal proximity and source coverage trade off in finite samples.

Table 12: Temporal-source diagnostic at $T_{\text{pilot}} = 20{,}000$. The target stream is the `recent_main` random-intervention log from 4/22–5/08. The `recent` prior is fit from the concurrent standard-policy source over the same date range, while `old_full` comes from the earlier standard-policy period. Entries are SNIPS replay reward lift versus cold start, positive values indicate improvement. 95% CIs are paired bootstrap intervals, p-values are Wilcoxon signed-rank.

Source size	Candidate	Δ_{SNIPS}	95% CI	p
1,000	`old_full`	-0.0072	$[-0.0153, +0.0009]$	0.160
1,000	`recent`	$+0.0007$	$[-0.0038, +0.0054]$	0.770
1,000	`shuffled_old`	-0.0149	$[-0.0241, -0.0070]$	0.010
10,000	`old_full`	$+0.0145$	$[+0.0108, +0.0183]$	0.002
10,000	`recent`	$+0.0081$	$[+0.0036, +0.0130]$	0.010
10,000	`shuffled_old`	-0.0058	$[-0.0116, +0.0006]$	0.160
100,000	`old_full`	$+0.0118$	$[+0.0074, +0.0170]$	0.002
100,000	`recent`	$+0.0121$	$[+0.0088, +0.0160]$	0.002
100,000	`shuffled_old`	-0.0189	$[-0.0270, -0.0105]$	0.006

B.7.3 HARD-CERTIFICATE DIAGNOSTICS

Table 13 reports certified upper bounds and hard-rule decision rates across the full source-size and pilot-size grid. Under the cold-relative hard rule, a candidate is accepted if $\overline{\mathcal{B}}^{(m)} \leq \underline{\mathcal{B}}^{\text{cold}}$ and rejected if $\underline{\mathcal{B}}^{(m)} > \overline{\mathcal{B}}^{\text{cold}}$; otherwise the outcome is INCONCLUSIVE. Across all 16 configuration cells, the certificate is INCONCLUSIVE: ACCEPT and REJECT rates are both zero.

The mechanism is straightforward. The floored lower bound $\underline{\mathcal{B}}^{(m)}$ equals zero for every warm candidate (the uncorrected value is negative throughout), and the floored cold lower bound $\underline{\mathcal{B}}^{\text{cold}}$ is likewise zero. The accept condition therefore reduces to $\overline{\mathcal{B}}^{(m)} \leq 0$, which is never satisfied for a positive upper bound. The reject condition reduces to $0 > \overline{\mathcal{B}}^{\text{cold}}$, which is also never satisfied. Thus every evaluated configuration falls into the inconclusive region under the realized bounds.

The warm upper bound $\overline{\mathcal{B}}^{(m)}$ is comparable to $\overline{\mathcal{B}}^{\text{cold}}$ at $s = 100$ ($\approx$22 versus $\approx$19), but grows substantially at larger source sizes, reaching $\approx$1133 at $s = 100{,}000$, $T_{\text{pilot}} = 5{,}000$. This growth is driven by the interaction between a stronger warm prior and the pilot geometry: larger source sizes increase the prior precision and the norm-conversion factor $\gamma^{(m)}$, so the certified uncertainty term $\gamma^{(m)}\rho_{\text{pilot}}$ can dominate the interval even when the warm prior improves downstream deployment. Increasing the pilot from $5{,}000$ to $20{,}000$ reduces these upper bounds, but not enough to make the hard rule decisive. The downstream policy results in Table 11 nonetheless show that the warm prior, `old_full`, is genuinely helpful at $s \geq 10{,}000$, confirming that the persistent inconclusiveness reflects the finite-pilot certificate rather than the absence of historical signal.

B.7.4 SOFT-VERSUS-WARM COMPARISON

Table 14 reports the paired difference between the soft hybrid policy and ungated warm start. Positive values indicate that SOFT improves SNIPS lift relative to WARM. The clearest takeaway can be seen at $s = 100{,}000$ where SOFT improves over ungated `shuffled_old` by $+0.0131$ ($p = 0.002$), reducing the SNIPS loss from -0.0189 to -0.0059 in the representative policy table. For the `old_full` prior, SOFT stays close to WARM; at $s = 100{,}000$, the difference is small and not significant (-0.0016, $p = 0.160$).

The smaller source rows are more volatile. At $s = 100$, SOFT is worse than WARM for `shuffled_old` (-0.0089, $p = 0.020$). At $s = 10{,}000$, the `shuffled_old` soft gain is positive ($+0.0039$), but the Wilcoxon test is not significant ($p = 0.232$). We therefore treat the robust soft-protection evidence as concentrated at $s = 100{,}000$, where the harmful null is clearly separated from cold and the soft improvement over warm is statistically reliable.

Table 13: Hard-certificate diagnostics for the KuaiRand source-size sweep at $T_{\text{pilot}} \in \{5{,}000,\ 20{,}000\}$. Entries report the certified warm-prior upper bound $\overline{\mathcal{B}}^{(m)}$. Accept and reject rates are zero in every cell; all outcomes are INCONCLUSIVE. Both warm and cold lower bounds are floored to zero throughout. $\overline{\mathcal{B}}^{\text{cold}}$ is shared across candidates for a given T_{pilot}: 19.3 at $T_{\text{pilot}} = 5{,}000$ and 20.8 at $T_{\text{pilot}} = 20{,}000$.

		$\overline{\mathcal{B}}^{(m)}$	
Source size	Candidate	$T_{\text{pilot}} = 5{,}000$	$T_{\text{pilot}} = 20{,}000$
100	`old_full`	22.5	24.0
	`shuffled_old`	22.2	23.7
1,000	`old_full`	27.6	28.4
	`shuffled_old`	28.1	28.9
10,000	`old_full`	261.7	110.2
	`shuffled_old`	257.1	105.7
100,000	`old_full`	1133.2	439.1
	`shuffled_old`	1124.3	430.2

Table 14: KuaiRand soft-versus-warm comparison at $T_{\text{pilot}} = 20000$. Entries report $\Delta_{\text{SNIPS}}(\text{SOFT}) - \Delta_{\text{SNIPS}}(\text{WARM})$, positive values indicate that SOFT improves over ungated WARM. Confidence intervals are paired bootstrap 95% intervals and p-values are Wilcoxon signed-rank tests.

Source size	Candidate	Soft − Warm	95% CI	p
100	`old_full`	-0.0015	$[-0.0065, +0.0034]$	0.625
100	`shuffled_old`	-0.0089	$[-0.0148, -0.0033]$	0.020
1,000	`old_full`	$+0.0019$	$[-0.0027, +0.0066]$	0.770
1,000	`shuffled_old`	$+0.0046$	$[-0.0024, +0.0130]$	0.375
10,000	`old_full`	-0.0001	$[-0.0019, +0.0012]$	0.695
10,000	`shuffled_old`	$+0.0039$	$[+0.0001, +0.0078]$	0.232
100,000	`old_full`	-0.0016	$[-0.0033, +0.0002]$	0.160
100,000	`shuffled_old`	$+0.0131$	$[+0.0087, +0.0179]$	0.002

B.8 SOFT-WEIGHT SENSITIVITY AND FIXED-WEIGHT COMPARISONS

B.8.1 SOFT-TEMPERATURE SENSITIVITY

The soft rule uses temperature, κ, in the INCONCLUSIVE regime. Certified ACCEPT and REJECT outcomes are clamped to the hard endpoints and are unaffected by κ. We run a small sensitivity sweep over $\kappa \in \{0.25, 0.5, 1, 2, 4\}$ on representative rows from the three experimental settings. The goal is not to tune κ but to determine whether the qualitative behavior of the soft rule is an artifact of the default $\kappa = 1$ choice we use for the experiments.

Table 15 reports the deployment effect of SOFT against cold start. The synthetic corrupted candidates considered in this search were almost always hard-rejected, with inconclusive rates of 0.005 for `flip_p=0.30` and 0.000 for `flip_p=0.40`. We omit them as κ has little opportunity to affect the deployed policy. The remaining rows include one helpful synthetic prior and helpful/corrupted or helpful/null cases from the two real-data settings where the hard rule is inconclusive and soft interpolation is active.

The qualitative behavior is stable over the tested range. The helpful priors in Synthetic and COVID remain beneficial for every κ, and `old_full` remains beneficial in KuaiRand. The corrupted COVID prior and the KuaiRand reward-permuted null remain harmful, but larger κ reduces the magnitude of harm by moving the soft rule more aggressively toward the cold side when the certificate score leans cold. We conclude that $\kappa = 1$ is not an outlier, but rather lies between the more gradual soft weighting induced by smaller κ and the more conservative, cold-leaning behavior induced by larger κ.

Table 15: Soft-temperature sensitivity. Entries report SOFT deployment effect versus cold start; positive values indicate improvement and negative values indicate harm. Synthetic and COVID entries are $-\Delta_{\text{regret}}$, while KuaiRand entries are SNIPS replay reward lift. The main experiments use $\kappa = 1$.

Setting	Candidate / regime	$\kappa = .25$	$\kappa = .5$	$\kappa = 1$	$\kappa = 2$	$\kappa = 4$
Synthetic	`flip_p=0.00`	$+3.735$	$+3.734$	$+3.728$	$+3.717$	$+3.694$
COVID	`10k_base`	$+79.333$	$+78.667$	$+78.967$	$+78.933$	$+79.867$
COVID	`10k_f7`	-254.667	-251.933	-246.067	-236.467	-183.700
KuaiRand	`old_full`, $s = 100{,}000$	$+0.011$	$+0.011$	$+0.010$	$+0.011$	$+0.006$
KuaiRand	`shuffled_old`, $s = 100{,}000$	-0.008	-0.010	-0.006	-0.007	-0.003

This sweep should not be interpreted as a deployment-time tuning procedure. In a real pre-deployment workflow, κ must be fixed before observing deployment outcomes, otherwise, the choice would leak information from the evaluation horizon into the policy. We use $\kappa = 1$ as a unit-temperature default and report the sweep only as a post hoc sensitivity check. A principled calibration rule for κ would need to depend only on pilot-side quantities, historical validation tasks, or externally specified risk tolerance, and is separate from the hard certificate's correctness guarantee.

B.8.2 RETENTION INTERPRETATION AND IMPLEMENTATION CHECK

For an INCONCLUSIVE decision, the warm and cold intervals overlap. Their midpoint difference is therefore no larger in absolute value than half the sum of their widths, which gives $\mathcal{S}^{(m)} \in [-1, 1]$ on this branch. Since the sigmoid is increasing,

$$\sigma(-\kappa) \le \omega^{(m)} \le \sigma(\kappa).$$

We can define a retained-weight floor $\varepsilon \in (0, 1/2)$ by

$$\varepsilon = \frac{1}{1 + e^{\kappa}}, \qquad \kappa = \log \frac{1 - \varepsilon}{\varepsilon}.$$

Under this definition, the inconclusive-branch weight lies in $[\varepsilon, 1 - \varepsilon]$. The certified ACCEPT and REJECT endpoints remain 1 and 0. This reparameterization describes how κ controls source retention when the evidence is ambiguous. It does not calibrate a probability of downstream harm or extend the hard certificate to intermediate weights.

We check the implementation using the policy outputs underlying the temperature sweep. For every seed, candidate, pilot size, and value of κ, we record the certificate branch, normalized score, and resulting source weight. The check covers $4{,}700$ policy evaluations across the synthetic, COVID, and KuaiRand conditions in Table 15. Of these, $1{,}995$ use a certified endpoint and $2{,}705$ use the inconclusive branch. Every inconclusive weight lies within its corresponding interval $[\varepsilon, 1 - \varepsilon]$, with a minimum observed distance of 0.0026 from either boundary.

B.8.3 FIXED SOURCE WEIGHTS

We next test whether a constant source weight reproduces the behavior of SOFT. The fixed-weight baseline interpolates the warm and cold sufficient statistics with $\omega \in \{0.25, 0.30, 0.50, 0.75\}$. It does not consult the certificate and never clamps to a hard endpoint. Table 16 reports $\omega = 0.30$, which is close to the weights selected by SOFT in the real-data comparisons. A/R/I gives the proportions of ACCEPT, REJECT, and INCONCLUSIVE decisions made by the certificate.

When every decision is INCONCLUSIVE, the fixed $\omega = 0.30$ blend closely matches SOFT on KuaiRand. The absolute differences are at most 0.0006 SNIPS, and none is statistically detected. The distinction appears when the certificate reaches REJECT. On the corrupted COVID prior, SOFT removes the source, while a fixed positive blend continues to retain it. This reduces the regret difference from $+324.800$ to $+22.333$ at $T_{\text{pilot}} = 1{,}000$ and from $+317.800$ to 0.000 at $2{,}000$. The comparison separates the value of hard rejection from the behavior of logistic interpolation on inconclusive cases.

Table 16: SOFT compared with a fixed $\omega = 0.30$ blend. COVID entries are regret differences relative to COLD. KuaiRand entries are SNIPS reward lifts.

Setting	Candidate	T_{pilot}	A/R/I	SOFT	Fixed	p
COVID	`10k_base`	1,000	0.00/0.00/1.00	−68.667	−68.700	0.796
COVID	`10k_base`	7,000	0.00/0.03/0.97	−81.833	−85.400	0.005
COVID	`10k_f7`	1,000	0.00/0.93/0.07	+22.333	+324.800	< 0.001
COVID	`10k_f7`	2,000	0.00/1.00/0.00	0.000	+317.800	< 0.001
KuaiRand	`old_full`	5,000	0.00/0.00/1.00	+0.0111	+0.0105	0.275
KuaiRand	`shuffled_old`	5,000	0.00/0.00/1.00	−0.0065	−0.0062	0.922
KuaiRand	`old_full`	20,000	0.00/0.00/1.00	+0.0103	+0.0105	0.695
KuaiRand	`shuffled_old`	20,000	0.00/0.00/1.00	−0.0059	−0.0062	0.625

B.9 PILOT RETENTION AND DECISION TIMING

The certification protocol compares the candidate warm prior and the cold baseline on a shared target-task pilot. The candidate is constructed before the pilot, and the pilot policy does not depend on that candidate. This common design isolates the source-trust decision from the candidate's later influence on deployment actions.

B.9.1 PILOT-INCLUSIVE INITIALIZATION

The held-out deployment suffix used in the main experiments separates evaluation outcomes from the data used for certification. The framework does not require discarding the pilot. Let

$$C_{\text{pilot}} = \sum_{t \le T_{\text{pilot}}} x_t x_t^\top, \qquad h_{\text{pilot}} = \sum_{t \le T_{\text{pilot}}} r_t x_t$$

denote its unregularized sufficient statistics. After the gate selects source statistics (A_g, b_g), the deployment learner can be initialized as

$$V_{\text{init}} = A_g + C_{\text{pilot}}, \qquad b_{\text{init}} = b_g + h_{\text{pilot}}, \qquad \hat{\theta}_{\text{init}} = V_{\text{init}}^{-1} b_{\text{init}}.$$

The pilot regularizer used to construct the certificate is not added again, and each pilot observation is incorporated only once. PILOT-ONLY uses $(A_g, b_g) = (I, 0)$, WARM+PILOT uses the candidate source statistics, and GATED+PILOT uses the source contribution selected by the gate.

Table 17 compares these three initializations on the same seeds, pilot prefixes, and deployment suffixes used in the main experiments.

Table 17: Pilot-inclusive comparisons measured relative to the original COLD policy. COVID and synthetic entries are regret differences, where lower is better. KuaiRand entries are SNIPS reward lifts, where higher is better.

Setting	Candidate	T_{pilot}	PILOT ONLY	WARM+PILOT	GATED+PILOT
COVID	`10k_base`	1,000	−12.467	−71.467	−62.533
COVID	`10k_base`	2,000	−26.267	−72.333	−58.833
COVID	`10k_f7`	1,000	−12.467	+219.600	−5.400
KuaiRand	`old_full`	5,000	+0.0131	+0.0159	+0.0163
KuaiRand	`old_full`	20,000	+0.0161	+0.0164	+0.0164
Synthetic	$\|\Delta\| = 2.0$	5,000	−4.031	+5.502	−4.031

Pilot-only learning does not uniformly dominate retaining source information. On the aligned COVID prior, GATED+PILOT improves over PILOT-ONLY by 50.1 and 32.6 regret units at pilot sizes 1,000 and 2,000. WARM+PILOT performs better on these aligned rows, which reflects the value of retaining a useful source. The corrupted COVID and synthetic rows show the complementary case. Combining the full warm prior with the pilot remains harmful, while gating removes most or all of that cost. On KuaiRand, the gain from gating over PILOT-ONLY decreases from 0.0032 SNIPS at $T_{\text{pilot}} = 5{,}000$ to 0.0003 at 20,000.

Pilot reuse does not change the hard guarantee in Theorem 4. That guarantee continues to order the original warm and cold prior errors. It does not certify an ordering of the pilot-updated initializations or their deployment regret.

B.9.2 PREFIX-BASED DECISIONS

Theorem 1 holds simultaneously over pilot prefixes up to a prespecified finite horizon. The certificate may therefore be evaluated after each prefix, and data collection may stop at the first ACCEPT or REJECT decision. No additional union bound over these prefixes is required. If the certificate remains INCONCLUSIVE at the planned stopping point, deployment can use the stated SOFT rule.

Revising source trust after the source has begun influencing action selection is a different online-transfer setting. Appendix B.10.2 compares pre-deployment gating with ARROW-CB, which updates source trust from deployment feedback.

B.10 ALTERNATIVE SOURCE-TRUST BASELINES

This section evaluates two additional pilot-based rules and an online-adaptation baseline. COVID outcomes are regret differences relative to COLD, where lower values are better. KuaiRand outcomes are SNIPS reward lifts relative to COLD, where higher values are better. Policy comparisons use the same seeds, pilot prefixes, and deployment suffixes. We use two-sided Wilcoxon signed-rank tests on paired outcomes.

B.10.1 ALTERNATIVE PILOT-BASED RULES

The point-estimate selector chooses WARM when the estimated warm-prior error is smaller than the estimated cold-prior error, and otherwise chooses COLD. It is evaluated on $5{,}259$ decisions from the synthetic and real-data sweeps. Every certified hard decision agrees with the point-estimate selector, so differences arise only when the certificate abstains. On the evaluated real-data rows, the selector chooses COLD in all 360 COVID evaluations and all 40 KuaiRand evaluations. Within the COVID cases where the certificate is INCONCLUSIVE, it matches the realized downstream ordering in 8.7% of evaluations. SOFT has 32.2 lower regret on average across these COVID settings, with $p < 0.001$.

The pilot-likelihood rule assigns the warm source its posterior model probability under a linear-Gaussian model with equal prior probabilities for warm and cold. We use the same noise scale as the pilot confidence radius and evaluate the rule on 40 KuaiRand cells formed by 10 windows, two candidates, and two pilot sizes. The likelihood weight falls outside $[0.01, 0.99]$ in 87.5% of these cells. At $T_{\mathrm{pilot}} = 20{,}000$, it selects the cold endpoint in every cell. Its SNIPS lift on `old_full` is then 0.0000, compared with $+0.0103$ for SOFT, with $p = 0.002$. At $T_{\mathrm{pilot}} = 5{,}000$, the corresponding values are $+0.0074$ and $+0.0111$, with $p = 0.049$. On `shuffled_old`, moving to the cold endpoint avoids the small loss retained by SOFT. The likelihood rule therefore produces a different value-retention tradeoff rather than uniformly improving over SOFT.

These comparisons also explain why a useful prior can remain INCONCLUSIVE. When the converted cold uncertainty radius exceeds $\|\hat{\theta}^{\mathrm{pilot}}\|_2$, the lower endpoint of the cold interval is zero. Since the warm upper endpoint is nonnegative, ACCEPT is then unavailable even when the prior contains useful downstream signal.

B.10.2 PRE-DEPLOYMENT GATING AND ONLINE ADAPTATION

We compare gated and ungated ARROW-CB on the corrupted COVID candidate `10k_f7`. Both policies use the same deployment suffix and 15 paired seeds. The gate removes the supervised source only when the certificate returns REJECT. The ARROW-CB implementation follows Algorithm 1 of Zhang et al. (2019) with $K = 2$, $\varepsilon = 0.0125$, and the eight-value source-weight grid used by the original method. The comparison uses $T_{\mathrm{pilot}} = 2{,}000$ and $T_{\mathrm{deploy}} = 2{,}000$.

Gating has lower observed regret through $T = 1{,}000$, but the differences at 100 and 250 rounds do not survive a Bonferroni correction over the five reported checkpoints. At $T = 2{,}000$, the contrast reverses sign and is not statistically detected. Over the same horizon, the ungated learner's mean bandit weight $\bar{\lambda}$ increases from 0.714 to 0.917. The pattern is consistent with pre-deployment gating reducing early exposure to a harmful source while online adaptation narrows the difference as

Table 18: Gated and ungated ARROW-CB on the corrupted COVID candidate `10k_f7`. Results use 15 paired seeds. Negative contrasts favor gating.

Rounds	Ungated	Gated	Gated − ungated	p	$\bar{\lambda}$
100	39.267	36.200	−3.067	0.021	0.714
250	93.933	89.400	−4.533	0.026	0.776
500	189.467	181.800	−7.667	0.169	0.811
1,000	368.733	359.400	−9.333	0.489	0.862
2,000	715.400	718.933	+3.533	0.679	0.917

feedback accumulates. It does not establish a statistically robust advantage or uniform dominance of either strategy.

B.11 Computational Resources

All experiments were run on a local CPU workstation. The machine used an Intel Core i7-14700F processor with 20 physical cores and 28 logical processors, and 32GB of RAM. No GPU acceleration was required for the reported experiments. The longest full experimental configuration sweep took just under one hour of wall-clock time on this workstation.

B.12 Broader Impact

Pilot-based prior certification can reduce harm from deploying misaligned warm priors. Its main risk is misuse as a substitute for live safety monitoring. An incorrect implementation, weak pilot, or unrepresentative target sample could give unwarranted confidence in a prior. The method, therefore, should be used as intended: a pre-deployment diagnostic alongside domain-specific auditing, privacy review, and post-deployment monitoring, not as a standalone safety guarantee.